STATISTICS & PROBABILITY

for the Australian Curriculum

Years 7&8

Helen MacGillivray & Peter Petocz

Consultants: Michael Evans and Peter Jones

ISBN 978-1-107-61421-5

Cambridge University Press

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/9781107614215

First published 2013

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 978-1-107-61421-5 Paperback

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

CONTENTS

ISBN 978-1-107-61421-5

ISBN 978-1-107-61421-5

ISBN 978-1-107-61421-5

Cambridge University Press

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 data sets. 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 978-1-107-61421-5

Introduction

Statistics and statistical thinking have become increasingly important in our society that relies more and more on information and demands for evidence. Consequently the need to develop statistical skills and thinking across all levels of education has grown. These skills are 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 of chance and skills 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 978-1-107-61421-5

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:

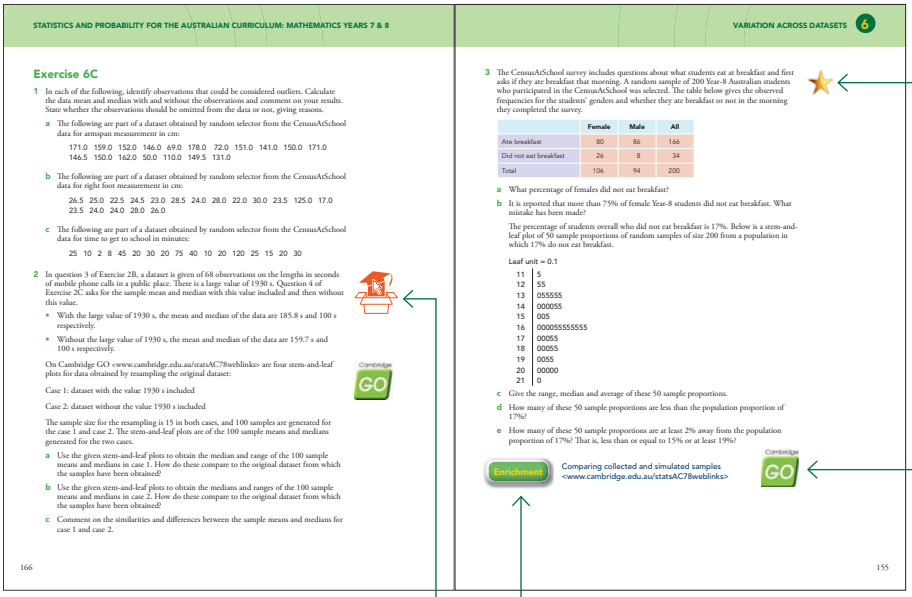
- What you will learn — a list of topics in the chapter
- Chapter introduction: this sets the scene by posing questions that will be addressed in the chapter
- Australian Curriculum linkage for the chapter
- Pre-test — a check of prior knowledge for the chapter
- Terms you will learn — a checklist of 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
- 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 978-1-107-61421-5

Explanation of icons in the textbook:



STATISTICS AND PROBABILITY FOR THE AUSTRALIAN CURRICULUM: MATHEMATICS YEARS 7 & 8

Exercise 6C

1 In each of the following, identify observations that could be considered outliers. Calculate the data mean and median with and without the observations and comment on your results. State whether the observations should be omitted from the data or not, giving reasons.

a The following are part of a dataset obtained by random selector from the CensusAtSchool data for armspan measurement in cm:

171.0 159.0 152.0 146.0 69.0 178.0 72.0 151.0 141.0 150.0 171.0
146.5 150.0 162.0 50.0 110.0 149.5 131.0

b The following are part of a dataset obtained by random selector from the CensusAtSchool data for right foot measurement in cm:

26.5 25.0 22.5 24.5 23.0 28.5 24.0 28.0 22.0 30.0 23.5 125.0 17.0
23.5 24.0 24.0 28.0 26.0

c The following are part of a dataset obtained by random selector from the CensusAtSchool data for time to get to school in minutes:

25 10 2 8 45 20 30 20 75 40 10 20 120 25 15 20 30

2 In question 3 of Exercise 2B, a dataset is given of 68 observations on the lengths in seconds of mobile phone calls in a public place. There is a large value of 1930 s. Question 4 of Exercise 2C asks for the sample mean and median with this value included and then without this value.

- With the large value of 1930 s, the mean and median of the data are 185.8 s and 100 s respectively.
- Without the large value of 1930 s, the mean and median of the data are 159.7 s and 100 s respectively.

On Cambridge GO <www.cambridge.edu.au/statsAC7&weblinks> are four stem-and-leaf plots for data obtained by resampling the original dataset:

Case 1: dataset with the value 1930 s included

Case 2: dataset without the value 1930 s included

The sample size for the resampling is 15 in both cases, and 100 samples are generated for the case 1 and case 2. The stem-and-leaf plots are of the 100 sample means and medians generated for the two cases.

a Use the given stem-and-leaf plots to obtain the median and range of the 100 sample means and medians in case 1. How do these compare to the original dataset from which the samples have been obtained?

b Use the given stem-and-leaf plots to obtain the medians and ranges of the 100 sample means and medians in case 2. How do these compare to the original dataset from which the samples have been obtained?

c Comment on the similarities and differences between the sample means and medians for case 1 and case 2.

166

VARIATION ACROSS DATASETS 6

3 The CensusAtSchool survey includes questions about what students eat at breakfast and first asks if they ate breakfast that morning. A random sample of 200 Year-8 Australian students who participated in the CensusAtSchool was selected. The table below gives the observed frequencies for the students' genders and whether they ate breakfast or not in the morning they completed the survey.

	Female	Male	All
Ate breakfast	80	86	166
Did not eat breakfast	26	8	34
Total	106	94	200

a What percentage of females did not eat breakfast?

b It is reported that more than 75% of female Year-8 students did not eat breakfast. What mistake has been made?

The percentage of students overall who did not eat breakfast is 17%. Below is a stem-and-leaf plot of 50 sample proportions of random samples of size 200 from a population in which 17% do not eat breakfast.

Leaf unit = 0.1

11	5
12	55
13	055555
14	000055
15	005
16	000055555555
17	00055
18	00055
19	0055
20	00000
21	0

c Give the range, median and average of these 50 sample proportions.

d How many of these 50 sample proportions are less than the population proportion of 17%?

e How many of these 50 sample proportions are at least 2% away from the population proportion of 17%? That is, less than or equal to 15% or at least 19%?

Enrichment Comparing collected and simulated samples <www.cambridge.edu.au/statsAC7&weblinks>

155

Gold star icon
Challenge question

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 on the website

Material for students on the website provided with the textbook:

- A digital copy of the textbook with note-taking enabled
- Enrichment activities linked from the text
- 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 978-1-107-61421-5

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 978-1-107-61421-5

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 978-1-107-61421-5

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 978-1-107-61421-5

Acknowledgements

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

Images: ©Shutterstock – 2012 Used under license from Shutterstock.com/Milen Kanev, p.1/ Noppasin, p.3/Kelvin Wong, p.7/Goodluz, p.8 (c-l)/vita khorzhevska, p. 8(c-r)/gururgu, p.8 (b-l)/ rangizzz, p.9/Radu Bercan, p.11/Vadim Ivanov, p.12/ M. Unal Ozmen(b-r) & vlad09 (b-l), p.14/ Planner, p.16/Amy Johansson, p.18/BortN66, p.20, p.80/Dmitry Berkut, p.21/ Michal Kowalski, p.26/Reeed, p.27/Sonja Foos, p.28/ FooTToo, p.34 /Aleksey Troshin, p.37/edhar, p.42/ Protasov AN, p.39/ Maxx-Studio, p.41/ Elena Elisseeva, p.43/ Andrey_Popov, p.47/ Peter Zachar, p.49/ phipatbig, p.49/ Ermolaev Alexander, p.54/ whitelook, p.56/ Barry Blackburn, p.56/ DUSAN ZIDAR, p.57/ Blend Images, p.52/ Pavel L Photo and Videop, 60/ shinobi, p.64/ Sebastian Duda, p.65/ Heike Brauer, p.66/ Jon Le-Bon, p.68/ Evgeny Murtola, p.69/Giraphics, p.71/Patrizia Tilly, p.73/ mrkornflakes, p.74/ Janelle Lugge, p.76/Delita, p.78/PILart, p.78/Dan Kosmayer, p.79/visivastudio, p.81/ Dmitrydesign, p.83/chevanon, p. 86/Volodymyr Baleha, p.88/ Sunny studio-Igor Yaruta, p.89/ Natursports, p.92/Benoit Daoust, p.95/ ejwhite, p.96 (l)/ littleny, p.96(r)/ Ignite Lab, p.99/ Oksana Kuzmina, p.100/voylodyon, p.101/withGod, p.104/Zlatko Guzmic, p.108/ostill, p.112/ pedrosala, p.121/ Sergey Nivens, p.122/CandyBox Images, p.123/Jarp2, p.124/ Wutthichai, p.126/ Elena Schweitzer, p.128/anaken2012, p.129/auremar, p.131, p.152/ luiggi33, p.133/majeczka, p.134/ Maxim Petrichuk, p.135/ supergenijalac, p.136/Kzenon, p.141/Pietus, p.142/Marta Kojadinovic, p.143/Maridav, p.144/gualtiero boffi, p.145/Kozoriz Yuriy, p.146/Denis Kuvaev, p.148/Sandra Gligorijevic, p.154/ Monkey Business Images, p.156/milaphotos, p.158/limpido, p.160/Mopic, p.162/Viorel Sima, p.167/maiwharn, p.170/BestPhotoStudio, p.171/HTU, p.172/Stephen Coburn, p.174/© Istock/pick-uppath, p.29/TokenPhoto, p.111(t); Australian National Maritime Museum Wikimedia Public Domain,p.111(b); https://en.wikipedia.org/wiki/File:Female_scarlet_robin.jpg, © Fir0002/Flagstaffotos/GNU Free Documentation License, Version 1.2, p.23; http://www.conkerstatistics.co.uk, used with permission, p.36; © paddynapper/Creative Commons Attribution-Share Alike 2.0 Generic license, p.62; © Ryan Lawler/Wikimedia Commons, p.60; © Alamy/Farlap, p.31/British Retail Photography, p.58; ©Yohan euan o4/Creative Commons Attribution-Share Alike 3.0 Unported license, p.70; ©Harris Morgan/Creative Commons Attribution-Share Alike 3.0 Unported license, p.72; "base image reproduced with the permission of TASMAP (www.tasmap.tas.gov.au) ©State of Tasmania", p.82; ©DTR/Creative Commons Attribution-Share Alike 3.0 Unported license, p.92; Wikimedia Commons, Public Domain, p.91, p.106, p.138; Brocken Inaglory/Creative Commons Attribution-Share Alike 3.0 Unported license, p.138(l); Source, ABS, licensed under a Creative Commons Attribution 2.5 Australia licence, p. 118; ©original artist/Search ID for0487/ Cartoonstock.com, p.149; © Corbis/Kevin Fleming, p.165.

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 978-1-107-61421-5

Cambridge University Press

CHAPTER 1

Investigating data

What you will learn

How often do people blink?

Blinking stops our eyes from drying out, by spreading tears across our eyeballs. With the help of eyelashes, blinking also removes dust and other small particles that might irritate the eye.

If you search for information on the rate at which people blink, you are likely to find a variety of figures, such as '10 times per minute' or 'approximately 20 to 25 times per minute'. Babies blink less because their eyes are smaller and they spend a lot of time sleeping. Children gradually increase their rate of blinking and teenagers reach the same rates as adults.

What affects our rate of blinking? If our eyes are focused on something, such as reading, our blinking might be slower. In contrast, stress and nervousness or fear may increase blinking. The numbers of blinks by politicians taking part in televised debates were counted and compared, and found to be much higher than 'the average'. The rate of blinking is different under different conditions.

How could we collect data to investigate blinking rates? We would need to decide the conditions under which we are going to observe blinking. Do we need to keep these conditions the same for each of our 'subjects'? If so, how? What questions or issues are we interested in investigating? <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Identify and investigate issues involving numerical data collected from primary and secondary sources **(ACMSP169)**

ISBN 978-1-107-61421-5

Pre-test

1 State whether these examples are categorical data, count data or measurement data.

- **a** The time it takes students to get to school
- **b** How long students can balance a book on their heads
- **c** Colours of cars
- **d** The number of times per minute a speaker says 'um'
- **e** Types of building materials
- **f** Brands of mobile phones
- **g** The number of pets per family
- **h** Distances between potholes along a road

2 Give some possible values for data collected in the following situations. In each case, say whether the data are categorical, count or measurement data.

- **a** Students are asked how many siblings (that is, brothers and sisters) they have.
- **b** People leaving a movie are asked to rate it.
- **c** In an experiment, how long people can hold their breath is recorded.
- **d** During storms, the numbers of lightning flashes are recorded.
- **e** In a sample of preschools, children's ages to the nearest month are recorded.
- **f** A sample of cricket ticket-holders are asked to name their favourite Australian cricketer.

3 A sample of students are asked which animal is their favourite for a pet. Which of the following graphs can be used to display these data and which are not appropriate?

a Column graph **b** Pie chart **c** Dotplot **d** Bar chart

4 A sample of students record the difference in their pulse rates before and after exercise. Which of the following graphs can be used to display these data and which are not appropriate?

a Column graph **b** Pie chart **c** Dotplot **d** Bar chart

5 A sample of students are asked how they come to school (car, bus, walk, train, cycle, other). Also recorded is whether a student is boy or girl. Name two ways these data could be presented in a report.

6 In a question on a survey, people are asked if they prefer a cat or a dog or neither as a pet. The responses are coded as 1 for a cat, 2 for a dog and 3 for neither.

- **a** What type of data are these?
- **b** What graph would you use to display these data?
- **c** What would be displayed in this graph?

Terms you will learn

categorical data
continuous variable
count data
data
experiment
experimental units
grouping
measurement data
observational units
observational study
ordinal variable
pilot study
primary data
quantitative data
random
grouping
measurement data
observational units
observational study
ordinal variable
pilot study
primary data
quantitative data
random
randomise
randomly representative data
recording sheet
secondary data
statistical data investigation process
statistical variable
subjects
survey
variation

ISBN 978-1-107-61421-5

1-1 What is the data investigation process?

How do statisticians investigate problems in the real world? They use the **statistical data investigation process**. This is how people who use statistics investigate questions in science, medicine, agriculture, business, engineering, psychology. It can be used anywhere that **data** need to be collected and where there is **variation**.

Statistical data investigation process: How real problems are tackled by statisticians and investigators conducting experiments, studies or surveys to obtain data for working statistically

Data: Information, facts, records and observations

Variation: The unpredictability of situations in which observations or measurements have different values or are not determined or not specified

Everything in nature involves variation. It happens for many reasons such as:

- people, animals, plants, materials and consumers, etc. are all different
- they don't behave or react in the same way
- conditions or circumstances change and cannot be made exactly the same all the time.

Because of variation, we need data – we need to make observations, to take measurements, to carry out **experiments**, to ask questions. Data helps us to see how much variation there is, and when it tends to be more or less, and what patterns there are in it.

Experiment: Data investigation in which investigators control conditions and measure the effect of these on some outcome(s) of interest

The data investigation process starts with the first thoughts about what questions to ask. It continues through planning, collecting and exploring data, to reporting on its features. There are number of ways of describing and summarising the process. One is 'Problem, Plan, Data, Analysis, Conclusion (PPDAC)'; another is 'Plan, Collect, Process, Discuss (PCPD)'. The process provides a practical framework for tackling real problems statistically. It consists of these steps:

- Initial questions such as 'what do we want to find out?'
- Identifying issues (what will affect the data we collect) and planning
- Collecting, handling and checking data

ISBN 978-1-107-61421-5

- Exploring and interpreting data in context (in the circumstances of collecting it)
- Considering new initial questions for a further investigation.

The last step makes it a cycle, which is a process that continues to a new starting point. This can be represented by a diagram like the following.

In the rest of this section we look at the initial questions, and issues and planning for an investigation.

LET'S START What do we want to find out?

Sometimes a data investigation starts with a specific question, sometimes an idea or belief, sometimes a problem, and sometimes just a general situation to be investigated. Here are some examples:

- Do girls tend to blink more than boys?
- You'll see more ants in the house when rain is due.
- How well can people estimate time? Does reading or other activity make a difference?
- How do cyclists and pedestrians behave on shared paths?

Think of a question, idea, problem or situation you would like to investigate.

Key ideas

An investigation needs to be planned:

- to collect data to investigate questions or issues.
- to help us to explore any interesting or important information that might turn up.

The first step is to decide what is going to be investigated, so we can decide what data to collect.

ISBN 978-1-107-61421-5

Below are some helpful questions to consider in this first step:

- What's of interest?
- What are we going to observe, record or measure?
- Is it possible or practical to collect the data we want?
- What conditions are we going to keep the same?
- Is there anything else we should observe or record in case it's useful?

Example 1: How often do people blink?

a Under the same conditions

Suppose we are interested in how often people blink under the same conditions. Our initial question is 'How often do people blink under the same conditions?'

Moving to issues and planning for this investigation, we need to ask how we collect data on this. Questions include:

- What conditions do we want?
- How can we keep conditions the same for all the people we observe?

We need to be able to observe people in the conditions we want and to count the number of times they blink. Staring at them while they are reading or watching TV might not be a good idea, as some people might notice us, and this would change the conditions. Also different people might notice us at different times, so keeping conditions constant would be very difficult.

We could perhaps count blinks for people being interviewed on TV but they'd need to be interviewed on the same topic!

Perhaps we could arrange it so that someone talks to the person being observed, while another person counts the number of blinks – without being obvious of course. We would need to have the conversation on the same topic, and keep the same people doing the talking and the observing. This could be thought of as mimicking an interview.

b Under different conditions

We may be interested in investigating the rate of blinking under different conditions. For instance we might want to compare the rate when someone is answering easy questions (e.g. what is your favourite food) to the rate when they are answering hard questions (e.g. add these three numbers in your head). This is much more difficult to plan. We would need to do the same type of thinking as in part **a** above for each set of conditions (easy or hard questions). We have the added problem that we would need to use the same subjects – that is, count blinks under each set of conditions for each person. And we would have yet another problem – do we give the same set of conditions first to each person each time? Or do we make the order **random**? We could do this by tossing a fair coin, or by using a spinner divided into two equal parts, to decide whether each person does the easy questions first or the hard questions first.

Random: Due to chance

Let's consider just one set of conditions as described in part **a**, with an 'interviewer' and an observer. What are some of the practical aspects of collecting the data? We need to keep the same person as the 'interviewer' and the same person as the observer –

ISBN 978-1-107-61421-5

otherwise we are changing the conditions. We should keep the 'interview' on the same topics, i.e. each person gets the same set of questions. And of course, the people we pretend to interview should not know what our real purpose is. We also must choose a time interval (for example, a minute or 30 seconds) in which to count the blinks, and make sure we stick to that interval.

What else should we record in case it turns out to be of interest? Certainly we should record whether the person being interviewed is male or female. Should we perhaps record if they wear glasses or not in case this is important? It won't take any extra effort. Perhaps we should also record the time of day, or what each subject has just been doing. But we need to be careful not to make the recording of the data too difficult because that might interfere with the core aim: counting the number of blinks in the chosen time interval under our chosen conditions.

Example 2: How do people clasp their hands?

If you do an internet search on 'clasping hands', you will find a number of sources saying that (almost) everybody tends to clasp their hands the same way each time, with either the left or right thumb on top. Some sources say it is determined by a person's genes (information that you inherited from your parents that helps make you who you are), and that the left thumb on top is 'dominant' (to use a term from genetics). There are also sources saying that it is a myth that it is genetic, or that it is not a simple case of a dominant gene. However, there is general agreement that it is difficult to do it the other way to what you normally do. Try it and see.

One source suggests that it is more common to have the left thumb on top when clasping hands. Let's plan a data investigation to investigate this. <www.cambridge.edu.au/statsAC78weblinks>

In many ways this is an easier investigation to plan than the one about blinking in Example 1 above. This is because we need only ask people to clasp their hands without thinking about it.

Is there anything else we should consider or collect and record? Recording whether the subject is male or female is an obvious one. Another possibility to consider is whether the person is left-handed or right-handed. And this raises the interesting question of how do we define right-handedness and left-handedness. For some people, the right hand is dominant in some tasks and the left is dominant in others. For an investigation in which left-handedness or right-handedness is recorded, the definition chosen must be quite clear and clearly stated. And everyone involved in the data collection must agree and be consistent. One possibility is that the choices are either right-handedness or not, that is, at least one activity being left-handed.

ISBN 978-1-107-61421-5

Exercise 1A

1 The topic for a data investigation is colours of cars. What are some practical issues in collecting data to investigate colours of cars? What else do you think should be recorded about each car?

2 The issue to be investigated is the amount of traffic on a main road near your school. Decide what you are going to observe, and any practical issues in your plans.

3 The question is 'how often does the word 'magic' appear in *Harry Potter* books?' It is decided to investigate how often it appears per page. What are some practical issues in this investigation, and what else could be recorded in case it is useful?

4 The situation to be investigated is how students get to school – that is, what form of transport they use.

- **a** What is a practical issue in this investigation no matter what school, region or country we are considering?
- **b** Consider investigating this in your school. What are some practical issues, and what else would you record?
- **c** Check the CensusAtSchool website to see how this question is asked in their questionnaires. <www.cambridge.edu.au/statsAC78weblinks>

5 The question to be investigated is 'how long can students balance a book on their heads?'

- **a** What are the practical issues relating to the book to be used and the actual balancing activity?
- **b** Name at least two other pieces of data that could be recorded for each attempt at balancing a book.

6 How many shots does it take to win a point in tennis? To investigate this requires a number of decisions to be made in planning the investigation. Give one example of possible decisions for the following:

- **a** What is going to be observed?
- **b** Under what conditions are the data going to be collected?
- **c** What else will be recorded?

How stretchy are jelly snakes?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

1-2 Types of data, variables and subjects

In statistics, it is very important to know what type of data you have. The main types of data are categorical, count and measurement.

Categorical data

In **categorical data** each observation falls into one of a number of distinct groups or categories. Such data are everywhere in everyday life. Some examples are:

- gender (male or female)
- direction on a road (right or left)
- type of dwelling (house, flat, room, caravan and so on)

Sometimes the categories are natural, such as with gender or direction on a road. Sometimes they are categories that we make up and describe such as type of dwelling.

Categorical data: Data that fall into categories that can be named or coded

Count data

Each observation in a set of **count data** is a count value or number. Count data occur in considering situations such as:

- the number of children in a family
- the number of vehicles passing in 2 minutes.

Count data: Record of a number of items, events, people and so on.

ISBN 978-1-107-61421-5

Measurement data

All **measurement data** need units of measurement. Observations are recorded in chosen units of measurement. Some examples of measurement data are:

- reaction time in seconds
- age in years
- weight in kilograms of Year 7 boys.

Count data and measurement data must be described by numbers, so together they are called **quantitative data**.

All measurement data are recorded to the *nearest decimal place* – which could be whole numbers or even tens! This is usually determined by what is possible with the measuring device, or what we choose. Think a bit more about measurement data. When we say that someone is 162 cm tall, we don't mean exactly 162. We mean the height is in between 161.5 and 162.5; perhaps we mean between 161.5 and 162.4 if we are rounding 0.5 up to the next whole number. This type of situation is true for all measurement data which is really reporting data in small intervals. Sometimes these intervals are very small, and sometimes not; it depends on the measuring device or on decisions made by the data collectors/reporters. Sometimes the standard practice used for recording the data is different – for example, ages in years are not usually rounded up – and so the convention used should always be reported.

Measurement data: Data which need units of measurement. Observations are recorded in the desired units of measurement

Quantitative data: Measurement or count data; the numerical values of the data are actual quantities

Statistical variables

When we collect or observe data, the 'what' we are going to observe is called a **statistical variable**. Statistical variables are described by words. When we consider types of data, we are also considering types of variables. So all of the above examples are examples of statistical variables, and the type of data is also the type of variable.

Measurement data and their associated measurement variables are examples of what are known in statistics as **continuous variables** because we observe their data in small intervals. This is important when we consider the graphs appropriate for data on continuous variables, and also how probabilities are estimated for, or assigned to, continuous variables.

Statistical variable: The 'what' we are going to observe when we collect or observe data

Continuous variables: Variables which take values in intervals – typically, continuous data have values given 'to the nearest …'

Subjects

When we collect or observe data, the records or observations are per person, or per time interval or per family or per car or per dwelling and so on. These are the **subjects** (or **observational** or **experimental units**) of the observations or records. In planning most data investigations the type of subject is obvious, but sometimes we can choose. For example, in Example 1, the observations are for two minutes for each person, but we could choose a different interval. If we are investigating advertisements on TV, we could take observations per advertisement or per advertisement break.

Subjects (or **observational** or **experimental units):** Individuals or objects or entities on which observations are made

ISBN 978-1-107-61421-5

LET'S START What are our subjects and variables and their types in our plan?

Consider Example 1 to investigate how often people blink. The plan is to count the number of times each person blinks in two minutes. We are also going to record the person's gender and whether they wear glasses or not. We will keep the 'interviewer' and observer the same. So the subjects are people (we will look more closely at which people in section 1-3).

We have three variables:

- number of blinks in two minutes, which is a count variable
- gender of subject, which is a categorical variable
- whether the subject is wearing glasses or not, which is a categorical variable.

If we ask them what activity they were doing just before the 'interview', we have another categorical variable. This needs some thought. It might be best for such a variable just to ask people what they were doing and decide later how to group and describe the categories.

In the plan of Example 2, what are the variables, what type are they, and what are the subjects?

Key ideas

Identification of the variables, their types and the subjects is very important in planning a data investigation, and is also very helpful in doing the planning. If we don't do this the planning is not complete, and the investigation itself might be messy and confusing.

- In planning, what are we going to observe, record or measure? These are our variables.
- Who or what are we going to collect our data on? These are our subjects.
- For categorical data, each observation falls into one of a number of categories which can be named.
- If the categories of a categorical variable are given numerical codes rather than names, this does not mean the numbers have any numerical meaning. It does not change the fact that the variable is categorical.
- Each observation in a set of count data is a count value. Count data always have a specified entity – such as a location, time or space interval, or group – within or for which the count is made.
- All measurement data need units of measurement. Observations are recorded in the units of measurement we want, and to a selected number of decimal places (or units or tens or hundreds etc.).
- A measurement variable is a statistical continuous variable. Measurement data are observed or recorded in intervals. The size of the interval depends on the measuring device or investigator's choice.
- If there are many possible categories for a categorical variable; investigators need to choose and clearly describe categories. Categories can be grouped together after data are collected, if this is appropriate or useful.
- A continuous or a count variable may be changed to a categorical variable by **grouping** possible values in broad groups or groups of unequal sizes.

Grouping: When **measurement data** or **count data** are placed in specified groups of values, or when some categories of a categorical variable are combined, observations are grouped together

ISBN 978-1-107-61421-5

Example 3: Which are the popular movies?

In a statistical investigation of this question, what are the subjects and what variables and measurements are appropriate?

The subjects are movies. The popularity of movies can be measured in a number of ways. The overall amount of money made by a movie – the gross sales or 'takings' of the movie – is one measure used. Other possible measures are the takings on the opening weekend and the run length (how long the movie runs for). Other aspects of interest in investigating this topic could be the type of movie, the rating (by critics) and the budget if this is known. Much of this information is available from the internet.

When collecting data from the internet, we need to know exactly what is meant by each of these variables, and where the data were collected.

The variables, their units if appropriate, and their types, could be:

- **Gross takings of each movie.** The units are usually US dollars. It is a measurement variable.
- **Takings of the movie on its opening weekend.** Again, the units are US dollars, and is a measurement variable.
- **Run length.** This is most likely to be in days, but might be in weeks. It is a measurement variable. This needs care in identifying – is it overall run length or maximum run length at one cinema?
- **Type of movie (comedy, action, horror, and so on).** This is a categorical variable. The categories need careful identification and description. Some of them might be combined when it comes to exploring the data.
- **Rating.** This is a categorical variable, but the order of the categories matters. Ratings are usually given numerically (for example 3 stars, 3.5 stars) but the numbers give an ordering of ratings from lowest to highest. This type of categorical variable is called an **ordinal variable** and occurs frequently in surveys collecting opinions. Note that the ratings do not measure some kind of distance; the 'distance' between a 3 and a 3.5 rated movie is not meant to be the same 'distance' as between a 3.5 and a 4 rated movie. Again this variable requires care in identifying – is it average rating across a number of critics, or the rating of a certain critic or critics' website?
- **Budget.** This is not usually as readily available as other data, but is a measurement variable, with units usually in US dollars.

Ordinal variable: Categorical variable for which the order of categories has meaning

ISBN 978-1-107-61421-5

Exercise 1B

1 A questionnaire for an investigation of households asks each one:

- What is the total number of people currently living there? (That is, it is their main place of residence.)
- What is the number of people in the household under 18 years of age?
- What is the gender and age in years of the oldest resident?

Identify the subjects, variables and their types in this investigation.

2 In each of questions 1 to 3 of Exercise 1A, identify the subjects and variables you chose in your planning. Identify the types of variables, giving (possible) categories for any categorical variables.

3 The distance between fingertips when arms are spread wide is called a span. Children's spans are approximately the same as their heights when they are small. Their spans gradually increase to more than their heights as they grow. Data are collected on spans (in cm) and heights (in cm) for boys and girls aged 4 to 18 years. Identify the subjects and variables. Identify the types of variables, giving (possible) categories for any categorical variables.

4 A statistical investigation of non-fiction books is carried out. Data are collected on their:

- price in dollars
- total number of pages
- topic
- whether their covers are hard or soft
- whether they have colour pictures or not.

Identify the subjects and variables. Identify the types of variables, giving (possible) categories for any categorical variables.

5 Coffee prices are being collected across a city. The prices of cappuccinos and flat whites are collected at cafes and restaurants in the centre of the city and in the suburbs. Identify the subjects and variables. Identify the types of variables, giving (possible) categories for any categorical variables.

Enrichment

How long can you hold your breath?
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

1-3 Collecting data

Recording sheets

To collect data, we need a **recording sheet**. Whether we write it by hand or use a computer spreadsheet, the design should be the same. So for Example 1 on investigating how often people blink, each subject has a row and each variable has a column in a recording sheet or spreadsheet. The first few lines of our recording sheet might look like the following:

Recording sheet: A table or spreadsheet to record data; each variable has a column and each subject (or experimental or observational unit) has a row

Subject	Gender	Glasses	Previous activity	Number of blinks per minute
Tom	Boy	No	Reading	14
Elise	Girl	No	On computer	13
Sophie	Girl	Yes	Watching TV	24

Note that the only reason for recording names is to be able to check data for errors, and to make sure that no subjects are recorded twice.

Even if you think you have identified your variables, subjects and the categories of any categorical variables, designing your recording sheet and putting in a few test entries will help to see if anything has been forgotten.

ISBN 978-1-107-61421-5

Pilot study

A **pilot study**, experiment or survey, is a trial run to check the plans for an investigation. These are the kinds of things that can be checked:

- The method of collecting the data will work
- The design of the recording sheet works
- The data can be collected under the planned conditions
- In an experiment, whether conditions need to be changed or better controlled
- In a **survey**, whether questions or situations need to be more carefully or clearly expressed or described
- In an **observational study**, when observations sometimes have to be made quickly, whether more help is needed
- Whether the plans need rethinking.

Pilot study: An initial trial of the investigation or preliminary experiment to check the practicalities of the planned collection

Survey: Asking questions of subjects with the data being responses; the subjects of a survey may be individuals, or groups such as companies, businesses or households

Observational study: Data investigations in which investigators observe subjects without altering or controlling conditions

Randomly representative data: Data obtained at random from a more general situation or population

Randomly representative data

Why are we collecting data? This question is at the heart of all statistical data investigations. The answer is because we want to find out information about a question or topic(s) or issue(s). That is, we want to be able to use our data to comment on a situation represented by our data.

Suppose our question is 'which sport is most popular among students in Year 7?'. We don't just ask students in one school. In another school or state or region the situation might be very different. If we want to ask 'which sport is most popular among students in our school?' and we don't have time to ask them all, we have to randomly choose students to ask. We won't get data representative of the whole school if we ask only the friends we play our favourite sport with. When we randomly choose subjects to collect the data, we can say that our data are **randomly representative** of the situation.

Suppose you want to test if fertiliser makes tomato plants grow faster. You divide your plants into two groups: group A gets fertiliser, group B gets none, and all other conditions are kept the same. You have to randomly allocate (give) test plants to each group. Your data would not be randomly representative if you put all the strong and healthy plants in group A and the smaller weaker ones in group B.

ISBN 978-1-107-61421-5

Cambridge University Press

You will see more of these important concepts as you learn more about statistics, but:

- when you collect data, try to collect it randomly within the conditions you have chosen <www.cambridge.edu.au/statsAC78weblinks>
- when you are exploring data or commenting on data, ask if the data can be assumed to be randomly representative with respect to the question(s) or issue(s) of interest.

Primary and secondary data

The data about movies referred to in Example 3 have been collected by other people or organisations and reported on the internet. This is an example of **secondary data**. Data we collect ourselves are called **primary data**. Notice that there are comments in Example 3 about finding out how the data are defined and/or collected. Whether we collect data ourselves, or use data collected by others, the same principles apply: the variables should be clearly identified, and exactly how, when and where the data are collected should be clear and fully reported.

Secondary data: Data collected by others

Primary data: Data collected by the investigators

LET'S START Clasping hands and folding arms

Consider Example 2 about investigating which thumb is on top in clasping hands. We've already considered that gender should also be recorded, and right-handedness or not might be of interest. People tend to fold their arms the same way each time – perhaps this could be related to handclasping. So we might also ask people to fold their arms and record which arm is on top. Now design the recording sheet.

Who are we going to ask? Because the way we clasp our hands is a characteristic that is particular to us, it does not tend to change, or to change with age. So trying to collect randomly representative data is not too difficult for this investigation, and any group of people can be assumed to be randomly representative. However, if we include folding arms, there is another consideration – what do we ask people to do first? Either we do the same for everyone or we **randomise** the order. What should we now include on our recording sheet? These practical questions are sometimes referred to as the practicalities of the investigation.

Randomise: To make random

Once the plan is done, including the design of the recording sheet and how subjects will be chosen as well as considering practical issues, carry out a pilot study. Does this tell us if there is anything that should be changed before carrying out the data collection?

The question of how many observations to collect is a big question in statistics. It takes very many observations to accurately estimate something like the proportion of people who have their left thumb on top in clasping their hands. This is why there is so much variation in the studies that have been done. <www.cambridge.edu.au/statsAC78weblinks>

Key ideas

This is a summary of things that you need to do and questions you may need to ask in planning an investigation.

- Choose the topic or issues to be investigated.
- Identify and describe the variables to be observed or recorded.

ISBN 978-1-107-61421-5

- Design the recording sheet.
- Work out how to choose subjects to be randomly representative of the situation in which we are interested.
- Consider the practicalities of collecting the data and decide exactly how to collect the data you want.
- Decide if there anything else you should observe or record in case it's useful.
- Modify the recording sheet if necessary.
- Have we forgotten anything?
- Should we do a pilot study? (The answer is yes, almost certainly!)
- What has our pilot study told us?
- Changes or problems suggested by the pilot study can then be allowed for, and the data collection plan carried out.

Example 4: How quickly does a vitamin tablet dissolve?

Some vitamin tablets are soluble. They dissolve in water, with fizzing and bubbling. A group of students decided to investigate how quickly they dissolve. They chose two brands of soluble vitamin C tablets having the same amount of vitamin C.

They decided to use three types of water (tap, filtered and mineral water). The tap water all came from the same tap, the filtered water came from the same source and the mineral water was all the same brand.

They decided to do the experiment with room-temperature water and cool water so they divided each type of water into two batches. One batch was kept in a room at 24°C and the other batch was stored in a cool place at 15°C. These were the experimental conditions.

The students decided to dissolve four tablets of each brand in each combination of experimental conditions. This meant altogether in their experiment they dissolved:

4 tablets × 2 brands × 3 types of water × 2 temperatures = 48 tablets

Because they were measuring the time the tablet took to dissolve, the students needed to decide how they would define when the tablets were completely dissolved. They chose to say it was dissolved when they could no longer see any solid bits of tablet. They used the same stopwatch, the same quantity of water each time, and the same person recorded the times in seconds.

Their subjects were tablets, and their variables and variable types were:

- brand of tablet – categorical
- temperature – measurement but because only two selected values, can be treated as categorical
- water type – categorical
- time to dissolve in seconds – measurement.

ISBN 978-1-107-61421-5

Their recording sheet looked like this (only the first few rows are shown):

Tablet	Brand	Water temperature	Water type	Time to dissolve (in seconds)
1				
2				
3				

After conducting their pilot study, they decided that the same two people should declare the tablet dissolved.

They chose the test tablets at random from the packets they'd bought, and they tested the combinations in random order. Even though they were careful in keeping other conditions constant, randomising the order ensures there was nothing else unintentionally introduced into the experiment. They did this by tossing a six-sided die. To choose which brand of tablet for the first test, they tossed the die, with an even number meaning brand A, and an odd number meaning brand B. The same rule is used to decide the other variable with two categories – temperature. For water type with three categories, throwing a 1 or a 2 means tap water, a 3 or a 4 means filtered water and a 5 or a 6 means mineral water. That means that of the 48 tablets in the investigation there were:

- 24 of each brand (48 tablets ÷ 2 brands)
- 24 tested at each temperature (48 tablets ÷ 2 temperatures)
- 16 tested in each water type (48 tablets ÷ 3 types of water).

For full details of ways of randomising the order of experiments, see Cambridge GO.

Exercise 1C

1 Data were collected at a bus station to investigate people's use of the lifts. The bus platforms can be accessed only from an overbridge, which can be reached by stairs or a lift for either direction. During the morning and evening peak times on two weekdays (Tuesday and Thursday), the choice of lift or stairs, the direction (up or down), and gender (male, female) of all passengers were recorded. At the end all the observations from all the recording sheets were combined in an overall spreadsheet.

a Give the column headings of this spreadsheet.

b Suggest at least one problem that may have shown up in a pilot study.

c What other information would you like to know about the data collection?

d Can we reasonably assume that the data is randomly representative of the situation being investigated? What further information is needed to be able to judge this?

ISBN 978-1-107-61421-5

2 In a study of usage of a new pedestrian and cycle bridge, the numbers of pedestrians and cyclists travelling in each direction were recorded for morning and afternoon periods of 15 minutes each, on each day of a week.

- **a** Describe the recording sheet, including the subjects and the variables.
- **b** Suggest any further information that should be provided about the data collection.
- **c** Suggest any practical problems.

3 During the study described in question 2, in the same time periods, a second data collection was made. Pedestrians and cyclists were selected at random, and their direction of travel, their speed and gender were recorded. Perform the steps **a**, **b**, and **c** described in question 2 on this second data collection.

4 In an experiment to investigate whether colour makes a difference to how long candles burn, candles of four different colours from one brand were chosen at random from a packet. Their lengths and diameters were measured. They were then placed upright in the same environment with no wind, lit and allowed to burn for 5 minutes before being extinguished. Their lengths were measured again.

- **a** What would the recording sheet look like?
- **b** What else could the experimenters do in carrying out their experiment to make sure no other effects (such as human factors or variables) were unintentionally included in the experimental conditions?

5 Your school wants to find out what families want for a school newsletter, including:

- whether they want one weekly or fortnightly
- whether they want to have the option of receiving it online as well as in print
- if they want each child to receive a copy or just the eldest child.

- **a** Write the questions for this survey, giving the choices of responses.
- **b** How do you suggest the survey be carried out?
- **c** Is there another question you'd like to include?

Enrichment

The ABS survey on Children's Participation in Cultural and Leisure Activities
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

Chapter summary

What is the data investigation process?

- Initial questions and issues
- Planning, designing
- Collecting, handling and checking data
- Exploring and interpreting data in context
- Considering new questions and issues to be investigated.

What's of interest?

- What to observe, record or measure?
- Can we collect the data we want?
- What conditions to keep consistent?
- Anything else we should observe or record?

Types of data and variables

- Categorical data
  - Observations fall into distinct named categories
  - Categories may be combined if appropriate
  - Numerical codes are just labels.
- Count data
  - Observations take count values
- Measurement data
  - Observations have units of measurement and a number of decimal places
  - Measurement variables are examples of continuous variables.

Subjects and observational or experimental units

- The people, objects or items on which data is collected
- Chosen or allocated as randomly as possible.

Recording sheet

- Each variable has a column
- Each subject or observational or experimental unit has a row.

Pilot study

- Trial collection plans
- Adjust plans if necessary.

Primary and secondary data

- Primary data are collected by investigators
- Secondary data are collected by others
- For both, must know or report *what*, *how*, *when*, *where*.

Multiple-choice questions

1 To investigate people's views on daylight saving, which of the following is **not** likely to be of interest in a survey?

A Their gender **B** Where they live
C What they had for breakfast **D** Their age

2 To investigate whether people jaywalk, which of the following is **not** likely to be of interest?

A Their gender **B** What pets they have
C Their occupation **D** Their age

ISBN 978-1-107-61421-5

3 To investigate a new plant species for growth, which of the following is **not** part of the data to be collected?

A The soil **B** Where it comes from

C Who is taking care of it **D** Its name

4 What type of variable is the number of goals scored in a football match?

A Categorical

B Count

C Measurement/continuous

5 What type of variable is the variety of a tree?

A Categorical

B Count

C Measurement/continuous

6 What type of variable is the age in years of a tree?

A Categorical **B** Count

C Measurement/continuous

7 In a poll, people are asked to choose from the following for their approval rating of a new mayor: 1 = strongly disapprove, 2 = disapprove, 3 = neutral, 4 = approve, 5 = strongly approve. What type of variable is the response to this question?

A Categorical **B** Count

C Measurement/continuous

8 What type of variable is the price per kilogram of apples?

A Categorical **B** Count

C Measurement/continuous

9 What type of variable is the number of accidents per week in a state?

A Categorical **B** Count

C Measurement/continuous

10 In collecting real-estate data for properties sold, its selling price, suburb, land size, number of bedrooms, bathrooms and car spaces are recorded, and whether it has a pool or not. What are the subjects of these data?

A Properties **B** Sales

C Suburbs **D** Buyers

Short-answer questions

1 In collecting daily data on the weather at a chosen location, three of the variables recorded are maximum temperature, wind direction and amount of rainfall.

a Give two other variables that may be of interest.

b State the types of all five variables.

2 The colours of some small sweets are often of interest because people like different colours.

a Below are two possible questions of interest. How would you suggest collecting data for each of these?

ISBN 978-1-107-61421-5

i The topic of interest is the allocation of colours by the manufacturer – that is, in what proportions are the different colours made?

ii The topic of interest is the number of blue sweets in the smallest packets of sweets produced by the manufacturer – that is, how many blue sweets are in each of the smallest packets of sweets?

b What are the subjects in each of the above?

3 Data on the usage of computer games by Year 7 students are to be obtained. It is decided to ask approximately how long (in hours) a student spends playing computer games in a week and the gender of the student.

a Give at least two other variables that may be of interest.

b State the types of all variables.

c Suggest how to choose a sample of students to be asked to obtain a reasonably randomly representative group.

4 Some music fans think that songs of some styles tend to be longer than those of other styles. To investigate this, they look at the top 100 songs on a particular chart, and collect the length of each song (in minutes and seconds), the genre of the song, and the nationality of the performer(s) and whether the song is by a solo artist or a band.

a State the types of all variables.

b Suggest no more than eight names for classifying song genre. Suggest groupings to change that to five.

c Suggest no more than four names for classifying nationality.

d Suggest a problem in using the top 100 songs on a particular chart.

5 Students decide to do an experiment to investigate the performance of three different paper plane designs. They decide to use two different types of paper for each design. Two students each made four planes for each design and each paper. These groups of four planes were divided into two, and thrown by two throwers in a large indoor space. The students recorded the flight time (in seconds), the distance travelled by the plane (in cm) and whether it landed upright or not.

a What was the total number of throws?

b Name the variables and state their types.

ISBN 978-1-107-61421-5

c As you can see, the students have designed this experiment very carefully and thoughtfully. What else should they do in carrying out the experiment to avoid introducing any unwanted effects?

d Show what the data recording sheet would look like by giving the headings and a few possible rows of data.

6 Real-estate data are collected on houses sold. For each house sold, its selling price, region, land size, numbers of bedrooms, bathrooms and car spaces are recorded, and whether it has a pool or not.

a State the types of all variables.

b It is decided to record numbers of bedrooms as <3, 3, 4, 5, >5, and bathrooms as 1, 2, 3, >3. How does this change your answers to part **a**?

c Show what the data recording sheet would look like (for part **a**) by giving the headings and a few possible rows of data.

Extended-response questions

1 The topic for a data investigation is family pets. Below are two possible main aspects to investigate. For each of these, identify practical issues and how you would handle them, and anything else you would choose to record.

a The number of pets a family has

b Whether they have cats and dogs as pets

2 Reactions can be measured in many different ways. A classic way is catching a ruler that is dropped. See also the CensusAtSchool questionnaire(s) to see how reactions are measured. This question in CensusAtSchool is an international common question that may be included in the CensusAtSchool questionnaires of the United Kingdom, Canada and New Zealand. <www.cambridge.edu.au/statsAC78weblinks>

a It is decided to measure reactions by catching a ruler that is dropped, and measuring the distance on the ruler where it is caught. List at least two practical considerations and at least two other pieces of data that should be recorded. Choose and name your variables, state their types, and show what the data recording sheet would look like by giving the headings and a few rows of possible data.

b Cambridge GO gives two internet games to test reaction time. Write at least two instructions for players to help provide consistent conditions for investigating people's reaction times. <www.cambridge.edu.au/statsAC78weblinks>

3 Find a news story that reports information from a survey.

a Does the report say how many observations were collected?

b Does the report say exactly how the data were collected?

c Does the report give all the questions that were asked?

d Do you think the report is justified in its statements?

e Do you think the report gives sufficient justification for its statements?

ISBN 978-1-107-61421-5

CHAPTER

2

Exploring quantitative data

What you will learn

How far do birds fly?

Birds are often studied by capturing, banding and releasing them so that their movements can be followed. One variable we can measure is the flight distance from where the bird is released after it is banded to where it first perches after it is released. Below are lists of these distance data obtained on two types of birds – a species of robin (a member of the flycatcher species in Australia) and a species of dove. The distances are in metres.

Robin

128.8	160.0	192.1	163.4	186.4	156.2	70.0	10.0
57.2	65.2	68.9	24.7	37.4	99.7	265.0	78.7
48.2	69.2	117.3	36.5	140.8	59.3	71.3	105.3

Dove

40.0	80.0	313.9	175.7	55.5	44.7	166.7	83.4
381.7	266.8	162.7	76.0	22.1	170.0	263.7	369.7
613.2	189.5	358.9	13.9	165.5	317.2	30.6	197.7
288.1	102.0						

What can we say about how far these two types of birds fly when they are first released? What can we see from just lists of numbers? How can we present these data and describe the information in the data?

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Construct and compare a range of data displays including stem-and-leaf plots and dotplots **(ACMSP170)**
- Calculate mean, median, mode and range for sets of data. Interpret these statistics in the context of data **(ACMSP171)**
- Describe and interpret data displays using median, mean and range. **(ACMSP172)**

ISBN 978-1-107-61421-5

Pre-test

1 Consider the example above on measuring flight distance after release.

- **a** Name the variable(s) and state whether they are categorical, count or continuous.
- **b** Name at least two pieces of information that we need to know about how the data were collected.
- **c** How would you improve the way the data are reported (but still keep them in a list)?

2 An investigation was conducted into whether petrol prices are affected by how close the petrol station is to the nearest town or city centre. On each day of the week for four weeks, investigators obtained the prices of two different brands of unleaded 91 and premium 98 petrol at outlets across a region. They also recorded the distance in kilometres and the direction in degrees from north of each outlet from the town or city centre.

- **a** What are the variables in this investigation?
- **b** For each variable, state whether it is categorical, count or continuous.
- **c** On the recording sheet or spreadsheet, what would the rows correspond to?
- **d** What practical aspects of obtaining these data do we need to consider and report on?

3 A pizza shop operates 7 days a week and the owner keeps records each day. These records include the numbers of three different types of pizza sold (Hawaiian, supreme and seafood), the daily takings for pizzas sold and the daily takings for drinks sold. The owner also records the daily maximum temperature for the region. She chose a three-week period for investigation of the data.

- **a** What are the variables in this investigation?
- **b** For each variable, state whether it is categorical, count or continuous.
- **c** On the recording sheet or spreadsheet, what would the rows correspond to?
- **d** Why do you think the place also recorded the daily maximum temperature?

4 State whether each of the following variables is categorical, count or continuous. If it is categorical, state whether it is an ordinal variable.

- **a** month of birth
  height
  weight
  year of school
  length of forearm
  position in family
  number of siblings
  left-handed (yes/no)
  wear glasses (yes/no)
  hours per week in part-time job
  accident proneness on a scale of 1–5
- **b** A student said that age is a count variable because you count the years. Why is this incorrect and what type of variable is age in years?
- **c** Another student collected times in minutes and described his variable as a discrete variable. Why is this incorrect, and why do you think he made this mistake?

Terms you will learn

average
centre
context of the data
dataset
dotplot
features of data
leaf (of a stem-and-leaf plot)
mean
median
mode
range
raw data
respondent
stem (of a stem-and-leaf plot)
stem-and-leaf plot
summary statistics
variability

ISBN 978-1-107-61421-5

2-1 Collecting and handling measurement data

In considering the statistical data investigation process, Chapter 1 gave examples of how to start an investigation, and how planning a data collection involves thought and work. Whether we are collecting our own data, or using secondary data, or reading a report of an investigation, we need always to remember what the whole statistical investigation process is. You may prefer to remember the diagram used in Chapter 1, or one of the abbreviations, either PCPD or PPDAC, or to refer to a description such as:

- considering initial questions that motivate an investigation
- identifying issues and planning
- collecting, handling and checking data
- exploring and interpreting data in context
- considering new questions and issues to be investigated.

Remember that this whole process is behind all statistical data investigations.

We have seen that planning for categorical data can involve decisions about categories, names or codes. Planning for count data involves choosing the subjects for the counts. These could be periods of time or space, or groups such as families.

As we have seen in examples and exercises in Chapter 1, measurement data can require a lot of thought and care. Exactly what and how is to be measured must be clearly identified so that the data collected are consistent – that is, the data are observations of what we say they are. In a report, or for secondary data, these also must be clear. The instrument for measuring and the method must also be clearly described and stuck to. For example, will the same person be doing the measuring? The units and accuracy of the measurements need to be chosen and quoted in any report. For example, are lengths in centimetres going to be measured correct to the nearest centimetre or to the nearest 0.1 centimetre?

LET'S START Deciding what to measure and how to measure it

The ABS survey mentioned in Chapter 1 on how children spend their spare time reports that the proportion of children (5–14 years) accessing the internet, either during school or outside of school hours, has increased from 64% in 2003 to 90% in 2012. But the survey did not ask for the amount of time spent using the internet. If you would like to find out how much time people spend on the internet, you need to decide what to measure and how to measure it.

A key choice is the period of time you ask about. You could ask **respondents** to give the amount of time they spent on the internet in the past two weeks. The ABS survey asked children how much time they spent watching TV in the past two weeks, so this would make comparison with their survey easier. Or you may decide to ask about the previous 24 hours. Or you could ask for an estimate, as in the CensusAtSchools questionnaire, which asks 'Estimate how many hours a week you usually spend doing these activities'.

Respondent: The individual answering the questions – the subject

ISBN 978-1-107-61421-5

You also need to decide the accuracy of the measurement of time spent on the internet. You could ask respondents to give the time to the nearest hour or the nearest 30 or 15 minutes. Do you want to ask about the time outside, or including, school hours? Are you going to survey through a questionnaire on paper or by asking respondents personally? If you are working in groups and asking respondents personally, the same person should do the asking, or it must be very clear exactly how they ask the questions. Ideally each questioner should be recorded.

You may decide to ask other questions such as gender (categorical), number of computers in the household (count), number of brothers and sisters, or siblings, (count). And even if you decide to keep the respondents' details private and report the data without disclosing any personal details, you still need to record either names or something to identify each person so you can check your data.

In some surveys this type of question sometimes provides intervals and asks respondents to choose one – for example, less than 2 hours, 2–5 hours, 5–10 hours, 10–15 hours, more than 15 hours. When we do this we are turning a continuous variable into a categorical variable. It might seem easier, but a lot of information may be lost. It is often no more difficult for a respondent to give a time to the nearest hour or half-hour.

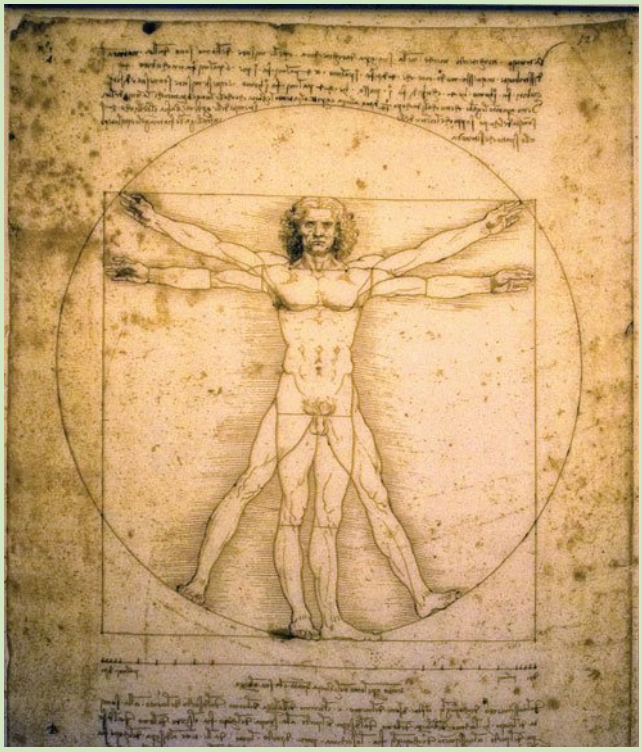
Key ideas

In planning to collect measurement data, or in using and interpreting measurement data collected by others:

- be very clear as to how the variable is defined and the details of exactly how the measurements are made
- make sure the units of measurement are clear, and the accuracy of the measurements (for example 'to the nearest…') are defined
- keep names or identifiers so that data can be checked, even if these are discarded later
- do not turn a continuous variable into a categorical variable without very good reasons (such as making questions less personal or much easier to answer).

ISBN 978-1-107-61421-5

Example 1: Vitruvian man, by Leonardo da Vinci



This is a famous drawing by Leonardo da Vinci and is named in honour of the ancient Roman architect Vitruvius who said that the proportions of the human figure (the ratio of certain body measurements, such as armspan, to other body measurements, such as height) should be used in architecture to design the proportions of buildings.

As shown by Vitruvian Man, it was thought that the armspan is equal to the height. However, modern research shows that armspan and height are approximately equal in children, but as you grow, your armspan becomes longer than your height, especially if you are male. Such measurements can be used to get the proportions of sleeve length and body length right. Sources for internet research on the topic are given on Cambridge GO. <www.cambridge.edu.au/statsAC78weblinks>

Working in groups, consider how to collect and check data on spans and heights. Write down what to measure, how accurately to measure it and what the practicalities are, what other information should be recorded. The following are suggestions.

- Use two people each holding an end of the tape measure to make the measurement, and a third person to record the measurement.
- Use the same tape measure for all the measurements.
- Measure the span and height for each subject.
- Record each person's age as well as gender, since internet research suggested that these may affect the data.
- Decide on an age range or do the measurements for just one age group.

ISBN 978-1-107-61421-5

- Choose a measurement accuracy. Measuring in millimetres can be difficult, so perhaps to the nearest 0.5 centimetre might be better. If we can measure to the nearest 0.1 centimetre (one decimal place), then our data will take values like 124.3, 126.0, 125.1; if we measure to the nearest 0.5 centimetre, these data will take values such as 124.5, 126.0, 125.0. Note that writing values such as 125 or 125.00 is not correct.
- Practicalities may include the following. For heights, people remove their shoes and stand with their back against the wall, with heels on the ground and firmly against the wall. Care must be taken with obtaining an accurate height to the top of the head. Similarly, for span, arms and hands must be fully extended and again care taken to measure from fingertip to fingertip.
- A trial should be carried out, especially if different people are performing the measuring.

Exercise 2A

1 In the introductory example of this chapter on the first flight distances of two species of birds, what *exactly* would have been measured?

2 In an experiment on how well people perceive time under various conditions, subjects were asked to estimate time periods such as 10 or 20 seconds. The time was measured from when the experimenter said 'Go!' until the subject said 'Stop!' when they thought the time period was ended.

a What would the experimenter need to be very careful to do?

b The estimates of time are recorded to the nearest 0.2 second. Give a few examples of possible observations.

c In part of the records, the following estimates for 10 seconds appeared: 9.60, 9.2, 10.4, 10.8, 98, 11, 1.24. What would you suggest doing with these observations?

3 Consider collecting data in your class on spans and heights as in Example 1 above.

a The following is how one person taking measurements started to record data:

Height: 154.2, 148.3, 145.5

Span: 155.6, 148.1, 145.8

What is wrong with recording data this way?

ISBN 978-1-107-61421-5

b Another person taking measurements started to record data as below:

Gender	Height	Span
Male	149.6	150.2
Female	152.5	152.8
Female	156.8	175.3

i What should be included in the recording sheet in case of problems?

ii What problem is there in the above data that needs the answer to part **i** would fix?

4 Seedlings are being checked each day to monitor their growth.

a What measurements do you suggest should be taken each day?

b One person started recording heights of seedlings in metres. Could this cause problems? If so, what?

c What units do you suggest should be used in recording the data?

d What practical issues are there, and what care do you suggest should be taken?

5 In an experiment to investigate the strength of different colours of a particular brand of balloon, each balloon was gradually filled with water while being held inside a large bucket. In this way, when the balloon burst, the water was collected in the bucket and measured. The amount of water therefore represented the strength of the balloon.

a What are some practical issues in obtaining good data from this experiment?

b The bucket has litres marked on its sides. Do you think this is sufficient for measuring the water? What units would you suggest using?

Enrichment

Measuring vision
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

2-2 Dotplots and stem-and-leaf plots

Dotplots

We first revisit dotplots, emphasising their use for measurement data. **Dotplots** can be used to present measurement data (data from a continuous variable) or count data. For count data dotplots are more useful when there are many different counts, for example, attendance at sporting matches. When there are only a few different values of count data, a column graph or bar chart may give a more useful picture.

In a dotplot, each dot represents a value in the set of data (called the **dataset**). The dots are placed above their value, which is read off from the horizontal axis.

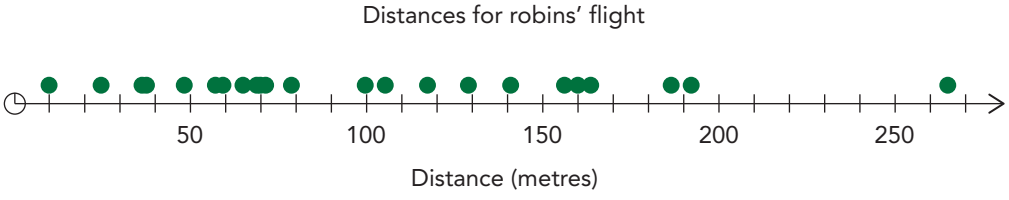
Recall the study of how far robins flew after release. Here are the data in metres, sorted from lowest to highest value:

10.0	24.7	36.5	37.4	48.2	57.2	59.3	65.2	68.9	69.2	70.0	71.3
78.7	99.7	105.3	117.3	128.8	140.8	156.2	160	163.4	186.4	192.1	265.0

Dotplots: A type of graph in which each dot corresponds to a given number of observations, a dot on the vertical scale represents an occurrence of a value in the data; the value is placed on the horizontal axis … *see glossary*

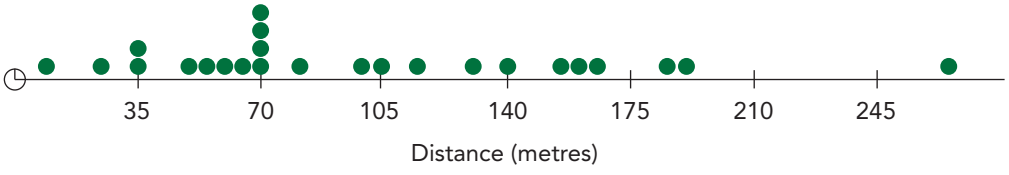
Dataset: The set of data

Here is a dotplot of this dataset:



Some of the dots are very close together and overlap each other, but no two values are the same.

Here is another dotplot of the same data. What is different about it?



It is also using dots, with one dot for each observation, but the data are rounded to the nearest 5 metres. So the values of 36.5 and 37.4 are rounded to be 35, the value of 48.2 is rounded to 50, and the values 68.9, 69.2, 70.0, 71.3 are all rounded to 70. This is the altered dataset:

10	25	35	35	50	55	60	65	70	70	70	70
80	100	105	115	130	140	160	160	165	190	190	265

Some values occur more than once. When this happens, dots are stacked in a column. The number of dots in a column of any dotplot tells us how many times that value occurred in the dataset.

There is now also space between the dots and they can be seen more easily.

ISBN 978-1-107-61421-5

Although we can see in the first plot that there are more observations in some regions than in others, the second plot gives us a better idea of the types and variation of values in the data.

This is a small dataset. But even large sets of measurement data can have no, or few, repeated values. Imagine what a dotplot of the first type above could look like for a dataset of more than 100 observations if they were all different values. It would be difficult to see how the data are behaving. Whether the data are rounded at all, or how much rounding is done, depends on the data. The idea of a dotplot is to be able to see the dots but also to represent the **raw data** in as much detail as possible.

Raw data: Original data, the data as recorded

The effect of rounding the data in the second dotplot is to put some of it into groups. The small amount of grouping of the data gives us a better picture of the data. We can quickly see that quite a number of the robins (10 out of 24) flew distances of approximately 35 to 70 metres; only two flew less than 35 metres, and all but one of the rest (11 out of 24) flew distances greater than 70 metres up to about 200 metres. One robin flew much further than the rest – more than 250 metres.

CAUTION
Grouping data to obtain a better picture does not mean we don't keep the original data! We *always* keep and use the original data for calculations and analysis.

Measurement data, and, more generally, data from a continuous variable, are almost always better displayed graphically using grouping because we get a better idea of where the data tend to be concentrated and where they tend to be more spread out. As data from a continuous variable are always recorded in intervals anyway – because we record to the nearest 'something' value – grouping is just using bigger intervals to give us a smoother and more informative picture of the data. Grouping measurement data provides a better picture of the data, but the more we group to obtain a smoother picture, the more information – the detail of the observations – we lose.

Stem-and-leaf plots

Stem-and-leaf plot: A plot for quantitative data that groups observations into intervals of equal lengths

One type of graph that groups the data to give a good picture but which keeps a lot of information is a **stem-and-leaf plot**. This graph groups the observations, but instead of using dots (or other symbols) to represent observations, the numbers themselves are used. The easiest way to see what a stem-and-leaf plot is, is to draw one.

ISBN 978-1-107-61421-5

LET'S START Drawing a stem-and-leaf plot

We will use the robins' first flight distance data again. We use the dataset as originally given to illustrate the process.

Robin

128.8	160.0	192.1	163.4	186.4	156.2	70.0	10.0	57.2	65.2	68.9	24.7
37.4	99.7	265.0	78.7	48.2	69.2	117.3	36.5	140.8	59.3	71.3	105.3

The smallest observation is 10.0 m and the largest is 265.0 m. We need to decide how many groups we want – how to divide up the data. Let's have the intervals for our groups of length 50 m each. So our intervals are 0–50, 50–100, 100–150, 150–200, 200–250, 250–300. Each interval goes up to, but doesn't include, the right-hand end. That is, an observation of 50.0 is in the second group, 50–100. Note that our first two intervals are actually 000–050, 050–100. Our largest place value in these data is the hundreds, so we start by writing down the digits of the hundreds of our grouping intervals in a column. This column is called the **stem**.

Stem (of a stem-and-leaf plot): Contains the digits in the highest place value of the observations

Stem	
0	← First group 000–050: digit of the hundreds is 0
0	← Second group 050–100: digit of the hundreds is 0
1	← Third group 100–150: digit of the hundreds is 1
1	← Fourth group 150–200: digit of the hundreds is 1
2	← Fifth group 200–250: digit of the hundreds is 2
2	← etc.

We add a column for 'Leaf'. Each **leaf** then contains the next digit to the right – that is, the tens of the observations. We then place the next digit in each number in the leaf appropriate to its value:

Leaf (of a stem-and-leaf plot): Contains the second highest place value of the observations; each observation is represented by a digit in a leaf

Stem	Leaf	
0		
0	5	← 57.2 has a stem of 0 and a leaf of 5 tens
1	2	← 128.8 has a stem of 1 and a leaf value of 2 tens
1	696	← 163.4 has a stem of 1 and a leaf value of 6 tens
2		← 160.0 has a stem of 1 and a leaf value of 6 tens
2		← 192.1 has a stem of 1 and a leaf value of 9 tens

Continuing in this way for the rest of the data gives us the first plot on the following page, with the observations entered as we come to them. We now put the numbers for the leaves in order to complete the plot, as shown in the right-hand plot:

ISBN 978-1-107-61421-5

Stem	Leaf
0	12343
0	576697657
1	2140
1	69685
2	
2	6

Stem	Leaf unit = 10
0	12334
0	556667779
1	0124
1	56689
2	
2	6

HINT
Note that either the leaf unit or the stem unit must be stated so that a reader knows where the digits go – that is, what the values of the observations are.

Note that we have lost the digit in the units and in the first decimal place. Effectively we have rounded all values to the ten below. We could instead keep two digits in the leaves, with commas between the observations, like so:

Stem	Leaf
0	10, 24, 36, 37, 48
0	57, 59, 65, 68, 69, 70, 71, 78, 99
1	05, 17, 28, 40
1	56, 60, 63, 86, 92
2	
2	65

From this we see that most of the robins flew less than 100 metres, but a reasonable number flew between 100 and 200 metres, while the one who flew over 260 metres really stands out. Also note that the last stem-and-leaf plot on the right with two digits and a comma for each observation tends to distort the picture and appears to show many more observations than there are. This type of stem-and-leaf is not recommended and is very seldom used.

HINT
Remember that if data values are not given exactly in a dotplot, they are rounded.

HINT
If all digits can't be shown in a stem-and-leaf plot, then the digits that cannot be shown are dropped. The data values are not rounded to the nearest digit that can be shown.

Key ideas

- A dotplot is a picture of a dataset of quantitative data, in which each dot represents an observation. If there are very many observations, it might be stated that each dot stands for two observations.
- Grouping measurement data into intervals can give a better picture of a dataset.
- A stem-and-leaf plot groups data into a number of equal-sized intervals that cover all the observed values. For each observation, the digit in the highest place value is in the stem and the digit in the next lowest place value is in the leaf. The number of digits in a leaf gives the number of observations in the interval covered by that leaf.

ISBN 978-1-107-61421-5

Example 2: Do people speed to go through amber lights?

To investigate how drivers approach green and amber traffic lights, observations were made at an intersection. At the time, traffic was free-flowing. For each vehicle travelling through the lights, the speed of approach was measured by the time in seconds that vehicles took to pass through a 50-m section just before the set of lights. A time of less than 3 seconds is a speed of more than 60 km/h.

HINT
60 km/h is 60 000 m in 3600 s, so 3600 s to travel 60 000 m. Hence the time in seconds to travel 50 m at 60 km/h is $50 \times 3600/60\,000 = 3$ s.

The dataset of 50 observations of drivers approaching amber lights was:

2.75	2.90	3.55	3.54	2.38	2.83	2.08	3.86	2.86	2.75	3.75	3.30	3.50
3.26	3.65	3.59	2.61	4.78	2.91	3.94	3.01	3.10	3.54	3.46	3.26	3.97
3.74	3.97	4.12	3.33	4.70	3.51	2.90	2.78	3.73	3.77	3.78	3.15	4.94
3.34	3.80	3.10	2.99	2.95	3.92	3.96	3.01	4.18	4.14	3.12		

Here is a stem-and-leaf plot of these 50 observations:

The digits in the stem are the units.

The stem is divided into intervals of 0.2 seconds.

So the first interval goes from 2.00 to 2.19.

The second interval goes from 2.20 to 2.39, and so on.

The digits in the leaves are the first decimal place of the observations.

Stem	Leaf unit = 0.1
2	0
2	3
2	
2	6777
2	8899999
3	001111
3	22333
3	4555555
3	677777
3	8899999
4	111
4	
4	
4	77
4	9

There is one observation in the interval 2.00 to 2.19; no observations in the interval 2.40 to 2.59; seven observations in the interval 2.80 to 2.99, with two having an 8 in their first decimal place, and five having a 9 in their first decimal place.

This shows that a number of drivers were going faster than 60 km/h as they approached the amber light.

ISBN 978-1-107-61421-5

Example 3: How often do people blink?

Example 1 of Chapter 1 discusses an experiment to investigate how often people blink when being interviewed for one minute. In another experiment, data are collected on how often students blink in two minutes while reciting a poem. The data consist of 50 observations:

28	5	26	17	49	51	30	49	39	83	105	67	70	32	67	26	54	44	46	75	63
50	77	8	16	38	34	46	95	32	14	36	96	87	56	70	66	16	18	36	24	10
52	32	48	57	30	64	34	16													

These data are count data but notice how much they vary and how many different values there are. Like measurement data, count data with lots of different values are better presented in graphs that group the observations. In this dataset there are 38 different values of the count variable (number of blinks in two minutes) and only a few that occur more than once. A column graph (or bar chart) does not give us an informative picture of data like these.

Below is a stem-and-leaf of these data.

The stem unit is 10. →

Stem	Leaf unit = 1
0	58
1	0466678
2	4668
3	00222446689
4	466899
5	012467
6	34677
7	0057
8	37
9	56
10	5

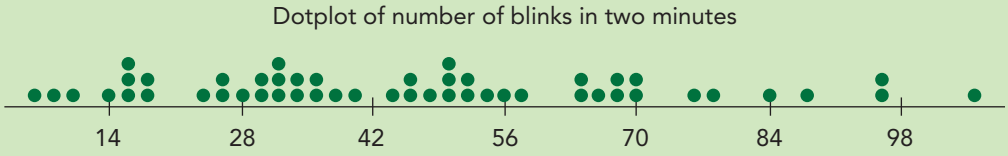
← The leaf unit is 1 because these are counts ranging from 5 to 105 blinks in two minutes.

The observations have been grouped into 11 intervals of length 10 each. →

We see, for example, that six people blinked from 40 to 49 times, with one blinking 44 times, two blinking 46 times, one blinking 48 times and two blinking 49 times.

The following page shows a dotplot of these data with only a small amount of grouping; the observations are grouped in pairs of odd and the following even number. Notice how the stem-and-leaf graph above provides a less bumpy picture while still showing that there is some clumping of the data, and still showing all the individual values. If we chose more intervals for the stem-and-leaf, for example 21 intervals of length 5 each, the stem-and-leaf would be much bumpier.

ISBN 978-1-107-61421-5



Exercise 2B

1 Consider Example 2 on the previous page.

- **a** What percentage of drivers were speeding (doing more than 60 km/h) in approaching the amber light?
- **b** What speed was the fastest driver doing?
- **c** What speed was the slowest driver doing?
- **d** Construct another stem-and-leaf plot using just 6 intervals.
- **e** Which stem-and-leaf plot do you think gives the better picture of the data? Why? (Note there is no right or wrong answer to this question!)

2 Below are 20 reaction times in seconds for one person in the Go for the Gopher game. These have been collected from two 'games' of 10 'pop-ups' of the gopher in each. <www.cambridge.edu.au/statsAC78weblinks>

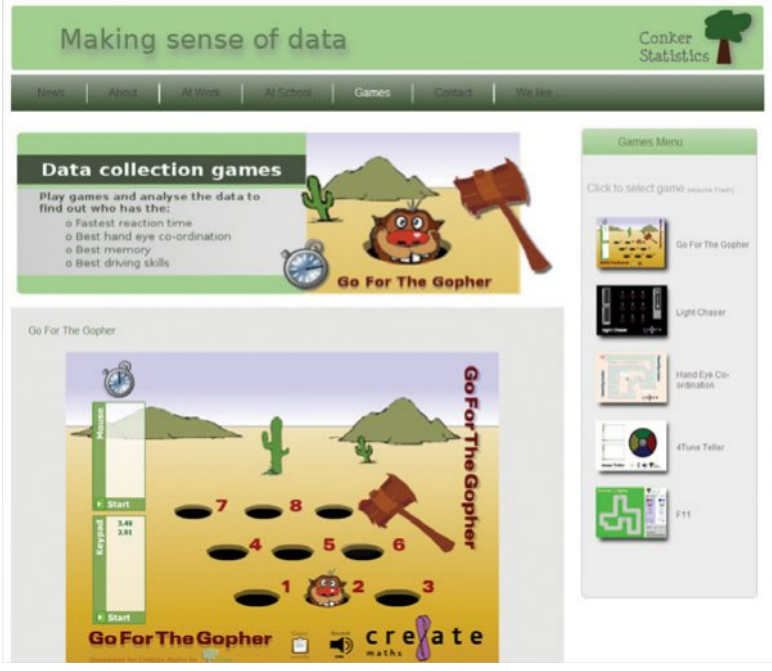
1.09	1.10	1.04	1.34	0.87	0.93	1.03	0.89	0.92	1.12
1.28	0.92	1.03	0.89	0.92	0.85	0.92	0.84	0.96	1.03

Construct a dotplot of these data, rounded to the first decimal place.

- **a** What value is assigned to 1.34 in this dotplot?
- **b** What value is assigned to 0.89?
- **c** What value have you assigned to 0.85?

Construct a stem-and-leaf plot of these data with intervals of length 0.05.

- **d** How many intervals do you have?
- **e** What is your leaf unit?



ISBN 978-1-107-61421-5

3 Below are 68 observations on the length in seconds of mobile phone calls in a public place.

380	65	25	45	150	120	70	355	95	360	150	120	60	35	200	50
1930	50	380	130	185	100	60	100	30	170	80	55	805	30	395	70
305	249	145	330	750	70	70	75	330	75	65	20	65	30	90	50
45	145	140	225	355	295	325	55	45	285	330	65	45	123	120	45
95	120	10	195												

a There is a very large observation. Remove it and then construct a stem-and-leaf plot of these data with 17 intervals.

b What is the leaf unit?

c What is the length of each interval?

d What percentage of people (in the original dataset) spoke for 6 minutes or more?

4 The graph below is a stem-and-leaf plot of the lengths in minutes of 128 music CDs.

Stem-and-leaf plot of CD lengths in minutes

Leaf unit = 1.0

2	4
2	
3	
3	77778899
4	001112233333444
4	566667777888899
5	1223333444
5	555555666678899
6	00000222234
6	55556777779
7	0011112222222334444444
7	55566667777788889999

a The shortest CD is recorded as being 24 minutes on the stem-and-leaf plot. Between what values would its length lie?

b How long is the longest CD in this dataset?

c What is the length of the intervals in this stem-and-leaf plot?

d Why are there no digits in the second and third leaves, with stem values 2 and 3?

e What percentage of the CDs are at least an hour and a quarter long?

Enrichment

Armspans and heights of Australian students in Years 7 and 8
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

2-3 Mean, median and range of quantitative data

The word *range* is used in statistics in the same way as in everyday speech, for example, 'the values of the observed reaction times range from 0.84 seconds to 1.34 seconds'. But in statistics we also use it to describe the distance between the smallest and the largest observations in quantitative data. So the **range** in the data of the above quotation is 1.34 – 0.84 = 0.5 seconds. The smallest observation is called the minimum, and the largest is called the maximum, so the range is (maximum – minimum). The range of the data gives a summary of how spread out the observations are. Another set of reaction times might range from 0.5 seconds to 1.45 seconds, so the range of these data is 0.9 seconds – the **variability** in the second dataset is greater.

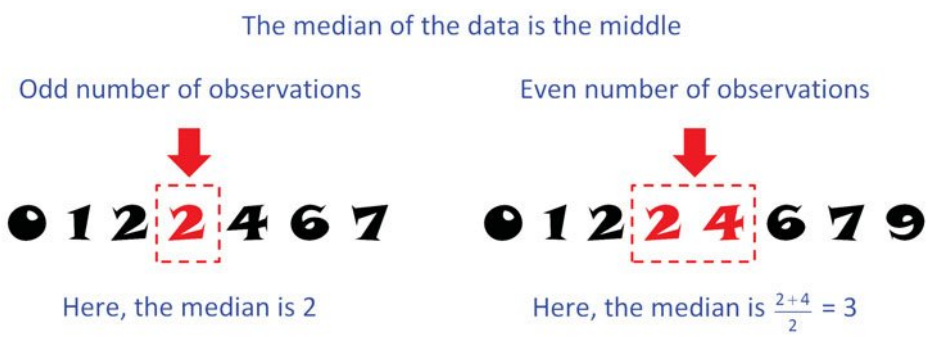
In any dataset, the minimum, maximum and range of the data give us some idea of the types of values, and, in particular, the overall spread of values. However, what can we say to the question 'how far do robins fly when first released?' or 'how often do people blink when speaking in public?' It would be useful to have some way of representing the general size of the data. Two such data quantities are the **average** or **mean** of the data, and the **median** of the data.

The average or mean of the data is obtained by adding all the values of the observations and dividing by the number of observations.

Range: The largest (maximum) value minus the smallest (minimum) value in the data

Variability: How greatly the data values differ from each other; what makes up variability depends on the situation

HINT
Remember, quantitative data are measurement data or count data.



The median of the data is the middle value – the value that has half the observations less than it in value, and half the observations greater than it in value.

If the number of observations is an odd number, then the median of the data is one of the values observed – the middle value with equal numbers of observations less than it and greater than it. If the number of observations is an even number, then the median of the data is taken as halfway between the two middle values so that again there are equal numbers of observations less than and greater than it.

Average or **mean:** Obtained by adding all the values of the observations and dividing by the number of observations

Median: The middle value; the value that has half the observations less than it in value, and half the observations greater than it in value

LET'S START Calculating some data means, medians and ranges

In the example at the beginning of this chapter, the minimum and maximum first-release flight distances for the robins are 10.0 metres and 265 metres, so the range of the data for the robins is 265.0 −10.0 = 255.0 metres. We see that this range is very large – the flight distances vary a lot.

ISBN 978-1-107-61421-5

The average or mean of the distances for the robins is obtained by adding up all 24 of the distances and dividing by 24. The total of the 24 distances is

128.8 + 160.0 + 192.1 + 163.4 +186.4 +156.2 + 70.0 + 10.0 + 57.2 + 65.2 + 68.9 + 24.7 + 37.4 + 99.7 + 265.0 + 78.7 + 48.2 + 69.2 + 117.3 + 36.5 + 140.8 + 59.3 + 71.3 + 105.3 = 2411.6 metres

Therefore the average or mean of these data is $\frac{2411.6}{24}$ = 100.5 metres

To find the median of the data, we need to first order them. We usually do this from smallest to largest, but it can be done from largest to smallest if preferred. And now we see another use for the stem-and-leaf plot, because this orders the data but keeps the values, making it much easier to find the value that has half the observations less than and greater than it. Although not all the place values appear in a stem-and-leaf, it is still easy to find the value of the median. Indeed, if we were looking for a useful way to order the data, we might have invented the stem-and-leaf!

For the robins' distances, there are 24 observations, so we need to find the 12th and 13th observations when they are arranged from smallest to largest. We use halfway between the 12th and 13th for the median of the data, because this will have twelve observations on each side of it. Looking at the stem-and-leaf plot of the robins' distances, we see that both the 12th and 13th observations have 7 in the leaf, so they are both seventy-something, and they are the 2nd and 3rd of the three observations in the 70s. Looking at the data, we see that they are 71.3 and 78.7. So the median of these data is $\frac{(71.3 + 78.7)}{2} = \frac{150}{2}$ = 75 metres.

To show what happens if we have an odd number of observations, let's suppose it is decided that the largest observation of 265.0 metres was not recorded correctly and that we should leave it out. Then we would have 23 observations, and the middle one would be the 12th (from the bottom) because it has 11 observations less than it, and 11 observations greater than it. We already know (see above) what the 12th observation is – it's 71.3 metres.

The median and the average are both useful in indicating where the data are centred. The median of the robins' distances has half the data less than it and half greater, so it provides a balance point for the numbers of observations. But the data less than it are rather 'squashed up', and the data greater than it are more variable. So the average gives us a balance point of the sizes of the observations.

ISBN 978-1-107-61421-5

Key ideas

For quantitative data (measurement or count data):

- the range of the data is the maximum value – the minimum value. The range gives the total spread of the data.
- the average or mean of the data is obtained by adding all the values of the observations and dividing by the number of observations
- the median of the data is the value that has half the observations less than it in value, and half the observations greater than it in value
- both the data mean and the data median indicate the general size of the observations and values about which the data are spread. That is, the data mean and the data median each give a measure of where the data tend to be centred.

Example 4: How far do doves fly?

For the data on first flight distances for the doves at the beginning of this chapter, there are 26 observations. The minimum and maximum are 22.1 m and 613.2 m, so the range of the distances for the doves is 613.2 – 22.1 = 591.1 m. As for the robins, the data are very variable with a large total spread as measured by the range.

The average or mean of the data is obtained by adding up the 26 observations and dividing by 26. This gives $\frac{4949.2}{26}$ = 190.35 m (to 2 decimal places).

To obtain the data median we need to order the data from smallest to largest. Because there are 26 observations, the data median is halfway between the 13th and the 14th so that there are 13 observations smaller and larger than it. Ordering from the smallest to the 14th smallest gives:

13.9 22.1 30.6 40.0 44.7 55.5 76.0 80.0 83.4 102.0 162.7 165.5 166.7 170.0

Hence the 13th and the 14th observations are 166.7 and 170.0, and the data median is $\frac{(166.7 + 170.0)}{2} = \frac{336.7}{2}$ = 168.35 m.

Suppose we have a stem-and-leaf of these data, as shown:

Leaf unit = 10

0	12344
0	5788
1	0
1	6667789
2	
2	668
3	11
3	568
4	
4	
5	
5	
6	1

The digits in the leaves are in the 10s, so the 13th and 14th observations on the stem-and-leaf are the third value of 160 and the first value of 170. If we have the original data, we can then locate the three observations that are in the 160s and pick out the largest, which is 166.7. Next we find the smaller of the two observations in the 170s, namely, 170.0.

If we do not have the original data, but only the stem-and-leaf, then the best we can do for the data median is halfway between 160 and 170, which is 165 m. However this is reasonably close to 168.35 m.

Also if we have only the stem-and-leaf, the best we can quote for the data range is 610 – 10 = 600 m, compared with 591.1 m.

ISBN 978-1-107-61421-5

For the data mean, using only the information in the stem-and-leaf, the total of the observations

$= (10 + 20 + 30 + 40 \times 2 + 50 + 70 + 80 \times 2 + 100 + 160 \times 3 + 170 \times 2 + 180 + 190 + 260 \times 2 + 280 + 310 \times 2 + 350 + 360 + 380 + 610)$

$= 4830$

giving an average of $\frac{4830}{26} = 185.77$ m, compared with 190.35 m using the original data.

Example 5: Taking care with data means and medians

In a survey, one of the questions asked respondents which brand of laundry detergent they usually bought. Five brand names were provided, with an 'Other' category also provided. In the data entry, these brands were entered as 1 to 5, with the 'Other' category entered as 6.

Clearly, it makes no sense to calculate an average or median for this dataset of numbers because the numbers do not represent any quantities. These numbers are simply codes, an alternative to entering the names of the brands, or to coding them by letters. The variable 'brand of laundry detergent' is a categorical variable, and it makes no sense at all to speak of an average brand. What is likely to be of interest is the most popular brand; we will see more about this in section 2-4.

In the same survey, there was this question, where responses were given numbers:

What is your response to recycled water being added to normal town water?

Please enter a number in the box: ☐

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

The responses to this question are numbers. Does it make sense to calculate an average or a data median?

Because the ordering of the numbers represents the ordering of amount of support, we can calculate an average to give us some idea of the support – the average amount of support. As we have seen, this type of variable is called an ordinal variable. But we must be very careful in interpreting this kind of average because these numbers do not represent a distance of some sort, just an ordering. The 'distance' between strongly disagree and disagree is not necessarily the same as the 'distance' between disagree and neutral. The main use of an average for this question would be to see which side of 3 it is, and which category it is closest to.

In addition, even if the same coding is used, the numbers cannot necessarily be compared across questions. For example, if the average 'amount of support' for adding recycled water is 3.9 and the average 'amount of support' for the current Australian cricket captain is 4.1, we can't say that there is more support for the current Australian cricket captain than there is for recycled water. This would be like saying that people are taller than their weight! We know that's nonsense because we can't compare units of height with units of weight. Similarly, when we give numerical codes for amount of support for different questions, it doesn't mean the responses to those questions can be compared.

ISBN 978-1-107-61421-5

The median of these data is even less useful because there are only 5 values so in practical situations there won't be a value that has half the observations on one side and half on the other.

And finally, one of the many statistical jokes about averages is:

> *The overwhelming majority of people have more than the average number of legs.*
>
> E. Grebenik

The variable, number of legs, is a count variable, and the numbers do have meaning as quantities, but it is clearly not only useless but misleading to calculate an average number of legs. There are only three possible values (0, 1, 2) and it would be far more useful to report the percentages of observations – that is, to report it as we do for categorical data.

CAUTION
Never calculate an average or median for coded data of a categorical variable. The codes just represent names or categories – they are not values or quantities.

CAUTION
Be careful when interpreting averages, medians and ranges for ordinal data or for count data with a very small number of different possible values.

Exercise 2C

1 **a** Calculate each of the following for the data in Example 2 on the time it took drivers to go through 50 m just before traffic lights when approaching amber lights.

- **i** The range of times
- **ii** The average time
- **iii** The median of the times.

b From your answers to part **a** calculate their average and median speeds.

c The age groups of the drivers were also recorded in this investigation. The first group were recorded as 1 (drivers approximately under 25 years), the middle group as 2 (approximately 25 to 50 years), and the third group as 3 (approximately over 50 years). An investigator used these numbers to obtain the average, median and range of ages of drivers. Is this meaningful? Why or why not?

2 Consider Example 3 above on how often people blink in two minutes while reciting a poem.

a Use the original data to calculate:

- **i** the average number of blinks in two minutes
- **ii** the median of the number of blinks in two minutes
- **iii** the range of the number of blinks in two minutes.

ISBN 978-1-107-61421-5

b If you do not have the original data and use only the stem-and-leaf plot provided in Example 3, what would you calculate for:

i the average number of blinks in two minutes?

ii the median of the number of blinks in two minutes?

c Also recorded in this experiment was the gender of each person, recorded as 1 for females and 2 for males. Why is an average of these numbers nonsense?

3 a In question 2 of Exercise 2B on the reaction times in the Go for the Gopher game, use the original data to calculate:

i the average reaction time

ii the median reaction time

iii the range of the reaction times.

b If you do not have the original data and use only the stem-and-leaf plot you obtained in that exercise, what would you calculate for:

i the average reaction time?

ii the median reaction time?

iii the range of the reaction times?

4 a In question 3 of Exercise 2B, there are 68 observations on the length in seconds, of mobile phone calls in a public place. Use the original data to calculate:

i the average length of the phone calls

ii the median length of the phone calls

iii the range of the lengths of the phone calls.

b Repeat part **a** omitting the largest value in the data.

c Pretend you do not have the original data and use only the stem-and-leaf plot you obtained in that exercise to calculate:

i the average length of the phone calls

ii the median length of the phone calls

iii the range of the lengths of the phone calls.

d Should you compare your answers in part **c** with those of part **a** or part **b**?

5 Refer to the plot in question 4, Exercise 2B titled 'Stem-and-leaf plot of CD lengths in minutes'. This is a plot of the lengths of 128 CDs.

a Use this plot to obtain the median and range of the lengths of those 128 CDs.

b The original data are in the Excel file called *CDs*. Use Excel (or another spreadsheet program) to show that the range, median and average of the lengths of the CDs are, respectively, 55 min, 59.5 min and 59.2 min (correct to 1 decimal place).

Enrichment

Armspans and heights of Australian students in Years 7 and 8
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

Cambridge University Press

2-4 Modes, myths and measures of centre

A **mode** is a value that occurs most often in a dataset. There may be more than one mode if different values occur the same number of times in a set of data. Let's look at some examples.

Mode: A data value that occurs most often; there may be one mode or many modes if different values occur the same number of times in a set of data

LET'S START Finding some modes

Look at the two dotplots in section 2-2 for the data on the flight distances of robins. The first dotplot did not round any of the observations. Every value occurs once only, so every observation is a mode! In the second dotplot, the observations are rounded to the nearest 5 metres, and there are 4 observations placed at 70 metres to the nearest 5 metres. Does that mean we can take 'the mode' to be 70 metres? Now look at the stem-and-leaf plots of the same data on pages 32 and 33. The observations are now grouped in intervals of length 50 metres. The stem-and-leaf plot with the commas has kept the digits in the units but all observations are different values, so all are 'modes'. The stem-and-leaf plot with leaf unit = 10 only contains the digits in the tens, and there are three observations in the 60s and three in the 70s. So are there two modes? Or we could look at the group with the most observations and say the 'mode' is the group between 50 metres and 100 metres. But even if we decide we'll consider a group of rounded values, the grouping depends not only on how much we group but also on our starting point.

Perhaps the dataset on flight distances for robins is tricky because there are only 24 observations. But the range of the data is so great, it is most likely that more observations will just give us more different values. And if we group them, the group that occurs most often will depend on how we group them.

It's clear that the idea of 'mode' is both vague and difficult for measurement data. So for what types of data is the concept of mode useful? Consider the variables in question 4a in the Pre-test of this chapter:

month of birth

height

weight

year of school

length of forearm

position in family

number of siblings

left-handed (yes/no)

wear glasses (yes/no)

hours per week in part-time job

accident proneness on a scale of 1–5

The most commonly occurring value in the data would be of interest and useful for: month of birth, number of siblings, wear glasses, accident proneness, left-handed or not, year of school, position in family. The data for all of these variables are either categorical or count with not too many different values.

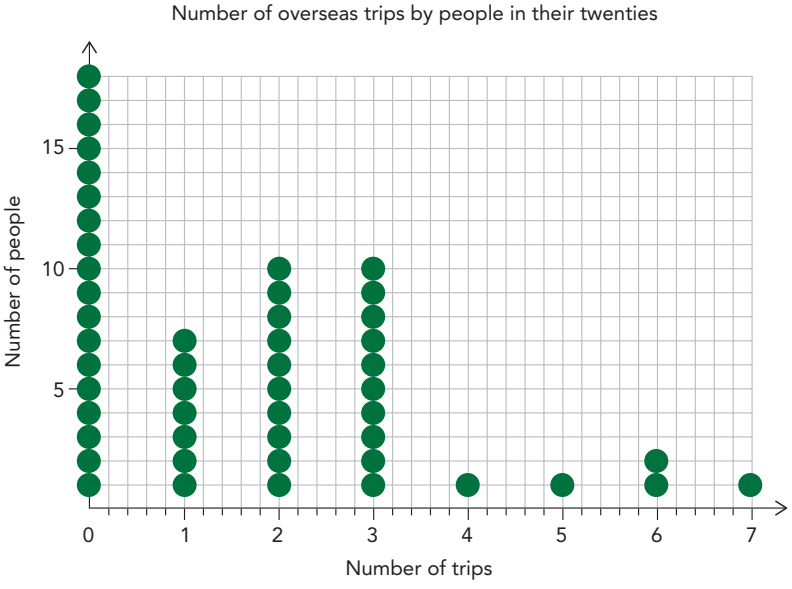
Means and medians of data are relevant only for quantitative data (that is, count or measurement data), and mostly for data with more than just a few different values. In

ISBN 978-1-107-61421-5

contrast, modes of data are mostly relevant for categorical data and quantitative data with only a few different values, and can be very difficult even to identify or choose for measurement data. So the first myth (or misconception) is that modes of data are relevant for measurement data – or, more generally, data on a continuous variable.

Another myth is that modes of data are like means and medians of data in that they indicate where the data are centred. For measurement data, there can be many modes – even as many as the number of observations! And if we group the data by rounding or by choosing groups, the modes depend on the choices of groups, and could be anywhere!

What if we consider a count variable with not many different values, so that the most commonly occurring value is of interest? Below is a dotplot of data from a survey of people in their mid-twenties, on the number of overseas trips made so far in their lifetime. We see that the mode is 0, but could we claim that it is the **centre** of the data? The median is 1 and the average is 1.7. It is not easy to choose to say where the centre of these data is. The mode is indeed useful in this example, but it is because there are not many different values in the data, and the mode is providing a *different* type of information from the mean and median.



Centre (of the data): Where the data tend to centred, or what the data tend to be spread around; a centre of the data also indicates the general size of the observations

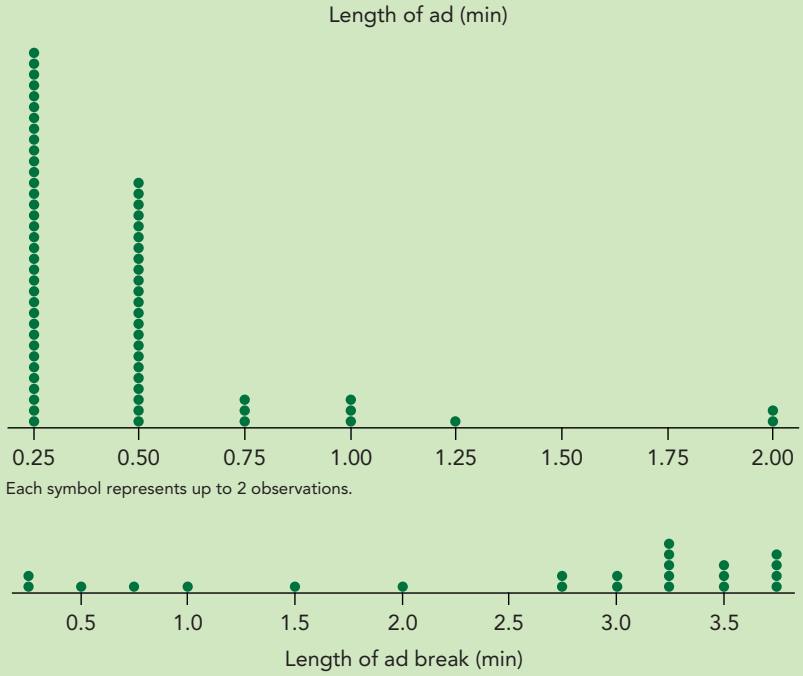
Key ideas

- A mode of a dataset is a value that occurs most often.
- There may be one or many modes in a dataset.
- For measurement data, modes may be difficult to choose or describe, unlike the mean and median. All the data values may occur once each, or modes might depend on how the data values are rounded or grouped.
- Modes are most useful for categorical data or for quantitative data that do not have many different values. In this case, a mode provides different information from the median and mean.
- A mode is often not a good indication of where the data tend to be centred.

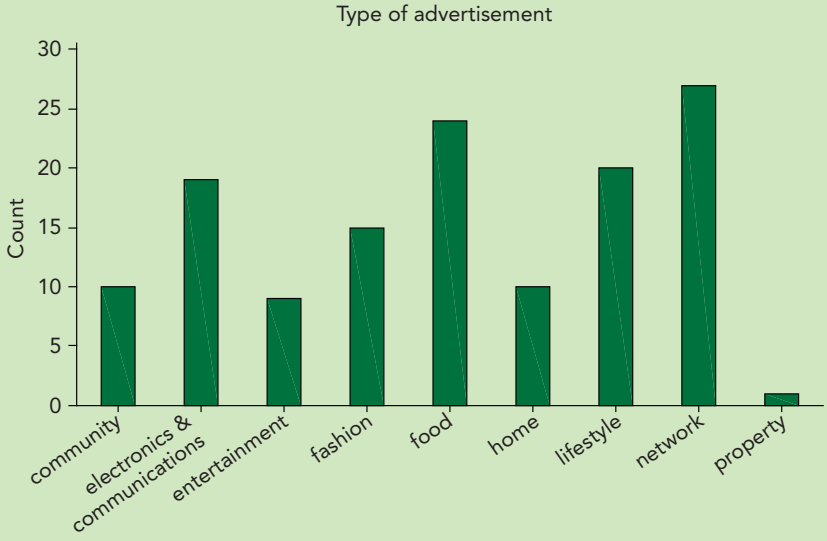
ISBN 978-1-107-61421-5

Example 6: Ads and ad breaks

The lengths of advertisements in minutes, the type of advertisement, and the lengths of advertisement breaks in minutes were recorded for a number of different television shows. The type of advertisement was classified under eight headings. (Note that these would need very careful identification.) Below are dotplots of the data for the two measurement variables and a column graph for the data for the categorical variable.



HINT
For large datasets, a dotplot may also use a dot to represent two (or even more for very large datasets) observations.



ISBN 978-1-107-61421-5

We see that there are only six different values for the lengths of individual advertisements. This is not surprising as advertisement space tends to be sold in specific lengths. So there is no problem in saying that the mode is 0.25 minutes (15 seconds), and we would be more likely to report this as follows:

'Advertisements are made as multiples of 15 seconds. The different lengths were 15, 30, 45, 60, 75, 120 seconds, with the most common length being 15 seconds.'

That is, we would tend to report this in terms of the different categories of fixed lengths of advertisements. In this case, the mode is useful because ads are made to specific lengths and there are only six different lengths. Because more than half of the ads are 15 s (71 out of 135), the median is also 15 s. So in this case the mode is the same as the median because there are so few different values. However, the average length of ad is 26.44 s. In this example, the mean and the median and the mode all contribute to describing the data.

There is more variation in the lengths of ad breaks, with eleven different values, although they are still in multiples of 15 s because they are made up of a number of individual ads. There is a single mode at 3.25 min or 3 min and 15 s, and the median is also 3.25 min. The average is 2.587 min which is approximately 2 min and 35 s. We see that twelve (out of 23) of the lengths are 3.25, 3.5 or 3.75 min, while the rest range from 30 s up to 3 min.

So for the data on lengths of ads and length of ad breaks, there is a mode and it is easily identified because of the small number of different values. For both these datasets, the mode is the same value as the median. This mode is useful because, although the data are quantitative, the data values fall into categories determined by multiples of 15 s and the mode tells us which of these categories occurs most often. The median and the mean are the quantities that indicate where the data are centred.

The third variable is the type of advertisement, and it is a categorical variable. For categorical variables, the only way to report the data is through frequencies or relative frequencies. The most commonly occurring category is 'network' – that is, ads for the TV network and programs on the network.

ISBN 978-1-107-61421-5

Exercise 2D

1 Consider Example 2 above on the time it took drivers to go through 50 m just before traffic lights when approaching amber lights.

- **a** Look at the original data. How many modes are there, what are their values, and how often do each of them occur?
- **b** Now look at the stem-and-leaf plot provided in Example 2. How many modes are there, what are their values and how often do they occur?

2 Consider Example 3 above on how often people blink in two minutes while reciting a poem.

- **a** Look at the original data. How many modes are there, what are their values, and how often do each of them occur?
- **b** Does the stem-and-leaf provided there give you different answers to part **a**? Why or why not?
- **c** Do you think the mode (or modes if there are more than one) give a good indication of where the data are centred? Compare with the mean and median obtained in question 2 of Exercise 2C.

3 In question 3 of Exercise 2B, there are 68 observations on the length in seconds of mobile phone calls in a public place. Omit the largest observation.

- **a** Using the original data, there is one mode. What is its value and how often does it occur?
- **b** Using the stem-and-leaf you obtained, how many modes are there and what are their values?
- **c** Where is the mode you obtained in part **a** in the stem-and-leaf plot?
- **d** Do you think any of the modes you obtained in parts **a** and **b** give a good indication of the centre of the data?
- **e** Compare the modes you found with the mean and median you obtained in question 4 of Exercise 2C.

4 Refer to the graph in question 4 of Exercise 2B titled 'Stem-and-leaf plot of CD lengths in minutes'. This is a graph of the lengths of 128 CDs.

- **a** Using the stem-and-leaf plot, how many modes are there and what are their values? How often does each value occur?
- **b** We know from question 5 of Exercise 2C that the mean is 59.2 min (correct to 1 decimal place) and the median is 59.5 min. Do you think the mode(s) you found in part **a** provide any useful information? If so, what?
- **c** The original data were recorded in minutes correct to 2 decimal places. Do you think it's possible to get different answers to those in part **a**? Why or why not?

How big is your fish?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

2-5 Commenting on data features

Features of data are various aspects of what the data look like and how they behave. You have probably noticed that we have already started to comment on features of data in this chapter. The data range, mean and median, and the most frequently occurring value or category provide information and represent some of the features of the data. These are examples of summaries of the data – often called **summary statistics** – which are used in describing features of the data.

We explore data through plots and summaries, using plots and summaries appropriate for the type of data – measurement, count or categorical. Even then, how we explore the data depends to some extent on the nature and behaviour of the data.

In reporting on the data and what we have found in our exploration, we comment on the main features of the data, but in the **context of the data** – the circumstances or 'story' of the data (what it tells us). Summaries such as data range, mean, median and the most frequently occurring value or category are useful but do not substitute for plots. They tend to be most informative when used together with plots in commenting on the data – but always within the 'story' of the data.

Features of data: Various aspects of what the data look like and how they behave

Summary statistics: Values calculated or obtained from data and used in describing features of the data

Context of the data: The circumstance or 'story' of the data – what the data are about

ISBN 978-1-107-61421-5

LET'S START Reporting on the doves

There are two datasets at the beginning of this chapter; one contains 24 observations on the distances flown by a species of robin on their first release after banding, and the other contains 26 observations on this first flight distance by a species of dove. Both of these datasets are small, so we need to be careful in our comments on them.

The distances vary a lot, with the range of the dove data being 591.1 metres, which is very large considering that the smallest distance is only 13.9 metres. Example 4 gives a stem-and-leaf for the doves' distances. This plot shows that there are three clumps of data – one in approximately 10–80 metres, another in approximately 160–190 metres, and the third in approximately 260–380 metres. There is also one very large distance at approximately 610 metres. The median is 168.35 metres, so half the doves flew less than this and half greater. The average distance flown was 190.35 metres, so the values of both the median and the average are in the middle clump of data in the stem-and-leaf plot.

HINT
If we only have the stem-and-leaf plot, the values of the range, median and average would have to be quoted as approximately 600 m, 165 m, 186 m.

Key ideas

- The data range, mean and median, and the most frequently occurring value or category are useful summaries of the behaviour of the data that can be used in commenting on features of the data.
- Plots are used to explore data and then present data, and comments on features of the data should refer to plots as well as to summaries.
- Comments on features of data must always be made in the context of the data.

Example 7: How often do people blink when speaking in public?

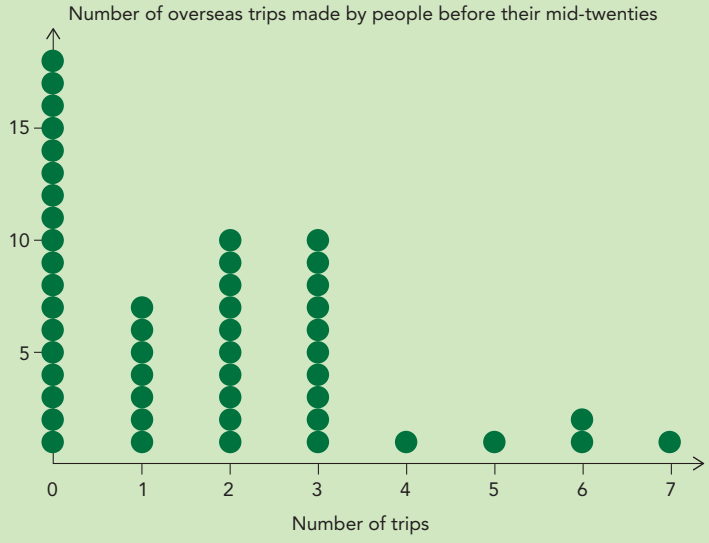
Example 3 gives 50 observations on how often students blink in two minutes while reciting a poem. A copy of the stem-and-leaf plot is shown at right.

How often students blink in two minutes while reciting a poem is very variable, with a minimum of 5 blinks, and a range of 100. The stem-and-leaf plot shows that a number of students do not blink much, blinking fewer than 20 times in two minutes. After 20 blinks, the number of blinks varies quite smoothly, with the most frequently occurring group being those who blink 30 to 39 times, with gradually fewer and fewer students blinking more until the maximum number of 105 blinks. The average number of blinks over all 50 students is 45.7 and the median is 45, so that half the students did blink more than 45 times in two minutes.

Stem	Leaf unit = 1
0	58
1	0466678
2	4668
3	00222446689
4	466899
5	012467
6	34677
7	0057
8	37
9	56
10	5

ISBN 978-1-107-61421-5

Example 8: How often do people travel oveseas before their mid-twenties?



The dotplot above from section 2-4 shows data from a survey of 50 people in their mid-twenties on the number of times they had travelled overseas so far in their lives. The range is 7, with the most commonly occurring number being 0 (18 out of the 50 people who responded). The average is 1.7 times, and the median is 1, so that half had travelled at least once and half at most once. However, a dotplot (or column graph or bar chart) shows that for those who had travelled overseas at least once, the most common number of trips is 2 or 3.

HINT
The median of data is easy to interpret – half the observations are less than it, and half are greater than it. Examples where medians are often quoted now include house prices in different districts, incomes and rainfalls.

Exercise 2E

1 Example 2 gives data from an investigation into the speed at which drivers approach traffic lights. The data given are on the time 50 vehicles took to travel 50m just before driving through amber lights. Questions 1 of Exercises 2B, 2C and 2D consider plots and summaries of these data. Use your work in these questions to comment briefly on the data.

2 Question 2 in Exercise 2B gives 20 reaction times in seconds for one person in the Go for the Gopher game. These were collected from two 'games' of 10 'pop-ups' of the gopher in each. This question and question 3 in Exercise 2C consider plots and summaries of these data. Use your work in these questions to comment briefly on the data.

3 Question 3 in Exercise 2B gives 68 observations on the length in seconds, of mobile phone calls in a public place. This question and questions 4 in Exercise 2C and 3 in Exercise 2D consider plots and summaries of these data. Use your work in these questions to comment briefly on the data.

ISBN 978-1-107-61421-5

4 Question 4 in Exercise 2B gives a graph of the lengths in minutes of 128 music CDs.

Stem-and-leaf plot of CD lengths in minutes

Leaf unit = 1.0

2	4
2	
3	
3	77778899
4	001112233333444
4	566667777888899
5	1223333444
5	555555666678899
6	00000222234
6	55556777779
7	0011112222222334444444
7	55566667777788889999

This question and questions 5 in Exercise 2C and 4 in Exercise 2D consider plots and summaries of these data. Use your work in these questions to comment briefly on the data.

5 The Enrichment question of Exercise 2B gives a random sample of 50 observations obtained from the ABS CensusAtSchool website, using Random Sampler. They are the responses to questions 9 (height in cm) and 11 (armspan in cm). The sample is taken from all the Year 7 and 8 students who completed the 2012 CensusAtSchool questionnaire in all Australian states and territories. This question and the Enrichment question in Exercise 2C consider plots and summaries of these data. Use your work in these questions to comment briefly on the data. <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

Enrichment

How big is your fish?
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

Chapter summary

Collecting and handling measurement data

- Give details of how measurements made
- Define units and accuracy of measurements
- Keep identifiers for checking
- Do not create unnecessary categories.

Graphing measurement data

- Grouping can help to see data behaviour
- Dotplots
  - Each dot represents given number of observations, usually one
- Stem-and-leaf plots
  - Data are grouped into equal-sized intervals
  - Digits in highest place value are in stem
  - Digits in next place value are in leaves
  - Number of digits in leaf gives frequency of observations.

Summaries of quantitative data

- Data range is overall spread = largest – smallest
- Average or data mean is sum of all observations divided by number of observations
- Data median has half the observations less than it and half greater
- Mean and median of data indicate general size and centre of data.

Modes

- A mode of a dataset is a value that occurs most often
- There may be one, two or many modes in a dataset
- For measurement data, modes are difficult to define or choose
- Modes are most useful for data with not many categories or different values
- Modes are generally not useful indicators of where data are centred.

Commenting on data

- Summary statistics are useful in commenting on data
  - Data range, mean and median, and the most frequently occurring value or category
- Refer to plots as well as summaries
- Always comment in the data context.

Multiple-choice questions

1 Data are collected on the prices for bottles of soft drinks. The data collected are the price, brand and flavour. What is missing that will prevent any useful analysis of these data?

A Name of collector **B** Name of shop
C Volume of bottle **D** Bottle material (plastic/glass)

2 The first few observations in a record of times between postings to an internet site are 1 h, 30 min, 5 min, 30 s. What is wrong with these observations?

A Who made the posting **B** Different accuracy of times
C What was in the posting **D** Different lengths of times

ISBN 978-1-107-61421-5

3 The first few observations in a record of times between phone calls to a pizza outlet are 3 min, 4.5 min, 5 min 20 s. What is missing in these observations?

A Who took the call
B Which pizzas were ordered
C The time of the call
D Information on accuracy of times

4 What is important for the dots in a dotplot?

A Their size
B Number of observations
C Their shape
D Their colour

5 In a report, the amount of time children read for pleasure in a fortnight was reported by how many in each of the following: 2 hours or less, 3–4 hours, 5–9 hours, 10–19 hours, 20 hours or more. What plot can we use for these data on amount of time?

A A column graph
B A stem-and-leaf plot
C A dotplot
D None of these

6 Using the report in question 5 above, what can we calculate for these data?

A Mean
B Range
C Median
D A mode

7 Consider the stem-and-leaf plot below of 25 observations. What is the median?

Leaf unit = 1.0

0	3
0	55688
1	1223445
1	66778
2	00111
2	56

A 15
B 1.5
C 15.5
D 12.5

8 Consider the stem-and-leaf plot below of 26 observations. What is the median?

Leaf unit = 0.1

0	3
0	55688
1	1223445
1	666778
2	00111
2	56

A 13
B 1.55
C 15.5
D 15

9 The mean of 1, 5, 6, 4, 3, 2, 1, 4 is

A 3.25
B 3
C 3.2
D 3.3

ISBN 978-1-107-61421-5

10 A report on a dataset said: 'The range of heights was 15 cm, with the minimum being 11 cm. The median height was 17 cm and the average was 18.5 cm.' What is missing?

A The mode
B Decimal place for the median
C The context
D Decimal place for the range

Short-answer questions

1 Data are collected on delivery time for pizzas from a place that caters for take-away or delivered pizzas. What must be carefully identified in the way the data are recorded?

2 For a city circle bus – a free bus service that does a continuous route within a city centre – data are collected on the times between buses. What must be decided before the data are collected?

3 The time difference in minutes between the actual and timetabled arrival of a number of planes at an Australian airport were observed, giving the following data.

41	3	8	2	–7	–2	3	5	2	7	25	12	9	2	–5	17
11	10	–14	16	3	26	7	25	15	15	–5	–5	30	3		

a What do the negative values mean?

b How do you think these data were collected?

c Construct a stem-and-leaf plot of these data with a leaf unit of 1.

d What is the length of the intervals in your stem-and-leaf plot?

e What percentage of these arrivals were later than as scheduled in the timetable?

4 The data collection of question 3 was repeated on another day, giving the following data (in min).

–8	9	8	–1	0	–1	–13	3	–2	–19	–17	–1	14	19	16	25	65	35
9	20	28	24	32	–5	5	7	0	20	15	–5	11	5	15	–5	22	

a Construct a dotplot of these data.

b Construct a stem-and-leaf plot of these data with leaf unit 1.

c What percentage of these arrivals were later than as scheduled in the timetable?

ISBN 978-1-107-61421-5

5 Refer to the graph on Cambridge GO titled *Time between buses (min)*. This is a graph of 62 times between city circle buses in a city.

a What is the length of the intervals in this stem-and-leaf plot?

b The time between buses is advertised as approximately 10 min. What percentage of these observations had the time between buses more than 10 min?

6 Refer to the graph on Cambridge GO titled *Stem-and-leaf of rainfall (mm)*. This is a graph of 215 monthly rainfalls at an Australian location.

a What is the length of the intervals in this stem-and-leaf plot?

b There are many zeros in the first leaf of this stem-and-leaf plot. Does this mean they are all months with no rainfall? If not, what would the rainfalls for each of these months lie between?

c What percentage of these months at this location had rainfall in excess of 200 mm?

d Now refer to the graph on Cambridge GO titled *Dotplot of rainfall (mm)*. This is a dotplot of the same 215 monthly rainfalls. There are 14 dots for 0 in this plot. Between what values do these rainfalls lie?

7 For the data in question 3 above, find:

a the range, mean and median of the data

b how many modes there are and their values.

8 For the data in question 4 above, find:

a the range, mean and median of the data

b how many modes there are and their values.

9 Refer to question 5 and the graph titled *Time between buses (min)* on Cambridge GO.

a What are the range and median of these data?

b How many modes are there and what are their values?

ISBN 978-1-107-61421-5

10 Refer to question 6 and the graph titled *Stem-and-leaf of rainfall (mm)* on Cambridge GO.

- **a** Use this graph to obtain the median and range of the data.
- **b** The months and years were also recorded in these data. If you numbered the months 1 to 12 from January to December, and you had a number of calendar years of monthly data, what would be the value of the average of the numbers for the months? Why is this meaningless?
- **c** Using the stem-and-leaf plot, how many modes are there and what are their values?
- **d** Compare the mode(s) you found in part **c** with the median obtained from the stem-and-leaf plot. Which do you think is a better indicator of where the data are centred?
- **e** The leaf unit in this stem-and-leaf plot is 10. Do you think a different stem-and-leaf plot might give a different answer to part **a** above? Why or why not?

11 Use your work in questions 3 and 7 to comment on the data.

12 Use your work in questions 4 and 8 to comment on the data.

13 Use your work in questions 5 and 9 to comment on the data.

14 Use your work in questions 6 and 10 to comment on the data.

Extended-response questions

1 Refer to the Enrichment question on Cambridge GO in Exercise 2A on measuring vision.

- **a** People can choose reading glasses of various strengths by self-testing at pharmacies. Research how people assess their vision in a typical pharmacy to choose reading glasses. What is the measure used in this system and what type of variable is being recorded?
- **b** Suggest a continuous variable that you could use (not as an expert) to measure quality of vision.

2 Refer to the dataset titled *Breakfast cereals* on Cambridge GO. It gives data on the weight, and amounts of carbohydrate, protein and dietary fibre (all in g per 100 g) for 66 packets of breakfast cereal. <www.cambridge.edu.au/statsAC78weblinks>

- **a** The amounts are of carbohydrate, protein and dietary fibre per 100 g of cereal. Why is it important to do this?
- **b** Use stem-and-leaf plots for these data.
- **c** Obtain the ranges, means and medians for these data.
- **d** Use the information in parts **b** and **c** to comment on the data.

ISBN 978-1-107-61421-5

CHAPTER

3

Chance

What you will learn

3-1 What is probability?

3-2 Equally likely outcomes

3-3 Equally likely lengths, areas or time periods

3-4 Probabilities of events proportional to 'size'

3-5 Investigating equally likely outcomes

Introduction to chance

Certain: Must happen

Impossible: Can never happen

Probability: A way of measuring chance, of seeing how **likely** any event is to happen

The American scientist and statesman Benjamin Franklin once said: 'The only things certain in life are death and taxes.' He may have exaggerated, but most aspects of life are not **certain**. They lie somewhere between **impossible** and certain, and we can measure how likely they are to happen using the idea of **probability**.

Some areas of life are built around probability. One of these is gambling games. In any gambling game, the outcome is uncertain and you can only determine your chances of winning using probability. It is often said that if you gamble you are most likely to lose, and the more often you gamble the closer you get to losing certainly!

Equally likely outcomes: The outcomes of an experiment are equally likely to occur

Many gambling games can be broken down into situations where there are a number of **equally likely outcomes**. A roulette wheel has 37 equally likely slots, numbered from 0 to 36. The chance that a particular number comes up is only $\frac{1}{37}$. If you buy tickets in a lottery, the winning number is randomly selected from perhaps 100 000 equally likely numbers, giving each ticket a $\frac{1}{100\,000}$ chance of being the winner. The equally likely aspect of such games allows us to investigate them mathematically.

AUSTRALIAN CURRICULUM

Statistics and probability

- Chance
- Construct sample spaces for single-step experiments with equally likely outcomes **(ACMSP167)**
- Assign probabilities to the outcomes of events and determine probabilities for events **(ACMSP168)**

ISBN 978-1-107-61421-5

Often, gambling is viewed very seriously by the punters involved. A 2005 news report told the story of the number 53 in a Venice lottery. The Italian national lottery is a type of lotto in which 50 numbers are picked, five numbers from 1 to 90 in each of 10 cities across the country. In the Venice lottery, the number 53 hadn't come up in the last 152 draws for almost two years. Many Italians were in a frenzy betting on 53 in Venice, convinced on each occasion that it failed to appear that it simply had to appear next time. Many people were financially ruined and, unfortunately, four people died in incidents related to this betting. <www.cambridge.edu.au/statsAC78weblinks>

So how can we measure our chances of winning in gambling? How does having equally likely outcomes help in this process? And how can we check that outcomes from a gambling game really are equally likely? In this chapter we will investigate these and other related questions.

Words you will learn

assign
assumption
balanced
certain
chance
compound event
continuous outcome
discrete outcome
equally likely outcomes
fair
frequency
impossible
likely
modelling
probability
simple event
subjective
symmetry
unlikely

Pre-test

1 Which of these numbers cannot represent a probability? (There may be more than one.)

a 0.5 **b** $\frac{1}{3}$ **c** –0.1
d 80% **e** 0 **f** 1.1

2 Which of these probabilities represents the more likely event?

a 0.2 and 2/5 **b** 33% and 0.06 **c** 0.15 and 0.085?

3 If an event is certain to happen, what is its probability?

4 Write these probabilities as decimals:

a 25% **b** $\frac{4}{5}$ **c** $\frac{1}{10}$

5 A coin is tossed 20 times and 12 heads are obtained. Is it likely that the coin is fair? Give a reason for your answer.

ISBN 978-1-107-61421-5

3-1 What is probability?

Some things in life are impossible. You couldn't fly up to the top of the nearest gum tree without any special equipment. Other things are certain. You were born several years ago and sometime in the future you will die. But most things are in-between. They have a **chance** of occurring but that chance is not so high that you would say they are certain.

Chance: Likelihood, possibility or probability

Probability is a way of measuring chance, of seeing how **likely** any event is to happen. An impossible event has a probability of 0, and a certain event has a probability of 1. Any event that is not impossible or certain has a probability somewhere between 0 and 1. If an event is likely to happen then its probability is closer to 1 than to 0. For instance, the probability that you finish high school with year 12 might be 0.9, as most people do complete year 12. If an event is **unlikely** to happen then its probability is closer to 0 than to 1. For instance, the probability that your next maths class will be cancelled may be 0.1, as most classes are not cancelled. An event that is equally likely to happen as not, will have a probability of $\frac{1}{2}$ or 0.5. For example, if you roll a **fair** die, the chance that it will land showing an even number will be 0.5.

Measuring quantities such as length, weight or time is direct. We can take a ruler to measure a length, a set of scales to measure a weight, and a stopwatch to measure a time. But if we want to 'measure' a probability, the process is not so easy.

There are three ways of considering the probability of an event:

- With a **frequency** approach, we can repeat a situation many times and see how often a particular event occurs.
- With a **modelling** approach, we can use our knowledge of the situation (such as **symmetry**) to make **assumptions** about probabilities.
- With a **subjective** approach, you can decide on a probability by considering your belief or idea about what will happen in the situation.

Likely: Probable or possible

Unlikely: Not likely, doubtful

Fair: Unbiased, balanced

Frequency: The number of times something happens

Modelling: Making a mathematical model of a situation

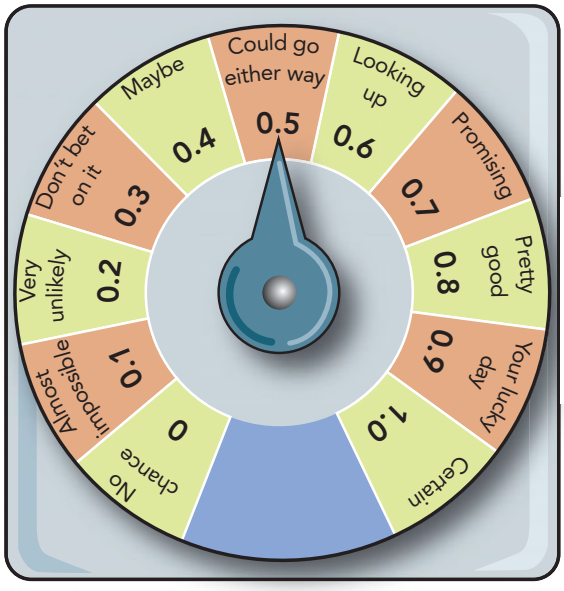
Symmetry: A situation that has repeated aspects that are the same

Assumption: Theory, guess or hypothesis

Subjective: From a particular person's point of view

ISBN 978-1-107-61421-5

None of these ways will always give you a 'correct' measurement of the probability. It isn't as easy as measuring a length, a weight or a time!



Key ideas

- Probability is a way of measuring chance.
- Values of probability are between 0 and 1.
- Probability can be estimated or given using frequency, modelling or belief.

Example 1: Tossing coins

a We toss a coin 10 times and get 5 heads. What can we say about the probability of heads? What if we tossed it more times and got 52 heads out of 100 tosses?

b Will we get a better measurement of probability if we toss a larger number of times?

c Explain why we are using the frequency approach to probability here? Could we use a modelling approach?

Solution

a If we get 5 heads in 10 tosses, we might want to say that the probability of heads is $\frac{5}{10} = 0.5$. But even if the coin is fair, in 10 tosses we could get a different number of heads. We could toss 10 times and get 3 heads, and we might not be happy to say that the probability of heads is 0.3. Maybe it would be better to toss more times. If we got 52 heads out of 100 tosses, we might want to say that the probability of heads is $\frac{52}{100}$, or 0.52.

ISBN 978-1-107-61421-5

b With more tosses, we are more likely to get a more accurate idea of the probability of heads. A South African statistician called John Kerrich spent time in a prison camp during the Second World War. To keep himself occupied he carried out some statistical activities. He tossed a coin 10 000 times and found that he got 5067 heads. For his coin (and his tossing) we estimate that the probability of heads is $\frac{5067}{10\,000} = 0.5067$.

c The previous results are based on tossing a coin many times, which is the frequency approach to probability. Alternatively, we could use the modelling approach. From the symmetry of a coin, we could say that 'heads' or 'tails' come up with a probability of 50% each. However, some coins are weighted so that they land heads more often, and some people are able to toss coins to get heads almost every time. So for a particular coin and the person tossing it, we may prefer to investigate the probability of getting heads using the frequency approach. This may give a better estimate of the probability.

Example 2: Probability of winning the game

Our netball team has a big game next weekend. How can we find the probability that our team will win?

Solution

This situation is quite different from the coin tossing example. We can play lots of games of netball, but we can't repeat next weekend's game lots of times! It will only happen once. Even if we play against the same opponents on another day, the game will be different. We may have different people in the team, the weather may be different, our preparation may be different, and we will know the result of the previous game that we played against the same team.

We can only estimate the chance that we will win using a subjective approach. This approach will be based on our knowledge of previous games that we have played against this team, beliefs about the strength and form of our team, and the opposition's, and the conditions for the match, for instance the weather, and how many supporters we expect to have at the game. We may decide that our team has a 70% chance of winning. But this is only an informed guess – other people may decide differently. For example, the opposition's coach may decide that we only have a 40% chance of winning.

Whichever way we estimate or measure chance, the probability is represented by a number from 0 to 1. The higher the number, the more likely the event is to occur. And we can represent the probability using fractions, decimals or percentages. So $\frac{1}{2}$, 0.5 and 50% all represent the chance of an event that is equally likely to occur or not – sometimes described by the phrase 'fifty-fifty'.

ISBN 978-1-107-61421-5

Example 3: Ways of showing probability

The table shows some events and their probabilities as fractions, decimals or percentages, and in terms of words. Fill in the missing cells in the table.

	Expression of probability			
Event	**fraction**	**decimal**	**percentage**	**informally in words**
You were born several years ago.	**a**	**b**	**c**	certain
Your team wins the district soccer cup.	$\frac{1}{5}$	**d**	**e**	**f**
It will rain tomorrow.	**g**	0.75	**h**	**i**
A tossed coin lands 'tails'.	**j**	**k**	**l**	fifty-fifty
You will live to be 100.	$\frac{1}{100}$	**m**	**n**	**o**
You will win an Olympic gold medal.	**p**	0.0001	**q**	**r**
You walk on the Moon without any special equipment.	**s**	**t**	0%	**u**

Solution

a 1 **b** 1.0 **c** 100% **d** 0.2 **e** 20% **f** unlikely

g $\frac{3}{4}$ **h** 75% **i** likely **j** $\frac{1}{2}$ **k** 0.5 **l** 50%

m 0.01 **n** 1% **o** quite unlikely **p** $\frac{1}{10\,000}$ **q** 0.01% **r** very unlikely

s 0 **t** 0 **u** impossible

Exercise 3A

1 Try these questions on different ways of representing probabilities.

a The probability that you toss three heads in a row is $\frac{1}{8}$. What is this probability as a decimal?

b The chance of getting your assignment finished on time is 0.8. What is this chance as a fraction?

c The probability that you will get a new computer for your birthday is 25%. What is this probability as a fraction and as a decimal?

2 Put these probabilities in order from least likely to most likely.

a 90% **b** 0.08 **c** $\frac{1}{5}$ **d** 0.1 **e** $\frac{1}{6}$ **f** 5%

3 In each case, decide which probability represents the event that is least likely:

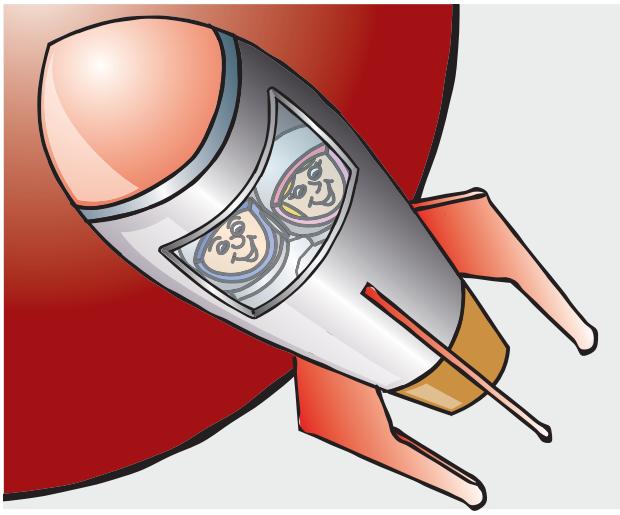
a 0.1, $\frac{1}{10}$, 1% **b** 0.025, $\frac{1}{8}$, 25% **c** 0.08, $\frac{1}{80}$, 18%

ISBN 978-1-107-61421-5

Cambridge University Press

4 Put these events in order of their chance, from least likely to most likely.

- **a** It will rain tomorrow.
- **b** You will visit the Moon during your lifetime.
- **c** You will finish high school after year 12.
- **d** It will rain for each of the next three days.
- **e** You will complete a university course.
- **f** You will visit the Moon and Mars during your lifetime.



5 You are practising your 3-point basketball shots from just outside the circle.

- **a** Out of your first 20 attempts, you make 12 successful shots. What is your probability of making a successful shot from just outside the circle?
- **b** Later you try 100 shots and are successful with 70 of them. What is your probability now? Has it changed from earlier? Why or why not?

Enrichment

Frequency, modelling or belief?
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

3-2 Equally likely outcomes

One situation in which we can easily **assign** probabilities is where there are a fixed number of **discrete outcomes** and they are all equally likely. If there are two outcomes that are equally likely, then each outcome will get a probability of $\frac{1}{2}$ or 0.5. For instance, if we toss a coin it can land showing heads or tails. If it is an ordinary coin and is tossed properly, these outcomes are equally likely, and so they will each have a probability of 0.5. If there are 10 outcomes that are equally likely, then each outcome gets a probability of $\frac{1}{10}$ or 0.1. If we ask a computer to give us a random number between 1 and 10, then there are 10 possibilities and each of them will have probability $\frac{1}{10}$. In each case, the total probability is 1, because something must happen, and this probability is shared equally between each of the outcomes.

Assign: Decide on particular values

Discrete outcomes: Can take only specific distinct possibilities

This is the simplest example of assigning probability using the modelling approach. Because the situation is **balanced**, we decide that all outcomes are equally likely. Such situations are often described using words such as 'fair' or 'balanced' – they imply that outcomes are equally likely. In order to apply the approach, we must describe the outcomes carefully and state clearly why we believe that they are equally likely. If we can do this, we get a situation where we can easily assign probabilities and use them to answer questions.

Balanced: All outcomes are equally likely

Key ideas

- If there are n equally likely outcomes, then each has probability $\frac{1}{n}$.
- We must describe outcomes carefully and say why we think they are equally likely.

ISBN 978-1-107-61421-5

Example 4: Rolling a fair die

What is the probability of getting a '6' when rolling a fair die? What about the probability of rolling a '1'?

Solution

A die is a cube with six faces, usually shown with 1, 2, 3, 4, 5 or 6 spots. When the die is rolled, the result of the roll is the face that is showing uppermost when it comes to rest. If the die is completely symmetrical, and is well shaken in a container before rolling, it is reasonable to assume that all six possible outcomes are equally likely, so we can assign the probability of $\frac{1}{6}$ to each outcome.

So the probability of rolling a 6 is $\frac{1}{6}$. More carefully, we can specify an event A = 'we roll a 6' and then write P(A) = $\frac{1}{6}$.

In this notation, P(A) means 'probability of A happening'.

We can specify other events, for instance B = 'we roll a 1', and since all the outcomes are equally likely, P(B) = $\frac{1}{6}$.

Example 5: Raffle tickets

We buy a single ticket in a raffle in which 100 tickets are sold. What is our chance of winning the first prize?

Solution

When the winning number for the raffle is selected, the ticket stubs are mixed up and one of them is drawn out randomly. If this process is carried out fairly, each number has the same chance of being selected. As there are 100 numbers, each of them will have a probability of 0.01 (or $\frac{1}{100}$) of becoming the winner. So our ticket will win with a probability of 0.01.

We should note that in each of these examples, we are making an assumption that all the outcomes are equally likely. For the first example, we assume that all the faces of the die are equally likely to turn up. This is indicated by our use of the term 'fair die'. What does the expression 'fair die' actually mean? It means that each of the faces is equally likely to turn up, in other words, the probability is the same for each face. In the raffle example, we assume that the winning number is selected in such a way that each number is equally likely to turn up. This is indicated by our statement that the winning number is 'randomly selected'. The expression 'randomly selected' means that each number is equally likely to be selected.

But how do we know that the die is fair? We might say that the die looks completely symmetrical; but we couldn't see from the outside if the die was 'loaded'. If a weight was put under the one spot, it would make the opposite face, the 6, more likely to turn

ISBN 978-1-107-61421-5

up. We can only check that the die is fair by rolling it many times and seeing how often each face turns up.

And how do we know that the winning number in the raffle is randomly selected? We might see how the process is carried out and notice that the ticket stubs are well mixed and the winning one is selected by a person who is not looking at them. But there are all sorts of ways in which the selection of the number could be 'fiddled'. If the stubs were put into a hat, one particular number could be pushed into the lining and later selected as the winner without looking. We can only check that the selection is occurring at random by doing it many times and seeing how often each number comes up.

If we roll a die 60 times and each number turns up 10 times, does that prove that the die is fair? Not necessarily, although it is a good indication. But even if we roll the die 60 times and get different frequencies for each of the numbers 1 to 6, that doesn't prove that the die is not fair. But the more different the frequencies are, the more suspicious we should be about the assumption that the die is fair.

Example 6: Frequencies rolling a die

Comment on these results of rolling a die 60 times.

a

Number obtained	1	2	3	4	5	6
Frequency	10	10	10	10	10	10

c

Number obtained	1	2	3	4	5	6
Frequency	12	6	8	11	14	9

c

Number obtained	1	2	3	4	5	6
Frequency	8	10	7	9	6	20

Solution

a This is what we would expect ideally, though we wouldn't be likely to actually see it. There is no evidence that the die is not fair (though we might suspect that the frequencies weren't actually obtained by rolling a die).

b This is a set of frequencies that could have easily come from actually rolling a fair die. The frequencies are not all the same, but they are not too different.

c This set of frequencies has a suspiciously high value for the 6. Maybe the die is weighted to favour the 6, or maybe the person rolling the die is doing something to try to get more 6s.

ISBN 978-1-107-61421-5

Exercise 3B

1 What is the probability of rolling a 4 with a fair die? What assumptions do you have to make to get this probability?

2 You have downloaded a music CD that has 14 tracks, and your iPod is set to 'shuffle', to play the tracks in a random order. What is the chance that the first track you hear was the first track on the original CD?

3 In the introduction to this chapter we referred to the Venice Lottery, where the number 53 had not turned up in almost two years of draws. If you select one number at random from the numbers 1 to 90, what is the probability that your selection is **not** 53?

4 When you pour drawing pins (thumb tacks) out of a box, they can land with their points facing upwards or downwards. Since there are only two possibilities, can we say that they are equally likely? How could you investigate whether the probability that a drawing pin lands point down was 0.5? How could you try to find the true probability?

5 A brand of soft drink is having a promotion. On the inside of the top of each bottle that you buy there is a coloured star – red, blue, yellow or green. Each time you collect a complete set of four colours, you get a prize. You and your friends decide to pool resources and over the course of two weeks you buy 40 bottles of the soft drink. What would be your comments if you got these distributions of colours:

a 12 red, 7 blue, 13 yellow and 8 green?

b 10 red, 10 blue, 10 yellow and 10 green?

c 13 red, 1 blue, 15 yellow and 11 green?

Enrichment

Is it a boy or girl?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

3-3 Equally likely lengths, areas or time periods

In the previous section we looked at situations where outcomes are categorical. Each outcome is one of a fixed number of distinct categories. We saw how to find the probability of each outcome if the categories are equally likely. But what can we do in cases where outcomes are **continuous**, such as points on a 10-kilometre road where an accident happens? Here each outcome is a value from a range of possibilities.

Continuous outcome: Can take any value in a specified range, e.g. height

Sometimes we want to talk about probabilities in situations where we are selecting a point from a length, an area or a time. For example, we might want to pick a point on a line, or a position on a map, or a time somewhere in a 2-hour period. In these cases we don't have a fixed number of separate outcomes, so we don't have an obvious way of applying the equally likely assumption.

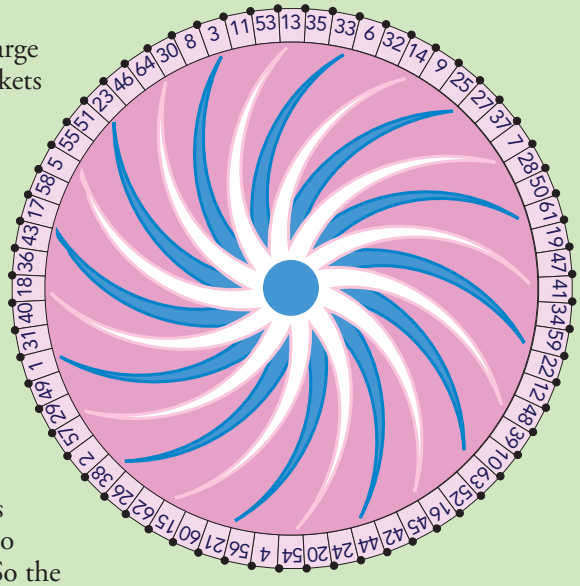
But we can divide the length, the area or the time into sections of the same size and use the equally likely model for each of the sections. There will be many different ways of doing this. For instance, we can divide a long stretch of road into 10-kilometre lengths, or into 2-kilometre lengths or into 1-kilometre lengths. Each choice will result in a different number of sections. Whatever our choice, we may need to investigate that the sections are really equally likely before we can use the model. We will give some examples that show how this works.

Key ideas

- We can divide continuous outcomes up into sections of the same size.
- We can use the equally likely model for the separate sections, but we may need to check whether the 'equally likely' assumption is reasonable.

ISBN 978-1-107-61421-5

Example 7: Wheel of Fortune



You can often see a Wheel of Fortune at shows or clubs. It is a large wheel, separated into numbered sections by pins. Numbered tickets are sold to raise money for some cause. The wheel is spun and a leather flap indicates which section is the winning section when the wheel stops spinning.

This Wheel of Fortune has 64 sections, and you can buy a ticket for any single number.

a What is your chance of winning if you buy a ticket for number 6?

b What is your chance of winning if you buy a ticket for number 64?

Solution

a There are 64 sections on the wheel. If it is well balanced it is equally likely to stop in any section, so it seems reasonable to assume that each section is equally likely to be the winner. So the chance of winning with ticket number 6 is $\frac{1}{64}$.

b The chance of winning with number 64 is the same, $\frac{1}{64}$, as long as the assumption of equally likely sections is true.

Example 8: Environmental surveys

A standard way of arranging an environmental survey is to divide an area into squares (usually 50 cm by 50 cm) called *quadrats*. Some quadrats are randomly selected and the plants growing in that quadrat are noted down. If an area 10 m by 10 m is being surveyed, what is the probability that a small Wollemi pine, the only one in the area, will be included in the first quadrat selected?

ISBN 978-1-107-61421-5

Solution

Since the area is 10 m by 10 m, it will be divided into 20 by 20 = 400 quadrats. If the quadrats are randomly selected, they will each be equally likely to be the first quadrat. So the probability of picking the first quadrat as the one with the Wollemi pine will be $\frac{1}{400} = 0.0025$.

Example 9: Goals in soccer

A professional soccer game lasts for 90 minutes, and we could choose to assume that goals are scored randomly during that time. If we know that the score in a game was 1-0, what is the chance that the only goal was scored in the last 10 minutes?

Solution

We can divide up the 90 minutes of the game into nine 10-minute periods. If goals are scored randomly, that means that they are equally likely to occur in any of the ten-minute periods. So the single goal of this game has a chance of $\frac{1}{9}$ of occurring during the last ten minutes.

In each of these examples, a length or area or time is divided into sections of the same size, and each of the sections is equally likely to contain the event that we are looking for. Notice the language that is used to indicate that this equally likely assumption is reasonable. The Wheel of Fortune is 'well balanced', the quadrats are selected 'at random', and the goals are scored 'randomly'. These phrases tell us that we can assume that the sections are equally likely.

If we are not sure of this assumption, we can repeat the process of selection many times and investigate the frequencies of each of the sections in the same way that we did earlier.

Exercise 3C

1 A spinner is divided into five equal sections, coloured red, yellow, green, blue and grey. If you spin the pointer, what is the probability that it comes to rest in the blue section?

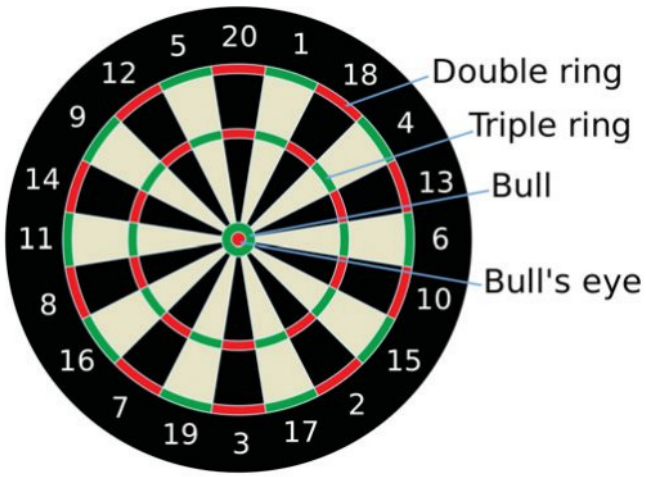
2 A mobile phone screen is divided into equal-sized regions for apps, four apps across and eight apps down. If you tap on the screen at random, what is the chance that you open the mapping app?

ISBN 978-1-107-61421-5

3 In a traffic accident study, a stretch of motorway 90 km long is divided into 10-km lengths. Records of accidents over the past year are checked to see where they occurred. What is the probability that the latest accident occurred in the middle 10-km section? What assumptions do you have to make to answer this question? How could you check these assumptions?

4 A dartboard is divided into 20 numbered sections scoring from 1 to 20, and also into circular regions. The green and red rings score double for the outer ring and triple for the middle ring. The bull's eye scores 50 and the bull scores 25. If you miss the board completely you don't score anything.

- **a** Are the numbered sections equally likely? How could you check this?
- **b** If you throw a dart and it lands in a black or white scoring section, what is the probability that you have scored 20?
- **c** If your throw lands in the outer ring, what is the chance that you have scored 38?
- **d** If you hit the triple ring, what is the chance that you score 33?



5 An aerial photo of an outback property is taken to check the location of a group of wild buffalo that have been destroying the plants. The photo shows an area 1 km by 2 km, and it is divided into squares 100 m by 100 m. The squares are examined carefully under magnification to find where the buffalo are resting. What is the probability that the animals are in the square at the top left of the photo? What assumptions do we have to make to answer this question, and how reasonable are they?

Enrichment

Catching the train
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

3-4 Probabilities of events proportional to 'size'

Up until now we have dealt with **simple events** in probability, such as rolling one die where we are interested in one outcome – that a certain number comes up. In some situations, the individual outcomes are equally likely, but we are interested in a **compound event** – one that is made up of several of these simple events or individual outcomes.

Simple event: Event consisting of a single outcome

Compound event: Event consisting of several individual outcomes

For example, when we roll a fair die we may want to find the probability of rolling a number that is greater than 4. The event 'rolling a number that is greater than 4' is a compound event that contains two individual outcomes, 'rolling a 5' and 'rolling a 6'. We can find the probability of this compound event by looking at the number of individual outcomes that make up the event as a fraction of all possible outcomes. In this case, there are two individual events that result in 'rolling a number greater than 4' out of the six possible outcomes, so the probability of rolling a number greater than 4 is $\frac{2}{6}$ or 0.333.

HINT
Remember, in probability, events can have different sizes, because they are made up of one or more outcomes. The more outcomes, the bigger the size of the event.

We can use the same approach with continuous outcomes if we are interested in events that have different sizes. For instance, on a map we could divide the land area of Australia into square centimetres. On our map, the total area of Australia is 76 square centimetres and the area of Western Australia is 25 square centimetres. If we pick a point in Australia at random, we can assume that each of the 76 square centimetres is equally likely to be selected. The probability of picking a point in Western Australia, in one of its 25 square centimetres, is $\frac{25}{76} = 0.329$.

In these sort of situations, we can find the probability of events by looking at the proportion of the equally likely outcomes that are contained in the event. We can extend the examples from the two previous sections to show how this works.

CAUTION
We are not talking here of a sequence of events, from several throws of the die. The compound event happens in one throw.

ISBN 978-1-107-61421-5

Key ideas

- In a situation with equally likely outcomes we can find the probability of a compound event by taking the proportion (fraction) of individual outcomes that make up the event:
  Probability of a compound event = $\frac{\text{number of outcomes that make up the event}}{\text{the total number of outcomes}}$
- In algebra this can be expressed as:
  If a compound event consists of k outcomes of the n equally likely outcomes, then its probability is $\frac{k}{n}$.
- The approach can also be used with continuous outcomes by dividing the outcomes into equal-sized sections.

Example 10: Raffle tickets again

Raffle tickets are being sold by your local sports group, and the first prize is a new mini tablet. One hundred numbered tickets were sold, 50 of them by Alice, 30 of them by Kim, and 20 of them by Chen. The ticket stubs are thoroughly mixed and one is picked out without looking to find the first prize winner. What is the probability that the winning ticket was sold by Kim?

Solution

Each of the hundred tickets is equally likely to be the winner, since the description shows that the winner was selected at random. As Kim sold 30 of the 100 tickets, the chance that the winning ticket was sold by her is found by looking at the proportion of tickets that she sold, $\frac{30}{100} = 0.3$.

Example 11: A jar of olives

A narrow-necked jar contains 3 green and 5 black olives some dressing). The jar is shaken and an olive is rolled out. What is the probability that it is green?

Solution

We can assume that each olive is equally likely to be rolled out from the jar, if it is well shaken and the olives are about the same size. Each individual olive has a chance of $\frac{1}{8}$ of being the first one shaken out. Since there are 3 green olives, the probability of getting a green one first is $\frac{3}{8} = 0.375$.

ISBN 978-1-107-61421-5

Example 12: Rolling two dice

If two fair dice are rolled, what is the probability of rolling a total of 7?

Solution

The basic outcomes from rolling two dice are shown by a pair of numbers, the number of spots on the first die and the number of spots on the second die. We can list the outcomes as:

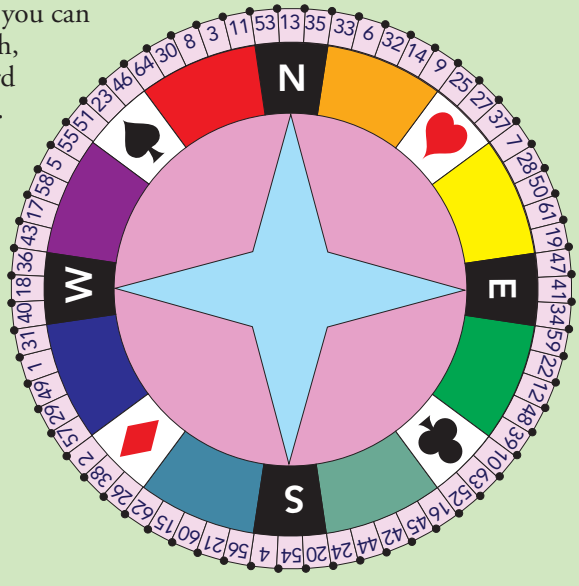
(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),
(5,1), (5,2), (5,3), (5.4), (5,5), (5,6),
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6).

Notice that it is good to show these outcomes in a systematic way. We have put all the outcomes with 1 on the first die in the first row, and with 1 on the second die in the first column, and so on with the other rows and columns. Notice also that even if the dice look exactly the same, there is still a 'first' die and a 'second die'.

Since the dice are fair, and if they are well shaken in a container before rolling, each of these outcomes will have the same chance, $\frac{1}{36}$. We can see that the outcomes (6,1), (5,2), (4,3), (3,4), (2,5) and (1,6) give a total of 7. Since there are 6 of these outcomes, the probability of rolling a total of 7 is $\frac{6}{36}$, or $\frac{1}{6}$.

Example 13: Wheel of Fortune

In this Wheel of Fortune, as well as buying tickets for numbers, you can buy a ticket for a colour or a point of the compass – north, south, east west – represented by the letters N, S, E, W or a playing card suit (♥ ♣ ♠ ♦). Each of these tickets wins with several numbers. What is the probability that you will win if you buy a ticket for yellow? What if you buy a ticket for south (S)?



Solution

There are 64 numbered sections on the wheel, and if it is well balanced and spun fairly fast then each of these will have equal probability of $\frac{1}{64}$. The yellow sector covers five numbers (7, 28, 50, 61 and 19), so the probability of winning with a ticket for blue is $\frac{5}{64}$. Since south (S) covers three numbers (20, 54 and 4), the probability of winning with south is $\frac{3}{64}$.

ISBN 978-1-107-61421-5

Example 14: Native forest

A map of part of a national park shows an area 10 km by 5 km. It is divided into squares with side 1 km, and 15 of these squares contain native forest. One of the squares is randomly selected for a detailed on-the-ground investigation. What is the probability that the selected square contains native forest?

Solution

The map is divided into 50 squares, each equally likely to be selected, since the selection is done randomly. Since there are 15 out of 50 squares that contain native forest, the probability of selecting one of them is $\frac{15}{50} = 0.3$.

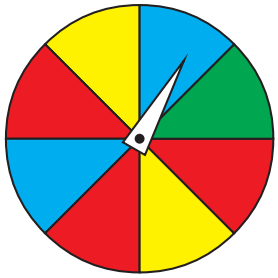
In general, the method shown in these examples is a good way of finding the probability of any compound event, as long as it is made up of outcomes that are (or can be assumed to be) equally likely. We count the number of outcomes that make up the event, and the total number of outcomes. The probability of the event is the fraction:

$$\frac{\text{number of outcomes that make up the event}}{\text{the total number of outcomes}}$$

ISBN 978-1-107-61421-5

Cambridge University Press

Exercise 3D



1 You roll a fair die.

a What is the probability that you get an even number?

b What is the probability that you will get a number divisible by 3?

2 A spinner has eight equal segments, 3 coloured red, 2 coloured blue, 2 coloured yellow and 1 coloured green. If you spin the pointer, what is the probability that it comes to rest in one of the blue sections?

3 You roll two fair dice.

a What is the probability that you get a total of 3 or less?

b What is the probability that you get a total of 10?

4 A bowl contains 25 jelly beans, 10 red, 12 green and 3 black, and we ask someone to mix the beans and select one without looking at the bowl. What is the probability that they pick a red jelly bean?

5 To start a game of Scrabble we randomly select one of the tiles. There are 100 tiles in the game, and they include 9 As. If we select an A, we will get the first go in the game. What chance do we have of selecting an A?

Enrichment

Selecting a random Australian
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

3-5 Investigating equally likely outcomes

We have pointed out that 'equally likely outcomes' is always an assumption. In some situations this assumption seems reasonable; for instance, in rolling a die or selecting a winning lottery ticket. In other situations, it is likely that the assumption doesn't work so well; for instance, in selecting a 'random number' between 1 and 10, or with road accidents along a stretch of highway. And there will be situations where we are not sure whether such an assumption is reasonable or not. For example, whether drawing pins land point up or point down, or goals scored during 10-minute segments of a soccer game.

The equally likely idea is a model that we use to find probabilities. The model and the probabilities that we get are only useful if the equally likely idea is realistic. So it is often important to check whether outcomes are equally likely, and we can do this by collecting some data from an experiment. The data let us compare reality with the model. If the results match the model then we conclude that it is useful. If the results are very different from the model, then we decide that it is not useful – outcomes are not equally likely. When we get extreme results it's easy to decide the model is not useful. It is harder to decide in borderline cases.

Key ideas

- The idea of equally likely outcomes is a very useful model for finding probabilities.
- We can find probabilities of compound events by finding the proportion of outcomes they include.
- Equally likely outcomes is always an assumption, so we may have to investigate it.
- To do this, we compare frequencies for each outcome with the equal frequencies that we expect if the model is correct.

ISBN 978-1-107-61421-5

Example 15: A jelly bean taste test

Can people determine the colour of jelly beans from taste alone? If we consider only two colours of jelly bean, say red and black, how can we investigate whether people can tell the difference between these two colours by taste?

CAUTION
Taste tests must always be done under close supervision of an adult with full consent of the subjects and their teachers.

Solution

Here is one possible way to carry out such an investigation. Volunteers (students in another class) are told that they will be given two jelly beans, one red and one black, in a random order, but they will not be able to see what the colour of each is. They have to taste each jelly bean and then tell us which one was black, the first or the second. They will either get this right or wrong. If they aren't able to tell which is which by taste, they will get the test correct 50% of the time. The two outcomes are equally likely. But if they can distinguish by taste they will get more correct than incorrect answers.

If we carry out the test with 30 students in a class, we would expect to get 15 correct and 15 incorrect tests if the colours can't be distinguished. If there were 17 correct and 13 incorrect results, we might think that this was still an indication that people could not tell the difference by taste alone. But if we found that 25 people got the correct result and 5 people got it wrong then we would decide that people can tell the difference in colour by taste alone.

There are some practical points to think about in this experiment. The way of hiding the colour from the volunteer must be reliable. We should make sure that the order of the jelly beans is randomly selected for each person, to avoid any possibility that if we talk about 'a red and a black jelly bean' people don't automatically think the first one they get is red and the second is black. Should we give people a practice run, and ask them to eat a red and a black jelly bean before they do the test? That might help them to make the correct choice. We could see if the number correct was higher this way. Should we test each person several times, or stick with one test per person? If they were told each time whether they had picked the black correctly or not, they might learn to do the test better, and this would increase the chance of success from the equally likely case.

ISBN 978-1-107-61421-5

Example 16: Selecting random numbers

Random selection is a very important idea in statistics, but it is not as easy to carry out as it first looks. How can we investigate how well people are able to select random numbers?

Solution

There are many possible ways, but here is one approach. We can visit a couple of other maths classes and ask everyone in each class to write down a random number between 1 and 100. If we group the numbers into tens (1 to 10, 11 to 20, ... 91 to 100) then each group of ten numbers should be equally likely to be selected. If we have asked 50 students, we would expect to get 5 numbers in the first ten, 5 in the next ten, and so on for each group of ten numbers. If the frequencies that we get are very different from these, that would imply that people have not chosen numbers as randomly as they thought – maybe people have picked obvious or favourite numbers (such as 100, or 50 or numbers between 1 and 10). But what would we say if we found these frequencies?

HINT
Remember, frequency refers to the number of times something happens.

Group	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	90–100
Frequency	11	2	3	6	8	4	3	1	3	9

The frequencies seem to show that each group of ten numbers is not equally likely.

ISBN 978-1-107-61421-5

Exercise 3E

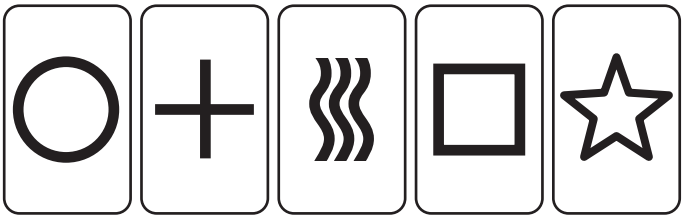
1 In the jelly bean taste test in Example 15, 40 students are tested and 22 correctly identify the black bean. What can you say about these results? What if 28 correctly identified the black bean?

2 Describe how you could carry out a study to check whether people are able to identify a black jelly bean from a group of three – one red, one green and one black. What are some of the practical points that you should be careful about?

3 Investigations into extrasensory perception (ESP) use a special set of cards each containing one of five symbols – circle, cross, waves, square or star. There are five cards of each type in the pack. These cards are called Zener cards after the psychologist who first designed them. An experimenter shuffles the cards, turns one of them over, and concentrates on that particular symbol. A subject (a person being tested for ESP) in another room writes down which symbol they think was selected.

a Is each type of card equally likely to be selected? Why or why not?

b What is the probability that the subject selects the correct symbol by chance?



4 Assume that the pack of Zener cards is shuffled after each selection, and the experiment is carried out 20 times.

a What would you think if the subject made 5 correct calls?

b What would you think if the subject made only 1 correct call?

c How many correct calls do you think would convince you that a subject did in fact have psychic powers?

Enrichment

A medical study on diabetes
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

Chapter summary

What is probability?

- Probability is a way of measuring chance
- Values of probability are always between 0 and 1
- Probability can be estimated or assigned using frequency, modelling or belief
- The frequency approach is based on repeating a situation many times to find the proportion of times that a particular event occurs
- The modelling approach is based on some assumption about the outcomes, e.g. that they are equally likely
- The belief approach uses subjective information about the situation.

Situations with all outcomes equally likely

- If there are n equally likely outcomes then each has probability $\frac{1}{n}$
- We must describe outcomes carefully and say why we think they are equally likely.

Equally likely lengths areas or time periods

- With continuous outcomes, we can divide them up into sections of the same size
- We can use the equally likely model for the separate sections
- We may need to check whether the assumption is reasonable.

Probabilities of events proportional to 'size'

- If outcomes are equally likely, we can find the probability of a compound event by taking the proportion of individual outcomes that make up the event
- If a compound event consists of k outcomes of the n equally likely outcomes, its probability is $\frac{k}{n}$
- With continuous outcomes divide the outcomes into equal sized sections.

Investigating the equally likely assumption

- Equally likely outcomes is a very useful model for finding probabilities
- Equally likely outcomes is always an assumption
- Investigate by comparing frequencies for each outcome with the equal frequencies that we expect if the model is correct.

Multiple-choice questions

1 The probability that a computer file is corrupted is 2.5%. As a decimal, this probability is:

A 2.5 **B** 0.25
C 0.025 **D** 0.0025

2 Which of these represents the smallest probability?

A $\frac{1}{20}$ **B** 2%
C 0.2 **D** $\frac{1}{2}$

3 We buy a single ticket in a class sweep in which 25 tickets are sold. What is the chance that we win the third prize?

A 25% **B** $\frac{1}{3}$
C $\frac{1}{25}$ **D** 0.033

ISBN 978-1-107-61421-5

4 We buy a ticket in a class sweep in which 25 tickets are sold. What is the probability that we don't win any of the three prizes?

A 0.7 **B** 0.67

C 0.12 **D** 0.88

5 A book is selected at random from a shelf that contains 4 novels, 3 textbooks and 2 dictionaries. What is the chance that a dictionary is selected?

A $\frac{2}{9}$ **B** $\frac{1}{3}$

C $\frac{1}{9}$ **D** $\frac{7}{9}$

6 A roulette wheel has 37 numbered sections. What is the probability of winning with a 'corner' – a bet on four adjacent numbers?

A $\frac{1}{2}$ **B** $\frac{4}{37}$

C $\frac{18}{37}$ **D** $\frac{33}{37}$

7 We have received a report that an accident has occurred on a straight 20-km stretch of highway, but we don't know where the accident has occurred. Assuming that the accident could have happened anywhere with the same chance, what is the probability that is more than 5 km from either end (so it will take the emergency services more than 3 minutes to get there)?

A 0.25 **B** 0.33

C 0.5 **D** 0.75

8 Two fair dice are rolled. What is the probability that the total obtained is 12?

A $\frac{1}{6}$ **B** $\frac{1}{12}$

C $\frac{1}{3}$ **D** $\frac{1}{36}$

ISBN 978-1-107-61421-5

9 A number is randomly selected from the whole numbers 1 to 50. What is the probability that it is divisible by 7?

A 0.14 **B** 0.033
C $\frac{1}{25}$ **D** $\frac{1}{7}$

10 A person rolls a fair die 30 times using a container to shake the die before each roll. The information below shows the number of times they obtained each of the results 1, 2, 3, 4, 5 and 6. Which set of results would you be least surprised to see?

A (5,5,5,5,5,5) **B** (6,5,4,5,5,4)
C (3,8,7,4,3,5) **D** (8,1,3,4,3,11)

Short-answer questions

1 You drop a piece of buttered toast from the breakfast table.

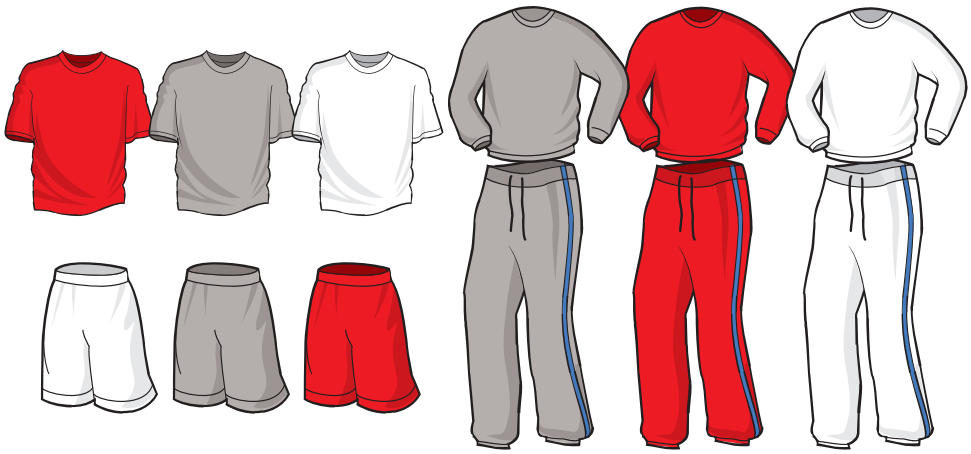
a List the possible outcomes of this experiment.

b Do you think that all of the outcomes are equally likely? How could you check?

2 A school sports team is choosing their uniform. There are three basic items – t-shirt, shorts, track suit – and to keep costs low, each of these will be in one of the school colours – grey, red and white.

a List the possible colour combinations for the uniform.

b If one of these is selected randomly, what is the chance that the t-shirt and shorts will be in the same colour?



3 On a TV quiz game, a contestant who has answered a question correctly selects one door from 16 numbered doors. Behind one door is the major prize – a car.

a What is the probability that she selects the major prize – the car?

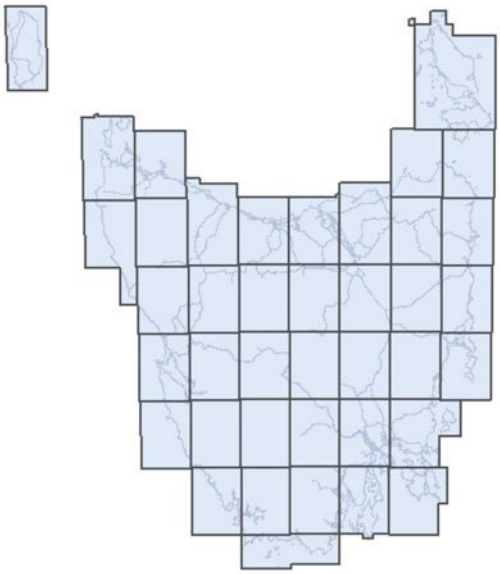
b What assumptions do you have to make to find an answer to this question?

4 There are five pairs of socks in a drawer, each pair a different colour. You put in your hand and select two socks without looking. What is the probability that you get a matching pair?

HINT
Imagine that you pick the socks out one at a time.

ISBN 978-1-107-61421-5

5 This diagram shows 40 TASMAP maps that cover Tasmania. One of these maps is randomly selected for a quality check.



Base image reproduced with the permission of TASMAP (www.tasmap.tas.gov.au) © State of Tasmania.

a What is the probability that the selected map is of the Freycinet Peninsula?

b What is the probability that the selected map shows one of the large northern islands (King, Flinders or Cape Barren)?

c What is the probability that the selected map includes no coastline at all?

Extended-response questions

1 Consider the 'experiment' of having a baby. What is the probability that it will be a boy? Find three different answers, using the three different interpretations of probability, and explain how you would justify these answers.

2 In 960 CE, Bishop Wibold of Cambrai (in northern France) worked out a way to allow his monks to gamble! He made a list of the 56 possible outcomes from rolling *three dice without regard to order*. Against each result he listed a virtue. For instance, against the result (3-3-5) he could have written honesty, and against (1-2-6) he could have written charity (his actual list has not survived). Each monk was allowed to pick up three dice, roll them on a table and then spend the rest of the day meditating on the virtue that corresponded to their result.

a Write down all the possible outcomes on the bishop's list. How will you make sure that you don't miss any possibilities?

b Do you think that each of these outcomes is equally likely? How could you investigate this question theoretically? How could you answer it practically?

3 Write three questions about probability using the material in this chapter – one multiple-choice question, one short-response question and one discussion question. Your fellow students should be able to answer the questions easily, but not too easily! Include answers, or for the third question, some suggested points for an answer.

ISBN 978-1-107-61421-5

Cambridge University Press

CHAPTER

4

Probabilities and language connectors

What you will learn

Introduction to probabilities and language connectors

Is there a relationship between the amount of soft drink that school students drink and their behaviour at school? Some studies have suggested that a diet high in sugar is associated with more aggressive behaviour. It may be that the extra sugar intake replaces healthier foods and leads to lack of essential nutrients, or it may be that abnormally low blood-sugar levels cause irritable and violent behaviour, as well as a tendency to consume more sugar. Ideas of probability can be used to investigate the problem.

An article in *The Sydney Morning Herald* reported on the results of a survey carried out with teenagers aged 14 to 18 in state schools in Boston, USA. Participants were asked many questions, including how many non-diet fizzy soft drinks they had drunk during the past week. Up to 4 cans consumed was considered 'low' and 5 or more cans was considered 'high'. They were also asked whether they had been violent towards their fellow students. This was defined as 'got into a physical fight with another child ... or pushed, shoved, slapped, punched, kicked or choked him or her, or attacked or threatened the other child with a weapon'. <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

From the results of the survey we can estimate the probability that one of these school students had a high soft-drink consumption: this probability

AUSTRALIAN CURRICULUM

Statistics and probability

- Chance
- Identify complementary events and use the sum of probabilities to solve problems **(ACMSP204)**
- Describe events using language of 'at least', exclusive 'or' (A or B but not both), inclusive 'or' (A or B or both) and 'and' **(ACMSP205)**
- Represent events in two-way tables and Venn diagrams and solve related problems **(ACMSP292)**

ISBN 978-1-107-61421-5

was about 0.3. We can also estimate the probability that a student had been violent towards their peers: this probability was about 0.44 overall, but was much higher, 0.57, for those students who drank more soft drinks.

How can we find the probabilities of single events, such as 'high soft-drink consumption' or 'violence towards peers' from such a survey? How can we find the probabilities of combinations of events, such as 'high soft-drink consumption and violence towards peers'? And how can we use such probabilities to investigate the possible link between diet and behaviour? With our answers, we can help to make our schools safer and more pleasant places to be in. But before we do that, we need to investigate how to work with probabilities of more than one event at a time, the topic of the current chapter.

Words you will learn

'A and B'
'A or B'
complementary events
disjoint
event
exclusive or
experiment (in probability)
inclusive or
intersection (on a Venn diagram)
mutually exclusive
sample space
two-way table
Venn diagram

Pre-test

1 Write these probabilities as decimals:

a 55% **b** $\frac{7}{8}$ **c** $\frac{3}{10}$

2 If an event is impossible, what numerical value is used to represent its probability?

3 An office Melbourne Cup sweepstake has 22 tickets, one for each of the horses. (In a sweepstake the tickets with the named horses are sold before the race and the money goes into a 'pot'. The person holding the ticket with the name of the horse that wins, wins the pot.) If you buy three tickets for horses randomly selected from the field, what is your probability of winning first prize?

4 You know that some treasure has been buried in a plot 50 m by 20 m. You divide the plot into square metres, pick five of them at random and excavate. What is the probability that you will find the treasure?

5 What would you say if a friend tossed a coin 20 times and obtained 18 heads?

ISBN 978-1-107-61421-5

4-1 Adding probabilities

We start with the definition of three important terms. These are technical terms in probability. The meaning of the words is slightly different from their normal English meaning.

An **experiment** in probability is any situation where we don't know what will happen. An experiment does not have to be something carried out by scientists. In probability, the term is used for any situation where the result is uncertain. An experiment could be something quite ordinary, such as rolling a die or choosing a colour from the rainbow.

Experiment (in probability): Any situation where the result is uncertain

Sample space: A list of possible outcomes

The **sample space** represents all the possible outcomes from an experiment. It could be shown as a list. For instance, {1, 2, 3, 4, 5, 6} represents the sample space for rolling a die, and {red, orange, yellow, green, blue, indigo, violet} represents the sample space for selecting a colour from the rainbow.

The sample space could be described in words instead. For instance {18 years or older} represents the sample space for the age of an Australian voter. Since the sample space shows all the possible outcomes, it has a probability of 1. This indicates that something from the sample space must occur.

An **event** is anything that can actually happen in an experiment. An event is part of the sample space, either a single outcome or a combination of outcomes. The probability of an event depends on the probability of the outcomes that it contains. If the outcomes are all equally likely, then the probability of an event is the proportion of the outcomes that it contains.

Event: Anything that can actually happen in an experiment

We can look at two events defined on the same sample space. For instance, 'rolling a 6' and 'rolling a number less than 4' are two events in the sample space for rolling a die. Also '25 years and older' and 'younger than 30 years' are two events in the sample space for the age of an Australian voter.

If the two events don't contain any points in common, they are called **mutually exclusive** (or **disjoint**). For instance, 'rolling a 6' and 'rolling a number less than 4' are mutually exclusive events. But '25 years and over' and 'under 30 years' are not mutually exclusive as they contain points in common, the ages 25, 26, 27, 28 and 29.

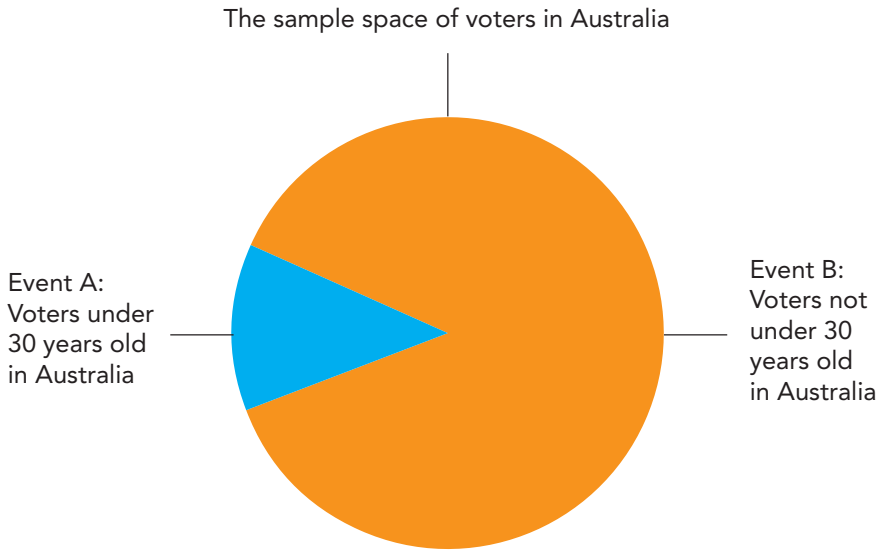
Complementary events are mutually exclusive events that together cover the whole sample space. For instance, 'rolling a number less than 4' and 'rolling a number 4 or greater' are complementary events in the sample space for rolling a die.

An easy way of getting two complementary events is to take an event and then take its opposite. For instance, 'younger than 30' and its opposite 'not younger than 30' (which is the same as '30 years and older') are complementary events for the age of an Australian voter. The probabilities of two complementary events add up to 1, since together they cover the whole sample space. This can be illustrated in a diagram:

Mutually exclusive (or **disjoint**) **events:** Two events that don't contain any outcomes in common

Complementary events: Two events that are disjoint and cover the whole sample space

ISBN 978-1-107-61421-5



Complementary events A and B are mutually exclusive and completely fill the sample space, which has a total probability of 1.

Key ideas

- An experiment in probability is any situation where we don't know what will happen.
- The sample space represents all possible outcomes of an experiment.
- An event is part of the sample space.
- Two events are mutually exclusive if they don't contain any outcomes in common.
- Two events are complementary if they are mutually exclusive and cover the sample space completely.

Example 1: Your future family

Stickers showing the members of your family can be bought at most newsagents and are very popular on the back window of cars. Suppose you have decided that, sometime in the future, you will have a family with three children.

a You don't know in advance whether you will get boys or girls, but can you write down the sample space?

b Assuming that each child is equally likely to be a boy or a girl, what is the probability that you will have children of both sexes?

c What is the probability that you get a majority of girls?

ISBN 978-1-107-61421-5

Solution

a You could show the sample space using three letters, b for boy or g for girl, for your three children in the order they arrive: {bbb, bbg, bgb, gbb, ggb, gbg, bgg, ggg}.

If we assume that each child is equally likely to be a boy or a girl, these 8 outcomes are equally likely, so they will each have probability of $\frac{1}{8}$. This may not be quite correct because, as we saw in the previous chapter, in some populations boys are slightly more likely than girls, but the model will be accurate enough for the moment.

b The event A = 'children of both sexes' contains six outcomes, bbg, bgb, gbb, ggb, gbg and bgg, so the probability that you get children of both sexes is $P(A) = \frac{6}{8} = 0.75$.

c A majority means more than half. The event B = 'you get a majority of girls' contains four outcomes, ggb, gbg, bgg and ggg, so the probability that you get a majority of girls is $P(B) = \frac{4}{8} = 0.5$.

Example 2: Mutually exclusive events

Give two examples of mutually exclusive events using the family sample space in the previous example.

Solution

If A = 'children of both sexes' and C = 'three boys', then A and C have no points in common, so they are mutually exclusive.

If B = 'majority of girls' and C = 'three boys', then B and C are mutually exclusive as they have no points in common.

Example 3: Complementary events

Give two examples of events that are complementary using the family sample space.

Solution

If A = 'children of both sexes', and D = 'all three children of the same sex', then A and D are complementary events. Between them, A and D contain all the outcomes, but they have no outcomes in common.

If B = 'majority of girls', and E = 'majority of boys', then B and E are complementary events. They have no points in common, but they contain all the outcomes, because with three children you must have a majority of boys or a majority of girls.

ISBN 978-1-107-61421-5

Exercise 4A

1 Give the sample space for each of these experiments:

a Plant a tomato seed **b** Serve in a tennis game **c** Drop a buttered piece of toast

2 In a small survey, three people were asked if they were in favour of building an extra runway at an airport. Write the sample space as S = {yyy, yyn, yny, nyy, nny, nyn, ynn, nnn}, where y = yes and n = no. Describe these events in words:

a ynn **b** nny, nyn, ynn **c** yyy, yyn, yny, nyy

3 For the airport survey in the previous question, write down the points in the sample space that make up each of these events:

a All three people were against the proposal.

b A majority were against the proposal.

c All three people gave the same answer.

4 A roulette wheel contains 37 slots numbered 0 to 36; 18 of them are red, 18 are black and one is green (number 0).

a Write down the sample space of colour for the result of one spin.

b If you bet on red, what is the probability that you will win?

c What is the complementary event to red coming up, and what is its probability?

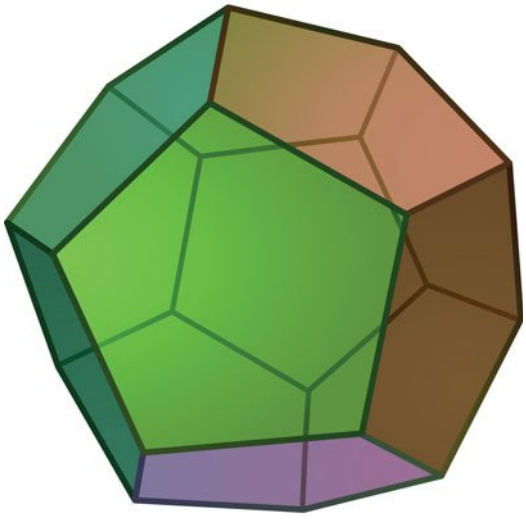
5 A dodecahedron is a regular (symmetric) solid with 12 faces in the shape of pentagons. A die is made from a dodecahedron with faces numbered 1 to 12.

a Explain why this die will be fair.

b Write down the sample space for a single roll of this die.

c What is the probability that the number obtained will be no more than 4?

d What is the complementary event and what is its probability?



Enrichment

Swimming finals
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

4-2 Combining events with 'and' and 'or'

When we look at two events in the sample space, we often want to find the probability that both events occur. For instance, we might ask: What is the probability that the Strikers score and the Wanderers score? Here we want to find the probability of 'A and B' – the word 'and' specifies that both events have to occur.

With two events in the sample space, we may want to find the probability that one of them occurs. For instance, we might ask: What is the probability that the Strikers score or the Wanderers score? Here we want to find the probability of 'A or B' – but what do we mean when we say that?

The expression 'A or B' can refer to 'exactly one of A and B', called the **exclusive or**. In this case, both events are not included. We would be interested in the probability that the Strikers score, or the Wanderers score, but not both teams.

Alternatively, 'A or B' can refer to 'at least one of A or B', called the **inclusive or**. We would be interested in the probability that the Strikers score, or the Wanderers score, or both teams score.

Exclusive or: Exactly one of the events occurs

Inclusive or: At least one of the events occurs

When we see the expression 'A or B' we have to be careful to decide which interpretation is correct, or which is more appropriate to use. 'Is the jelly bean red or black?' uses the 'exclusive or', as the bean can't be both red and black at the same time. 'Can you play the clarinet or the flute?' uses the 'inclusive or', and you would answer 'yes' if you could play both instruments. But sometimes, the expression 'A or B' can be interpreted in either way, and then we must be clear about which one we choose.

If A and B are complementary events, then 'A or B' can only have one meaning, as A and B can't occur together. For instance, 'Is the baby a boy or a girl?' is easy, as these are the only two possibilities for one baby, and so P(boy or girl) = 1. And as they are complementary events, P(boy and girl) = 0, if we are talking about only one baby.

ISBN 978-1-107-61421-5

Key ideas

- The phrase 'A and B' specifies that both of the events A and B occur.
- The phrase 'A or B' can mean 'exactly one of A and B' occurs ('exclusive or').
- The phrase 'A or B' can mean 'at least one of A and B' occurs ('inclusive or').
- We have to be careful to specify our interpretation of 'A or B'.
- If A and B are complementary events, P(A or B) = 1 and P(A and B) = 0.

Example 4: Athens Olympics

The first modern Olympic Games took place in Athens, Greece in 1896, following a gap of more than 1500 years in which there had been no Olympic Games. There were 177 officially registered competitors, of which 98 – more than half – were from the host country, Greece. First prizes (in the form of silver medals) were awarded to 41 athletes, including 10 Greeks. From the official records of the Games, we select one of the athletes at random for a historical investigation.

a What is the chance that our selected athlete was Greek? What is the chance that he won a first prize? (Note that all the athletes were men. Women were first allowed to participate at the following Olympic Games in Paris, 1900.)

b What is the chance that our selected athlete was Greek and won a first prize?

c What is the chance that he was Greek or he won a first prize?

Solution

a Let us use G = 'Greek' and F = 'first prize'. Since we select an athlete at random, each of the 177 is equally likely to be selected. So P(G) = $\frac{98}{177}$ = 0.554, and P(F) = $\frac{41}{177}$ = 0.232.

b Since there were 10 Greek first prize winners, P(G and F) = $\frac{10}{177}$ = 0.056.

c We want to find P(G or F), and it seems reasonable to use the 'inclusive or', to find the probability that the athlete is Greek, or won a first prize, or both. There were 88 Greek athletes who didn't win a first prize, 31 first prize winners who were not Greek, and 10 Greek first prize winners – a total of 129 people. So P(G or F) = $\frac{129}{177}$ = 0.729.

If we interpreted the question using the 'exclusive or', we would find the probability that the athlete was Greek or that he won a first prize, but not both. Here P(G or F) = $\frac{(88 + 31)}{177}$ = 0.672. This interpretation doesn't seem as good.

ISBN 978-1-107-61421-5

Example 5: Further investigation of the Athens Olympics

Use the information in the previous example with an athlete selected randomly from the official competitors at the Athens 1896 Olympic Games. Define two events that are complementary. Find the probability that one of these events occurs, and find the probability that both of them occur.

Solution

Taking the event G = 'Greek', we can find the complementary event by taking its negation, that the athlete is not Greek: we could write V = 'Visitor'. Then G and V are complementary events since they don't contain any people in common, and they include all people.

The probability that one of the events occurs is P(G or V). In this case, the 'exclusive or' is appropriate, and P(C or V) = 1, since every competitor was either Greek or a visitor. The probability that both events occur is P(G and V) = 0, since no competitor was Greek and a visitor.

Exercise 4B

1 Which interpretation of 'or' is more appropriate in each of these situations? Explain your answer in each case.

- **a** A loaf of bread contains wheat or rye flour.
- **b** We would like to win first or second prize in a lottery in which we have bought a ticket.
- **c** The school musical is looking for people who can act or sing.
- **d** Either Norths or Easts will win the grand final this season.

2 A letter is selected at random from the 26 letters of the English alphabet. Define three events: A = 'a vowel (a,e,i,o,u) is chosen', B = 'a consonant is chosen', C = 'a letter from a to g (a,b,c,d,e,f,g) is chosen'. Find:

a P(A and C) **b** P(A or B) **c** P(A or C)

3 You have to choose two elective subjects to study next year: the choices are literature, art, history, music, French or Mandarin.

- **a** Write down the sample space for your choices, assuming that the order of choice is not important.
- **b** Write down two events that are mutually exclusive and use them to illustrate the 'exclusive or'.
- **c** Write down two events that are not mutually exclusive and use them to illustrate the 'inclusive or'.

ISBN 978-1-107-61421-5

4 At the Athens 1896 Olympic Games, 14 of the 177 participants were from the USA, and 8 of them were amongst the 41 athletes who won first prizes. One of the participants is selected at random.

a What is the probability that the participant was from the USA and they won a first prize?

b What is the probability that the participant was from the USA or they won a first prize?

5 An industrial spy uses the internet to send a message to her controller. The words of the message don't matter – the important thing is exactly when the message was sent. (The time of sending appears at the top of the email when it is received). If the message was sent in the first half of the hour it means 'investigation going well'. If it was sent in the first or last quarter of the hour it means 'I have discovered something interesting'. If it was sent during the first five minutes of any of the quarter hours it means 'I am in danger'.

a What can you conclude from a message sent at 3:22?

b What can you conclude from a message sent at 3:32?

c Give a time that means her investigation is going well and she has discovered something interesting.

d Give a time that means she has discovered something interesting but she is in danger.

Enrichment

Australia's Prime Ministers
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

Cambridge University Press

4-3 Two-way tables of words and data

A survey of Year 7 students carried out during Friday classes asked each student whether they watched any television on the previous day and whether they did any homework on the previous day. We can summarise the results in a **two-way table**:

Two-way table: A table that can be used to summarise data on frequencies of two events; can be used to estimate probabilities

	Did not do homework	Did homework	Total
Did not watch TV	7	13	20
Watched TV	18	12	30
Total	25	25	50

We can use the information to estimate the probability of various events. The total row and column give us information about individual events. For instance, if A is the event 'watched TV', then:

$$P(A) = \frac{30}{50} = 0.6$$

estimates the probability that a randomly selected student in Year 7 watched TV on the previous day.

If B is the event 'did homework', then

$$P(B) = \frac{25}{50} = 0.5$$

estimates the probability that a student in Year 7 did any homework on the previous day.

The numbers in the middle cells of the table give us information about combinations of events. For instance, P(A and B) represents the probability that a student watched TV

ISBN 978-1-107-61421-5

and did homework. We can estimate this probability as $\frac{12}{50} = 0.24$, since 12 out of the 50 students did both of these things.

We can ask another question: 'What is the probability that a student watched TV or did homework?' We recognise this as P(A or B), which is a question that could have two meanings (it is ambiguous). If we use the 'inclusive or', we are asking about at least one of these events – watching TV or doing homework or both – then we estimate the probability as

$$\frac{(18 + 13 + 12)}{50} = \frac{43}{50} = 0.86.$$

If we use the 'exclusive or' we are asking about exactly one of these events – watching TV or doing homework but not both. Then we estimate the probability as

$$\frac{(18 + 13)}{50} = \frac{31}{50} = 0.62.$$

The information in the table comes from a survey of the 50 students in Year 7 who were at school on the day of the survey. So the results that we get are estimates of probabilities for a student selected at random from this group. The results may not be accurate for all the students at the school, or for Year 7 students generally. And the results might be different if the survey was carried out on another day. For instance, if the survey was taken on Monday then 'the previous day' would be on the weekend. We should keep practical information such as this in mind whenever we are working with statistical information.

Key ideas

- A two-way table of frequencies can be used to show information about two variables.
- The frequencies can be used to estimate probabilities of individual events or combinations of two events for the group represented.

Example 6: Road deaths in Australia

The table shows the number of deaths on the roads in Australia during 2012, classified by sex and road user group of the victim (the 'biker' group includes motorbike riders, passengers and cyclists).

	Vehicle driver	Passenger	Pedestrian	Biker
Male	463	123	117	233
Female	151	136	56	23

a How many people died on the roads in Australia during 2012?

b What is the probability that a road victim in 2012 was female? What is the complementary event and its probability? Comment on the value.

c What is the probability that a victim was a female and a biker, and what is the probability that a victim was a male and a biker? Comment on the difference between these values.

ISBN 978-1-107-61421-5

d Consider the probability that a victim was female or a pedestrian. Which interpretation of 'or' would be more appropriate here? What is the value of this probability? What is the complementary event and its probability?

Solution

a The overall total of the frequencies in the table is 1302 people, representing the total number of deaths. (In fact, there were also 3 more people with missing information who were not able to be included in the table.)

b There were 366 female victims, so the probability is $\frac{366}{1302} = 0.281$. The complementary event is that the road victim was not female, that is, that the road victim was male; there were 936 male victims. This event has probability $\frac{936}{1302} = 0.719$, or equivalently $1 - 0.281 = 0.719$. There were about two-and-a-half times as many male victims as female victims.

c The probability that a victim was a female biker is $\frac{23}{1302} = 0.018$, and a male biker is $\frac{233}{1302} = 0.179$. There were more than ten times as many male as female biker victims.

d It seems more reasonable to use the 'inclusive or' to find the probability that a victim was female, a pedestrian, or both. There were 366 females, and 117 male pedestrians, so this probability is $\frac{(366 + 117)}{1302} = 0.414$. The complementary event is that 'a victim was a male and a vehicle user'. The probability of this event is $1 - 0.414 = 0.586$.

Exercise 4C

1 A survey of secondary school students found that 23 out of 35 smartphone users and 15 out of 40 users of other phones were generally happy with the features of their phone.

a Show this information in a two-way table.

b What is the probability that a survey respondent had a smartphone and was unhappy with the features?

c What is the probability that a survey respondent didn't have a smartphone or was unhappy with the features of their phone?

d What is the complementary event of the event described in part **c**?

2 Triskaidekaphobia is a fear of the number 13. Many people suffer from this fear, and some are especially worried if Friday falls on the 13th day of the month. One theory is that this can be traced back to the arrest and subsequent killing of the Knights Templar on Friday 13 October 1307. These knights patrolled the roads of Jerusalem and the Holy Land to protect pilgrims. For the year 2015, this table allows you to assess your chances that you are indeed triskaidekaphobic.

	13th	Other date	Totals
Friday	3	49	52
Other day	9		
Totals	12		365

ISBN 978-1-107-61421-5

a Fill in the missing frequencies in the table.

b If a day is selected at random during the year 2015, what is the probability that it will be the 13th of the month?

c What is the probability that it will be the 13th day of the month and a Friday?

d What is the probability that it will be the 13th day of the month or a Friday? Would you use the 'inclusive or' or the 'exclusive or' to answer this question? Explain why.

3 An investigation of fatal road traffic accidents classified each by the speed limit at the site of the crash and how many vehicles were involved. The following table summarises the results.

	Single vehicle	Multiple vehicles	Totals
90 km/h or less	6	7	13
100 km/h or more	8	5	13
Totals	14	12	26

For an accident selected at random from the investigation:

a What is the probability that multiple vehicles were involved?

b What is the probability that a single vehicle was involved and the speed limit was 90 km/h or less?

c What is the probability that multiple vehicles or a higher speed limit was involved?

4 High GI diets are associated with a number of health problems. (GI means 'glycaemic index'.) A medical study investigated the effect during pregnancy of eating a low GI diet or a high GI diet on the chances of giving birth to a baby that was 'large for gestational age' (LGA). This means they were above a certain weight at birth, which can cause complications during birth. The following results were obtained:

	Not LGA	LGA	Totals
Low GI	31	1	32
High GI	20	10	30
Totals	51	11	62

ISBN 978-1-107-61421-5

a For a mother randomly selected from this group, what was the probability of giving birth to a LGA baby?

b For a mother randomly selected from this group, what was the probability of being on a low GI diet and giving birth to a LGA baby?

c For a mother randomly selected from this group, what was the probability of being on a high GI diet and giving birth to a LGA baby?

d What can you conclude from the answers to **b** and **c**?

5 A study examined the relationship between survival of patients after heart surgery 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 after surgery. The table below shows the data collected.

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

a What is the probability that a patient had a pet?

b What is the probability that a patient survived for at least a year after surgery?

c What is the probability that a patient had a pet and survived at least a year after surgery?

d What is the probability that a patient died or did not have a pet?

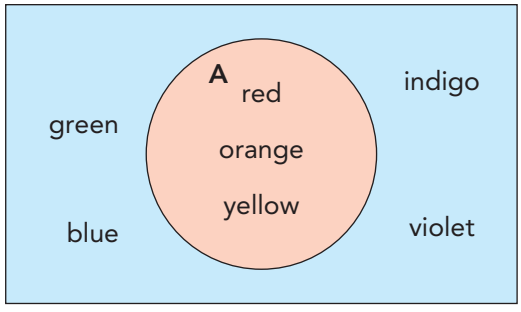
Fizzy drinks and aggression
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

4-4 Venn diagrams

A **Venn diagram** is useful way to show a sample space and events. The rectangle represents the sample space, and curves (usually circles) are used to represent events. The simplest case is if we are dealing with only one event. For example, here is a Venn diagram showing the sample space for a colour selected at random from the rainbow, and the event A that the colour was 'warm' (which we define as red, orange or yellow). The individual outcomes are shown in the rectangle (including the circle) and a circle specifies the event.

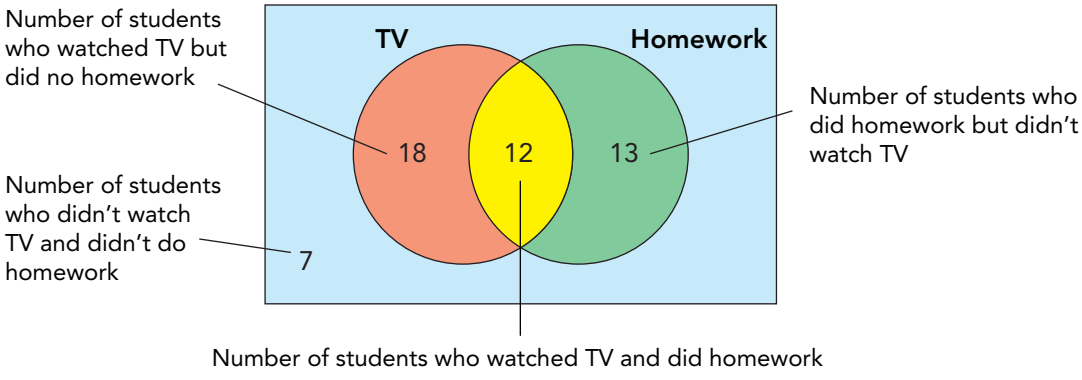
Venn diagram: A diagram that shows data as circles, possibly interlocking, inside a rectangle; can also be used to show events



Since each of the colours is equally likely, the probability for each is $\frac{1}{7}$. The probability of getting a 'warm' colour is $\frac{3}{7}$, since the event A contains 3 of the 7 colours.

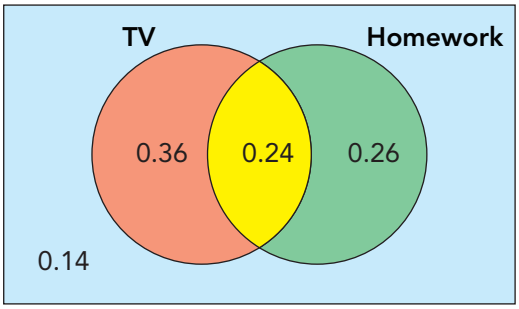
But a Venn diagram is much more useful for two events. Each event is represented by a circle, and the **intersection** (overlap) of the circles shows that both events occurred. Here is the Venn diagram for the TV and homework survey from the previous section, with explanatory labels.

Intersection (on a Venn diagram)**:** Area of overlap that shows where more than one event has occurred

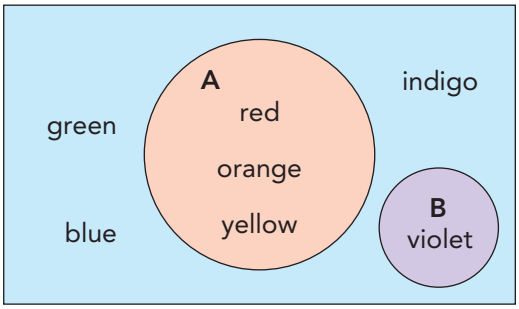


ISBN 978-1-107-61421-5

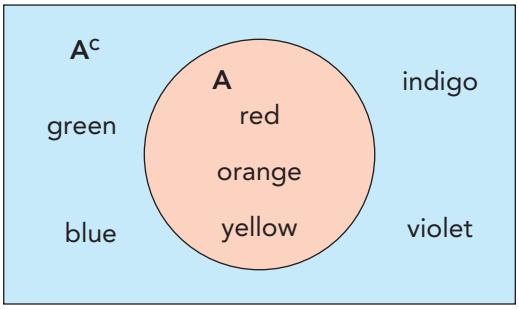
We could replace the numbers in the sections with the corresponding probabilities. We can then check that the probability of the sample space is 1, which is the sum of the individual probabilities.



On a Venn diagram, mutually exclusive events are shown by circles that don't intersect. For instance, this diagram shows the event A = 'selecting a warm colour', and the event B = 'selecting violet'.



These events are mutually exclusive. Complementary events are shown as the inside and the outside of a circle. For instance, this diagram shows the event A = 'selecting a warm colour' and its complement A^C = 'selecting a colour that is not warm'.



It is often easier to work with a sample space and events if we can show them in a visual way such as a Venn diagram. With two or even three events, a Venn diagram can be very useful – but it becomes much more complicated if there are more events.

ISBN 978-1-107-61421-5

Key ideas

- A Venn diagram can be used to show events (as circles) in a sample space (a rectangle).
- The relationship between two events can be shown by the intersection (or non intersection) of the circles.

Example 13: Blood types

Blood types were discovered in 1901 by the Austrian doctor Karl Landsteiner during early experiments with blood transfusion. Blood contains substances called antigens. Blood type depends on whether a person's blood has A-antigen and/or B-antigen.

- In Australia, 49% of people have neither of these antigens, so their blood type is called type-O.
- 38% have only A-antigen and their blood is called type-A.
- 10% have only B-antigen so their blood is called type-B.
- 3% have both antigens so their blood is called type-AB.

a Show this information in a Venn diagram.

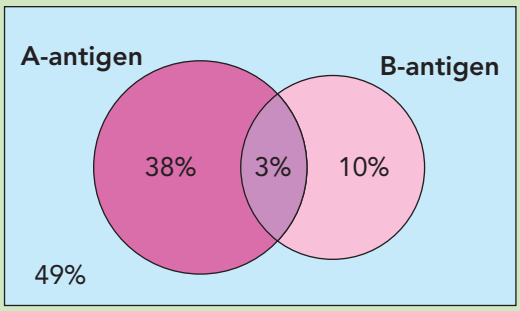
b An individual with blood type-A cannot receive blood from anyone who has B-antigens. If a person with blood type-A needs a transfusion, what percentage of blood donors would be compatible?

c A patient with type-O blood cannot receive a transfusion from anyone with A-antigens or B-antigens in their blood. What interpretation of 'or' is being used here?

d What percentage of the population is unable to donate blood to a patient with type-O blood?

Solution

a The Venn diagram has circles for A-antigen and for B-antigen. Type-O is represented by the region outside the two circles, and type-AB by the region inside both circles. Type-A is represented by the rest of the larger circle and type-B is represented by by the rest of the smaller circle.



ISBN 978-1-107-61421-5

b A type-A patient can receive blood from anyone who is not inside the B circle, that is, any type-A or type-O person. This is 38% + 49% = 87% of the population.

c The 'inclusive or' has to be used here. A type-O patient can't receive blood from anyone with A-antigens, or B-antigens, or both.

d The only people who can donate to a patient with type-O blood are other type-O people – 49% of the population. Therefore 51% of the population is unable to donate to a patient with type-O blood.

Exercise 4D

1 A student in Year 8 is selected randomly to be the junior school student representative at an education evening. We define two events, A = 'the student has blond hair' and B = 'the student has blue eyes'. Show this information in a Venn diagram. On separate copies of this diagram:

- **i** Shade the regions representing these events.
- **ii** Describe the events in words.

a A

b A and B

c A or B using the 'inclusive or'

d B^C, the complement of B

2 For a science degree at the University of Technics, a first-year student must select four subjects from a list that includes Statistics and Computing.

a What is the sample space for selecting Statistics and/or Computing?

b Show this sample space in a Venn diagram.

c Shade the region that represents the event 'select Statistics or Computing'. Which interpretation of 'or' will be more useful here?

ISBN 978-1-107-61421-5

3 Two snapdragon seeds are planted in a pot. The seeds grow into flowers that are white (w), pink (p) or red (r). The sample space can be written as S = {ww, wp, wr, pp, pr, rr}.

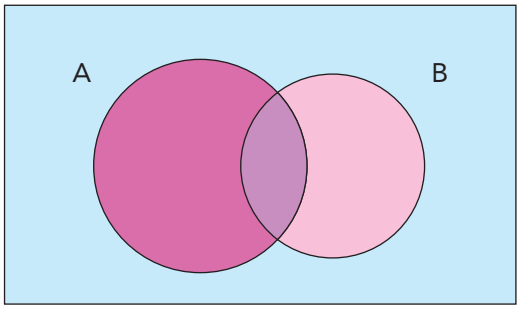
a Show this sample space in a Venn diagram.

b Mark on your diagram the event B = 'both plants had flowers of the same colour'.

c Mark on your diagram the event C = 'at least one of the plants had white flowers'.

d How would you describe the event 'B or C' if the 'inclusive or' was used?

4 A number is randomly selected from the numbers 1 to 50. The Venn diagram below shows two events, A = 'the number is divisible by 3' and B = 'the number is divisible by 5'.

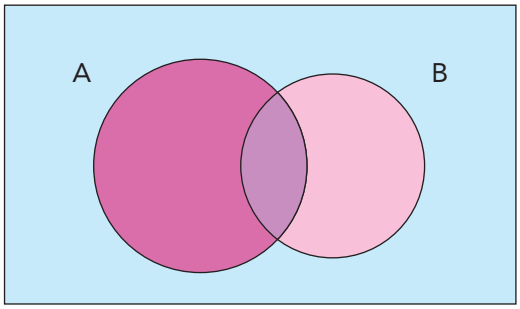


a Which numbers would make up the event 'A and B'? What is P(A and B)?

b Which numbers would make up the event 'A or B' if you use the 'exclusive or'? Describe the event in a way that is not ambiguous and find its probability.

c Which numbers would make up the event B^C, the complement of B? Describe this event in words and find its probability.

5 The Venn diagram showing two generic events, A and B, is given below.



a Shade the region representing the event 'A^C and B^C' and describe the event in words.

b Shade the region representing the complement of 'A or B', which can be written as '$(A \text{ or } B)^C$', using the 'inclusive or'. Describe the event in words.

c What general rule can you conclude from this?

ISBN 978-1-107-61421-5

6 The Venn diagram was invented in the late 1800s by the British mathematician and philosopher John Venn. A related diagram is called the Carroll diagram (or Lewis Carroll's Square), and it was invented around the same time by the Reverend Charles Lutwidge Dodgson. Here is a Carroll diagram:

	Prime	Not prime
Even	2	4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26
Not even	3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41	1, 9, 15, 21, 25, 27, 33, 35, 39, 45, 49

a What else is Reverend Charles Lutwidge Dodgson known for? Why is the diagram he invented called a Carroll diagram rather than a Dodgson diagram?

b Explain how a Carroll diagram works, using the example above.

c In what way is a Carroll diagram different from a Venn diagram?

Lewis Carroll

Enrichment

Three alphabets
<www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

ISBN 978-1-107-61421-5

Chapter summary

Adding probabilities

- An experiment in probability is any situation where we don't know what will happen
- The sample space represents all possible outcomes of an experiment
- An event is part of the sample space
- Two events are mutually exclusive (or disjoint) if they don't contain any outcomes in common
- Two events are complementary if they are mutually exclusive and they cover the sample space completely.

Combining events with 'and' and 'or'

- The phrase 'A and B' specifies that both of the events A and B occur
- The phrase 'A or B' can mean 'exactly one of A and B' occurs ('exclusive or')
- The phrase 'A or B' can mean 'at least one of A and B' occurs ('inclusive or')
- We have to be careful to specify our interpretation of 'A or B'
- If A and B are complementary events, P(A or B) = 1 and P(A and B) = 0.

Two-way tables of words and data

- A two-way table of frequencies can be used to show information about two variables
- The frequencies can be used to estimate probabilities of individual events or combinations of two events for the group represented.

Venn diagrams

- A Venn diagram can be used to show events (as circles) in a sample space (a rectangle)
- The relationship between two events can be shown by the intersection (or non intersection) of the circles.

Multiple-choice questions

1 A balanced coin is tossed three times and the sample space is written out as S = {hhh, hht, hth, thh, tth, tht, htt, ttt}. Which of these outcomes is **not** included in the event 'a majority of tails was obtained'?

A htt **B** tht
C hht **D** ttt

2 Consider the event that consists of the outcomes hht, hth, thh, tth, tht and htt. Which of these describes the event in words?

A 'a majority of heads' **B** 'fewer than 3 heads'
C 'all three tosses gave the same result' **D** 'not all tosses gave the same result'

3 A fair die is rolled and the event A is defined as 'a 6 is obtained'. Which of these events is not mutually exclusive with A?

A 'even number' **B** 'a 3'
C 'a prime number' **D** 'a number less than 4'

4 A fair die is rolled and the event B is defined as 'a prime number is obtained'. Which of these events is complementary to A?

A '4, 5, 6' **B** '1, 3, 5'
C '1, 4, 6' **D** '2, 4, 6'

ISBN 978-1-107-61421-5

5 An American roulette wheel contains 38 slots. Half of the numbers 1 to 36 are coloured red and the other half are coloured black. The 0 and 00 are coloured green. If you bet on green, what is the probability that you will win?

A $\frac{1}{19}$ **B** 0

C $\frac{18}{19}$ **D** $\frac{9}{19}$

6 Using the information in question 5, what is the complementary event to 'green'?

A 'red' **B** 'red or black'

C 'red and black' **D** 'black'

The Year 7 students at a school are classified by gender and whether they play competitive sport or not. A student is selected at random to be interviewed for the school magazine. Use this information to answer questions 7–10.

	Sport	Not	Totals
Boy	47	3	50
Girl	48	7	55
Totals	95	10	105

7 What is the probability that the selected student will be a girl who plays competitive sport?

A $\frac{95}{105}$ **B** $\frac{55}{105}$

C $\frac{48}{105}$ **D** $\frac{102}{105}$

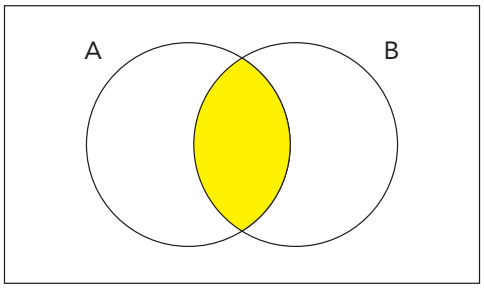
8 What is the probability that the selected student will be a boy or a competitive sport player?

A $\frac{47}{105}$ **B** $\frac{47}{95}$

C $\frac{57}{95}$ **D** $\frac{98}{105}$

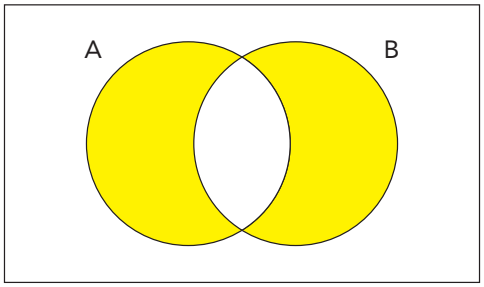
ISBN 978-1-107-61421-5

9 What combination of events is shown in the yellow part of this Venn diagram?



A 'exactly one of A and B'
B 'both of A and B'
C 'neither of A and B'
D 'at least one of A and B'

10 What combination of events is shown in the yellow part of this Venn diagram?



A 'exactly one of A and B'
B 'both of A and B'
C 'neither of A and B'
D 'at least one of A and B'

Short-answer questions

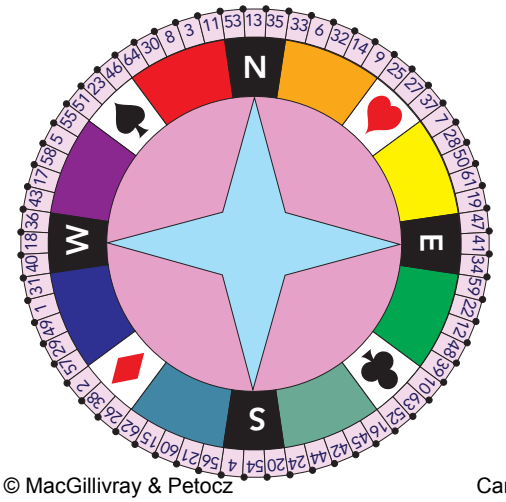
1 A box contains four coins of denominations \$2, \$1, 50c, 20c. Two coins are taken randomly from the box, one after the other.

a Write down the sample space for this experiment.

b If event A = 'the \$2 coin is amongst those selected' what is the complementary event?

c What is the probability that the coins selected are enough to pay for a bus fare of \$2.20?

2 A Wheel of Fortune is shown opposite. As well as 64 numbered sections, it has sections marked by **colours**, **letters** for points of the compass (N, S, E, W), or playing card **suits** (spades ♠, hearts ♥, diamonds ♦, clubs ♣). Each of these covers several numbers. In this question we are concerned only with the colours, letters and suits – you can buy a ticket for any of them. The wheel is spun and comes to rest with a leather flap indicating the winning colour, letter or suit.



ISBN 978-1-107-61421-5

a What is the sample space for these tickets?

b What is the probability that a suit wins?

c What is the probability that a letter or a suit wins? What interpretation of 'or' are you using here?

d What is the complementary event to 'a letter or a suit wins'? What is its probability?

3 'Bill shock' is the term for an unexpectedly high bill for your mobile phone usage, often resulting from a difficulty in understanding (or even knowing) the telephone company's rules. If this happens to you, you can pay the bill or challenge it – either by yourself or using a company that specialises in dealing with bill shock cases. Of 380 people who had registered with such a company, 220 challenged the bill themselves and 170 of them were successful in getting it reduced. The other 160 were supported by the company, which was successful in 155 cases – though for each they took one-quarter of the reduction as a fee.

a Show the sample space of outcomes for the people who registered with the company.

b What is the probability that a randomly selected person from the 380 who registered received a reduction in their bill?

c What is the probability that a randomly selected person from the 380 who registered was supported successfully by the company?

4 A random sample of 900 people were asked in a national survey whether Australia should become a republic. Here are the results:

	Yes	No	Undecided
Male	145	260	22
Female	125	305	30
Not recorded	8	4	1

a What is the probability that a person was in favour of Australia becoming a republic?

b What is the probability that a respondent was male and had a definite opinion?

c What is the probability that a respondent's gender was not recorded or they had not made up their mind about the question?

5 In a class of 32 students at the beginning of a maths lesson, 8 of them have a textbook and a calculator, 12 of them have a textbook but no calculator and 3 of them have neither textbook nor calculator.

a Show this information on a Venn diagram and fill in the missing number.

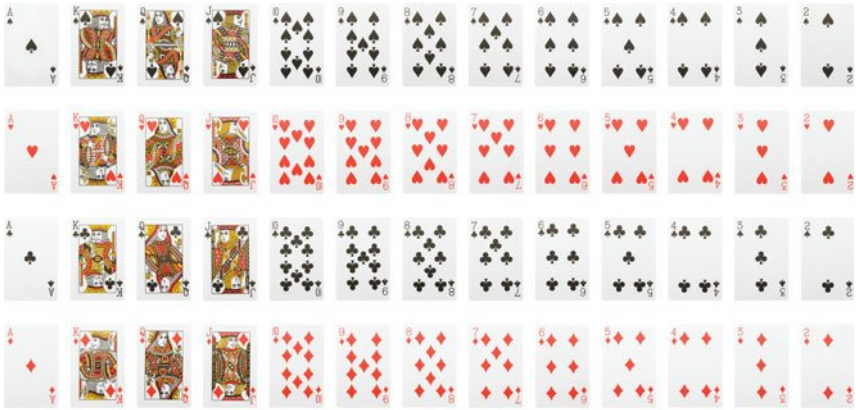
b What is the probability that a randomly selected student in this class has either a textbook or a calculator? What interpretation of 'or' are you using?

c What is the complementary event to the event in part **b**, and what is its probability?

ISBN 978-1-107-61421-5

Extended-response questions

1 A standard deck of playing cards contains 52 cards, consisting of four 'suits' – spades and clubs, which are black, diamonds and hearts, which are red – each with thirteen 'values' – 2, 3, ..., 10, jack (J), queen (Q), king (K) and ace (A = 1). This is shown in the picture.



The deck is shuffled and the top card is turned over. What is the probability that it is:

a red and a 10?

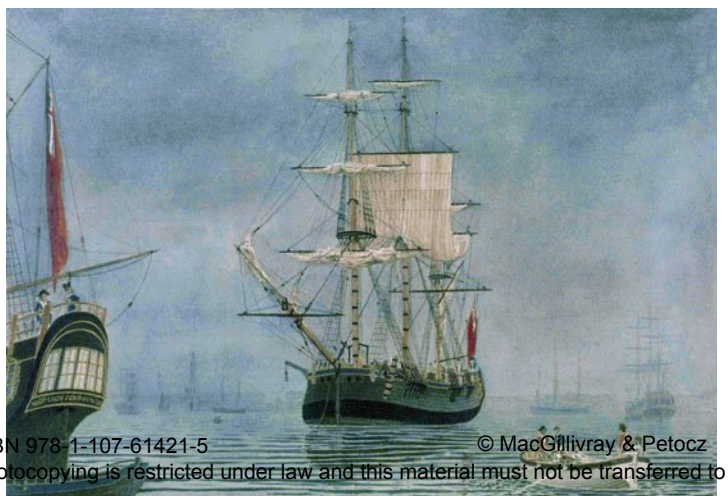
b a spade or a heart?

c a diamond or an ace?

In each case explain how you got your answer.

d How many shuffles does it take to randomise a deck of cards?
Using 'riffle shuffling', in which you divide the pack and interleave the two halves randomly, up to 5 shuffles leaves much of the original information in the pack. But if you use 7 or more riffle shuffles, you produce something quite close to randomness (and further shuffles don't help).
How could you check that 7 shuffles produces a random ordering of the cards?
<www.cambridge.edu.au/statsAC78weblinks>

2 In 1778, the British Government sent a fleet of ships containing convicts, together with their gaolers and administrators to start a penal colony in Botany Bay (Sydney). The table on the next page gives information about the crime and the term of transportation of the 778 convicts on this First Fleet. (Some of the original categories have been combined to simplify the table.)



ISBN 978-1-107-61421-5

Crime	7 years	14 years	Life	Total
Larceny	295	2	3	300
Theft apparel	125	0	0	125
Animal theft	57	2	3	62
Burglary	54	4	3	61
Assault highway robbery	47	4	3	61
Felony	30	1	1	32
Robbery	24	0	0	24
Highway robbery	21	0	0	21
Assault robbery	18	0	2	20
Return transport	6	1	12	19
Housebreaking	12	2	3	17
Receiving stolen goods	3	8	0	11
Fraud	3	0	1	4
Forgery	1	1	1	3
Armed robbery	2	0	0	2
Assault	2	0	0	2
Burglary theft apparel	2	0	0	2
False impersonation	1	0	1	2
Perjury	1	0	1	2
Other crime	6	0	2	8
No crime listed	2	0	0	2
Total	712	25	41	778

a The most common crime was larceny. What is larceny? What is the probability that a convict selected at random from the First Fleet was convicted of larceny?

b The most common sentence was for 7 years. What is the probability that a randomly selected convict was sentenced to 7 years transportation?

c What is the probability that a randomly selected convict was sentenced to a term of 7 years for larceny?

d If you select one of these convicts randomly for a historical study, what is the chance that you get a person convicted of larceny or sentenced to transportation for 7 years? Explain why it is reasonable to use the 'inclusive or' here.

e How do your answers for parts **a**, **b** and **c** relate to your answer for part **d**?

3 Write three questions about probability using the material in this chapter – one multiple choice question, one short-response question, and one discussion question. Your fellow students should be able to answer the questions easily, but not too easily! Include answers, or for the third question, some suggested points for an answer.

ISBN 978-1-107-61421-5

CHAPTER

5

Collecting useful data

What you will learn

Newsflashes

Here are some examples of the types of claims we see in the media, in news, or in reports. The first comes from a real newspaper article, and the other two are fictional but typical. <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

> 'UK study found that 97 per cent of parents thought it was OK for teenagers below the age of 16 to own a mobile phone.'
>
> Source: www.couriermail.com.au

> 'New battery lasts longer'

> 'Boys use mobiles more than girls but girls talk longer'

The basis for all claims and information like the above is data – the word *data* even means information! But how can we go from data to claims? Were all UK parents asked? Not all batteries could ever be tested. And similarly, not all the boys and girls in the world or even one country could have been observed now, let alone for any claim that might extend beyond now.

It is natural for humans to collect or observe information, to combine bits of information, and then be tempted to generalise. *Generalise* means to make a general statement about what we think happens in that situation. Statistics is the science that enables us to collect or observe useful data that can be used to make a statement that goes further than the actual data we have obtained. Statistics is also the science that enables us to know what we cannot say and what we must say.

AUSTRALIAN CURRICULUM

AC

Statistics and probability

- Data representation and interpretation
- Investigate techniques for collecting data, including census, sampling and observation **(ACMSP284)**
- Explore the practicalities and implications of obtaining data through sampling using a variety of investigative processes **(ACMSP206)**

ISBN 978-1-107-61421-5

Cambridge University Press

Pre-test

1 For an assignment, a student obtains data for the previous year for 100 developing countries. The data consisted of population size, average income per person, birth rate as number of births per 1000 people, and number of doctors per 1000 people.

- **a** What are the subjects (or observational units) of these data?
- **b** Name the variables in this study.
- **c** State whether each variable is categorical, count or continuous.
- **d** What plot would you suggesting using to explore or present the following?
  - **i** The average income per person
  - **ii** The birth rate
  - **iii** The population size
- **e** How many observations would there be in each of the plots in part **d**?

2 For each of the following variables, state whether the variable is categorical, count or continuous. If it is categorical, state if it is ordinal.

- **a** Amount of pocket money/allowance in a week
- **b** Head circumference (in cm)
- **c** Favourite television show
- **d** Last movie seen
- **e** Importance of sport ('very', 'somewhat', 'not very')
- **f** Number of pieces of fruit eaten per week.

3 Suppose that in a survey, 1000 people are randomly selected and asked approximately how much television they watch in a week (in hours), whether they play a sport, what their favourite recreation is, and whether they think there should be more or less televised sport.

- **a** Name the variables in this study.
- **b** State whether each variable is categorical, count or continuous.
- **c** Give two other variables which you think must be included in this survey, giving reasons.
- **d** What are the types of variables you have suggested for part **c**?
- **e** A pilot survey is conducted. What do you think might arise in the pilot that will help in carrying out the big survey?

4 City planners want to investigate whether people obey the pedestrian crossing lights in a city centre. Observations are made at two sets of pedestrian lights. During each cycle of the lights (that is, green, flashing red, red), the number of people who start crossing against the flashing red or against the red are counted.

- **a** What is the main variable of interest and what type is it?
- **b** What other variable is mentioned above and what type is it?
- **c** What are the subjects of this investigation?
- **d** Give two other variables which you suggest should be included in this investigation, giving reasons.

Terms you will learn

blinding
case-control study
census
closed questions
cluster sampling
double-blinding
experimental investigation
observational study
open questions
placebo
placebo effect
polling
random numbers
random sample
randomisation (in experiments)
response rate
sample of data
sample size
sample survey
sampling plan
stratified sampling
table of random digits

ISBN 978-1-107-61421-5

5-1 Census or sample?

In the preceding chapters we have seen examples of the planning for many different statistical data investigations. We have seen the importance of:

- identifying topics or issues or questions of interest
- identifying the variables on which data are to be obtained – or have been obtained in the case of secondary data
- identifying the subjects of the investigation
- designing the recording sheet
- carrying out a pilot study or experiment.

We have also seen the importance of identifying the types of variables in our investigation. This helps in designing our recording sheet, choosing and clearly describing our categories for categorical variables, and choosing our units of measurement and amount of accuracy for measurement variables. It is particularly important in collecting measurement data to carefully describe (or find out) exactly how the measurements are (were) made. Understanding types of variables is also essential in choosing appropriate graphs, plots and summaries for our data. Useful summaries we've seen so far are most frequent category, mean, median and range.

Understanding the type of data collection being carried out is also important in understanding how the data can be used, and whether and how we can use the data to comment on a more general situation.

ISBN 978-1-107-61421-5

What is a census?

To most people the word *census* means the survey carried out every few years to count the number of people in the country – its population. A survey form is sent to every household or dwelling asking how many people stayed there on a certain date, as well as other information. National censuses obtain information used by businesses, industries and governments. They help in understanding what is happening across a country, in allocating and building resources, and in planning for the future. But the name **census** can be applied to any data collection where the aim is to collect information about every member of a population.

Census: A data collection, in which the aim is to collect information about every member of a population

Samples of data

If we do not collect all possible observations, we say we have a **sample of data** (or a 'data sample'). The general meaning of the word *sample* is a portion, piece, or segment that is representative of a whole. In statistics, when we have a sample of data it means that more observations could have been taken – sometimes on every member of a population, and sometimes without limit, observing for ever and ever!

The number of subjects (or observational units) in a sample is called the **sample size**. This is also called the **number of observations** no matter how many variables you have – it is the number of rows in a recording sheet or spreadsheet. For example, if you interview 50 people and record their age, gender, favourite food, favourite school subject and favourite sport, your sample size is 50 and you have 50 observations and 5 variables – you do not have 250 observations!

Sample of data: A set of observations for which many more observations could have been taken

Sample size (or **number of observations**): The number of subjects in a sample

Randomly representative samples

We can comment on the data in a sample just as it is, but we almost always want to use a sample of data as information about a more general situation or population. In chapter 1 we introduced the idea of randomly representative data. This means data that are, or can be considered to be, a random set of observations obtained in circumstances that are representative of the general situation or population. If we want to collect data so that we can comment on a general situation, then the data must be collected randomly from that situation. This is often called a **random sample**.

Random sample: A set of randomly representative data

On the other hand, if data are collected to investigate certain topics or issues, or have been collected by someone else, then careful thought must be given to how the data were collected. What situation or population is it reasonable to say that the data are randomly representing? A particular dataset might be considered to be randomly representative for some questions or issues, but not for others.

Suppose you want to collect data on shoe size of Year 8 students. It is lunchtime and you see a group of Year 8 students in the computer lab, so you ask them. They are likely to be reasonably representative of Year 8 students' shoe sizes because there is no reason to suppose that certain shoe sizes are more likely to make students want to play on computers in their free time. On the other hand if you want to collect data about Year 8 students playing computer games, this group is unlikely to be representative of Year 8 students.

ISBN 978-1-107-61421-5

LET'S START What to collect?

A school wants to find out the parents' views on the school open day. The school decides to send a simple survey form to each household of their students, asking only about choices of dates and times for the open day. This is a census because it surveys all members of the target population – parents at the school. But it is a only a census if all respond, so the school does a follow-up to ensure that all do.

The school then decides to ask some less simple and more open questions of a sample of parents: what type of activities and displays they would like to see, what types of stalls, how much students, parents and teachers should be involved, and how widely it should be advertised. How should they choose the parents for the sample? Should they just ask representatives of the Parents' and Friends' Association? Should they ask a sample of the parents of the senior students? Should they choose a random sample of all students and ask their parents? It depends on whose views they want the sample to randomly represent. If they ask a sample of parents of senior students, then there's no basis for saying their comments apply to all parents. But if they want to only get information on what parents of senior students think, then this is acceptable.

Key ideas

- A census is a set of data in which all possible observations are collected. Almost always this refers to data collected on every member of a population.
- A sample of data, or a data sample, is a set of observations where more observations could have been taken, sometimes without limit.
- To be able use a sample to comment on a more general situation or population, the data must be obtained randomly in circumstances that are representative of the more general situation or population with respect to the issues or questions of interest.

ISBN 978-1-107-61421-5

Example 1: National censuses

Australia conducts a national census approximately every five years. It is called the Census of Population and Housing, carried out by the Australian Bureau of Statistics. <www.cambridge.edu.au/statsAC78weblinks> The date of the sixteenth Australian Census was 9 August 2011. The seventeenth is in August 2016. Everyone needs to complete, or be included on, a Census form.

The word *census* comes from the Latin, *censere*, which means 'to rate', and an essential and first aim of a country's census is to count – total number of people and numbers in different groupings. This is partly why a census is of the whole population.

It is very important for nations to have accurate census data. National censuses aim to obtain population data not only for vital information for future planning and strategies, but also to guide further data collections. The quality of Australia's census data is highly regarded internationally. What can go wrong in collecting census data? There are many challenges:

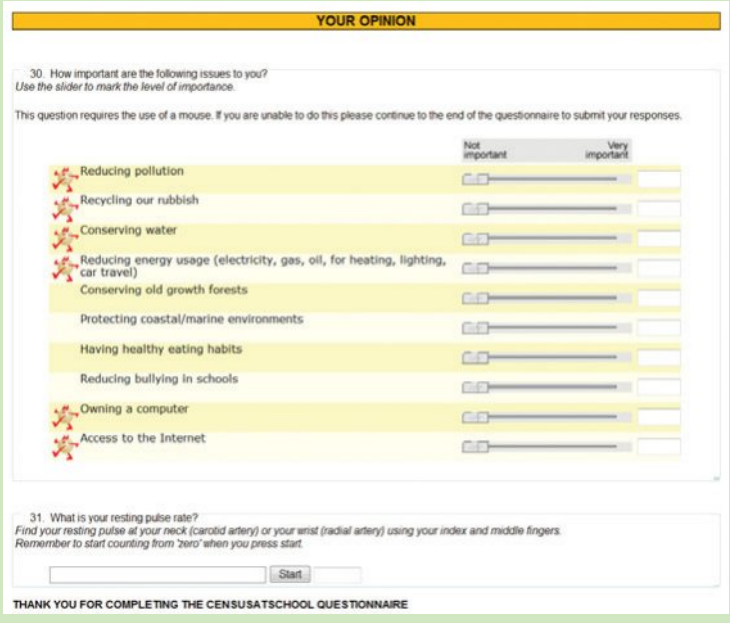
- ensuring everyone is reported and reported on one and only one census form
- ensuring every census form is completed and returned
- omissions
- accidental errors
- errors due to language or understanding difficulties
- deliberate errors.

National offices of statistics use many sophisticated statistical techniques to estimate and cross-check for errors, and to allow for the types of challenges outlined above.

ISBN 978-1-107-61421-5

Example 2: CensusAtSchool

CensusAtSchool was originally developed by the Royal Statistical Society Centre for Statistical Education for the Office for National Statistics in the United Kingdom. <www.cambridge.edu.au/statsAC78weblinks>



YOUR OPINION

30. How important are the following issues to you?
Use the slider to mark the level of importance.

This question requires the use of a mouse. If you are unable to do this please continue to the end of the questionnaire to submit your responses.

	Not important	Very important
Reducing pollution		
Recycling our rubbish		
Conserving water		
Reducing energy usage (electricity, gas, oil, for heating, lighting, car travel)		
Conserving old growth forests		
Protecting coastal/marine environments		
Having healthy eating habits		
Reducing bullying in schools		
Owning a computer		
Access to the Internet		

31. What is your resting pulse rate?
Find your resting pulse at your neck (carotid artery) or your wrist (radial artery) using your index and middle fingers.
Remember to start counting from "zero" when you press start.

Start

THANK YOU FOR COMPLETING THE CENSUSATSCHOOL QUESTIONNAIRE

The project, originally a one-off, was linked to the UK population census of 2001. It has now developed into a dynamic and ongoing initiative running in a number of countries. The original CensusAtSchool questionnaire in 2000 consisted of a single sheet with simple questions covering information about pupils, their households and their school life. Over 2000 primary, secondary and special schools registered for the project and over 60 000 children took part. Since then many other countries have joined the project, adapting to suit local culture and traditions, including Australia, New Zealand, Canada, South Africa, Ireland, Korea, Japan and USA. In 2007 an International Committee was established. <www.cambridge.edu.au/statsAC78weblinks>

When developing the CensusAtSchool questionnaire for Australia, the Australian Bureau of Statistics (ABS) drew on experience from world-wide CensusAtSchool projects as well as consulting experts, students and teachers. The CensusAtSchool questionnaire asks questions about students' everyday lives, experiences, opinions and interests, without being too invasive. New data are available once the questionnaire is closed, but data from previous Australian questionnaires can be accessed at any time of the year through the Random Sampler or prepared samples. Internationally, the CensusAtSchool Random Data Selector web facility gives access to many CensusAtSchool databases. You can take random samples of the raw data collected from CensusAtSchool in the UK, Canada, Australia, New Zealand and South Africa. Sample sizes allowed are up to 200 for UK, Canada, New Zealand and Australia, and 500 from the South African database. <www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

To ensure the CensusAtSchool complies with these aims, the questionnaire does not ask personal questions and you do not have to participate. In addition, student names are not recorded and Student Access Numbers are not part of the data and all data are stored securely.

Is it a census? It's called CensusAtSchool because it aims to collect data from all the students in participating classes. But unlike a national census, it is not compulsory to participate.

The Random Data Selector and related resources are two of the main reasons for the popularity of CensusAtSchool. It provides random samples of data for students to explore to see the variation within and across random samples and within and across countries.

Exercise 5A

1 In each situation below, state whether the observed data is a census or a sample.

a A teacher asks all the students in a class where they would like to go for their maths excursion.

b A media research organisation phones a random selection of phone numbers and asks the respondents for their opinions on a reality TV show.

2 In each situation below, state whether the observed data is a census or a sample.

a A sports commentator collects data on all the Australian cricket captains for the past 50 years.

b A health expert measures the pulse rates of a random selection of gym users before and after their gym session.

3 In each situation below, what are the subjects and what is the sample size?

a Twenty people have their pulse rates measured every kilometre during a 20-kilometre bike ride

b Weekly height measurements of twenty rows of seedlings with 10 seedlings in each row are taken for 3 weeks.

c The prices of three different sizes of a brand of soft drink collected from 30 different shops and other types of outlets scattered across a region

d Fifty companies fill in a survey of twenty questions on their usage of energy.

ISBN 978-1-107-61421-5

4 In each case, briefly comment on whether you think the available sample data can be taken as randomly representative of a larger group or more general situation and, if so, identify the group or situation.

- **a** The heights of students in your class
- **b** Whether each student in your class can curl their tongue or not
- **c** The opinions of students in your class on how much time people under the age of 16 years should spend on the internet
- **d** The opinions of people who vote in an online survey attached to a news story on a website
- **e** The opinions of people who are phoned in the afternoon through a random selection of home phone numbers
- **f** The responses to a survey of randomly selected customers, both in person and online, of a chain of large stores, on how often they buy from the store, what started them buying and whether they prefer to shop in person or online
- **g** The responses in the same survey as in part **f**, on their approval or not of large-scale wind farms for Australia.

Enrichment

Can we research it?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

5-2 Experimenting

Statistical data investigations can be classified into three broad groups: experiments, observational studies and surveys. Experiments can be further classified into those that take place in a laboratory, and field experiments that occur outside a laboratory. In an **experimental investigation**, investigators control conditions and measure the effect of these on some outcome(s) of interest. Conditions are controlled, by deciding which ones to fix and which ones to vary, and choosing how to vary these.

Experimental investigation: Data investigation in which investigators control conditions and measure the effect of these on some outcome(s) of interest

The word *experiment* is used in any situation in which the outcome is uncertain. In Chapters 3 and 4, the situations are very simple to introduce the use of assumptions to obtain probabilities. In data investigations, data are collected on the outcomes of experiments, and we are interested in how the data vary and how they are affected by the experimental conditions.

Randomisation

Randomisation in experiments is essential to being able to interpret data from an experiment. In randomised experiments, either conditions are allocated to subjects at random, or subjects are allocated to conditions at random. For example, suppose we want to compare two new medications for clearing up and preventing a certain kind of rash. The investigators have a group of volunteers. They need to allocate each volunteer randomly to product A or B (for example, by tossing a fair coin or die or some other means of choosing at random between two outcomes). After some specified time (e.g. two weeks) they compare each volunteer's skin after using the medication with their skin before using the medication. If the investigators don't allocate the volunteers randomly – or worse, if the volunteers are allowed to choose which product – any differences we see might be due to the investigators' or the volunteers' choices!

Randomisation (in experiments): Process of allocating conditions to subjects at random, or subjects to conditions at random

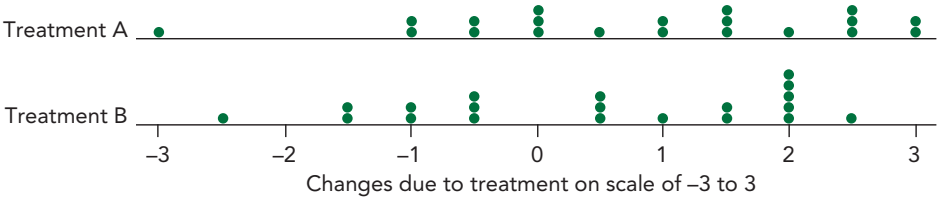
Can we design this experiment better? Yes, there are a number of ways. The subjects should not know which product they're using in case it affects the way they use it or

ISBN 978-1-107-61421-5

their judgement of it. This method is called **blinding**. In some health experiments, even the investigators do not know which subject has which treatment.

Blinding: When the subjects in an experiment do not know which product or treatment they've been allocated

Another aspect is that the subjects' skins might vary so much that any difference between the treatments might be hard to detect. The dotplot below shows such a possibility when changes are marked on a scale of –3 (much worse) to 3 (much better), in steps of 0.5, with 0 representing no change.



The variation across subjects is huge. There are two ways of avoiding this situation:

- Each subject could use one product for two weeks, then the other product. In this case, we would randomise the order in which the products are used: this becomes the randomisation.
- Subjects could be grouped in pairs so that each pair had similar skin and state of their skin at the beginning of the trial. In this case, each member of a pair would use a different product, and products A and B would be randomly allocated within each pair.

Because the products are allocated randomly or the products are used in a random order, if the investigators do see a difference between the effects of the two products, then they can say it is due to the products.

Note that it is very likely that there will be considerable variation in the data. Judging if there is sufficient difference between the effects of the two products is not a topic for now. But it is the randomisation in experiments that allows us to say that an effect was caused by a certain condition or combination of conditions.

ISBN 978-1-107-61421-5

Placebo

It is common that when someone expects something to do them good, they feel better after using or taking it – even if the 'something' has no active ingredients. This is called the **placebo effect** (*placebo* means 'I please'). It is important in many health and social science experiments because sometimes there can be an effect just because the subject, and even the investigators, expect there to be one. A **placebo** treatment is when a 'pretend' treatment is given but actually no treatment is given.

In the experiment above, the placebo effect was not considered because it is a direct comparison of two products. If the investigators did want to use a placebo to see if either treatment has any effect, they would have to use a substance known to have no beneficial or harmful effect on skin.

Placebo effect: Effect due just to being on a treatment or being in an experiment

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

LET'S START Designing a taste test

Can people tell the difference between low-fat and full-fat yoghurt just by the taste? Obviously they will need to be comparing the same flavour of yoghurt. We could choose a brand and a flavour that has both versions (low fat and full fat) and design an experiment. Each subject will taste both yoghurts. The subjects must not know which one they are tasting, and the order of tasting must be randomised for each subject. Perhaps the subjects should also be blindfolded (in addition to the experiment being blinded!) in case they might be able to see a difference between the yoghurts. The person or people giving the spoons of yoghurt to the subjects probably should not know which is which in case they subconsciously react to the subject's choice – no matter how slight the reaction might be! This is called **double-blinding**. So the yoghurt portions could be prepared by one group with random allocation of the letters A and B to the low-fat and full-fat portions – keeping a record of which is which in each pair. The prepared portions are then given to another group who will feed them to the blindfolded subjects who say what they think each portion is. Is there anything else we can randomise? We should randomly choose the subjects for the taste test. For example, by pulling names out of a box containing the names of

Double-blinding: Occurs in an experiment when investigators who are interacting with subjects do not know which products or treatments have been allocated to different subjects

ISBN 978-1-107-61421-5

everyone in the class, we could randomise the order in which the subjects do their taste test. We do this not because we expect this might make a difference, but just in case – and because we are trying to randomise everything we can.

So we've randomised as much as we can in this experiment. What else could we do? Maybe we should record something about the subjects just in case it helps us to interpret the data after we've collected it. Perhaps we should record each subject's gender and ask them how often they eat yoghurt. Note that we're not controlling their gender or whether they are yoghurt-eaters; we are just recording data in case it's useful because once the experiment is done, we may not be able to go back and find out this information.

Key ideas

- In an experiment, investigators fix some conditions and allow other conditions to vary. This enables them to measure the effect of the varying conditions on some outcome(s) of interest.
- Randomisation is very important in experiments because it allows us to say that observed effects were caused by the varying conditions.
- Randomisation can be done in a number of ways, such as random allocation of conditions to subjects or of subjects to conditions, or random ordering, or both random allocation and random ordering.
- Other procedures are also used in the design of experiments to avoid or minimise unwanted effects, or to prevent variation of subjects from hiding effects.

Example 3: Murphy's law and toast

'Murphy's law' has been a topic of many sayings and jokes over many years. It typically states that 'if anything can go wrong, it will go wrong'. Murphy's law as it applies to toast refers to the popular opinion that dropped toast is more likely to land butter-side (or other topping) down because that is a worse outcome than toast-side down. Another version of the saying is that 'the more expensive the floor covering, the more likely toast is to land topping-side down'.

To investigate Murphy's law as it applies to toast requires choices:

- What will be kept constant in the experiment
- What will vary
- How to randomise the treatments.

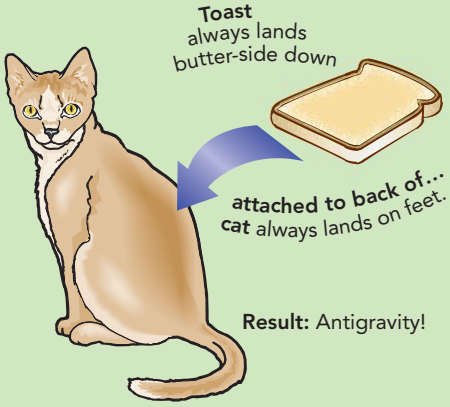
Suppose it is decided to use two floor coverings – carpet and tiles – and two toppings on toast – margarine and jam. The outcome of interest is whether the dropped toast lands topping-side up or down. How can toast be dropped in this experiment? It needs to be like real life so perhaps there should be someone who holds the toast on a plate and someone who bumps the plate making the toast fall. Unless it is decided to take into account different people doing the holding and bumping, the same people should do this for the whole experiment. To be really 'real', the floor covering should be cleaned between each drop.

ISBN 978-1-107-61421-5

So there are four possible combinations – tiles and butter (TB), tiles and jam (TJ), carpet and butter (CB), and carpet and jam (CJ). How many of each should we do? For an experiment with categorical outcomes (such as topping down or up), quite a lot of observations are needed. The sample size is the total number of pieces of toast dropped.

The randomisation is done through randomising the order, so a four-sided die or a spinner with four equal segments could be used to say which combination for successive pieces of toast.

HINT
If you research Murphy's law and toast on the internet, you will find that there has been quite a lot of interest – both serious and not so serious – in investigating this topic and making jokes.



Example 4: Vitamin C and colds

A lot of medical research has been done on the claim that taking vitamin C supplements helps to prevent colds or to reduce how bad they are when you catch one. Conducting an experiment on this requires a lot of care because of the human elements involved – the subjects and the investigators.

Because catching colds and how bad they are depends on aspects such as individuals' health, age and immunity, it would be best to have pairs of subjects, with the subjects in each pair having similar age and various medical criteria. Then one member of each pair would be randomly chosen to receive vitamin C supplements for some time, while the other would receive a placebo – something that would taste the same but contain no vitamin C. Of course it would be essential that the subjects didn't know which they were taking. Then the subjects would need to be observed for a while to see how many colds they caught and how bad they were. Could the investigators depend on the subjects to report? Probably not, particularly in judging how bad a cold is. The person/people observing the subjects and judging how bad a cold is also should not know who was taking the vitamin C. This is because knowing might subconsciously influence them – and it could be either way! As with the yoghurt taste-test, the

ISBN 978-1-107-61421-5

person/people who kept the records would give the coded tablets to a 'middle' person who would give them to the subjects and report back on their colds.

As you can see, experiments that involve humans may need to be quite complex. And of course in research in medicine and other areas involving humans or animals, it's often not possible to give treatments and see what happens. Often researchers have to research the histories of people who are sick and compare with the histories of people who aren't sick, but without being able to conduct an experiment, it takes a lot of very, very good science and statistics to be able to discover causes.

HINT
If you research vitamin C and colds on the internet, you will find that there has been a lot of research in investigating this topic – and the results may surprise you!

Exercise 5B

1 In an experiment to compare two types of liquid fertiliser for seedlings, 20 seedlings of the same species were divided into two groups of 10. Each group was assigned to a fertiliser, and the same amount of fertiliser was applied daily to the seedlings.

- **a** What measurements need to be taken?
- **b** What should be kept constant?
- **c** What should be randomised?
- **d** It is decided to also investigate whether to apply the fertiliser in the morning or evening. How should this be included in the experiment?

2 In a field experiment, the effects of two amounts of the same fertiliser are to be investigated on two different types of wheat. The fertiliser is ploughed or dug into the ground before the wheat is planted. A field is divided into four blocks so that the ground is similar within each block. Each block is then also divided into four sections.

- **a** How should the types of wheat and amounts of fertiliser be assigned to the blocks and sections?
- **b** What should be randomised?
- **c** What is the sample size?

3 It is decided to investigate if listening to music affects concentration – for better or for worse! Volunteers are called for, and each subject is asked to type a piece of text while listening to music and while not listening to music. The length of time it takes to complete the typing and the number of mistakes are recorded.

- **a** What should be randomised?
- **b** Do you think the same piece of text should be used for the typing while listening and not listening to music? Why or why not?
- **c** Name at least two pieces of information about the subjects that should be recorded in case they are useful in analysing the data.
- **d** Suggest a different way of carrying out this experiment. What would be randomised in your suggested way?

ISBN 978-1-107-61421-5

4 Does the colour of a party candle affect its rate of burning? Will the rate be affected by whether they are quite vertical or at an angle? An experiment is set up to investigate these questions. This is the way the rate of burning is measured:

1 Measure the length of each candle before lighting.

2 Light each candle carefully the same way.

3 Allow each candle to burn for a specified amount of time.

4 Extinguish carefully (not by blowing!).

5 Measure the length again.

The difference in length is the amount burnt in the fixed time and represents the rate of burning, which can be compared across candles.

Packets of party candles have four colours and the same number of each colour. In the experiment, the candles of one packet are burnt vertically, and the candles of the other packet are burnt at a slight angle.

a What is wrong with the design of this experiment? Why?

b What should be done and what should be randomised?

Enrichment Which can you see?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

5-3 Surveys

A survey involves asking questions. The data are the responses. The subjects of a survey may be individuals, or groups such as companies, businesses or households. A survey may be a census (if everyone in the population you want to know about is surveyed) or a sample if they are not. In a **sample survey**, subjects are chosen by a **sampling plan** that is designed to be randomly representative of the larger population about which the investigators wish to gather information. Often the word 'sample' is omitted, and by 'survey' we usually mean a sample survey not a census.

Sample survey: Survey for which the subjects are chosen by a sampling plan

Sampling plan: Plan designed to be randomly representative of a larger population about which the investigators wish to find or state information

A survey might seem to be the simplest type of data investigation to carry out, but in many ways it is the most difficult, involving a wide range of complications. This is why there are companies that specialise in surveys, and why designing a survey is such a large part of the work of government statistical offices.

The first challenge is choosing a sample that is randomly representative. Surveys can be difficult to design because they depend on people (or groups) answering questions. Investigators are dependent on reaching those chosen to be in the sample, then obtaining a response, and then on the respondents answering all the questions as accurately and truthfully as they can. So these aspects must also be thought about in designing a sampling plan.

Survey questionnaires

Designing questions takes a lot of skill and practice. We must know exactly how people are interpreting questions because if we don't, the data are useless. For example, in survey of people not born in Australia, asking 'How long is your residence in Australia' could produce answers in metres – the length of their house! This is why a pilot of a survey is absolutely essential.

Also having a person to ask the questions (i.e. an interview) may be better than asking questions on paper or online, but exactly the same questions must be asked in the same order. Interviews can also be more expensive. Questions must not be too difficult or awkward to answer. There must not be so many questions that people give up or can't be bothered answering. Briefly explaining the benefits of a survey can encourage cooperation. Designing survey questions is a balance between getting the information desired and getting accurate, or indeed any, information.

ISBN 978-1-107-61421-5

Stratified sampling

The challenges of choosing a sample that is randomly representative, reaching those chosen and obtaining a useable response, tend to increase with the size of the overall population of interest. In the Let's Start activity in section 5-1 (choosing a sample of parents of students in a school), it is fairly simple to choose a random sample of all parents, or of a section such as parents of senior students.

Another approach is to choose separate random samples within levels – say, parents of junior students (Years 7–8), middle students (Years 9–10), and senior students (Years 11–12). Here the population we are investigating is already divided into groups. Because these groups are like layers, which are also called strata, this form of sampling is called **stratified sampling**. We sample within all the strata that we define. If we wanted to survey parents of a number of schools, we could use stratified sampling, choosing random samples of parents within each school. What about choosing a sample of all the parents of all schools in Australia!

Stratified sampling: When random samples are taken within specified groups (called strata)

Cluster sampling

Suppose we want to survey parents from many schools across a region. Another possibility is to choose schools at random from all the schools in the region, and then survey all, or a sample of, the parents of Year 8 students in those schools. This is called **cluster sampling**.

Cluster sampling: A random sample of groups is chosen and all the subjects in these groups are surveyed

Non-respondents

What do we do about non-responses because those who do respond might be different from those who don't respond? Non-respondents are often contacted again, but the extent and nature of the non-responses must be taken into account in reporting and reading survey results.

Online and phone-in surveys

The worst possible way to pretend to conduct a survey is the phone-in or the online survey. You will see statements below such online opinion surveys such as 'These polls are not scientific and reflect the opinion only of visitors who have chosen to participate'. In other words, the subjects are not only just those who have visited the site but only those who chose to participate! It is the same with phone-in surveys – the opinions are only those of people who decide to phone in. Such surveys are so bad their results are useless. The reason websites and radio programs do such surveys is for viewer or reader participation or gratification. Unfortunately they can be completely misleading as surveys.

HINT
Once again we see that it is essential for any investigation to report exactly how data were collected.

ISBN 978-1-107-61421-5

LET'S START How much do you recycle?

We want to find out how aware Year 8 students are of recycling and to what extent they recycle.

Who are we going to survey? If we just survey the Year 8 students in one school, who will that sample be randomly representative of? Perhaps we could take a region and for each school, choose a random sample of Year 8 students. This would be a stratified sample. We can do the random sampling within a school by choosing however many **random numbers** we want from 1 up to the total number of Year 8 students in a school. We could then use those numbers to choose the students to be surveyed from an alphabetical list of all students.

Random numbers: Numbers chosen at random from a given range

Now, what are we going to ask? Consider the question 'Do you recycle as much as you can?' This is far too ambiguous – it has no reference point, and depends on your understanding of what can be recycled and what 'much' is and the interpretation of 'you'. Perhaps we could ask some questions such as:

1. 'Do you check items to see if they are recyclable?'
2. 'What do you recycle when you are not at home?'
3. 'What do you and your family recycle at home?'
4. 'Do you think you and your family could recycle more than you currently do?'
5. 'What types of items do you think you and your family could recycle more of?'

Question 1 could offer these options to tick: always, usually, sometimes, never. Question 4 could also offer options. Questions 2, 3 and 5 could have a list of types of items to be ticked (allowing as many ticks as applicable) plus 'Other'; these are called **closed questions**. Or they could be **open questions**, allowing people to list whatever they wish. Respondents might miss possibilities in responding to open questions, but investigators might miss aspects with closed questions. So open questions are often used in planning or pilots of surveys.

Closed questions: Questions with given responses as choices in a survey

Open questions: Questions in a survey that allow people to respond as they wish

Should we ask the questions in person or by paper or online? If we ask in person, it will tend to reduce non-responses but the respondent might be influenced by trying to impress – or the opposite! If we survey by paper or online, then we should give a deadline and ask the non-responders again – nicely – just after the deadline. Surveys never have 100% response, and you may be surprised by how low the percentage is that is regarded as a good response. **Response rate** depends on many factors – how much the people surveyed are involved in the topic, whether they will see the results, as well as how quick and easy it is to answer the questions. The response rate should always be reported.

Response rate: The number of responses divided by the number of people asked to respond, usually expressed as a percentage

ISBN 978-1-107-61421-5

Key ideas

- A sample survey involves asking questions of subjects chosen by a sampling plan that is designed to be randomly representative of a larger population about which the investigators wish to collect information.
- Practicalities of sampling plans include being able to reach the selected subjects and trying to reduce non-responses.
- Questions must be clear, and as quick and easy to answer as possible.
- Exactly how the data were collected, and the response rate, must always be reported.

Example 5: The famous name Gallup in polling

The name *Gallup* is world famous in **polling**, so much so that people sometimes call a poll a Gallup poll even when it's run by an organisation other than the Gallup. George Gallup (1901–84) was a journalist in the USA, but was more interested in finding out opinions, so he moved into market research. In 1935 he set up his own company, The Institute of Public Opinion. Throughout his whole life, he emphasised questioning everything, and it is said that his son commented that his father was always practising polling on them.

Polling: Process of surveying people, usually to ask opinions

Gallup made an enormous splash in 1936 when he correctly predicted the result of a USA presidential election. He also said that the prediction of a certain magazine's poll would be wrong, and he predicted almost exactly what their poll would say! The poll had been very successful in predicting previous US presidential elections. The magazine mailed out millions of simple surveys, but their mailing list came from their subscribers, car registrations and telephone subscribers. Gallup did many polls of only about 2000 people at a time, but they were randomly selected across all types of Americans. He also sent his pollsters to ask people personally. To predict the outcome of the magazine poll, he simply chose a random sample of about 3000 from their lists.

Some people believe that Gallup's success was mostly due to his sampling scheme and others say it was due to the very low response rate for the magazine poll – it was probably a combination. In 1944, Gallup's predictions were somewhat out because soldiers overseas couldn't be polled. In 1948, Gallup's prediction of the presidential winner was incorrect, which he put down to stopping polling too soon. Gallup and his organisations constantly researched and improved polling methods, including both sampling methods and designing questions. Some of the questions he posed in the 1930s are still standard questions today in political polling. But Gallup made his fortune through researching ad campaigns and TV shows.

Exercise 5C

1 Schools often send information home, or request information back via the eldest child in a family when there is more than one child at the school.

a From a statistical point of view, why do you think schools do this?

b What are the subjects (in the statistical sense) of the information/data being sent/requested by the school?

ISBN 978-1-107-61421-5

Cambridge University Press

2 In the USA, as in a number of other countries, voting is not compulsory.

a Why does this add an extra difficulty to the pollsters' job of estimating the support for political candidates?

b A standard question in political polling is 'If the election were held today, whom would you vote for?' What is another question that would be needed in political polls in countries such as the USA?

3 A newspaper included a yes-or-no survey question, 'Do you think there is too much sport on television?' and asked readers to email their answers to the newspaper.

a Do you think the results represent the opinions of all readers of the newspaper?

b The newspaper placed the question in their sports pages. How do you think this will affect the responses?

c An alternative question is 'The amount of sport on television is (please choose one of the following): too little about right too much'.
Why is this a better question?

d What is a problem with both of these questions?

4 Below are three possible questions in a survey about smoking. Which should be used in a survey? Why?

1 'Do you believe smoking should or should not be banned near entrances to public buildings?'

2 'Considering that research has shown that exposure to cigarette smoke is harmful, do you think smoking should be banned near entrances to public buildings or not?'

3 'Do you agree that, considering it is not against the law to smoke, smoking should be banned near entrances to public buildings?'

5 A local government wants to find out if there is support for using more of the region's money to build more cycle paths.

a What is the population of interest?

b How could they conduct this survey?

ISBN 978-1-107-61421-5

Cambridge University Press

c The survey is conducted by interviewing Saturday morning shoppers. What is a problem with this?

d A local cycling club volunteers to do the survey by door-knocking. What is a problem with this?

6 To obtain ratings for a newly released movie, movie-goers are asked for their rating as they come out from the movie.

a The polling is done after a weekend afternoon session in two cinemas. What is a problem with this?

b What would you suggest instead of, or as well as, part **a**?

c People often go to movies in groups. What does a person doing the polling need to be careful to avoid?

d How should a person doing the polling choose who to ask amongst the movie-goers?

e Does a poll conducted amongst movie-goers as above represent the popularity of a movie? Why or why not?

f How could the popularity of a movie be measured?

g Internet sites on movies often ask people to give ratings for movies. For what population is this rating relevant?

Random numbers
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

5-4 Observing

In an **observational study**, investigators observe subjects without altering or controlling their behaviour in any way. There are many different types of investigations that can be called observational studies. These are some examples:

- Observing the behaviour of drivers, pedestrians, shoppers and suchlike.
- Observing other types of human behaviour
- Observing animals
- Collecting data on prices and sales figures
- Investigating company data
- Investigating sports data.

Observational study: Data investigation in which investigators observe subjects without altering or controlling conditions

Look at the examples and exercises in Chapter 2. How many of these are observational studies?

Below are just some examples of observational studies:

- Speed at amber traffic lights
- Children' spans and heights
- Lengths and weights of fish caught
- Delivery times for pizzas
- Time between city circle buses
- Time between postings on an internet site.

But notice that all of these involve choice of conditions of some type, for example:

- Where and when will we choose to observe drivers?
- Where and when will we choose to observe the times between city circle buses?

ISBN 978-1-107-61421-5

- When and where are the fish data collected?
- Which pizza delivery outlet(s) and which internet site(s) will we observe, and when should we observe these?

Almost all observational studies involve at least some aspects of design similar to experimental studies. However, because we are not controlling the conditions, as in an experiment, we usually cannot say if anything caused what we see in the data. We can investigate variation and features of the data, and we can say that there appear to be connections, but we need to be very careful in generalising, and extra careful in commenting on possible reasons. It is always very important to know and describe exactly how data were collected, but this is especially true in observational studies. What population or general situation can we say our data randomly represent? In planning and describing an observational study, we should include as much as possible of what we know about the situation, what we can control and what we can't, and randomise where possible.

The disadvantages of observational studies are the question of what do our data randomly represent, and that we can't investigate causes of what we see. The advantage though is that we are collecting data in natural settings.

In many practical, real-life circumstances, we are limited to observational studies. For example, in areas such as medicine and psychology, it is often neither ethical (morally right or wrong) nor practical to carry our experiments. To investigate diseases or psychological conditions, researchers often identify a group without the disease or condition, who are as similar as possible to those who do have the disease or condition. These are called **case-control studies**. This is one reason why twin registers are so valuable to researchers! As you can imagine, it's important to be aware of, and record data on, variables that might help understand what's happening in the data. Many big results in science and medicine have come about because clever and careful researchers 'noticed something'.

Case-control study: Observational study in which researchers identify a group of people without a disease or condition, who are as similar as possible to those who do have the disease or condition

LET'S START Do drivers really stop at stop signs?

Drivers are meant to come to a complete stop at stop signs before moving off. A complete stop is 0 km/h – do drivers do this? Mostly? Very few? Perhaps the location of the stop sign matters. Perhaps gender or age or both matter. Perhaps red cars don't come to a complete stop.

This has to be investigated by an observational study as we want to know what real drivers do in real situations, so it would be best also if they don't realise they're being observed. We should choose a number of stop signs. What time of day should we choose? If there is a lot of traffic, drivers are more likely to have to stop, whereas we're interested in whether they choose to stop (completely).

ISBN 978-1-107-61421-5

Perhaps the weather or day of the week makes a difference. No matter what size the study is, we need to make choices. If we are able to take observations at the same time at two different locations, then weather and day are constant; if not, it is possible that the weather and the day might have some effect. The least we can – and must – do is to carefully describe the circumstances. We can then say to what extent we think our data apply to more general situations, and a reader or listener can decide whether they agree with our assumptions or not.

Without weather and day, our variables above are stop completely/not stop completely (categorical), driver gender (categorical), driver age group (categorical). If we can, we should record type of vehicle, and, if we wish, colour. Observational studies require as much thought and care as experiments and surveys.

Key ideas

- In an observational study, investigators observe subjects without altering or controlling the subjects' behaviour.
- Most observational studies involve at least some choice or design of conditions.
- Observational studies collect data in natural settings, but almost always cannot investigate causes; experiments are usually needed to be able to say that something caused something else.
- It is very important to describe exactly how observations were made, so that any assumptions about what the data can be taken to represent, can be explained and judged.

Example 6: Who uses their mobile most? Who talks for longer?

How are we going to collect data on how often boys and girls use their mobile phones and how long they talk (rather than text or play games, and so on). One possibility is that we could survey them and ask them for information. But this is not the type of information people necessarily record, have access to, or remember. We could ask them to keep daily records for perhaps a week, but would they change what they usually do? However, this is one way to proceed.

Another possible way is to take observations in public places – in which case our comments can only apply to the types of public places we observe. We could take observations in public transport, and/or in places where people stay in one place for at least a while – perhaps parks. Perhaps we could record length of conversations and total length of other usage. They will need to be in a specified interval of time, or, given the practical challenges, something like per 10 minutes. What else should we record: gender, approximate age, time of day, as well as location?

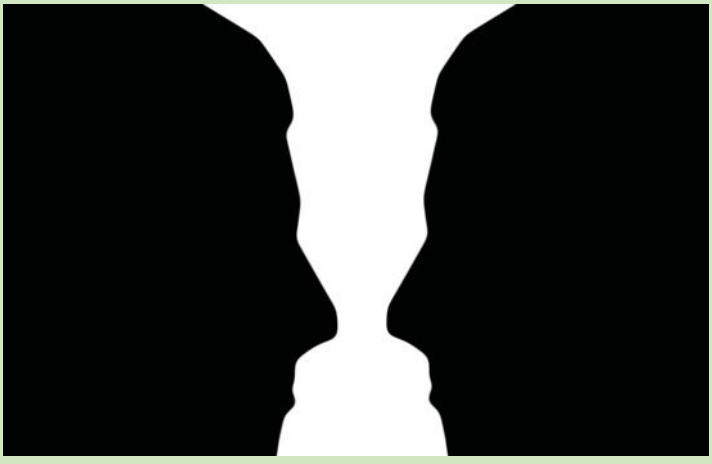
No matter how we decide to proceed, this topic for investigation has challenges. Careful description of how the data were collected will be essential in any report.

ISBN 978-1-107-61421-5

Example 7: Is it an experiment, a survey or an observational study?

The classification of data investigations into experiment, survey or observational study is not meant to be exact, and many real-life investigations are a mixture. But considering what's involved in them helps to see what is needed in their planning and design, and reporting on them, and what their statistical possibilities and limitations are.

Consider the Enrichment question in Exercise 5B on Cambridge GO, investigating the pictures seen in optical illusions. The emphasis in that question is on the experimental conditions of what to say to subjects and on randomisation. But part **c** asks what other variables should be recorded, and it could be seen as an observational study. But we are also choosing people to ask so perhaps it is a survey. It doesn't matter what we call it, and what is needed to be considered in each of these broad types of investigations, helps us in planning. Like surveys, we need to try to choose subjects randomly, and to phrase our questions carefully and the same for all. Like experiments, we need to consider what to control and what to randomise. Like observational studies, we need to record variables that might matter, to describe clearly and thoroughly exactly how the data were collected, and take care in justifying any general comments we might make.



ISBN 978-1-107-61421-5

Exercise 5D

1 Some studies have claimed that left-handers have a shorter life expectancy than right-handers.

- **a** Why is it not possible to carry out an experiment on this?
- **b** Can you think of a problem with past data for such studies?

2 Levels of pollution are constantly monitored in some cities.

- **a** Why are such studies observational?
- **b** What should be kept constant in such studies?
- **c** Apart from the actual level of pollution, name at least two other variables that should also be recorded.

3 It is wished to investigate the number of words toddlers (aged two years) can say.

- **a** Why must this be an observational study?
- **b** What aspects of a survey apply to this study?
- **c** Suggest at least two other variables that should be recorded in case they matter.

4 There have been some claims that bald-headed men have more body hair than men who are not bald.

- **a** Why is it not possible to carry out an experiment on this?
- **b** What must be considered and is not mentioned in this claim?
- **c** Suppose the data of a properly conducted study does provide some evidence for this claim. Does this mean that going bald produces more body hair?

5 The difference between the arrival and scheduled arrival times of trains at a central station are studied to investigate if trains on some routes tend to be delayed.

- **a** Why is this an observational study?
- **b** Apart from day of the week and scheduled time of arrival, what other variables should be recorded which might help in interpreting these data?

How long are your leaves?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

Chapter summary

Census

- Comes from the Latin *censere* – to rate
- Almost always refers to data collected on every member of a population
- National censuses are of vital importance to governments, businesses, industries.

Sample data

- A set of observations such that more observations could have been taken, sometimes without limit
- If used for comments on a general situation or population, must be obtained randomly with respect to issues of interest.

Experiments

- Investigators control conditions, fixing some and allowing others to vary
- Randomisation allows us to say that observed effects were caused by the choice of conditions
- Randomisation may be random allocation of subjects or of conditions or of order of conditions.

Sample surveys

- Involves asking questions of subjects chosen by sampling plan
- Sampling plan must be randomly representative of a larger population
- Practicalities include reaching selected subjects and reducing non-responses
- Questions must be clear, and as quick and easy to answer as possible
- How the data were collected, and the response rate, must always be reported.

Observational studies

- Subjects are observed without altering or controlling their behaviour
- At least some design involved, but causes cannot be investigated
- Exactly what data and how it was collected, must be reported.

Multiple-choice questions

1 Aeroplane passengers arriving in Australia from overseas complete arrival forms. What are the data from these forms?

A A census of arrivals by aeroplane **B** A census of all arrivals

C A random sample of arrivals **D** A sample survey

2 A random sample of the CensusAtSchool data for 50 Year-8 NSW students is selected. With respect to heights, who is it likely to be randomly representative of?

A All students who did CensusAtSchool **B** Year 8 NSW students

C Only Year 8 students in one school **D** Year 8 Australian students

3 A random sample of the CensusAtSchool data for 50 Year-8 NSW students is selected. With respect to opinions on exams, who is it likely to be randomly representative of?

A All students who did CensusAtSchool **B** Year 8 NSW students

C Only Year 8 students in one school **D** Year 8 Australian students

ISBN 978-1-107-61421-5

4 To compare two sunscreens, 50 people were chosen and paired by age and skin colouring. Each sunscreen was used by one person in each pair. What is the sample size?

A 50 **B** 100

C 25 **D** Depends on the results

5 In question 4 above, what should be randomised?

A Allocation of subjects to pairs **B** Allocation of sunscreens within pairs

C How long they spend in the sun **D** Numbering of pairs

6 A survey of 50 people consisted of 5 questions, with each question having 3 parts. All 50 responded. What is the sample size?

A The number who answered every question **B** 250

C 50 **D** 750

7 A phone survey uses a **table of random digits** to select phone numbers within the 02 (NSW) area code and calls are made in the evening. What population is being randomly sampled?

A NSW residents who eat at home **B** NSW residents with listed numbers

C NSW residents with land lines **D** NSW residents who don't own mobiles

Table of random digits: List of random numbers from 0 to 9, which can be used to obtain a set of random numbers in any range

Short-answer questions

1 A country conducts its national census by leaving census forms in letterboxes and asking households to complete them and leave them in the letterboxes to be collected on a certain day. If the forms are not in the letterboxes on that day, the collectors leave another form. What are some serious problems with this procedure?

2 Data are collected on the top 100 songs in two music charts in one year. The variables observed are lengths of songs, nationality of artists, genre of song, and country in which the song was produced.

a Are these data a census or a random sample?

b For which, if any, of these variables, could the data be reasonably taken as randomly representative of all songs played in that year?

c The total number of songs over the two lists is 180. How could this happen?

ISBN 978-1-107-61421-5

Cambridge University Press

3 For each of the following questions, are the available data appropriate to investigate the research question?

a Available data: Salaries for a random sample of males and females in the public service

Research question: Are women paid less than men who are in equivalent jobs?

b Available data: Pulse rates for males and females in a large secondary school

Research question: Do teenage males and teenage females have different pulse rates?

c Available data: The extent of a destructive pest on trees of a certain species randomly sampled in a region

Research question: Should there be a national program to destroy the pest?

d Available data: Opinions on whether recycled water should be included in drinking water collected from a random sample of people attending a festival

Research question: Does a majority of people support including recycled water in drinking water?

4 A famous simple but important experiment was conducted by the leading and pioneering statistician Ronald Fisher who reported it in his 1925 book *Statistical Methods for Research Workers*. A woman at Fisher's place of work commented that she could tell from the taste of a cup of tea whether the tea was poured after the milk or whether the milk was poured in after the tea. Fisher gave the woman 8 cups of tea, 4 with milk in first, and 4 with milk added after. <www.cambridge.edu.au/statsAC78weblinks>

a What had to be randomised in this test?

b What else would have required care?

5 An experiment is to be conducted to investigate the setting properties of jellies of different colours. Batches of jellies of different colours of one brand are prepared in small cups and placed in refrigeration to set. Each batch is removed from refrigeration after a specified time chosen by the investigators, so that each batch corresponds to an increasing amount of set time. The cups are turned upside down so that the jellies fall out. The amount of set is measured by the maximum spread of the jelly – so more spread means less set.

a What would need to be kept constant in this experiment?

b A pilot study showed that there was a practical problem with the refrigeration part of the experiment. What do you think that may have been?

ISBN 978-1-107-61421-5

6 Two methods for memorising information are to be compared.

a In one experiment, each subject uses both methods. What would need to be randomised?

In another experiment, subjects are paired, with each pair using the two methods – one for each pair member.

b What would need to be randomised?

c Suggest two variables that could be used to pair subjects.

7 An internet survey asks visitors if they are likely to shop online this Christmas. Apart from the voluntary nature of the survey, what is another problem with this?

8 A survey on gym usage is conducted by a short interview with users arriving at a gym. The interviews are conducted at the gym on the same day of the week over three weeks.

a Name two problems with this way of conducting the survey.

b Suggest an alternative way of conducting the survey by interview.

c Suggest an alternative way of conducting the survey by paper.

d Assuming the survey by interviewing was designed well, could the results be used for general comments on gym usage? If not, how could they be used by other gyms wanting to carry out their own surveys?

9 The transport authority wants to survey train users on a particular route.

a One method could be to choose a train at random and give a survey form to all passengers on that train. What type of sampling is this?

b Another method could be to randomly choose 20 passengers from each train on that route in a day. What type of sampling is this?

10 A teacher wants to investigate the amount of TV watched by students and their performance on exams. The teacher chooses students randomly and interviews them to find out how much TV they watched during the previous term, and their performance on the end of term tests.

a What else should he ask the students?

b Suppose he collects and analyses these data appropriately, and finds that higher-performing students generally do not watch much TV. Why can't he say that watching too much TV decreases performance? And why is the study seriously flawed?

ISBN 978-1-107-61421-5

11 A burger chain introduces a new meal deal and after a month, analyses sales data for the past two months to see the impact of the new meal deal.

a Is this an experiment or an observational study?

b Apart from sales of the new meal deal, what other sales records would they need to consider?

c Suggest how the two months could be compared.

Extended-response questions

1 Below is an extract from a report in 2011 by the Cancer Council of Australia.

Cools go cool on tanning

National Skin Cancer Action Week: November 20-26, 2011

• 15% fall in teens who prefer a tan since 2003–2004
• Only 12% of teens believe a tanned person is more healthy.

Young Australians are changing their attitudes towards tanning with fewer seeking the bronzed look than ever before, according to new Cancer Council research released today (22 November 2011).

Cancer Council's latest National Sun Protection Survey conducted in summer 2010–11, shows the preference for a suntan among 12 to 17 year olds has steadily dropped, down to 45% since the previous surveys (51% in 2006–07 and 60% in 2003–04).

Source http://www.cancer.org.au/news/media-releases/media-releases-2011/new-research-teens-go-cool-on-tanning.html

The National Sun Protection Survey of 2010–11 was the third survey of this type; the first was conducted in 2003–04. The study is funded by the Cancer Council Australia and the Australian Government through Cancer Australia. Trying to obtain accurate and consistent information about people's sun protection habits is very challenging because people themselves can be quite inconsistent in behaviour of this type, and it may depend on a number of factors as well as on interpretation of questions. The 2006–07 survey

ISBN 978-1-107-61421-5

reached respondents through random phone dialling and phone interviews were conducted on Monday and Tuesday evenings during summer. The interviews focused on weekend behaviour in summer, and also recorded whether the person was an adolescent or an adult, and in which state they lived.

a Why do you think questions are asked about a specific weekend?

b Suggest two reasons why the phone interviews were done on Monday and Tuesday evenings.

The Cancer Council's recommendation for applying sunscreen is to 'slop on' SPF30+ broad-spectrum sunscreen 20 minutes before going out and reapply every two hours. They comment that people often do not use enough sunscreen, and recommend at least a teaspoon of sunscreen to each arm, leg, front of body and back of body and at least ½ teaspoon to the face (including the ears and neck).

c For what ages and sizes do you think the Cancer Council is suggesting these amounts?

d Suggest how to conduct an experiment to compare two sunscreens on how quickly such an amount of sunscreen is absorbed. Use arms only.

2 It is wished to investigate drying a towel in three different brands of clothes dryers. Two types of towel are available, one washing machine and three dryers. Each dryer has two temperature settings as well as timer settings. Towels are rinsed and spun dry in the same washing machine on the same setting. The measure of amount of drying is difference in the weight of the towel (there are appropriate scales available to measure this). Three different timer settings are used.

a How many different combinations of experimental conditions are there?

b A different towel is used for each combination of conditions. If a towel is taken out and measured after the shortest time setting, why is it decided not to put that towel back in until the next timer setting?

c It is decided to do two observations for each combination of experimental conditions. How many towels of each type are needed?

d What would the pilot experiment help in deciding?

e Dryers are still warm after a time setting is complete. How cool should a dryer be before it is used for the next towel?

f Towels will need to be placed in dryers as soon as the spinning cycle is finished. Why is this?

g How should this experiment be randomised?

ISBN 978-1-107-61421-5

CHAPTER

6

Variation across datasets

What you will learn

6-1 Categorical data and variation of proportions

6-2 Quantitative data and sampling variation

6-3 Variation of sample mean and sample median

Data and variation are everywhere!

Data are everywhere, and wherever there are data, there is variation. Variation is natural to everything in our world – to all life, all nature, and to everything humans and nature make or do. People are constantly trying to explain variation, whether in medicine, science, psychology, economics, weather, agriculture – everything that relates to our lives and our world. Why do some people get ill and others not? Why do some people react to foods or drugs and others not? Why do some weather patterns develop into major weather events and others not? Are males better at reading maps than females? Is it harder to concentrate while listening to music? In asking such questions, people are trying to discover reasons for variation. But even if we can find reasons for some of the variation, we understand that variation due to chance is an essential part of our world. Variation due to chance is even a strength in natural development. And in trying to find reasons for variation, we have to allow for variation due to chance.

Why are there so many opinion polls?

Just as we will never be able to predict everything, we also cannot observe or measure everything all the time. Our data are almost always sample data – just a part of what could be observed or measured or asked if we could do it non-stop forever. So samples of data from the same situation will vary simply because they are samples. Even national census data are not exact.

AUSTRALIAN CURRICULUM

Statistics and probability

- Data representation and interpretation
- Explore the variation of means and proportions of random samples drawn from the same population **(ACMSP293)**
- Investigate the effect of individual data values, including outliers, on the mean and median **(ACMSP207)**

ISBN 978-1-107-61421-5

National offices of statistics use many sophisticated statistical techniques to estimate and cross-check for errors, and to 'allow for' omissions, errors, mistakes and misunderstandings.

Data summaries also vary

Wherever there are data, graphs and summary statistics are used to present the data.

Summary statistics, such as sample proportions, means and medians, are used to estimate more general quantities. But, because they are calculated from samples of data, they also vary because the samples vary. This type of variation is called **sampling variation**. This variation can occur in random samples from the same population, or can be obtained by repeating the same experiment or even by taking observations of the same situation. We need to understand and allow for it when we use data summaries. In this chapter you will see something of how much samples and their proportions, means and medians can vary.

HINT
Remember that summary statistics are quantities calculated from data to represent aspects of the data.

Sampling variation: Variation across different data samples obtained under the same conditions

PRE-TEST

1 The table below gives data from 200 randomly chosen school students on their main method of travelling to school.

	Car	Bus	Walk	Cycle	Other	Total
Male	35	30	10	10	7	92
Female	53	28	20	3	4	108
Total	88	58	30	13	11	200

a What is the sample size?

b What is the relative frequency (expressed as a percentage) of:

i girls who walked?

ii boys who travelled by car?

iii those who travelled by bus who are boys?

c Do we have enough information to be able to say of what population or larger group these data are randomly representative? If not, what do we need to know?

d Another survey asked school students how they travelled to school this morning?

i What is an advantage of this question compared to the one asked?

ii What is a disadvantage of this question compared to the one asked?

ISBN 978-1-107-61421-5

2 The table below gives data from 200 randomly chosen school students in Years 7 and 8 in NSW and Victoria on their favourite take-away food.

	Chips/ Fries	Hamburgers Kebabs/ Wraps	Pizza/ Pasta	Rice/ Noodles	Chicken/ Fish	Salads/ Fruit	Other	Total
NSW	22	15	23	10	21	6	18	115
Vic	15	18	12	13	11	3	13	85
Total	37	33	35	23	32	9	31	200

a What is the relative frequency (expressed as a percentage) of:

i NSW students who said chicken or fish?

ii pizza/pasta lovers who live in NSW?

iii Victorian students who said chips or fries?

b Do we have enough information to be able to say of what population or larger group these data are randomly representative? If so, what?

3 The following are data on the times to complete a concentration task (in seconds) for 25 school students.

36 42 59 36 41 42 50 42 35 38 27 52 62 81 42
35 30 30 29 40 36 31 30 37 36

a Draw a stem-and-leaf plot of these data.

b Calculate the mean and median of these data.

4 Two programs designed to increase fitness of school students are tested by selecting 50 students at random. These students were then grouped into similar pairs with respect to lifestyle, and each member of a pair did one of the programs. Fitness was measured before the program, halfway through and at the end of the program. What is the sample size?

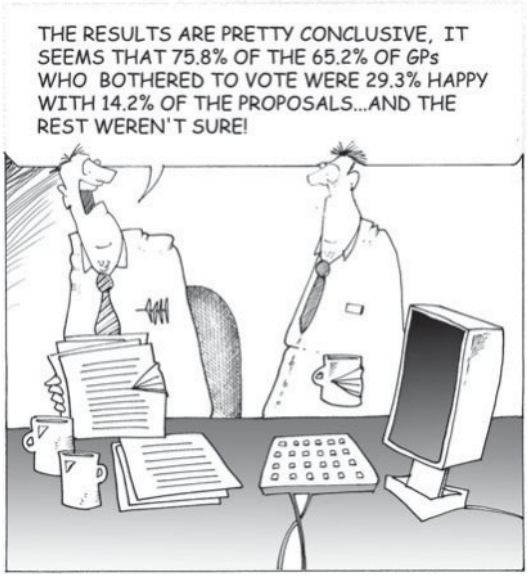
Terms you will learn

estimates
outliers
randomly generated data
resampling
sample mean
sample median
sample proportions
sampling variation
sampling with replacement
sampling without replacement
simulation

ISBN 978-1-107-61421-5

6-1 Categorical data and variation of proportions

The types of data summaries used depend on the type of data. For categorical data, proportions (or relative frequencies) are the key quantities calculated from data. And as you see if you look at any form of media, percentages are everywhere!



Relative frequencies obtained from a sample of data are called **sample proportions**. Sample proportions can be expressed as fractions, decimals or percentages. Percentages are most commonly used, and they are constantly being quoted or reported. A popular joke is that '99% (or whatever percentage you'd like to quote) of statistics are made up on the spot'.

Sample proportions: Relative frequencies obtained from a sample of data; term usually used when a relative frequency is being used to estimate a population proportion … *see glossary*

Percentages are most commonly quoted from surveys. There seems to be no end to the fascination for people of opinion, lifestyle, marketing and the many surveys in so many areas of government, industry, business and research. Percentages are also often quoted from areas such as medical, health and psychology. For example, 'x% of people who take this medication may suffer side effects'.

We have already seen in Chapter 5 something of how surveys and experiments require care so that the sample data can be considered to be a random sample. But even if this care is taken, there will still be variation across samples of data simply because they are samples. So how much do sample proportions vary across random samples collected from the same population or under the same circumstances?

LET'S START Collecting samples on clasping hands

In Example 2 of Chapter 1, we considered that when people clasp their hands, they usually have the same thumb on top each time, and find it very difficult to clasp their hands so that the other thumb is on top. If you do an internet search, you will find a number of websites discussing this. There are debates, but there is general agreement that left thumb on top is more common.

Cambridge GO

In one study by students, a sample of 203 students gave 116 with the left thumb on top, that is, a sample proportion of 0.57 correct to two decimal places, or 57%. This is an **estimate** of the proportion of the population who have left thumb on top when they clasp their hands. But how good an estimate is it?

Estimates: Quantities calculated from sample data to be used as approximations of quantities relevant to the more general situation or population of which the sample data can be taken as being randomly representative

ISBN 978-1-107-61421-5

Question

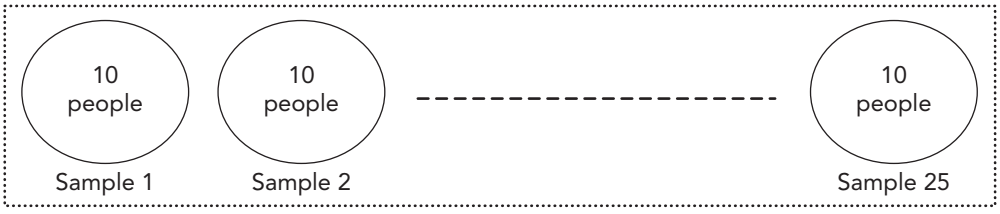
If we asked lots of groups of people, what sort of variation would we find for the sample proportion across the groups?

Investigating this question

Let's suppose 57% is in fact the proportion of the whole population who have left thumb on top when clasping hands. Let's call these people Ls. On the companion website is a table of **randomly generated data**. The table has five hundred 0s and 1s randomly generated such that the chance of a 1 each time is 0.57. It's as if we had a coin with 0.57 as the chance of a head and we tossed it 500 times. The table is arranged in 25 rows of 20 observations in each row so that it's easy to work with.

Randomly generated data: Data constructed artificially by some mechanism so that the data are produced under some specified rules of chance across all possible values

We can use this table to 'collect' a number of groups of people from this population. Let's first get 25 groups with 10 people in each group, as illustrated by the diagram below.



Calculate the proportion of 1s in the first 10 observations in each row. This gives you 25 observations of sample proportions of 10 people. Plot these 25 sample proportions using a stem-and-leaf plot. Notice that the sample proportions range from 0.3 to 0.9! The median – the 'middle' of the sample proportions – is 0.6.

Repeat the procedure for the second 10 observations in each row. How do these compare with the first group of 25 sample proportions? Notice that the sample proportions again range from 0.3 to 0.9 but the stem-and-leaf plots look quite different. The median is again 0.6.

ISBN 978-1-107-61421-5

Now repeat the above for each whole row of 20, giving 25 sample proportions of groups of 20 people. The sample proportions of 20 people range from 0.4 to 0.85, again with a median of 0.6.

All of these sample proportions come from a population with 57% of Ls. Are you surprised by how much variation there is in the sample proportions? Notice that the sample proportions are slightly less variable for the samples of size 20 than for those of size 10. But yes, we need big samples to get reasonable estimates of proportions!

Key ideas

- For categorical data, a sample proportion is the number of observations in a category divided by the total number of observations.
- If a sample of data is randomly representative of some population, then sample proportions estimate the corresponding population proportions.
- Sample proportions can vary greatly across random samples collected in the same situation or from the same population, especially for smaller sample sizes.

Example 1: What happens if we ask 1000 people?

Let's continue considering how many Ls (people who have left thumb on top when clasping hands) there are in random samples of people if there are 57% of Ls overall. Let's see what can happen if we ask lots of people, say 1000. That is, in the whole population, 57% of people have their left thumb on top when they clasp hands. In one study, 100 groups, with each group having 1000 people, were randomly generated. For each group, the percentage of people with left thumb on top (the Ls) was calculated.

Question

How much did these 100 percentages vary?

Investigating this question

Below is a stem-and-leaf plot of the percentages of Ls in 1000 people in 100 random samples.

Leaf unit = 0.10

53	0
54	0
55	000000000000000
56	00000000000000000000000
57	0000000000000000000
58	000000000000000000000000000
59	000000
60	00000000

ISBN 978-1-107-61421-5

We see that in these 100 samples, the percentages with left thumb on top ranged from 53% to 60%. This is certainly much less variable than in the samples of 20 people, but still remarkably variable considering each percentage is of 1000 people! The median of these is 57%. So the sample proportions that are trying to estimate 57% are centred on it.

Clearly we have to be very careful in reporting percentages, and we need a lot of data to be able to accurately estimate proportions.

CAUTION
Don't forget to always report how many observations were collected, and how they were collected.

Example 2: How many teens prefer suntans? Let's simulate.

Extended-response question 1 at the end of Chapter 5 reports that the Cancer Council's latest National Sun Protection Survey shows the preference for a suntan among 12 to 17 year olds has dropped to 45%.

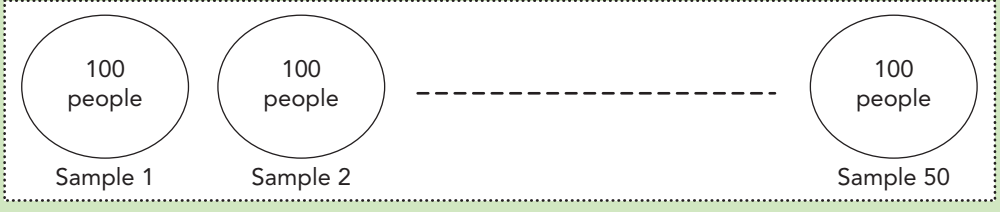
If this is true, how can we investigate the variation of the percentage who prefer a suntan in different samples of 12 to 17 year olds?

Solution

The solution is to use a computer to produce random data. This is called **simulation**. Simulation is when computers are programmed to mimic a real situation and generate data that can be taken as coming from that real situation.

We can assume the population proportion is 45% and use Microsoft Excel to simulate data from this population. Suppose we want 50 samples, each of size 100 as illustrated in the diagram below.

Simulation: When computers are programmed to mimic a real situation and generate data that can be taken as coming from that real situation

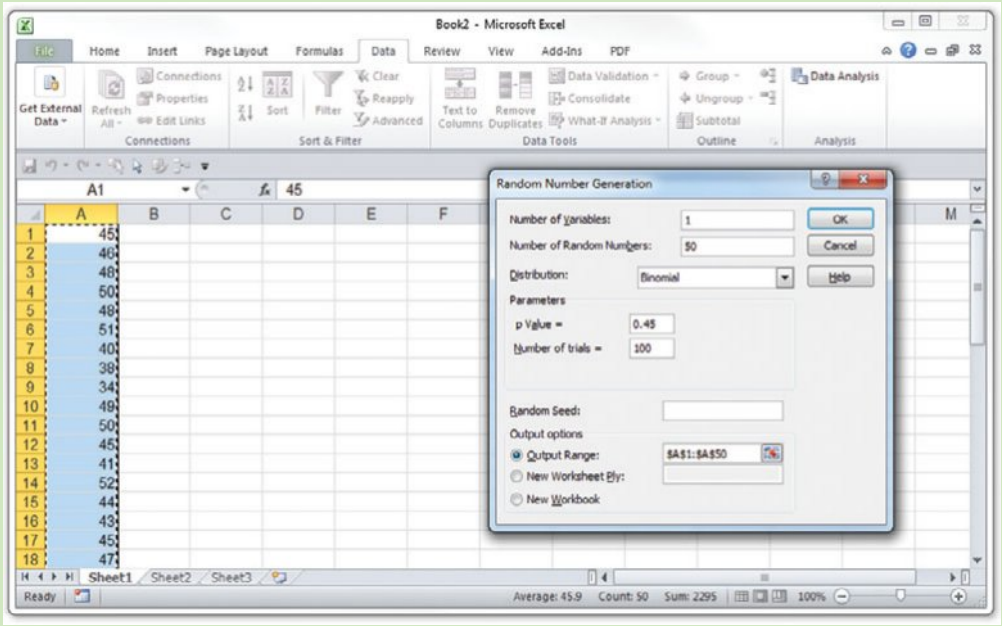


Follow these steps to generate your own random data:

Cambridge GO

1. In Excel, you need to use the Data Analysis command. In Excel 2010 it is on the Data tab of the ribbon. Follow the instructions on Cambridge GO or search Excel Help for 'Load the Analysis Toolpak'.
2. Click Data Analysis and choose Random Number Generation.
3. In the Random Number Generation box, make the settings shown in the following screen capture and click OK.

ISBN 978-1-107-61421-5



The output will consist of a set of numbers out of 100, giving you 50 simulated percentages of 100 people with left thumb on top. Use a stem-and-leaf plot to look at how much these vary.

Here is an example – yours won't be the same because we're generating random data. The median is 45% (as we would expect) but the sample proportions range from 32% to 58%.

HINT
If you want a different sample size (e.g. 200), you will need to divide your output by the sample size (e.g. 200) to get your set of simulated proportions.

Stem-and-leaf of 50 proportions of samples of size 100; population proportion = 45%

Leaf unit = 1.0

3	2
3	45
3	
3	99
4	00011
4	2222333
4	4444455555
4	6666667777
4	9999
5	0000
5	23
5	4
5	6
5	8

ISBN 978-1-107-61421-5

Exercise 6A

1 **a** In the simulation reported in Example 2 above, how many of the 50 simulated samples had sample proportions greater than 45%?

b Use Excel as in Example 2 to simulate your own simulated samples, each of size 100, and count how many of your simulated samples have sample proportions greater than 45%.

2 In a study on pedestrian behaviour at an inner-city intersection with traffic lights and pedestrian lights, observations were made during morning, early afternoon and evening on all pedestrians crossing the intersection. Among the variables recorded for each pedestrian were gender and whether the pedestrian started crossing when the pedestrian lights were showing green, flashing red, or steady red. The table below gives the observed frequencies for the combination of these two variables.

	Female	Male	All
Flashing	21	33	54
Green	546	618	1164
Red	13	29	42
Total	580	680	1260

a What percentage of female pedestrians crossed against the red or flashing red?

b What percentage of those crossing against the red were male?

c The percentages of females and of males crossing against the red were 2.24% and 4.26% respectively. Below are two stem-and-leaf plots, each of 50 sample proportions of random samples of size 500 of data from populations with population proportions of (1) 2.24% and (2) 4.26% respectively.

(1) Leaf unit = 0.1

0	4
0	6
1	024
1	666668888888
2	0000002222244444444
2	666688888
3	0002
3	8

(2) Leaf unit = 0.1

2	24
2	
3	00444
3	6666688888
4	000222222244444
4	66888888
5	002244
5	68
6	00

i What are the ranges, medians and averages of these sample proportions?

ii How many of the 50 sample proportions in (1) are greater than 3%?

iii How many of the 50 sample proportions in (2) are less than 3.5%?

iv How many of the 50 sample proportions in (2) are less than the largest of the sample proportions in (1)?

v A claim is made that males tend to cross against the red lights at pedestrian crossings more than females do. What do you think of this claim?

ISBN 978-1-107-61421-5

3 The CensusAtSchool survey includes questions about what students eat at breakfast and first asks if they ate breakfast that morning. A random sample of 200 Year-8 Australian students who participated in the CensusAtSchool was selected. The table below gives the observed frequencies for the students' genders and whether they ate breakfast or not in the morning they completed the survey.

	Female	Male	All
Ate breakfast	80	86	166
Did not eat breakfast	26	8	34
Total	106	94	200

a What percentage of females did not eat breakfast?

b It is reported that more than 75% of female Year-8 students did not eat breakfast. What mistake has been made?

The percentage of students overall who did not eat breakfast is 17%. Below is a stem-and-leaf plot of 50 sample proportions of random samples of size 200 from a population in which 17% do not eat breakfast.

Leaf unit = 0.1

11	5
12	55
13	055555
14	000055
15	005
16	000055555555
17	00055
18	00055
19	0055
20	00000
21	0

c Give the range, median and average of these 50 sample proportions.

d How many of these 50 sample proportions are less than the population proportion of 17%?

e How many of these 50 sample proportions are at least 2% away from the population proportion of 17%? That is, less than or equal to 15% or at least 19%?

Comparing collected and simulated samples
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

6-2 Quantitative data and sampling variation

The examples and exercises of Chapter 2 were full of variation in quantitative data. Indeed, Chapter 2 introduced ways to present, summarise and comment on features of the variation in quantitative data.

Where does variation come from? Differences across people, plants, animals and birds, and conditions such as weather and traffic, can all contribute to variation. Differences due to 'natural' variation can mean that individuals – whether people, animals or plants – can differ in countless ways. We have already seen examples in the earlier chapters such as how often people blink, how long they take to do a task, their pulse rates, blood pressure and opinions, the growth rates of plants and the flight times of birds.

Example 6 of Chapter 3 gives just three of many possible samples of data from rolling a fair die 60 times. We are used to the variation we get in rolling a fair die a number of times because that is what so many board games and other games of chance are based on. In this section, in real situations, we explore the variation that is possible across random samples of data on quantitative variables. This is called **sampling variation**. And because the random samples vary, any quantities calculated from them will also vary. So while we are exploring variation across samples, we will also start to consider the **sample means** (averages), **sample medians** and ranges of the data.

Sampling variation: Variation across different data samples obtained under the same conditions … *see glossary*

Sample mean: Average of values in a sample

Sample median: Median of values in a sample

LET'S START Misuse of an express computer lab

The problem

One computer lab in a large school is meant to be for express use, with students requested to restrict their time on a computer to 15 minutes, but there is no automatic cut-off at 15 minutes. Records are not currently kept of usage, but, because of student complaints, the usage is being investigated to consider the possibility of extending the time limit but with better control. For example, a warning 5 minutes before an automatic cut-off is applied. An unobtrusive observational study was conducted and 150 times on computers in the lab were collected.

ISBN 978-1-107-61421-5

Below is a stem-and-leaf of these 150 times in minutes. As you can see there are many times greater than 15 minutes, with a small number even 2 hours or more.

HINT
Why was it important to conduct this study unobtrusively?

Leaf unit = 1.0

0	234455556666667788999999999
1	000000111111112222223333333344444455555555555557889
2	0001344445555788
3	0000023455566999
4	12235577899
5	045556
6	015578
7	00579
8	0027
9	08
10	0
11	6
12	0069

The questions

If students observed a smaller sample of times, for example 10, what impression would students get? How variable would samples of 10 be? How variable would the means, medians and ranges be?

How to investigate

To investigate these questions, we can use the 150 times as our 'population' of times and generate random samples from these 150. We can use Microsoft Excel or other statistical software to generate random samples from the 150 times – see Example 4 for how to do this in Excel. Alternatively, we could write the 150 times on pieces of paper and choose pieces of paper at random (see Example 3) but this is very tedious and time consuming. If we take the 150 times as being our 'population', and sample randomly, then each observation in the 150 must be equally likely to be chosen each time. So repeated values in the sample have a greater chance of being chosen. For example, if there are three observations in the original dataset with the same value, that value is three times more likely to be chosen in resampling than a value that occurs only once in the original dataset.

In statistics, this is called **resampling**. It is a very important technique for exploring sampling variability in the situation in which our original data were collected.

On the next page are just four possible random samples of size 10 that could be observed by waiting students, obtained by resampling the data. Although it would be best to also look at these samples using stem-and-leaf plots, you can see the different impressions waiting students could get! Also notice how much the mean, median and range vary.

Resampling: Taking random samples of data from a set of previously observed data; each observation in the dataset is equally likely to be chosen in the resampling … *see glossary*

ISBN 978-1-107-61421-5

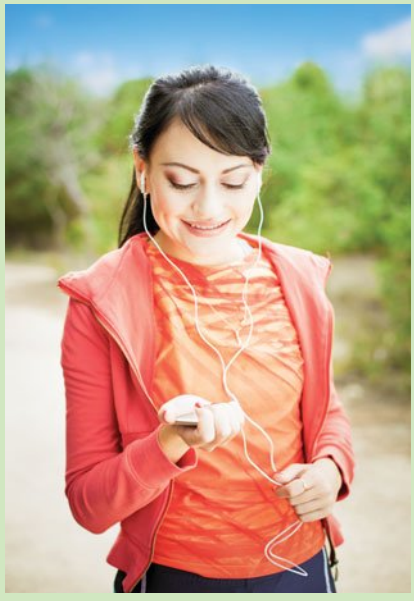
Sample	Data in samples obtained by resampling										Mean	Median	Range
Sample 1	100	129	7	12	116	25	9	45	18	47	50.8	35	122
Sample 2	20	13	8	14	36	30	14	14	24	55	22.8	17	47
Sample 3	13	13	11	12	11	15	116	25	42	67	32.5	14	105
Sample 4	6	6	6	20	36	9	15	8	47	20	17.3	12	41

Key ideas

- Random samples of data collected under the same circumstances or from the same population can vary greatly from each other; this is called sampling variation.
- Quantities calculated from data, such as data means, medians and ranges also vary across samples.
- Understanding and allowing for sampling variation is very important in statistical investigations, analysis and interpretation.
- A way of exploring possible variation in the ways a dataset was collected is to resample – to randomly sample the set of data as if it was a population of values.

Example 3: Random shuffling of songs

A student put most of her favourite songs on her iPod and often uses random shuffle in listening to songs. She has a total of 198 songs available to be listened to. If she listens to 20, there are billions of different possible groups of 20 songs she could hear without any repeats. The actual number has 28 digits before the decimal – it is greater than the number given by 13 followed by 26 zeros. This is called **sampling without replacement**. If she decided she didn't mind how often she heard a song and used repeat shuffle, you can imagine how many possible groups of 20 there would be – that is called **sampling with replacement**.



Let's consider a random selection of 10 songs from just 50, whose lengths in seconds are given below. The mean of these data is 233.32 s, the median is 232 s, and the range is 129 s.

267	241	306	228	271	203	201	238	231	197	238	233	265
294	299	234	230	224	242	243	257	197	253	226	242	290
212	207	206	224	204	202	270	177	213	216	247	212	262
225	206	199	264	248	185	207	185	252	258	235		

Sampling without replacement: Selecting observations at random from a limited group of values or items so that as each value or item is selected, it is not available for subsequent selections

Sampling with replacement: Selecting observations at random from a limited group of values or items so that as each value or item is selected, it is replaced and is available for subsequent selections

ISBN 978-1-107-61421-5

Here are the lengths in just two samples, each of 10 songs, randomly selected without replacement:

243 185 247 238 224 203 231 177 224 262
248 252 199 262 241 248 204 247 225 199

HINT
The word resampling means sampling from a large set of data with replacement.

The means, medians and ranges of these two samples are:

Sample 1: mean = 223.4 s, median = 227.5 s, range = 85 s

Sample 2: mean = 232.5 s, median = 244 s, range = 63 s

And here are the lengths in just two samples, each of 10 songs, selected with replacement:

270 207 206 264 204 212 257 185 306 233
238 233 258 290 224 213 235 201 238 177

The means, medians and ranges of these two samples are:

Sample 1: mean = 243.4 s, median = 222.5 s, range = 121 s

Sample 2: mean = 230.7 s, median = 234 s, range = 113 s

These are examples of the variation we can get in random sampling. You can generate such samples yourselves by writing the 50 lengths on pieces of paper, placing the 50 pieces of paper in a container, and selecting pieces of paper at random from the container, either with or without replacement each time.

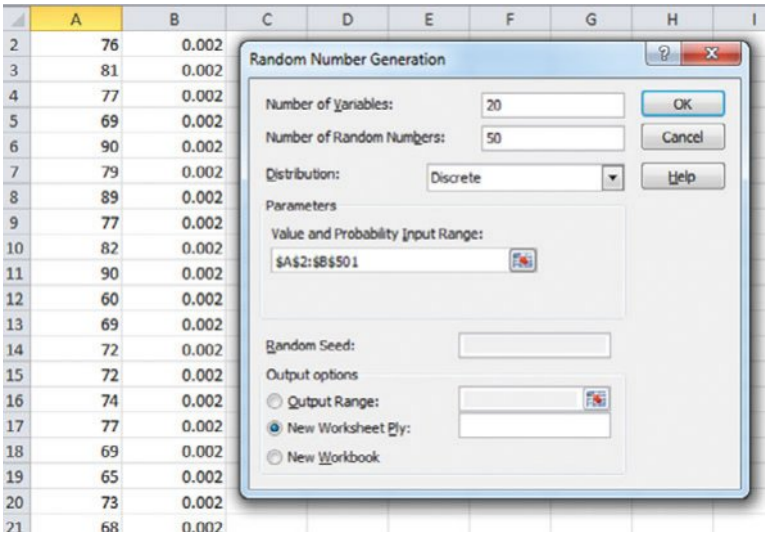
Example 4: Resampling using Microsoft Excel

Cambridge GO

This example can be found on Cambridge GO.

Exercise 6B

1 Refer to Example 4. A dataset consisting of the resting pulse rates of 500 Year-8 students was entered into Excel column A, with the heading *Pulse rate*. Column B consisted of 500 values, all = 0.002, with the heading *Probability*. In Data Analysis, under Tools, Random Number Generation was chosen. The number 20 was entered into Number of Variables. The number 50 was entered into Number of Random Numbers. In Distribution, Discrete was chosen, and in Value and Probability Input Range, $A$2:$B$501 was entered.



Cambridge GO

ISBN 978-1-107-61421-5

a How many samples are being generated?

b What is the size of the samples being generated?

c Is the sampling with or without replacement?

2 Consider the above Let's start section on the misuse of an express computer lab.

a Use the given stem-and-leaf to show that the sample median of the 150 times in the dataset is 17.5 min.

b Use stem-and-leaf plots to graph the four samples of size 10 obtained above by resampling. Use a leaf unit of 10.

c Check the values of the averages, medians and ranges of each of these four samples of size 10.

d For each sample, what proportion of the sample has values of at least 20 min?

e The average of the 150 original observations is 30.6 minutes (correct to 1 decimal place). Comment on the comparison amongst the four samples and also compared with the original data.

3 Question 4 in Exercise 2B refers to a dataset of the lengths in minutes of 128 CDs. This dataset is available in the Excel file called *CDs*. <www.cambridge.edu.au/statsAC78weblinks>

Cambridge GO

Below are 4 samples, each of size 15, obtained by randomly resampling from the 128 values.

59	77	67	51	47	78	55	37	73	58	74	72	53	43	60
71	58	78	65	43	70	56	70	60	47	74	78	43	59	56
65	72	52	44	37	74	60	45	53	65	46	49	64	37	57
72	24	60	57	62	73	43	65	43	46	44	41	37	76	78

a Use stem-and-leaf plots to graph these samples.

b Obtain the medians, averages and ranges of each of these samples.

c Compare the summary statistics obtained in part **b** for the samples with those of the original dataset of 128, given in question 5 of Exercise 2C.

d What proportions of each of the four samples is greater than the median of the original data (59.5 min)?

ISBN 978-1-107-61421-5

4 A drop-off/pick-up zone in a city has a time limit of 2 min. Observations were taken throughout a day on how long vehicles stayed in this zone, giving 200 observations in seconds. These data are available on the companion website in an Excel file called *Drop-off*. Below is a stem-and-leaf of the 200 observations. <www.cambridge.edu.au/statsAC78weblinks>

Leaf unit = 10

0	00000000000000000000000000000000000000000000000000000011111111111111+
0	22222222222222233333333333333
0	4444444455555
0	666666666677777
0	8888999
1	001111
1	2
1	45
1	7
1	9
2	01
2	
2	5
2	
2	8
3	00000000000000000000

HINT
The + sign in the first leaf means that there are more observations in this group than have been graphed.

a Why do you think there are so many observations of 300 seconds?

b Below are 3 samples, each of size 20, obtained by randomly reampling from the 200 values.

Sample 1: 7 53 49 4 87 5 6 90 5 53 12
10 50 5 5 216 36 4 26 300

Sample 2: 90 8 69 82 36 8 216 8 153 82
10 300 11 32 9 4 70 30 110 9

Sample 3: 62 108 300 12 8 12 300 31 83 6
197 65 6 11 300 46 4 112 21 300

i Use stem-and-leaf plots to graph these samples.

ii Obtain the medians, averages and ranges of each of these samples.

iii Compare the summary statistics obtained in part ii for the samples with those of the original dataset which are average = 62.81 s; median = 19.5 s; range = 297 s.

iv What proportions of each of the three samples have values greater than the median of the original data (19.5 s)?

Enrichment

How tall are Year 8 students?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

6-3 Variation of sample mean and sample median

The examples and exercises in section 6-2 illustrate the variation we can get in data in different samples collected under the same circumstances or from the same population. When we randomly generate smaller samples of data by resampling an original dataset, we can see in graphs such as stem-and-leaf plots how they can represent the original dataset. But the examples and exercises of section 6-2 illustrate that the variation in the values of the sample means, medians and ranges can be quite large; this is a worry when a mean or median or range are quoted from data. How do we know how much variation there could be in the values if another dataset in the same circumstances or from the same population, had been randomly selected?

The examples and exercises of section 6-2 also illustrate that even if the chance of getting some rather extreme values is small, when a random sample does include them, they can have a very great effect on summary statistics. Extreme values are sometimes called **outliers**, but how extreme does a value have to be to be called an outlier, and should we do anything about these values? There are no fixed or simple rules for saying if an observation is an 'outlier'. If an observation seems to be unusually far away from the rest of the data, it should be checked to see if there are any reasons for it to be considered to be different. In this section we focus on how means and medians vary, and the effects of extreme values. The possibilities of extreme values in samples are also considered.

Outliers: Extreme values in a dataset – that is, very large or very small values – that are considered to be unusually far away from the rest of the data

LET'S START How long do students spend in an express computer lab?

Let's look again at the data of the lengths of times students spent at the computers in the express computer lab. The 150 observations are plotted in a stem-and-leaf plot in Let's start in section 6-2, and question 2 of Exercise 6B tells us the median is 17.5 minutes and the mean is 30.6 minutes. So half of the students spent between 2 minutes and 17.5 minutes, but the half who spent more than 17.5 minutes, spent between 17.5 minutes and

ISBN 978-1-107-61421-5

129 minutes! This is why the mean is so much greater than the median. The median gives us the value that half the observations are greater than and half less, but it doesn't matter what the values are – just that the median splits them into two groups. But the data mean is affected by the actual values of all the observations.

Effect of extreme values on data mean and median

The data mean uses the actual values of the observations, so is particularly affected by very large or very small observations. The median is affected only by whether they're small or large. In this dataset, we have a much greater spread of values above the median, with a few students spending quite a long time at the computers. So these contribute quite a lot to the total time and therefore to the total/150 = data mean.

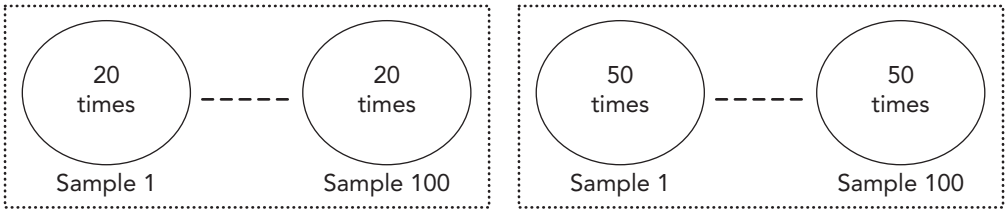
The total of the 150 times is 4589 minutes (so mean = $\frac{4589}{150}$ = 30.593). What happens if we omit the largest time? Looking at the stem-and-leaf, we see that this is 129. So the total of the other 149 times would be 4589 − 129 = 4460 and the mean = $\frac{4460}{149}$ = 29.93 minutes.

What happens to the median (17.5 minutes)? Again looking at the stem-and-leaf, we see that 17 must be the 75th observation, because for 150 observations, the median is halfway between the 75th and 76th. So for 149 observations, we now need the 75th observation, which is 17 minutes.

So the mean has decreased slightly and so has the median. The effect of omitting large values in this dataset will not be greatly different for the mean and median because there are observations spread all along up to the largest ones. But Example 5 below is quite different.

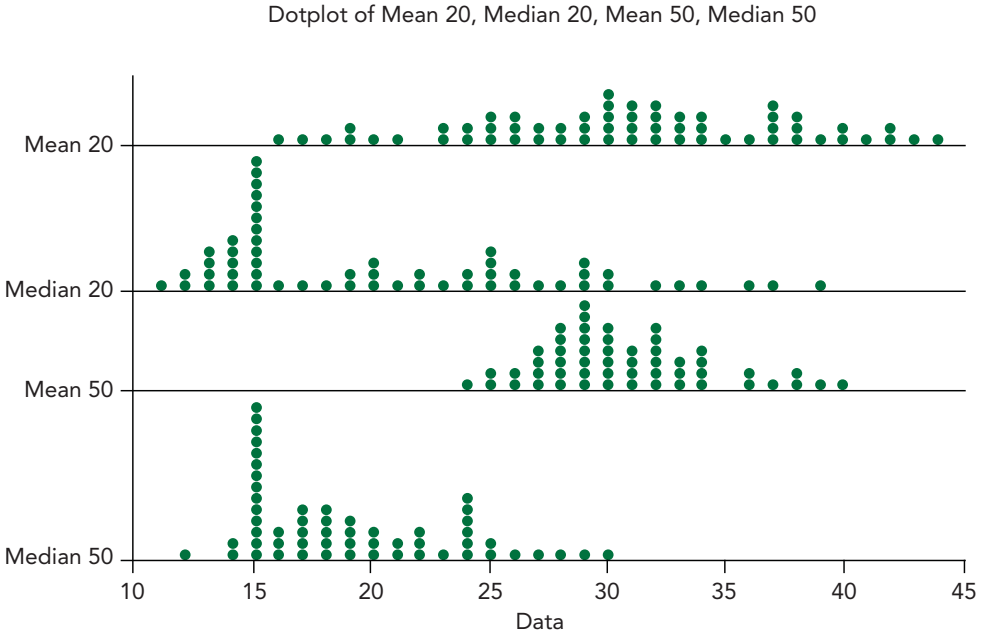
Sampling variation of data mean and median

If we resample from these data, how much will the sample mean and median vary across the samples? And what difference will it make if we take small or larger samples? To get some idea, two lots of 100 random samples were obtained by resampling these data (the full dataset of 150). The first group of 100 are samples of size 20, and the second group of 100 are samples of size 50. The diagram below illustrates this.



For each of the samples, the mean and median were calculated. Below are dotplots of the 100 means and 100 medians of these two lots of 100 random samples.

ISBN 978-1-107-61421-5



Each symbol represents up to 2 observations.

There is a lot to notice in these dotplots.

- The values of the means and the medians vary considerably, but there's less variation for the samples of size 50 than for the samples of size 20.
- We can see that the values of the means are centred around the mean of the original dataset (30.6) and the values of the medians are centred around the median of the original dataset (17.5). The sample means are trying to estimate 30.6, and the sample medians are trying to estimate 17.5.
- We can also see the effect of the larger values in the original dataset – the values spread out from 17.5 up to 129. Because some samples have more of these larger values in them than others, both the means and the medians in the samples of size 20 vary a lot. The medians in particular have a lot of variation among the larger values.

The stem-and-leaf plots of these two groups of 100 means and 100 medians are available on Cambridge GO. <www.cambridge.edu.au/statsAC78weblinks>

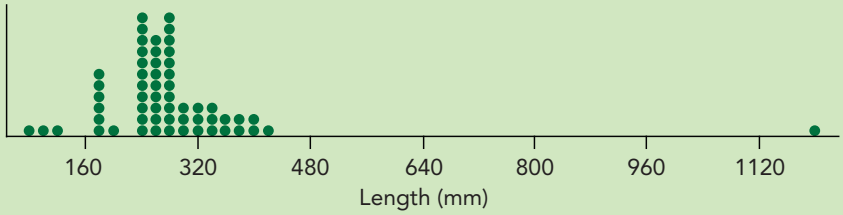
Key ideas

- Extreme data values, both large and small, tend to affect the data mean more than the data median. The data median is just affected by whether they are small or large, but the data mean is affected by their actual values.
- Sample means and sample medians can vary greatly across random samples collected under the same circumstances. The variation is greater in smaller samples than in larger ones.
- In resampling, the variation of the sample means tends to be centred around the mean of the original dataset, and the variation of the sample medians tends to be centred around the median of the original dataset.

ISBN 978-1-107-61421-5

Example 5: Is that a shark?

The Enrichment question of Exercise 2D on Cambridge GO considers data from a fishing expedition. The full dataset from that expedition contained 58 observations. A dotplot of the lengths in mm (to the nearest mm) of fish caught is given below.



We can see there is one very large fish at more than 112 cm! It was a reef shark measuring 120 cm.

Question

Should we include the shark or leave it out?

Solution

The mean of the data including the reef shark is 278 mm; so the total is 278 × 58 =16 124 mm. Removing the shark reduces the total to 14 924 mm with a mean of $\frac{14924}{57}$ = 261.83 mm.

The median of the data including the shark is 255 mm. This value is halfway between the 29th and the 30th observations ordered from smallest to largest. So either there are at least three observations of 255 mm or perhaps the 29th is 250 and the 30th is 260. Without the full data, we don't know. If we remove the shark, the median is the 29th observation. So it might stay the same or it might decrease slightly. In fact, the 29th observation is 250 and the 30th is 260, so the median without the shark is 250 mm.

This example is different from the computer lab times, because there is one very different observation that is much larger than the rest, and we know why it is different. Including the reef shark is going to make it difficult to investigate and analyse the rest of these data because the reef shark is so different from all the other fish, even though it is officially a fish. In this case it would be better to omit this observation, clearly explaining that it is a shark and so is very different.

CAUTION
Never omit observations unless there are clearly explained good reasons. Sometimes in exploring data, you might temporarily leave an observation out of a graph or a calculation to see what happens. Always explain what you are doing and why.

ISBN 978-1-107-61421-5

Exercise 6C

1 In each of the following, identify observations that could be considered outliers. Calculate the data mean and median with and without the observations and comment on your results. State whether the observations should be omitted from the data or not, giving reasons.

a The following are part of a dataset obtained by random selector from the CensusAtSchool data for armspan measurement in cm:

171.0 159.0 152.0 146.0 69.0 178.0 72.0 151.0 141.0 150.0 171.0
146.5 150.0 162.0 50.0 110.0 149.5 131.0

b The following are part of a dataset obtained by random selector from the CensusAtSchool data for right foot measurement in cm:

26.5 25.0 22.5 24.5 23.0 28.5 24.0 28.0 22.0 30.0 23.5 125.0 17.0
23.5 24.0 24.0 28.0 26.0

c The following are part of a dataset obtained by random selector from the CensusAtSchool data for time to get to school in minutes:

25 10 2 8 45 20 30 20 75 40 10 20 120 25 15 20 30

2 In question 3 of Exercise 2B, a dataset is given of 68 observations on the lengths in seconds of mobile phone calls in a public place. There is a large value of 1930 s. Question 4 of Exercise 2C asks for the sample mean and median with this value included and then without this value.

- With the large value of 1930 s, the mean and median of the data are 185.8 s and 100 s respectively.
- Without the large value of 1930 s, the mean and median of the data are 159.7 s and 100 s respectively.

On Cambridge GO <www.cambridge.edu.au/statsAC78weblinks> are four stem-and-leaf plots for data obtained by resampling the original dataset:

Case 1: dataset with the value 1930 s included

Case 2: dataset without the value 1930 s included

The sample size for the resampling is 15 in both cases, and 100 samples are generated for the case 1 and case 2. The stem-and-leaf plots are of the 100 sample means and medians generated for the two cases.

a Use the given stem-and-leaf plots to obtain the median and range of the 100 sample means and medians in case 1. How do these compare to the original dataset from which the samples have been obtained?

b Use the given stem-and-leaf plots to obtain the medians and ranges of the 100 sample means and medians in case 2. How do these compare to the original dataset from which the samples have been obtained?

c Comment on the similarities and differences between the sample means and medians for case 1 and case 2.

ISBN 978-1-107-61421-5

d For case 1:

- **i** what proportion of the 100 sample means lie between 160 and 200 inclusive?
- **ii** what proportion of the 100 sample means lie between 170 and 190 inclusive?
- **iii** what proportion of the 100 sample medians lie between 80 and 120 inclusive?
- **iv** what proportion of the 100 sample medians lie between 90 and 110 inclusive?

e For case 2:

- **i** what proportion of the 100 sample means lie between 140 and 180 inclusive?
- **ii** what proportion of the 100 sample means lie between 150 and 170 inclusive?
- **iii** what proportion of the 100 sample medians lie between 80 and 120 inclusive?
- **iv** what proportion of the 100 sample medians lie between 90 and 110 inclusive?

f What is of interest in parts **d** and **e**?

g If samples of size 50 were obtained by resampling of these data, what would you expect to see in stem-and-leaf plots of the sample means and medians?

HINT
Remember that with a leaf unit of 10, the mean value of 185.8 s in case 1 would be presented in the stem-and-leaf plot as 180 s.

3 On Cambridge GO are four stem-and-leaf plots for the data in the dotplots in the Let's start example of section 6-3. One hundred sample means and sample medians were obtained by resampling the 150 observations of times spent by students in the express computer lab (in minutes). Two different sample sizes have been used in the resampling – there are 100 samples of size 20 and a second group of 100 samples of size 50. <www.cambridge.edu.au/statsAC78weblinks>

a Use the given stem-and-leaf plots to obtain the medians and ranges of the 100 sample means for the samples of size 20 and for the samples of size 50 and comment on these.

b Use the given stem-and-leaf plots to obtain the medians and ranges of the 100 sample medians for the samples of size 20 and for the samples of size 50 and comment on these.

c Comment on the similarities and differences between the sample means and medians.

ISBN 978-1-107-61421-5

d Consider the sample means.

- **i** For the samples of size 20, what proportion of the 100 sample means lies between 25 and 35 inclusive?
- **ii** For the samples of size 50, what proportion of the 100 sample means lies between 25 and 35 inclusive?
- **iii** For the samples of size 20, what proportion of the 100 sample means lies between 29 and 31 inclusive?
- **iv** For the samples of size 50, what proportion of the 100 sample means lies between 29 and 31 inclusive?

e Consider the sample medians.

- **i** For the samples of size 20, what proportion of the 100 sample medians lies between 12 and 22 inclusive?
- **ii** For the samples of size 50, what proportion of the 100 sample medians lies between 12 and 22 inclusive?
- **iii** For the samples of size 20, what proportion of the 100 sample medians lies between 16 and 18 inclusive?
- **iv** For the samples of size 50, what proportion of the 100 sample medians lies between 16 and 18 inclusive?

f What is of interest in parts **d** and **e**?

g If samples of size 100 were obtained by resampling of these data, what would you expect to see in stem-and-leaf plots of the sample means and medians?

HINT
Remember that with a leaf unit of 1, the mean value of 30.6 min would be presented in the stem-and-leaf plot as 30 min, and the median value of 17.5 would be presented as 17 min.

Enrichment

How does the average height of students in your class compare?
<www.cambridge.edu.au/statsAC78weblinks>

ISBN 978-1-107-61421-5

Chapter summary

Sample proportions

- A sample proportion is the number of observations in a category divided by the total number of observations
- For randomly representative data, a sample proportion estimates the corresponding population proportion
- Sample proportions can vary greatly across random samples collected in the same circumstances or from the same population
- Simulation of many samples can help in exploring the variation of sample proportions.

Sampling variation in quantitative data

- Random samples of data collected under the same circumstances or from the same population can vary greatly from each other
- Understanding and allowing for sampling variation is very important in statistical investigations, analysis and interpretation
- Sampling variation in the circumstances of a dataset can be explored by resampling from the dataset.

Sampling variation of means and medians

- Quantities calculated from data, such as data means, medians and ranges, are also subject to sampling variation
- The sampling variation of sample means and medians depends on the data circumstances, but is greater in smaller samples than in larger ones
- In general the sample mean and median are trying to estimate the mean and median, respectively, in the situation or population from which the data have been collected.

Extreme values

- Extreme data values, both large and small, tend to affect the data mean more than the data median
- The data median is just affected by whether they are small or large
- The data mean is affected by the actual values because the total of all the observations is affected
- Data values should not be removed without good reason.

Multiple-choice questions

1 Using Excel to simulate 100 samples, each of size 150, to investigate how much the sample proportion could vary, the number 150 is entered into:

A Number of variables **B** Number of random numbers
C Number of trials **D** None of these

2 Using Excel to simulate 100 samples, each of size 150, to investigate how much the sample proportion could vary, the number 100 is entered into:

A Number of variables **B** Number of random numbers
C Number of trials **D** None of these

ISBN 978-1-107-61421-5

3 The table below gives the numbers of people observed in morning, lunch and evening periods in a city mall and wearing different types of footwear.

	Footwear type				
Time of day	**Casual exposed**	**Casual enclosed**	**Fitness**	**Formal**	**All**
Afternoon	81	278	39	57	455
Lunch	133	183	76	25	417
Morning	117	314	279	102	812
Total	331	775	394	184	1684

The percentage of people observed at lunchtime and wearing formal footwear is:

A 1.5% **B** 6%

C 13.6% **D** 11%

4 Using the table of question 3 above, the percentage of people in the morning observed to be wearing fitness footwear is:

A 16.6% **B** 70.8%

C 23.4% **D** 34.4%

5 Using the table of question 3 above, the percentage of people wearing casual exposed footwear who were observed in the afternoon is:

A 4.8% **B** 17.8%

C 19.7% **D** 24.5%

6 Refer to Example 4 on Cambridge GO. In using the Random Number Generation in Excel to generate 50 samples, each of size 25 from a dataset with 450 observations, the number 25 is:

A The Number of Variables **B** The column size in the Probability Input Range

C The Number of Random Numbers **D** None of these

7 Refer to Example 4 on Cambridge GO. In using Random Number Generation in Excel to generate 50 samples, each of size 25 from a dataset with 450 observations, the number 450 is

A The Number of Variables **B** The column size in the Probability Input Range

C The Number of Random Numbers **D** None of these

ISBN 978-1-107-61421-5

Short-answer questions

1 In a study on a vending machine, the following data on people's choices were observed.

	Chocolate	Other food	Drink	All
Female	42	71	129	242
Male	38	53	196	287
Total	80	124	325	529

a It is reported that more than 50% of females select chocolate from a vending machine. What mistake has been made?

b The percentage of people overall who chose a drink from the machine was approximately 61%. To the right is a stem-and-leaf plot of 50 sample proportions (as percentages) of random samples of size 250 of data from a population with population proportion of 61%.

i Give the range and median of these 50 sample proportions.

ii What percentage of these 50 sample proportions are less than the population proportion of 61%?

iii What percentage of these 50 sample proportions are at least 2% away from the population proportion of 61%? That is, less than or equal to 59% or at least 63%?

Leaf unit = 0.1

53	66
54	
55	6
56	0048
57	6
58	48
59	2666
60	444448
61	6
62	00004448888
63	222666
64	0000448
65	6
66	0
67	2
68	04

ISBN 978-1-107-61421-5

2 The following data on people's types of usage of city gardens during lunchtimes were observed.

	Cut through	Eating	Fitness	Relax	Study	Other	All
Female	34	46	19	60	16	30	205
Male	78	32	29	56	20	37	252
All	112	78	48	116	36	67	457

a What percentage of females were eating or relaxing?

b What percentage of males were studying?

c The percentage of people overall who were using the city gardens for cutting through was approximately 25%. To the right is a stem-and-leaf plot of 75 sample proportions (in %) of random samples of size 400 of data from a population with population proportion of 25%.

i Give the range and median of these 75 sample proportions.

ii What percentage of these 75 sample proportions is at least the population proportion of 25%?

iii What percentage of these 75 sample proportions is at least 2% away from the population proportion of 25%? That is, less than or equal to 23% or at least 27%?

Leaf unit = 0.1

20	2255
21	57
22	0255577
23	0002225577
24	02225555555557777
25	0002255555777
26	00022257
27	002577
28	000022
29	2
30	5

ISBN 978-1-107-61421-5

3 In a study on delivery times for pizzas conducted by a pizza chain, 575 observations of delivery times in minutes were collected over a week. The average was 21.5 min, the median was 20 min and the range was 51 min.

Below are four samples, each of size 20, obtained by randomly resampling from the 575 values.

23	16	19	28	24	6	21	19	41	36	25	39	17	25	19	21	18	11	17	22
26	41	26	7	17	32	12	14	27	30	18	21	16	36	13	35	16	11	22	20
18	22	24	28	19	16	14	26	17	22	13	23	46	36	35	20	18	22	23	28
13	17	18	12	23	19	30	14	32	22	11	25	24	17	15	32	17	22	24	14

a Use stem-and-leaf plots to graph these samples.

b Obtain the means, medians and ranges of each of these samples.

c Compare the summary statistics obtained in part **b** for the samples with those of the original dataset.

d What proportion of each of the four samples is greater than the mean of the original data?

4 In each of the following, identify observations that could be considered outliers. Calculate the data mean and median with and without the observations and comment. State whether the observations should be omitted from the data or not, giving reasons.

a The following are part of a dataset on pizza delivery times in minutes:

30	15	6	22	33	19	33	30	22	28
20	16	22	22	29	27	26	15	20	30

b The following are part of a dataset on speeds of cyclists in km/h on a bike path:

33	28	22	20	22	19	7	23	22	24	28	37	27	34	21	29

c The following are part of a dataset obtained by random selector from the CensusAtSchool data for height in cm of Years 7 and 8 students:

181.5	171.0	157.5	156.0	170.0	1.0	158.0	161.0
123.0	146.0	150.0	170.0	1.5	1.5	161.0	

5 A dataset of 625 observations was collected on the speed in km/h of cyclists on a bike path shared with pedestrians as part of a study into whether speed limits should be placed on cyclists on this path, and, if so, what they should be. The average speed in the dataset is 26 km/h, which is also the median speed.

a On Cambridge GO is a dotplot of the 625 speeds. There is an outlier. Identify it, and recalculate the mean of the 624 observations without it. Do you think the median will change?

b On Cambridge GO are stem-and-leaf plots of the means and medians of 75 samples each of 25 observations obtained by resampling from the 625 observations. Use the given stem-and-leaf plots to obtain the sample medians and ranges of the 75 sample means and medians for the samples of size 25. Comment on your results.

c Consider the sample means.

i What proportion of the 75 sample means lies between 25 and 27 inclusive?

ii What proportion of the 75 sample means lies between 25.5 and 26.5 inclusive?

ISBN 978-1-107-61421-5

d Consider the sample medians.

i What proportion of the 75 sample medians lies between 25 and 27 inclusive?

ii What proportion of the 75 sample medians is equal to 26?

e What effect do you think the outlier identified in part **a** is having on the sample means and medians, if any?

f If samples of size 50 were obtained by resampling of these data, what would you expect to see in stem-and-leaf plots of the sample means and medians?

Extended-response question

The data of Short-answer question 3 of the 575 delivery times in minutes of pizzas was resampled. The sample means and medians for 100 samples each of size 25 were obtained. Then the sample means and medians of 100 samples each of size 50 were obtained. On Cambridge GO are four stem-and-leaf plots of the 100 sample means and medians of each of these two groups of 100 samples.

Use these stem-and-leaf plots to explore the variation of sample means and medians in resampling from these 575 delivery times.

ISBN 978-1-107-61421-5

Glossary

Chapter 1

Categorical data: Data that fall into categories that can be named or coded

Continuous variables: Variables which take values in intervals – typically, continuous data have values given 'to the nearest …'

Count data: Record of a number of items, events, people, etc.

Data: Information, facts, records, observations

Experiment: Data investigation in which investigators control conditions and measure the effect of these on some outcome(s) of interest

Experimental units (or **subjects,** or **observational units**): Individuals or objects or entities on which observations are made

Grouping: When **measurement data** or **count data** are placed in specified groups of values, or when some categories of a categorical variable are combined, observations are grouped together

Measurement data: Data which need units of measurement. Observations are recorded in the desired units of measurement

Observational units (or **subjects** or **experimental units**): Individuals or objects or entities on which observations are made

Observational study: Data investigation in which investigators observe subjects without altering or controlling conditions

Ordinal variable: Categorical variable for which the order of categories has meaning

Pilot study: An initial trial of the investigation or preliminary experiment to check the practicalities of the planned collection

Primary data: Data collected by the investigators

Quantitative data: Measurement or count data; the numerical values of the data are actual quantities

Random: Due to chance

Randomly representative data: Data obtained at random, from a more general situation or population

Randomise: To make random

Recording sheet: A table or spreadsheet to record data; each variable has a column and each subject (or experimental or observational unit) has a row

Secondary data: Data collected by others

Statistical data investigation process: How real problems are tackled by statisticians and investigators conducting experiments, studies or surveys to obtain data for working statistically

Statistical variable: The 'what' we are going to observe when we collect or observe data

Subjects (or **observational** or **experimental units**): Individuals or objects or entities on which observations are made

Survey: Asking questions of subjects with the data being responses; the subjects of a survey may be individuals, or groups such as companies, businesses or households

Variation: The unpredictability of situations in which observations or measurements have different values or are not determined or not specified

Chapter 2

Average (or **mean**): Value obtained by adding all the values of the observations and dividing by the number of observations

Centre (of the data): Where the data tend to centred, or what the data tend to be spread around; a centre of the data also indicates the general size of the observations

Context of the data: The circumstance or 'story' of the data – what the data are about

ISBN 978-1-107-61421-5

Dataset: The set of data

Dotplot: A type of graph in which each dot corresponds to a given number of observations, a dot on the vertical scale represents an occurrence of a value in the data; the value is placed on the horizontal axis; usually just one observation per dot, unless there are too many observations, in which case the number of observations represented by each dot is stated

Features of data: Various aspects of what the data look like and how they behave

Leaf (of a stem-and-leaf plot): Contains the second highest place value of the observations; each observation is represented by a digit in a leaf

Mean: Value obtained by adding all the values of the observations and dividing by the number of observations

Median: The middle value; the value that has half the observations less than it in value, and half the observations greater than it in value

Mode: A data value that occurs most often; there may be one mode or many modes if different values occur the same number of times in a set of data

Range: The largest (maximum) value minus the smallest (minimum) value in the data

Raw data: Original data, the data as recorded

Respondent: The individual answering the questions – the subject

Stem-and-leaf plot: A plot for quantitative data that groups observations into intervals of equal lengths

Stem (of a stem-and-leaf plot): Contains the digits in the highest place value of the observations

Summary statistics: Values calculated or obtained from data and used in describing features of the data

Variability: How greatly the data values differ from each other; what makes up variability depends on the situation

Chapter 3

Assign: Decide on particular values

Assumption: Theory, guess or hypothesis

Balanced: All outcomes are equally likely

Certain: Must happen

Chance: Likelihood, possibility or probability

Compound event: Event consisting of several individual outcomes

Continuous outcome: Can take any value in a specified range, e.g. height

Discrete outcome: Can take only specific distinct possibilities

Equally likely outcomes: The outcomes of an experiment are equally likely to occur

Fair: Unbiased, balanced

Frequency: The number of times something happens

Impossible: Can never happen

Likely: Probable or possible

Modelling: Making a mathematical model of a situation

Probability: A way of measuring chance, of seeing how **likely** any event is to happen

Simple event: Event consisting of a single outcome

Subjective: From a particular person's point of view

Symmetry: A situation that has repeated aspects that are the same

Unlikely: Not likely, doubtful

Chapter 4

'A and B': Represents the situation where both A and B occur

'A or B': Can be interpreted as 'at least one of the events occurs' (**inclusive or**) or as 'exactly one of the events occurs' (**exclusive or**)

Complementary events: Two events that are **disjoint** and cover the whole **sample space**

Disjoint: Two events that don't contain any outcomes in common

ISBN 978-1-107-61421-5

Event: Anything that can actually happen in an experiment

Exclusive or: Exactly one of the events occurs

Experiment (in probability): Any situation where the result is uncertain

Inclusive or: At least one of the events occurs

Intersection (on a Venn diagram): Area of overlap that shows where more than one event has occurred

Mutually exclusive (or **disjoint**): Two events that don't contain any outcomes in common

Sample space: A list of possible outcomes

Two-way table: A table that can be used to summarise data on frequencies of two events; can be used to estimate probabilities

Venn diagram: A diagram that shows data as circles, possibly interlocking, inside a rectangle; can also be used to show events

Chapter 5

Blinding: When the subjects in an experiment do not know which product or treatment they've been allocated

Case-control study: Observational studies in which researchers identify a group without a disease or condition, who are as similar as possible to those who do have the disease or condition

Census: A data collection, in which the aim is to collect information about every member of a population

Closed questions: Questions with given responses as choices in a survey

Cluster sampling: A random sample of groups is chosen and all the subjects in these groups are surveyed

Double-blinding: Occurs in an experiment when investigators who are interacting with subjects do not know which products or treatments have been allocated to different subjects

Experimental investigation: Data investigations in which investigators control conditions and measure the effect of these on some outcome(s) of interest

Number of observations: The number of subjects in a sample

Observational study: Data investigation in which investigators observe subjects without altering or controlling conditions

Open questions: Questions in a survey that allow people to respond as they wish

Placebo effect: Effect due just to being on a treatment or being in an experiment

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

Polling: Process of surveying people, usually to ask opinions

Random numbers: Numbers chosen at random from a given range

Random sample: A set of randomly representative data

Randomisation (in experiments): Process of allocating conditions to subjects at random, or subjects to conditions at random

Response rate: The number of responses divided by the number asked to respond, usually expressed as a percentage

Sample of data: A set of observations for which many more observations could have been taken

Sample size (or **number of observations**): The number of subjects in a sample

Sample survey: Survey for which the subjects are chosen by a sampling plan

Sampling plan: Plan designed to be randomly representative of a larger population about which the investigators wish to find or state information

Stratified sampling: When random samples are taken within specified groups (called strata)

ISBN 978-1-107-61421-5

Cambridge University Press

Survey: Data obtained from questions asked of subjects; subjects of a survey may be individuals, or groups such as companies, businesses or households

Table of random digits: List of random numbers from 0 to 9, which can be used to obtain a set of random numbers in any range

Chapter 6

Estimates: Quantities calculated from sample data to be used as approximations of quantities relevant to the more general situation or population of which the sample data can be taken as being randomly representative

Outliers: Extreme values in a dataset – that is, very large or very small values – that are considered to be unusually far away from the rest of the data

Randomly generated data: Data constructed artificially by some mechanism so that the data are produced under some specified rules of chance across all possible values

Resampling: Taking random samples of data from a set of previously observed data; each observation in the dataset is equally likely to be chosen in the resampling; data obtained by resampling can be considered randomly representative of the original dataset

Sample mean: Average of values in a sample

Sample median: Median of values in a sample

Sample proportions: Relative frequencies obtained from a sample of data; term usually used when a relative frequency is being used to estimate a population proportion; can be expressed as fractions, decimals or percentages

Sampling variation: Variation across different data samples obtained under the same conditions; can occur in random samples from the same population, or obtained by repeating the same experiment, or obtained by taking observations of the same situation

Sampling with replacement: Selecting observations at random from a limited group of values or items so that as each value or item is selected, it is replaced and is available for subsequent selections

Sampling without replacement: Selecting observations at random from a limited group of values or items so that as each value or item is selected, it is not available for subsequent selections

Simulation: When computers are programmed to mimic a real situation and generate data that can be taken as coming from that real situation

ISBN 978-1-107-61421-5

Answers

Chapter 1

Pre-test

1 **a** Measurement **b** Measurement
c Categorical **d** Count
e Categorical **f** Categorical
g Count **h** Measurement

2 **a** Possible values: 0, 1, 2, 3…. Count
b Possible values: very poor, poor, average, good, very good. Categorical
c 21 s, 15 s, 24 s (to nearest second). Measurement
d 20, 100, 35. Count
e 58, 62, 61. Measurement
f Clarke, Watson, Steve Waugh. Categorical

3 Column graph, pie chart and barchart are all appropriate; dotplot is not. If a dotplot used, it is being used as a column graph.

4 Dotplot is only appropriate one.

5 Side-by-side column graph, two-way table

6 **a** Categorical
b Column graph or pie chart or barchart
c Frequencies or relative frequencies of the responses 1, 2, 3. The pie chart would most likely display relative frequencies as percentages.

Exercise 1A

1 There are many colours so the investigators would need to take care to decide how to classify colours before starting to collect data. Even with this, it may be advisable to collect data in pairs. Make of car could probably also be recorded, and possibly type (e.g. Mazda, sedan). Recording data on moving cars might be difficult and prone to errors, so perhaps collecting data in carparks might be easier.

2 There are many decisions to be made for this. Amount of traffic needs to be recorded as counts in a selected length of time, for example, in 2 or 5 minutes. So the length of time needs to be chosen, and at least two people are needed – one as timekeeper and the other as counter. If the road is busy, it might be necessary to record traffic in each direction separately. The time of day and day of the week will also make a difference, so it is necessary to decide whether to focus on peak traffic or peak and non-peak, and whether weekday or weekend or both. Also are all vehicles to be counted, including motor bikes, buses, trucks, etc? If traffic is of interest, then yes.

3 If it is desired to include all the books, then pages have to be chosen at random from across the books. Perhaps it might be best to record which book, which edition and whether UK or USA, and choose pages at random for each book, for example, using a random number generator or tables of random digits. Perhaps it could also be recorded whether the word is being used as an adjective or a noun.

4 **a** Students may use different forms of transport on different days, or more than one form every day.
b This will depend on the school. Considerations of age, distance from school, whether siblings attend the same school, gender, outside school activities, etc. may need to be discussed. As there are many possible variables to consider, thought is needed as to what and how to survey.
c The CensusAtSchool questionnaire on www.abs.gov.au/censusatschool asks 'What is the main method of travel that you usually use to get to school? Select one only.' It gives seven alternatives plus Other. The questionnaire asks many other questions, some of which are relevant in considering how students travel to school, including the international questions of age, year level, how long it takes to get to school, gender, where you live.

5 **a** What is the size and weight of the book? Is it hard cover or soft cover?
Is the person to stand or sit still or to walk? Is the person to do both, and, if so, in what order? Is the person to do anything else while balancing the book? Note that the more complications that are included, the more challenging it is to design the investigation.
b Gender, age, main sport played

6 **a** Are we going to count the serve(s)? Probably should as a first serve fault affects the tactics. Are we going to consider singles or doubles? Men's or women's?
b Are we going to focus on matches between specific players or choose points at random from random

ISBN 978-1-107-61421-5

matches? The latter is much more challenging, but would need to be done if we are considering tennis in general, and even then, it would be tennis at a certain level (e.g. televised tennis matches).

c This depends on what we choose in parts a and b above. As well as singles or doubles, men's or women's, specific players or not, we might choose to record who is playing/serving, the surface, which set, whether the server is leading or not.

Exercise 1B

1 Subjects are households. Variables and their types are:
- number of residents; count
- number of residents under 18 years of age; count
- gender of oldest resident; categorical
- age of oldest resident; measurement/continuous.

2 a (Q1 of Exercise 1A) Subjects are cars. Possible variables and their types are:
- colour; categorical; possible categories: white, black, grey/silver, cream/yellow/gold, blue, green, red, other.
- make; categorical; possible categories: Toyota, Mitsubishi, Mazda, Ford, Holden, Hyundai, Other.

b (Q2 of Exercise 1A). Subjects are time intervals. Possible variables and their types are:
- number of vehicles per time interval; count
- time of day; categorical; possible categories: morning peak, off-peak, evening peak
- direction; categorical; possible categories: inbound, outbound
- day of week; categorical; possible categories: Monday, Tuesday, etc.

c (Q3 of Exercise 1A). Subjects are pages. Possible variables and their types are:
- frequency of word 'magic' per page; count
- book; categorical; categories are the names of the books
- usage of word; categorical; possible categories: noun, adjective.

3 Subjects are children. Variables and their types are:
- span in cm; measurement/continuous
- height in cm; measurement/continuous
- gender; categorical; categories are: male, female
- age in years; measurement/continuous.

4 Subjects are non-fiction books. Variables and their types are:
- price in $; measurement/continuous
- number of pages; count
- topic; categorical; possible categories are: history, sport, gardening, cooking, biography, travel, other
- cover; categorical; categories: hard, soft
- pictures; categorical; categories: colour, not colour.

5 Subjects are coffee outlets. Variables and their types are:
- cappuccino price in $; measurement/continuous
- flat white price in $; measurement/continuous
- type of outlet; categorical; categories are: café, restaurant
- location of outlet; categorical; possible categories: centre, suburb.

Exercise 1C

1 a

Day (Tues/Thurs)	Peak time (morning/afternoon)	Direction (up/down)	Choice (lift/stairs)	Gender

b If very busy, may need separate collectors for each direction for lift and stairs. Also difficult to record people as they get out of the lift, so having one person record those getting in at the bottom and other recording those who get in at the top would be best. If still not able to record everybody, may need to sample.

c Did they record everyone? Day(s) of week and exact time(s) and date(s) of data collection. Any local information of relevance.

d Local information on type and location of bus station, and generally who does it service. Also weather on day(s) of collection and whether the day(s) of collection were fairly typical.

2 a, b, c Subjects are 15-minute periods.

Period	Day	Time of day (morning/afternoon)	Direction (inbound/outbound)	Number of pedestrians	Number of cyclists

Note that the time period is not a variable but is there for reference and checking data. The starting and finishing times should be reported. As the

ISBN 978-1-107-61421-5

focus is on numbers of users, all users should have been counted, so it should be reported how this was ensured. Dates, times and local information should be given.

3 a, b, c Subjects are users.

Day	Time of day (morning/ afternoon)	Direction (inbound/ outbound)	Type of user (cyclist/ pedestrian)	Speed	Gender

As above, the starting and finishing times should be reported. As the focus is on users, and the data are on a sample of users, it should be reported how users were chosen. As above, dates, times and local information should be provided.

4 a

Candle	Colour	Initial length	Initial diameter	Length after burning

b Same person doing the measuring, and same person doing the lighting. Randomise the order in which the candles were used. Take a few diameter measurements along candle and average to allow for any variation along candle.

5 a

- How frequently would you like the school newsletter to appear? Please choose one.
  Weekly Fortnightly Other (please specify)
- Would you like to receive the newsletter online? Please choose one.
  Yes, as well as in print Yes, instead of in print
  No, just in print Don't mind
  Do not want to receive newsletter.
- Would you like each child to receive a copy, or just the eldest child in each family?
  Each child Eldest child Don't mind
  Do not want children to receive newsletter.

b Sent home with each child with envelope for returning and request that one survey be filled in per family.

Chapter summary

Multiple-choice questions

1 C **2** B **3** C **4** B **5** A

6 C **7** A **8** C **9** B **10** B

Short-answer questions

1 a Minimum temperature, humidity (maximum or average or at specified time).

b Maximum and minimum temperatures, rainfall, humidity are all measurement/continuous. Wind direction usually reported as categorical, e.g. NE, SW etc.

2 ai, bi Buy a number of packets, and record colour for each sweet. Subjects are sweets and variable is colour – categorical.

aii, bii Buy a number of these smallest packets and record number of blue sweets in each. Subjects are packets, and variable is number of blue sweets – count.

3 a Number of computers in household, sport played, favourite pastime.

b How long – measurement/continuous; gender – categorical; number of computers – count; sport played – categorical; favourite pastime – categorical.

c Use student IDs and random number generator or table of random digits. If no student IDs, number students on roll and use random numbers.

4 a Length of song – measurement/continuous; genre of the song – categorical; nationality of the performer(s) – categorical; solo artist or a band categorical.

b

Rock
Alternative Rock
Hard Rock
Dance
Electronic
Indie
Pop
Rhythm and Blues

Rock
Alternative/Hard Rock
Dance/Electronic/Indie
Pop
Rhythm and Blues

c Australian/NZ, British (UK, Ireland), US, other.

d Top 100 may not be randomly representative of songs with respect to length.

5 a 48

b Time – measurement/continuous; distance – measurement/continuous; landing position – categorical; design – categorical; paper – categorical; maker – categorical; thrower – categorical

ISBN 978-1-107-61421-5

c Randomise order of throwing planes.

d

Time (s)	Distance (cm)	Design	Paper	Maker	Thrower
10	53	A	B	Fred	Julie
8	45	C	B	Sonia	Julie
12	49	C	A	Sonia	Harry

6 a Selling price – measurement/continuous; region – categorical; land size – measurement/continuous; number of bedrooms – count; bathrooms – count; car spaces – count; pool – categorical.

b Bedrooms and bathrooms change to be categorical because of combining count values into categories.

c

Selling price	Land size	Region	Bed-rooms	Bath-rooms	Car spaces	Pool
$520 000	706 sq m	A	3	1	1	No
$865 000	850 sq m	C	4	2	2	Yes
$425 000	550 sq m	B	3	1	1	No

Extended-response questions

1 a Need to define a family, and what a pet is, and decide whether to collect information on all pets, and if so, how much. For example, pets could be defined as being living, and currently living in the household. Each household could be asked to give the total number of pets, the number of people in the household under 18 years of age, the type of 'main' pet (the pet regarded as most belonging to everyone), and the age in years (human) of this pet.

b Could ask for number of currently living dogs and cats in household, number of people living in household, and number of people in the household under 18 years of age. Could ask for the type of 'main' pet (the pet regarded as most belonging to everyone), and the age in years (human) of this pet.

2 a Need to decide how high to hold ruler above person's hand before dropping. Need to decide if subjects are to use their dominant hand or each hand in turn – if the latter, need to keep same order (in terms of dominant hand) or randomise. If choose simpler experiment of asking people to use their dominant hand only, the possible variables and their types are: whether caught or not – categorical; distance from top of ruler when caught – measurement/continuous; gender – categorical; age – measurement/continuous; dominant hand (left/right) – categorical. Could also choose to record if the subjects wear glasses or not – categorical.

Caught	Distance from top of ruler (cm)	Gender	Age	Hand used	Glasses
Yes	20	F	15	R	No
Yes	15	M	24	L	Yes
No		M	51	R	No

b In 'Go for the Gopher', participants can use either the hammer or the keypad with numbers. Either all participants should use the same method, or each participant uses both, either in the same order, or randomised. In addition, if numbers are used, all participants should use either the numbers on the RHS of the keyboard or the numbers along the top – probably best to use those on the RHS. It should be decided whether each participant is allowed a trial. As some may have played a similar game, it is probably best to give everyone a trial run.

Chapter 2

Pre-test

1 a Flight distance – continuous; bird species – categorical

b Is the distance the straight line distance from point of release to point of first landing? How was this distance measured? Was the same measuring device used for all measurements?

c List data in order from shortest to furthest distance and show in a table with one column for each bird.

2 a Price of unleaded 91, price of premium 98, day, week, brand, distance, direction. Outlet may be a variable if the same outlets were used each time.

b Price of unleaded 91, price of premium 98, distance and direction in degrees are all continuous. Day, week and brand are categorical.

c If outlets randomly chosen each time, rows are outlets. If same outlets used, rows correspond to outlets and days.

d The actual dates, any public holidays, the time of day the prices were collected, whether they were collected by observation or phoning. If the same outlets were used each day, they need naming.

3 a Number of pizzas sold of each type (Hawaiian, supreme and seafood), the daily takings for pizza sold, the daily takings for drink sold, day, week, daily temperature.

ISBN 978-1-107-61421-5

b Number of pizzas sold of each type – count; daily takings for pizza sold, daily takings for drink sold, daily temperature – continuous; day, week - categorical.

c Days

d In case there's a relationship between the temperature and people buying pizzas.

4 a month of birth – categorical
height – continuous
weight – continuous
year of school – ordinal
length of forearm – continuous
position in family – ordinal
number of siblings – count
left-handed (yes/no) – categorical
wear glasses – categorical
hours per week in part-time job – continuous
accident proneness on a scale of 1–5 – ordinal

b Age is time and is a measure; it is continuous. If it is rounded down to give whole numbers of years, this is a decision as to how to record age.

c Time is a continuous variable and has units of measurement. If the times were correct to the nearest minute, the students may have thought that whole numbers might represent counts.

Exercise 2A

1 The straight line distance between the point where the bird started flying and the point where it first landed after this flight

2 a Starting the stopwatch exactly when saying go, and stopping exactly when the subject said stop.

b For 10 s guess: 9.8 s, 11.2 s, 10.6 s, 12.0 s

c 9.60 should be written 9.6. 11 should be checked that it was 11.0 and then written 11.0. 98 and 1.24 are clearly wrong in some way – the collector could be asked, or the recording sheet checked, to see if they are meant to be 9.8 and 12.4.

3 a The heights and spans must be written as pairs or in corresponding columns.

b i Names or identifiers so the data can be checked

ii The span of 175.3 for the third person seems wrong and would need to be checked.

4 a Height of seedling from soil level to highest point at the same time each day; possibly also number of leaves.

b More likely to make a mistake in either reading or recording or converting

c cm

d Seedlings might not tend to grow straight, and are also easily blown, so measuring away from wind would be important. Also the vertical height is needed.

5 a Making sure all the water is collected in the bucket, and keeping the water flow slow and fairly constant. Holding the balloon also needs care as some water might run down/across the arms.

b No – would need more accurate measurement; litres with decimal or mL would be best.

Exercise 2B

1 a $\frac{13}{50}$ = 26% of drivers were speeding (doing more than 60 km/h) in approaching the amber light.

b The smallest time is 2.08 s, so the speed is $\frac{50}{2.08}$ m/s. Use 2 s for ease of calculation. In km/h this is $\frac{50}{1000}$ km per $\frac{2}{3600}$ h. This is $\frac{(50 \times 3600)}{(1000 \times 2)}$ = 90 km/h! Using 2.08 s gives 87 km/h.

c The largest time is 4.94 s. This gives a speed of $\frac{(50 \times 3600)}{(1000 \times 4.94)}$ = 36.4 km/h

d

Leaf unit = 0.10

2	03
2	67778899999
3	001111223334
3	5555556777778899999
4	111
4	779

e The given stem-and-leaf shows how variable the times are, particularly for the slower and faster drivers; the stem-and-leaf above shows where most of the drivers' speeds are concentrated.

2 a 1.3 **b** 0.9

c Need to decide if values ending in 0.05 go up or down. As those ending in 0.00 will stay there, the usual convention is to round those ending in 0.05 up. So 0.85 is assigned to 0.9.

ISBN 978-1-107-61421-5

Leaf unit = 0.010

8	4
8	5799
9	22223
9	6
10	3334
10	9
11	02
11	
12	
12	8
13	4

d 11 **e** 0.01

3 a The 1930 s observation is removed.

Leaf unit = 10

0	122333344444
0	555556666667777778999
1	00222223444
1	55789
2	024
2	89
3	02333
3	556889
4	
4	
5	
5	
6	
6	
7	
7	5
8	0

b 10

c 50 s

d 6 min = 360 s. There are 7 people for more than 360 s, so $\frac{7}{68}$ = 10.3%.

4 a 24.0 up to 24.9

b between 79.0 and 79.9 min

c 5 min

d Because there are no observations between 26 and 36 min

e $\frac{20}{128}$ = 15.625%

Exercise 2C

1 a i 4.94 – 2.08 = 2.86 s

ii 3.44 s (to 2 dp) **iii** 3.46 s

b i 50 m in 3.44 s is $\frac{50}{1000}$ km in $\frac{3.44}{3600}$ h. So average speed is $\frac{(50 \times 3600)}{(1000 \times 3.44)}$ = 52.33 km/h

ii 50 m in 3.46 s is $\frac{(50 \times 3600)}{(1000 \times 3.46)}$ = 52.02 km/h

c The numbers do not represent quantities. The difference between 1 and 2 is not the same as the difference between 2 and 3.

2 a i 45.7 **ii** 45 **iii** 100

b The leaf unit is 1 so there have been no digits omitted from the stem-and-leaf and the average and median are the same as in part a.

c The numbers do not represent any quantities – an average 'gender' is clearly nonsense.

3 a i 0.9985 s **ii** 0.9450 s

iii 1.34 – 0.84 = 0.5 s

b The leaf unit is 0.01, so no digits are omitted in the stem-and-leaf, and the average, median and range are as in part a.

4 a i 185.8 s **ii** 100 s **iii** 1920 s

b i 159.7 s **ii** 100 s **iii** 795 s

c i 157.3 s **ii** 100 s **iii** 790 s

d part b

5 a Range = 79 – 24 = 55 min. Median is 59.5 min.

Exercise 2D

1 a Seven: 2.75, 2.9, 3.01, 3.10, 3.54, 3.97 s. Each occurs twice.

b One, at 3.5; it occurs 6 times.

2 a Two; 16 and 32; each occurs 3 times.

b Same because the leaf unit is 1 and no digits have been omitted.

c The mean and median of the data are 45.7 and 45 – very similar. 16 and 32 are very different to each other and to the mean and median. 32 could be considered as not too far from the centre of the data but 16 is very small. And note that we really need the median to decide what the approximate centre of the data is.

ISBN 978-1-107-61421-5

3 **a** 45 s; occurs 5 times.

b Two: 60 s, 70 s; occur 6 times each.

c In the first (lowest values) group

d Not 45 s. 60 or 70 s are in the 50 s-wide interval with the most observations – 50–99 s – so that interval can be regarded as the most frequently occurring group.

e Mean is 159.7 s and median is 100 s. These plus the interval (50–99 s) of most frequently occurring lengths, all give slightly different information.

4 **a** Two modes: 72, 74 min; each occurs 7 times.

b The modes provide little information without the stem-and-leaf plot to show that the data are fairly evenly spread. The mean and median are less than the two 'modes'. The 72 and 74 values are at least in the interval that has slightly more than the others, but it is very likely that a different arrangement of groups would give a different most-likely interval.

c The stem-and-leaf plot has a leaf unit of 1, so the digits in the 2 decimal places have been omitted (not rounded). If the data were rounded to the nearest minute, there may or may not be different modes to those in part a.

Exercise 2E

1 The times range from 2.08 s to 4.94 s, giving a range of 2.86 s. The stem-and-leaf plot shows that the most common interval for times is between 3.5 and 3.9 s. The average time is 3.44 s and the median is 3.46 s. In terms of speed, this means that the speeds range from 36.4 km/h to 87 km/h, with a median speed of 52.02 km/h. That is, half of the drivers were doing 52.02 km/h or more km/h; 26% of the drivers were doing more than 60 km/h.

2 The reaction times range from 0.84 s to 1.34 s, giving a range of 0.5 s. The stem-and-leaf plot shows that the reaction times are variable, with a group less than 0.93 s and another group more than 1.03 s. The average is just under 1s (0.9985) and the median is also just under 1 sec at 0.945 s. So both the mean and median are in between the two groups of reaction times. These two groups may correspond to different regions of the game as some of the numbers may be faster to hit than others.

3 The data range from 10 s to a very large and isolated value of 1930 s. As the next largest phone call length is 800 s, it was decided to omit the value of 1930 s as being too different to the rest. The stem-and-leaf plot shows that there are still two unusually large call lengths of 750 and 800 s, while the most common interval is 50 to 99 s. This is probably fairly typical of phone conversations, with a few long ones and many small ones. The average length of calls is 159.7 s and the median is 100 s, considerably lower than the average because half the calls are shorter than 100 and half greater than 100 s but spread from 100 up to 800 s.

4 The stem-and-leaf plot shows that the data are fairly evenly spread from 37 min to 79 min, with one unusually short CD of 24 min. The interval 70–74 min is the most common, but only just more common than the interval 75–79 min. The median is 59.5 min, so half of the CDs are longer than 59.5 min, which is close to an hour.

5 There are three spans that are clearly mistakes: 40 cm, 70 cm, 74 cm. Because the data cannot be checked, these observations were omitted. The heights have a range of 37 cm, with an average height of 159.3 cm and a median height of 160 cm. The spans have a much greater range of 84 cm with an average of 157 cm and a median also of 160 cm. The difference of span-height was taken and the stem-and-leaf shows that most of these are clustered around their median of 0, with all but one ranging from –1.9 to 1.4 cm. There is one difference of –56 cm which indicates that this observation may also have an incorrect span measurement. The average difference with this observation left in, is –2.3 cm. (Omitting the observation with a difference of –56 cm will bring that average closer to 0.)

Chapter summary

Multiple-choice questions

1 C **2** B **3** D **4** B **5** A

6 D **7** A **8** B **9** A **10** C

Short-answer questions

1 Does it include making or just delivery? When is the start of the time taken? What is the endpoint of the time taken to be? How accurate are the times to be – to the nearest minute?

2 The definition of time between – is it time between arrivals? Time between departures? Or time between departure of one and arrival of the next? At which stop/s will the times be recorded?

ISBN 978-1-107-61421-5

3 a That the plane was early

b By someone recording the times on the board when the plane was designated as landed, and comparing with the arrival times as scheduled in the timetable (not as shown on the board as expected because that takes account of the real time situation)

c

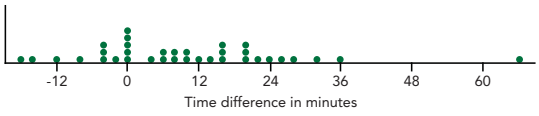
Leaf unit = 1.0

-1	4
-0	7555
-0	2
0	2223333
0	57789
1	012
1	5567
2	
2	556
3	0
3	
4	1

d 5 min

e There are 30 observations and 6 of them are negative, that is, early. So $\frac{24}{30}$ = 80% of the arrivals were later than as timetabled.

4 a



b

Leaf unit = 1.0

-1	973
-0	85552111
0	003557899
1	145569
2	002458
3	25
4	
5	
6	5

c There are 35 observations, with 13 less than or equal to 0, so $\frac{22}{35}$ = 62.9% of these arrivals were later than as scheduled in the timetable.

5 a 1 minute **b** $\frac{8}{62}$ = 12.9%

6 a 20 mm

b No. The rainfalls would be between 0 and 9.9 mm.

c $\frac{11}{215}$ = 5.1%

d 0 mm and 4.4 mm

7 a Range of 55 min. Mean = 8.63 min, and the median = 7 min

b One mode of value 3 min

8 a Range of 84 min. Mean = 9.43 min, and the median = 8 min

b Two with values –5 and –1 (occurring 3 times each)

9 a Range = 13.3 – 0.9 = 12.4 min
Median is 7.15 min

b There are 10 values that each occur twice in the stem-and-leaf plot. The values are: 3.9, 5.3, 5.5, 6.3, 7.7, 8.5, 8.6, 8.7, 8.8

10 a Median = 40 mm; range = 490 mm

b It would be 6.5; meaningless because the numbers do not stand for quantities.

c The value 20 mm occurs 27 times in the stem-and-leaf plot.

d Because the stem-and-leaf has omitted the units, the median tells us that half of the rainfalls are less than 49 mm; 47 of the rainfalls are between 0 and 19 mm, and 47 are between 20 and 39 mm, so the mode of 20 can be taken as representing the interval 0–39 which is the most commonly occurring interval. The median is a better indicator of centre here.

e It would give us something in the 40s because the 108th observation was amongst the digits of 4 in the leaf.

11 Arrivals ranged from 14 min early to 41 min late. On average, the planes were 8.63 min late with half of them arriving more than 7 min late, and 80% arriving after the scheduled timetabled time. The most commonly occurring interval of arrivals was between 0 and up to 5 min late.

12 Arrivals ranged from 19 min early to 65 min late. On average, the planes were 9.43 min late with half of them arriving more than 8 min late, and 34.3% arriving after the scheduled timetabled time. Most arrived between 8 min early and 9 min late.

ISBN 978-1-107-61421-5

13 The times between buses ranged from 0.9 min to 13.3 min. Half the times between buses were less than 7.15 min. Between 8 and 9 min was the most frequent interval of times between buses. 12.9% of the times between buses were more than 10 min.

14 The daily rainfalls ranged from 0 to 490 mm, with approximately half of them more than 40 mm. Only 5% of them were greater than 200 m. There were 47 rainfalls in each of the intervals 0–19 mm, and 20–39 mm, and only 6 daily rainfalls more than 280 mm.

Extended-response questions

1 a For buying reading glasses in pharmacies a chart is used with a number of lines of printed letters of decreasing size similarly to the Snellen chart. People are instructed to stand approximately 33 cm back from the chart and read the lines from largest to smallest until they cannot read most of the letters. Beside each line is a number ranging in half sizes from 1 to 3.5 – that is, there are 6 lines of printing, with the increasing from smallest to largest as the numbers go from 1 to1.5 to 2 to 2.5 to 3 to 3.5. These numbers give a measure of the strength of reading glasses required. These are intended only for people who do not require expert testing of their eyes but who need help for reading. The variable is categorical, but order matters so it is ordinal.

b A possibility is walking a person forward, starting from quite a distance until they can read a line of printing, and measuring their distance from it.

2 a We can't look at amounts of carbohydrate, protein and dietary fibre over different cereals unless they are in the same weight of cereal.

b

Stem-and-leaf of Protein in g per 100 g
Leaf unit = 0.10

0	555
0	666666666677777777
0	888888889999999999999999
1	00001111111
1	2
1	455
1	
1	999
2	11
2	3

Stem-and-leaf of Carbohydrate in g per 100 g
Leaf unit = 1.0

4	77
5	1
5	555569
6	2333334
6	67777999
7	00022223444
7	58888999
8	0001113334
8	5555666668888

Stem-and-leaf of Fibre in g per 100 g
Leaf unit = 0.10

0	11111111112222222222222334
0	56666788888899999
1	0001111222224444
1	5578
2	0
2	77
3	
3	
4	
4	6

c Protein: range = 23.1 – 5.4 = 17.7 g per 100 g; mean = 10.07 g per 100 g; median = 9.3 g per 100 g

Carbohydrate: range = 88.5 – 47.8 = 40.7 g per 100 g; mean = 73.4 g per 100 g; median = 74.2 g per 100 g

Fibre: range = 46.1 – 1.1 = 45 g per 100 g; mean = 8.68 g per 100 g; median = 8.3 g per 100 g

d Protein: The amount of protein per 100 g of cereal ranges from 5 to 23 g per 100 g, but for most cereals it lies between 5 and 11, with a few scattered over the remainder of the range. Half the cereals have less than the median of 9.3 g per 100 g. The average amount is 10.01 g per 100 g.

Carbohydrate: The amount of carbohydrate per 100 g of cereal ranges from 47 to 89 g per 100 g, with more cereals having large amounts. Half the cereals have less than the median of 74.2 g per 100 g. The average amount is 73.4 g per 100 g.

ISBN 978-1-107-61421-5

Dietary fibre: The amount of fibre per 100 g of cereal ranges from 1 to 46 g per 100 g, but for most cereals it is less than 14, with half the cereals having less than the median of 8.3 g per 100 g. The average amount is 8.68 g per 100 g. There are a small number with higher fibre, including one with very high fibre.

Chapter 3

Pre-test

1 c, f

2 **a** $\frac{2}{5}$ **b** 33% **c** 0.15

3 1

4 **a** 0.25 **b** 0.8 **c** 0.1

5 If it is fair we should ideally get 10 heads, but 12 isn't so much larger, so it could be fair.

Exercise 3A

1 **a** 0.125 **b** $\frac{4}{5}$ **c** 0.25, $\frac{1}{4}$

2 f, b, d, e, c, a

3 **a** 1% **b** 0.025 **c** $\frac{1}{80}$

4 f, b, d, a, e, c (other orders are possible, as long as f comes before b, d comes before a, and e comes before c)

5 **a** $\frac{12}{20}$ = 0.6 using frequency

b 70% = 0.7 using frequency, seems to have increased a little, which is not surprising due to the extra practice.

Exercise 3B

1 $\frac{1}{6}$, making the assumption that the die is fair, i.e. the faces are equally likely to turn up

2 $\frac{1}{14}$ **3** $\frac{89}{90}$

4 No, two possibilities doesn't mean that they are equally likely; could investigate experimentally by pouring out lots of drawing pins and counting how many fall each way; the more pins you pour, the closer you get to the 'true' probability but you can never know if/when you have got there!

5 **a** Seems reasonable if the colours are equally likely

b Seems fake, even if the colours are equally likely

c Gives you evidence that the colours are not equally likely (though actually the company hasn't said that they are)

Exercise 3C

1 $\frac{1}{5}$ **2** $\frac{1}{32}$

3 $\frac{1}{9}$, assuming that accidents are equally likely along the motorway

4 **a** Yes, since they have equal area – but it may be that people are more likely to shoot their darts into a particular area of the board; we could check by looking at the results of many random throws.

b $\frac{1}{20}$

c $\frac{1}{20}$ (double-19)

d $\frac{1}{20}$ (triple-11)

5 There are 10 by 20 = 200 squares; if we assume that the animals are equally likely to be in any of the squares (and completely in one square) then the probability is $\frac{1}{200}$; this may be unrealistic, as we might aim the photo towards where we think they are likely to be.

Exercise 3D

1 **a** $\frac{3}{6} = \frac{1}{2}$ **b** $\frac{1}{3}$ **2** $\frac{2}{8} = 0.25$

3 **a** $\frac{3}{36} = \frac{1}{12}$ **b** $\frac{1}{12}$

4 $\frac{10}{25} = \frac{2}{5}$ **5** $\frac{9}{100} = 0.09$

Exercise 3E

1 22 out of 40 seems consistent with guessing, so our overall conclusion would be that people can't identify black by taste; 28 out of 40 seems high enough that it provides evidence that (some) people are able to do this

2 Give them beans of the three different colours in a random order and ask them to say whether 1, 2 or 3 was the black; whether to give them practice beans beforehand, how to decide on random order for the three beans, how to record the results.

3 **a** Yes, as there are five of each type in the pack.

b $\frac{1}{5}$

ISBN 978-1-107-61421-5

4 a This is only one better than chance (4 correct), so no evidence of ESP.

b This is worse than chance, but not a huge amount worse – again, no evidence of ESP.

c Ten correct guesses might give you evidence that the subject knew which card was being selected (at least half the time), and 15 would give very strong evidence for this – but there may be another explanation for the success rather than psychic powers.

Chapter summary

Multiple-choice questions

1 C **2** B **3** C **4** D **5** A

6 B **7** C **8** D **9** A **10** C

Short-answer questions

1 a {falls buttered-side up, falls buttered-side down}

b Not necessarily (and common wisdom says that it will fall buttered-side down!); check by dropping the toast many times

2 a Three colours for t-shirt, shorts and track suit in order {ggg, ggr, ggw, grg, grr, grw, gwg, gwr, gww, rgg, rgr, rgw, rrg, rrr, rrw, rwg, rwr, rww, wgg, wgr, wgw, wrg, wrr, wrw, wwg, wwr, www}

b $\frac{9}{27} = \frac{1}{3}$

3 a $\frac{1}{16}$

b That the contestant selects the number at random, which may not be correct as there are 'favourite' numbers (such as 3) that people are more likely to select.

4 $\frac{1}{9}$, when you pick the second sock there are nine left in the drawer, one of which is the pair of the first one you selected.

5 a $\frac{1}{40}$

b $\frac{2}{40} = 0.05$

c $\frac{12}{40} = 0.3$ (although some of the more southern maps may show a small amount of coastline, which would reduce the probability).

Extended-response questions

1 Using the frequency idea, we can find the probability of giving birth to a baby boy by looking at records of a large number of births. See Exercise 3B question 6 for John Graunt's value of 0.517 from 17th century London.

Using the model approach, we could argue that since there are two possibilities, boy or girl, the probability of a boy is $\frac{1}{2}$ by symmetry.

In a particular situation, say if your older sister is pregnant, the subjective approach may be appropriate. If your grandmother has done a ring test and said that the baby will be a boy, your subjective probability may be a value close to 1.

2 a {111, 112, 113, 114, 115, 116, 122, 123, 124, 125, 126, 133, 134, 135, 136, 144, 145, 146, 155, 156, 166, 222, 223, 224, 225, 226, 233, 234, 235, 236, 244, 245, 246, 255, 256, 266, 333, 334, 335, 336, 344, 345, 346, 355, 356, 366, 444, 445, 446, 455, 456, 466, 555, 556, 566, 666}, use a systematic approach such as having the three numbers in non-decreasing order.

b No; theoretically, an outcome with three different numbers, such as 123, could come up in several ways on the three dice (actually, six different ways) while an outcome such as 333 could come up in only one way; practically, we could roll three dice many times and check whether all outcomes seemed to come up roughly the same number of times (this would require a large number of rolls).

3 There are many possible questions, and you have lots of examples in the chapter!

Chapter 4

Pre-test

1 a 0.55 **b** 0.875 **c** 0.3

2 0 **3** $\frac{3}{22}$ **4** $\frac{5}{1000} = 0.005$

5 Maybe she knows how to toss to get a particular result; maybe the coin is not fair; maybe the results are just chance from a fair coin.

Exercise 4A

1 a {plant grows, plant doesn't grow}

b {first serve ace, first serve in but not ace, second serve ace, second serve in but not ace, double fault}

c {lands buttered-side up, lands buttered-side down, (caught and eaten by the dog)}

ISBN 978-1-107-61421-5

Cambridge University Press

2 **a** First respondent was the only one in favour

b Exactly one person in favour

c Majority in favour

3 **a** nnn

b nnn, nny, nyn, ynn

c nnn, yyy

4 **a** {red, black, green}

b $\frac{18}{37}$

c 'black or green', $\frac{19}{37}$

5 **a** The die is symmetric and regular.

b {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

c 'no more than 4' = 1, 2, 3, 4, so probability is $\frac{4}{12} = \frac{1}{3}$

d 'more than 4' or 'at least 5', probability $\frac{2}{3}$

Exercise 4B

1 **a** Inclusive or (both flours could be used in the same loaf)

b Exclusive or (we can only win one prize with one ticket)

c Inclusive or (they would be happy with people who can do both)

d Exclusive or (there will be only one winner)

2 **a** $\frac{2}{26}$

b 1 (they are complementary events)

c $\frac{10}{26}$ (using inclusive or)

3 **a** {lit & art, lit & hist, lit & mus, lit & Fre, lit & Man, art & hist, art & mus, art & Fre, art & Man, hist & mus, hist & Fre, hist & Man, mus & Fre, mus & Man, Fre & Man}

b 'choose two languages' and 'include literature in your choice'

c 'include art in your choice' and 'include French in your choice'

4 **a** $\frac{8}{177}$ **b** $\frac{47}{177}$

5 **a** Investigation going well **b** In danger
c 3:12 **d** 3:47

Exercise 4C

1 **a**

	Happy with features	Not happy	Totals
Smartphone	23	12	35
Other phone	15	25	40
Total	38	37	75

b $\frac{12}{75}$ **c** $\frac{52}{75}$

d 'had smartphone and was happy with features'

2 **a** 304, 313, 353 **b** $\frac{12}{365}$ **c** $\frac{3}{365}$

d $\frac{61}{365}$ using inclusive or, 13th or Friday or both seems to be what is required

3 **a** $\frac{12}{26}$ **b** $\frac{6}{26}$ **c** $\frac{20}{26}$ using inclusive or

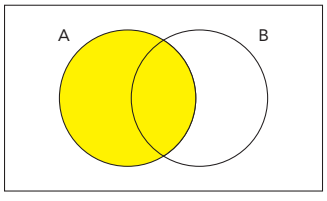
4 **a** $\frac{11}{62}$ **b** $\frac{1}{62}$ **c** $\frac{10}{62}$

d Chance of LGA baby and high GI diet is ten times the chance of LGA baby and low GI diet, so it seems that the low GI diet is better.

5 **a** $\frac{53}{92}$ **b** $\frac{78}{92}$ **c** $\frac{50}{92}$ **d** $\frac{42}{92}$

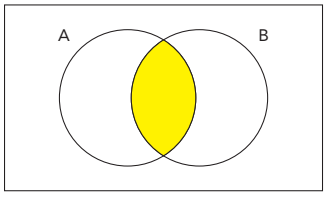
Exercise 4D

1 **a**



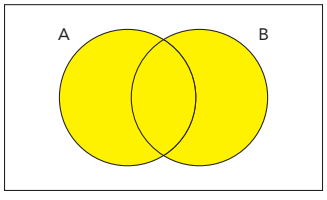
The student has blond hair.

b



The student has blond hair and blue eyes.

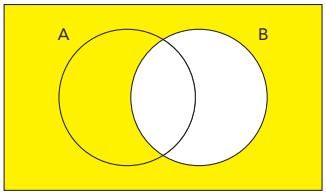
c



The student has blue eyes or blond hair or both.

ISBN 978-1-107-61421-5

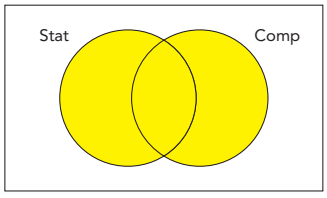
d



The student does not have blue eyes.

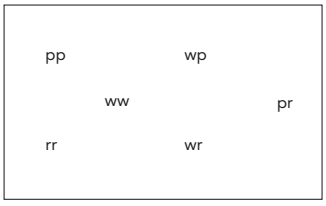
2 a {Statistics and Computing, Statistics only, Computing only, neither Statistics nor Computing}

b, c



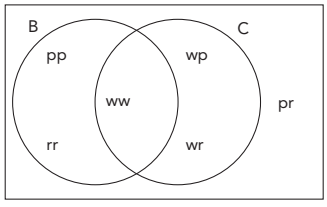
Using the inclusive or, Statistics or Computing or both

3 a



Separate outcomes shown

b, c



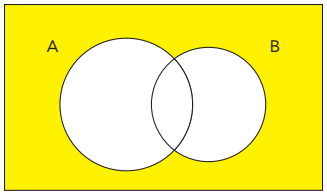
d 'both plants had flowers of the same colour or at least one white flower was obtained or both', or alternatively, 'everything except one pink and one red flower'

4 a 15, 30, 45, $\frac{3}{50} = 0.06$

b 3, 5, 6, 9, 10, 12, 18, 20, 21, 24, 25, 27, 33, 35, 36, 39, 40, 42, 48, 50, 'a number divisible by 3 or 5 but not by 15', $\frac{20}{50} = 0.4$

c 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18,19, 21, 22, 23, 24, 26, 27, 28, 29, 31, 32, 33, 34, 36, 37, 38, 39, 41, 42 ,43, 44, 46, 47, 48, 49, 'a number that is not divisible by 5', $\frac{40}{50} = 0.8$

5 a



'not in A and not in B'

b Same diagram as in part a, 'outside the area represented by a or B or both'

c This demonstrates that 'A^C and B^C' is the same as '(A or B)^C' (inclusive or)

6 a Dodgson wrote *Alice in Wonderland* and various other books and poems under the pseudonym of Lewis Carroll.

b A Carroll diagram classifies numbers or objects into whether they have or don't have a property A, and whether they have or don't have a second property B.

c The Carroll diagram has four regions: A and B, A but not B, B but not A, and neither A nor B. The Venn diagram for two events A and B has the same four regions: A and B is the intersection of the circles, and neither A nor B is the region outside both circles. So the two diagrams work in a similar way. People usually use the Venn diagram, but sometimes the Carroll diagram is easier.

Chapter summary

Multiple-choice questions

1 C **2** D **3** A **4** C **5** A

6 B **7** C **8** D **9** B **10** A

Short-answer questions

1 a {$2 $1, $2 50c, $2 20c, $1 50c, $1 20c, 50c 20c}

b '$2 coin not selected' **c** 0.5

2 a {orange, yellow, green, light blue, mid blue, dark blue, purple, red, N, S, E, W, spades, hearts, diamonds, clubs}

b $\frac{12}{64}$ **c** $\frac{24}{64}$, exclusive or **d** 'a colour wins', $\frac{40}{64}$

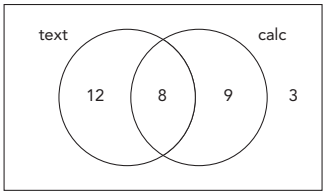
3 a {self and win, self and lose, company and win, company and lose}

b $\frac{325}{380}$ **c** $\frac{155}{380}$

4 a $\frac{278}{900}$ **b** $\frac{405}{900}$ **c** $\frac{65}{900}$, inclusive or

ISBN 978-1-107-61421-5

5 a



b $\frac{29}{32}$, 'inclusive or'

c 'neither textbook nor calculator', $\frac{3}{32}$

Extended-response questions

1 a $\frac{2}{52}$, only 10 diamonds and 10 hearts

b $\frac{26}{52} = 0.5$, exclusive or since the events are disjoint

c $\frac{16}{52}$, any diamond (13) plus the three other aces

d Riffle shuffle 7 times then turn over the top card, note down the card, repeat many times, look at frequencies for each card

2 a Larceny is the taking and carrying away of another person's property, $\frac{300}{778} = 0.386$

b $\frac{712}{778} = 0.915$

c $\frac{295}{778} = 0.379$

d $\frac{717}{778} = 0.922$, inclusive or since it is more reasonable to consider larceny or 7 years or both

e 300 + 712 – 295 = 717, the numbers for larceny plus the numbers for 7 years 'double count' the numbers for both, so that has to be subtracted to get the numbers for larceny or 7 years

3 There are many possible questions, and you have lots of examples in the chapter!

Chapter5

Pre-test

1 a Developing countries

b (i) population size, (ii) average income per person, (iii) birth rate as number of births per 1000 people, and (iv) number of doctors per 1000 people.

c (i) count; (ii)–(iv) continuous (because the counts in (iii) and (iv) are rates)

d Dotplots or stem-and-leaf plots can be used for all three.

e 100

2 a Continuous **b** Continuous

c Categorical (not ordinal)

d Categorical (not ordinal)

e Categorical: ordinal **f** Count

3 a (i) Amount of television watched per week in hours, (ii) play a sport (yes/no), (iii) favourite recreation, (iv) opinion on TV sport.

b (i) continuous; (ii)–(iv) categorical

c Gender, age because these might affect the variables in (i); possibly their favourite sport.

d Gender is categorical, age is continuous (unless age groups are used in which case it's categorical), favourite sport is categorical.

e The question 'whether they play a sport' is a bit vague and may be altered.

4 a Number of people who start crossing against the flashing red or against the red; count

b Location of lights; categorical

c Cycles of lights – that is, sets of green, flashing red, red

d Time of day, day of week because how busy it is and types of users may affect variable in part a.

Exercise 5A

1 a Census **b** Sample

2 a Census **b** Sample

3 a People: 20 **b** Seedlings: 200

c Outlets: 30 **d** Companies: 50

4 a Heights of Year 8 students in Australia

b All Australians as it does not change with age; possibly everyone

c Year 8 students similar to your class

d Only of those who see it and choose to vote

e People who are at home in the afternoon (and answer their phones and agree to be surveyed)

f The customers of that chain of stores

g The customers of that chain of stores

ISBN 978-1-107-61421-5

Exercise 5B

1 **a** Daily heights, including initial height

b Conditions including amount of water, sun, shade, etc. Also time of day when heights measured

c Allocation of seedlings to fertiliser

d Divide seedlings into 4 groups (randomly), with the 4 conditions being fertiliser A, morning application; fertiliser B, morning application; fertiliser A, evening application; fertiliser B, evening application.

2 **a** Denote wheat type by A, B; fertiliser amount by 1, 2. In each block, A1, A2, B1, B2 each should be assigned to one of the sections.

b Randomise the assigning to sections within each block.

c 16

3 **a** The order of music/not music for each subject

b No, because the subjects will have become a bit familiar with it.

c Whether they usually listen to music while working; age; gender; possibly their music preference. Note that their basic expertise in typing is measured within the experiment.

d Subjects could be paired to be similar as to age and gender and typing expertise, and then type the same piece, one with and one without music. Randomise within each pair.

4 **a** Any differences between the angles might be to do with differences between the packets.

b Half the candles from each packet to be assigned to vertical and half to angle, with same number of each colour in vertical and angle. That is, if there are 6 of each colour in a packet, take a candle from packet A, randomly assign to vertical or angle; continue until correct number of each colour in vertical and angle. Candles should be lit by same person; candles should be extinguished by same person.

Exercise 5C

1 **a** Because they don't want more than one lot of information per family

b Families of students at the school

2 **a** Because people who give their opinion but don't vote will distort the poll predictions.

b If the election were held today, would you vote?

3 **a** No, because it's a voluntary survey. Also not all readers will have email

b Higher No vote than should be if placed elsewhere in paper

c Because it includes 'about right' which was not an option in a simple yes/no. A No vote could have meant it's about right or that they want more.

d Which sport? Some people might think there's too much of some but not of others. Also which is their favourite sport?

4 Question 1 because it is a neutral question, where questions 2 and 3 are leading respondents in a certain direction.

5 **a** Ratepayers in the region

b Sending survey forms out with rate notices or phoning a random sample of households and asking to speak to the person who pays the rates

c They are not necessarily a random sample of ratepayers.

d It's hard to say no to more cycling paths to a cyclist!

6 **a** Only reaches people who tend to go to afternoon sessions.

b Sample people leaving different session times at more than one cinema

c Avoid asking a group – ask only one person of a group

d Have a set of random numbers, e.g. from 1 to 50, and arrange them from smallest to largest and ask those people as they come out. However if random numbers are too close, perhaps multiply them by 2 to get sufficient gap between. Alternatively, give these people a simple piece of paper asking for a rating and provide a box close by.

e No, because these are people who have already chosen to see the movie. High ratings might encourage more to see it.

f Popularity is measured by numbers of tickets sold; approximately by takings.

g Those who read the internet sites on movies and who answer their surveys

ISBN 978-1-107-61421-5

Cambridge University Press

Exercise 5D

1 **a** Because we can't assign people to be left-handers or right-handers

b In previous times, people who naturally wrote left-handed were forced to change. So in older age groups there will be more who became right-handers than in younger age groups.

2 **a** We can't randomly assign weather and traffic conditions.

b Location and time of day of measurement

c Weather conditions, day of week, traffic conditions

3 **a** Whatever conditions we may be interested in, we can't randomly assign these to toddlers.

b Choosing toddlers at random

c Gender, position in family, educational level of parents, amount of contact with other children

4 **a** We can't randomly assign men to be bald or not.

b Age of man

c No, because this is not an experiment.

5 **a** We are observing trains without controlling the conditions.

b Length of route or number of stations or both; number of passengers (approximate from records).

Chapter summary

Multiple-choice questions

1 A **2** B **3** B **4** A **5** B

6 C **7** C

Short-answer questions

1 There will clearly be many forms that are not returned or not filled out properly. And there may be some households that return two forms.

2 **a** A census of the top 100 songs in two charts in that year

b It's probably OK to assume that the lengths of the songs are randomly representative of all the songs played that year, but the others may be more associated with being in the top 100.

c Some of the songs are in both charts.

3 **a** No, because the data are for one particular sector of employment. Even though it is broad it is different from other sectors.

b Yes for teenagers in that age group, but there's no representation of the oldest teenage group.

c No, because only one region has been sampled. This data might be helpful in indicating what could be sampled elsewhere.

d This depends on the festival, but generally no, because people attending a festival are not a random sample of people.

4 **a** The order in which the lady was asked to taste the teas.

b Same amount of milk, same strength and temperature of tea; and that person giving the lady the tea or watching should not know which was which.

5 **a** The amount of jelly in each cup; temperature in fridge; number of jelly specimens in fridge at one time

b Temperature varies slightly over the interior of a fridge, so jelly specimens would need to be placed where temperature is much the same. Also the number of jellies placed in a fridge to set may affect the setting time, so the number being set in a fridge at one time could not be too many and needs to be constant.

6 **a** The order in which the methods are used

b Assigning of methods within pairs

c Age, gender

7 People who are online users are more likely to shop online.

8 **a** Need to avoid interviewing same people. Even if different people, they are not a random sample of the gym users, as different types of people tend to prefer different times.

b Choose sessions at random throughout a week and choose people at random who attend those sessions.

c Choose gym members at random from their membership numbers and mail/email surveys to them. If the gym has a large number of casual users, would need to combine this with giving random casual users a survey form.

ISBN 978-1-107-61421-5

d No, because gym usage may be quite different across gyms. However, the results, including responses to any open questions, could help in designing surveys for other gyms.

9 a Cluster sampling

b Stratified sampling

10 a He needs to know what students' previous performances have been, and whether their TV viewing habits have changed. He can't compare students' performances with respect to TV watching or anything else unless he has some way of matching their capabilities.

b No. Not only is it an observational study but he has no comparison of them without any TV watching or when they all watch the same amount. The study is badly planned.

11 a An observational study because, although they've introduced a new meal deal, they are not controlling conditions.

b They would need to look at overall sales to see if the new meal deal has added to overall or just moved sales from other products. They could look at sales of specific products that could be considered direct alternatives to the new meal deal.

c To compare the two months, they could compare daily sales for the same days of the week as there are probably patterns over days.

Extended-response questions

1 a Because it's easier for people to answer about a specific weekend than in general, and also because the investigators will know what the weather was like on that weekend in the region where the people live.

b As close to the weekend as possible; people are more likely to be home on Monday and Tuesday evenings.

c Fully or nearly fully grown people so teenagers and adults

d For a random sample of teenagers and adults of both genders and a variety of ages, rub in recommended amount of sunscreen A on one arm, and sunscreen B on the other arm. Randomise which arm has which sunscreen – just in case there's a difference between subject's left and right arms. Measure time until absorbed, with what is meant by absorption to be carefully identified and described before the experiment starts – probably through a pilot experiment. Then take differences in times for each subject.

2 a $2 \times 3 \times 2 \times 3 = 36$

b Because we wouldn't know if any differences due to time in dryer might involve some differences between individual towels.

c 36 so 72 observations altogether

d Timer settings

e Cool to the same starting temperature each time

f Because they will start to dry by themselves and we need the starting point for each towel to be the same when it's put in the dryer.

g We need 36 of each type of towel, so assign towels to conditions randomly.

Chapter 6

Pre-test

1 a 200

b i 18.5% **ii** 38% **iii** 51.7%

c No, because it just says 'school students'. We don't know from which type or larger group of school students, these 200 were randomly chosen.

d i An advantage is that it is easier to answer and we are not depending on interpretations of 'main method'.

ii A disadvantage is that it may be an unusual morning for some students. Another disadvantage is that some students may have used two (or more) different types of transport and won't know which to say.

2 a i 18.3% **ii** 65.7% **iii** 17.6%

b We know the sample was taken from Years 7 and 8 students in NSW and Vic. We don't know when the data were taken.

ISBN 978-1-107-61421-5

Cambridge University Press

3 a

Stem-and-leaf of concentration times
Leaf unit = 1.0

2	79
3	0001
3	55666678
4	012222
4	
5	02
5	9
6	2
6	
7	
7	
8	1

b Mean = $\frac{1019}{25}$ = 40.76 s. Median = 37 s

4 25 because there are 25 pairs and results are being collected per pair

Exercise 6A

1 23

2 a $\frac{34}{580}$ = 5.9% **b** $\frac{29}{42}$ = 69%

c i For (1), range is 3.4%, median = 2.2%, mean = $\frac{107.8}{50}$ = 2.156%.
For (2), range is 3.8%, median = 4.2%, mean = $\frac{213.8}{50}$ = 4.276%

ii 2 **iii** 7 **iv** 12

v Even when the population percentages are 2.24% and 4.26%, can get a lot of variation in proportions even in as many as 500 people, so need to be careful in making any claims.

3 a 24.5% = $\frac{26}{106}$

b Have calculated $\frac{26}{34}$, which is % of those who did not eat breakfast who are female

c Range is 9.5%, median = 16.5%, mean = $\frac{818.5}{50}$ = 16.37%

d $\frac{30}{50}$ = 60%

e $\frac{(17 + 10)}{50}$ = 54%

Exercise 6B

1 a 20 **b** 50 **c** With replacement

2 a Need halfway between the 75th and 76th observations: these are 17 and 18 min so the median is 17.5 min

b Leaf unit = 10 in all stem-and-leaf plots below.

Sample 1

0	0011
0	2
0	44
0	
0	
1	01
1	2

Sample 2

0	01111
0	2233
0	5

Sample 3

0	111111
0	2
0	4
0	6
0	
1	1

Sample 4

0	000001
0	223
0	4

d For sample 1, $\frac{6}{10}$ = 0.6. For sample 2, $\frac{5}{10}$ = 0.5.
For sample 3, $\frac{4}{10}$ = 0.4. For sample 4, $\frac{4}{10}$ = 0.4

e The means, medians and ranges of the four samples have a lot of variation among them as well as being very different to the original dataset. It seems that a lot of variation is possible in such tiny samples.

3 a The leaf unit is 1 in all the stem-and-leaf plots below.

Sample 1

3	7
4	3
4	7
5	13
5	589
6	0
6	7
7	234
7	78

Sample 2

4	33
4	7
5	
5	6689
6	0
6	5
7	0014
7	88

ISBN 978-1-107-61421-5

Sample 3

3	77
4	4
4	569
5	23
5	7
6	04
6	55
7	24

Sample 4

2	4
2	
3	
3	7
4	1334
4	6
5	
5	7
6	02
6	5
7	23
7	68

b

Sample	Mean	Median	Range
1	60.27	59.00	41.00
2	61.87	60.00	35.00
3	54.67	53.00	37.00
4	54.73	57.00	54.00

c The mean, median and range of the 128 observations are 59 min, 59.5 min, 55 min. We see that the means and medians of the four samples of size 15 vary around these values but not excessively. However the ranges are more variable – this is not surprising as the range depends on what happens to be the minimum and the maximum in the small sample obtained by resampling.

d For sample 1, $\frac{7}{15}$ = 47%; for sample 2, $\frac{8}{15}$ = 53%; for sample 3, $\frac{6}{15}$ = 40%; for sample 4, $\frac{7}{15}$ = 47%.

4 a This is 5 min so perhaps there is someone checking after 5 min.

b i The leaf unit in each of the stem-and-leaf plots below is 10.

Sample 1

0	0000000011234
0	55589
1	
1	
2	1
2	
3	0

Sample 2

0	00000011333
0	67889
1	1
1	5
2	1
2	
3	0

Sample 3

0	0000111234
0	668
1	01
1	9
2	
2	
3	0000

ii

Sample	Mean	Median	Range
1	51.2	19.0	296.0
2	66.9	34.0	296.0
3	99.2	54.0	296.0

iii The ranges are very similar. Because there are so many of the original observations with the minimum and maximum values, these have a good chance of being included in any resampling. Both the means and the medians vary a lot – again because there are so many of the small and large values in the original data.

iv Because the leaf unit of the stem-and-leaf plots is 10, we can only see how many are at least 20, but this will give us the correct number because the observations are in whole numbers. For sample 1, $\frac{10}{20}$ = 50%; for sample 2, $\frac{12}{20}$ = 60%; for sample 3, $\frac{13}{20}$ = 65%

Exercise 6C

1 a The observations 69 cm, 72 cm and 50 cm are all considerably less than the other observations. The mean and median of all the data are 136.61 cm and 149.75 cm respectively. Without these three observations, the mean and median are 151.2 cm and 150 cm respectively. So without them, the mean has been increased by almost 15 cm but the median has hardly changed. Yes, the observations should be removed if they can't be checked as they are not reasonable values for armspans and are almost certainly mistakes or mis-measurements.

b The observation of 125 cm is an outlier. The mean and median of all the data are 30.28 cm and 24.3 cm respectively. Without this observation, the mean and median are 24.7 cm and 24 cm respectively. So without this observation, the mean has decreased by almost 6 cm but the median hardly changed. Yes, the observation should be removed if it can't be checked as it is obviously a mistake.

ISBN 978-1-107-61421-5

c The observation 120 min is an outlier. Note that although 2 and 8 are small, there is a value of 10 so these observations are small but cannot really be called outliers. The mean and median of all the data are 30.3 min and 20 min respectively. Without this observation, the mean and median are 24.7 min and 20 min respectively. So without this observation, the mean has decreased by almost 6 min but the median is unchanged. The observation may be a mistake as 120 min is a long time to take to get to school, but it is possible – it's not clear whether it should be omitted or not.

2 a For (1): the median of the sample means is 160 s and the range is 420 – 100 = 320 s; the median of the sample medians is 100 s and the range is 300 – 50 = 250 s. So the (re)sample medians are spread around a centre of 100 which is exactly equal to the median of the original data. The middle of the (re)sample means is slightly less than the mean of the original data. Also, the (re)sample means have a greater range than the (re)sample medians – they are varying more than the (re)sample medians.

b For (2): the median of the sample means is 160 s and the range is 290 – 80 = 210 s; the median of the sample medians is 100 s and the range is 220 – 60 = 160 s. So the (re)sample medians are again spread around a centre of 100 which is exactly equal to the median of the original data. The (re) sample means are also now spread around a centre of 160 s which is exactly the mean of the original data. Again the (re)sample means vary more than the (re)sample medians.

c Leaving out the very large value of 1930 s from the original dataset has reduced the variation in the means and medians of samples taken from the larger dataset. Whether the value of 1930 is in or not, the sample medians are centred on the median of the original dataset (100 s). However, the sample means are more affected by whether this outlier is in or not; without it, the sample means are also centred on the mean of the bigger dataset from which the smaller samples are taken.

d For case (1):

i 25% **ii** 15% **iii** 48% **iv** 21%

e For case (2):

i 45% **ii** 24% **iii** 47% **iv** 23%

f In situation (1), the means of the resamples of size 15 are centred around 160 s, whereas the mean of the original dataset is 185.8 s, and not many of them are within 20 or 10 s of this original mean. In contrast, in situation (2), the means of the resamples of size 15 are centred around 160 s and many more are within 20 or 10 s of this value. Also these proportions are very similar to those for the medians of the resamples, which are much the same for situations (1) and (2).

g Would expect them to be less spread out

3 a For the samples of size 20, the median of the 100 sample means is 31 min and the range is 43 – 15 = 28 min.

For the samples of size 50, the median of the 100 sample means is 30 min and the range is 39 – 24 = 15 min.

b For the samples of size 20, the median of the 100 sample medians is 18 min and the range is 39 – 11 = 28 min.

For the samples of size 50, the median of the 100 sample medians is 17 min and the range is 30 – 12 = 18 min.

c Similarities: the sample means are centred around the mean of the original dataset (30.6 min) and the sample medians are centred around the median of the original dataset (17.5 min). The variations of the sample means and the sample medians are both much less for the samples of size 50 compared with the samples of size 20.

Contrast: about the only contrast is that the range of the sample means for the samples of size 50 is smaller than the range of the sample medians of these samples.

d For the sample means:

i 61%

ii 92%

iii 22%

iv 36%

e For the sample medians:

i 64%

ii 80%

iii 6%

iv 28%

ISBN 978-1-107-61421-5

f For both the sample means and the sample medians, there are more close to the values of the original mean (30.6 min) and median (17.5 min) in the samples of size 50 than 20. For both sample sizes, there are more sample means close to the mean of the original than there are sample medians close to the median of the original – that is, the sample means tend to get closer to the original mean than the sample medians to the original median.

g Both would be centred even more closely around the original mean and the original median.

Chapter summary

Multiple-choice questions

1 A **2** B **3** A **4** D **5** D

6 C **7** B

Short-answer questions

1 a $\frac{42}{80}$ has been calculated – this is the % of chocolate choosers who were female.

b i Median = 62.2% (halfway between 62% and 62.4%) and range = 68.4% – 53.6% = 14.8%.

ii $\frac{20}{50}$ = 40% **iii** $\frac{(10 + 18)}{50}$ = 56%

2 a $\frac{(46 + 60)}{205}$ = 51.7% **b** $\frac{20}{252}$ = 7.9%

c i Median is the 38th observation and = 24.7%, range = 30.5% – 20.2% = 10.3%

ii 35/75 = 46.7% **iii** $\frac{(16 + 14)}{75}$ = 40.0%

3 a The leaf unit in each of the plots below is 1.0.

Sample 1

0	6
1	1
1	6778999
2	11234
2	558
3	
3	69
4	1

Sample 2

0	7
1	1234
1	6678
2	012
2	667
3	02
3	56
4	1

Sample 3

1	34
1	67889
2	0222334
2	688
3	
3	56
4	
4	6

Sample 4

1	1
1	23
1	445
1	777
1	89
2	
2	223
2	445
2	
2	
3	0
3	22

b Sample 1: mean = 22.35 min, median = 21 min, range = 35 min

Sample 2: mean = 22 min, median = 20.5 min, range = 34 min

Sample 3: mean = 23.5 min, median = 22 min, range = 33 min

Sample 4: mean = 20.05 min, median = 18.5 min, range = 21 min

c The samples vary, but the 4 sample means are not too far from the original mean of 21.5 min and the 4 sample medians are within 2 min of the original median of 20 min. The 4 sample ranges are less than the original range but are still fairly wide.

d For samples 1, 2 and 4, $\frac{9}{20}$ = 45%; for sample 3, $\frac{12}{20}$ = 60%

4 a The value 6 min is a possible outlier. The mean and median of all the observations are 23.25 min and 22 min. Without the 6 min, the mean and median are 24.16 min and 22 min. Although 6 min is considerably smaller than the other observations, it is not necessarily a mistake and should not be omitted without a reason.

b The observation of 7 km/h is a possible outlier. The mean and median of all observations are 24.75 km/h and 23.5 km/hr. Without the 7 km/h, they are 25.93 km/h and 24 km/h. Although the 7 km/h is much smaller than the others, it is valid as a possible speed. Whether to omit or not depends on whether we want to cover all users or focus on most users.

ISBN 978-1-107-61421-5

Cambridge University Press

c The observations of 1 cm, 1.5 cm, 1.5 cm are clearly outliers even though there are three of these. The mean and median of all observations are 127.3 cm and 157.5 cm. Without these three observations, the mean and median are 158.75 cm and 159.5 cm. These observations are clearly mistakes and should be omitted.

5 a The observation of 7 km/h is the outlier. The mean of the observations without it = $\frac{(26 \times 625 - 7)}{624}$ = 26.03 km/h. So the mean has hardly changed because there are so many observations and 7 is not that far away from the rest. The median will change from being the 313th observation from the top (largest) to halfway between the 312th and 313th from the top so will either not change or change by 0.5 – most likely not change.

b The median of the 75 sample means is 25.9 km/h and the range is 28.4 – 23.3 = 5.1 km/h. The median of the 75 sample medians is 26 km/h and the range is 28 – 23 = 5 km/h. So both the sample means and the sample medians are centred around the mean and median of the original dataset. Their variation is also similar, varying by up to 5 km/h.

c For the sample means:

i $\frac{61}{75}$ = 81.3% **ii** $\frac{35}{75}$ = 46.7%

d Consider the sample medians.

i $\frac{61}{75}$ = 81.3% **ii** $\frac{26}{75}$ = 34.7%

e There don't appear to be any effects.

f They would be even more closely concentrated around 26 km/h with smaller ranges.

Extended-response question

For the samples of size 25, the sample means vary from 18.8 min to 26.9 min with a range of 8.1 min. The median of the sample means is 21.5 min so the sample means are centred on the mean of the 575 observations which are being resampled. The sample medians vary from 16 to 24 min with a range of 8 min. The median of the sample medians is 20 min, so the sample medians vary around their centre, which is the median of the 575 observations being resampled.

For the samples of size 50, the sample means vary from 18.5 min to 24.4 min with a range of 5.9 min. The median of the sample means is 21.5 min so the sample means are centred on the mean of the 575 observations which are being resampled. The sample medians vary from 17 to 23 min with a range of 6 min. The median of the sample medians is 20.5 min, so the centre of the sample medians is close to the median of the 575 observations being resampled.

So for both sample sizes, the sample means and medians vary but are centred on the mean and median respectively of the 575 observations being resampled. The variation of the sample means and medians is less in samples of size 50 compared with samples of size 25, but the variation is still approximately 6 min.

For the samples of size 25, 13% of the sample means are more than 2 min away from the original mean of 21.5 min, and 10% of the sample medians are more than 2 min away from the original median of 20 min.

For the samples of size 50, these percentages are unchanged: 13% of the sample means are still more than 2 min away from the original mean of 21.5 min, and 10% of the sample medians are more than 2 min away from the original median of 20 min. This would not necessarily be the same if we took another 100 samples of size 50 from the original 575 observations.

ISBN 978-1-107-61421-5

ISBN 978-1-107-61421-5

Cambridge University Press

ISBN 978-1-107-61421-5

Cambridge University Press