CHAPTER

1

Quantitative data across groups

What you will learn

How well can people judge distances?

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

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

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

AUSTRALIAN CURRICULUM

Statistics and probability

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

ISBN 978-1-107-65599-7

Cambridge University Press

Pre-test

Terms you will learn

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

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

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

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

b Give the types of each of these four variables.

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

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

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

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

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

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

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

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

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

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

b What population was being sampled?

c What type of sampling is this called?

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

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

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

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

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

ISBN 978-1-107-65599-7

1-1 Investigating to compare

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

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

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

Planning and collecting data

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

Cambridge GO

ISBN 978-1-107-65599-7

Planning a primary data collection

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

Using secondary data

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

LET'S START Judging distances

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

ISBN 978-1-107-65599-7

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

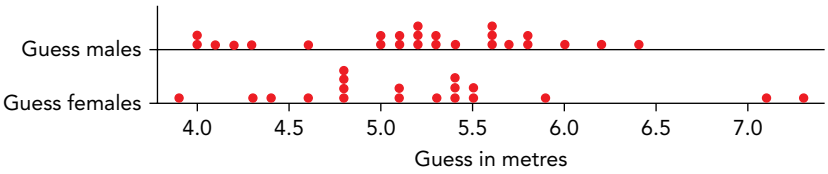
Males

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

Females

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

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



Back-to-back stem-and-leaf plots

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

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

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

ISBN 978-1-107-65599-7

Leaf unit = 0.1

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

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

Key ideas

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

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

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

ISBN 978-1-107-65599-7

Density at reefs < 20 km from coast

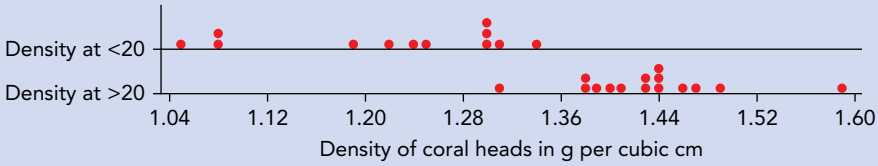
1.34 1.22 1.31 1.05 1.08 1.08 1.24 1.19 1.30 1.25 1.30 1.30

Density at reefs > 20 km from coast

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

Question: How do the densities compare?

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



Leaf unit = 0.01

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

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

Exercise 1A

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

a What measurements should the investigators take in this experiment?

ISBN 978-1-107-65599-7

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

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

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

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

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

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

Into Tasman Sea

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

Into Pacific Ocean

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

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

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

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

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

Lengths of Indie songs

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

Lengths of Alternative rock songs

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

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

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

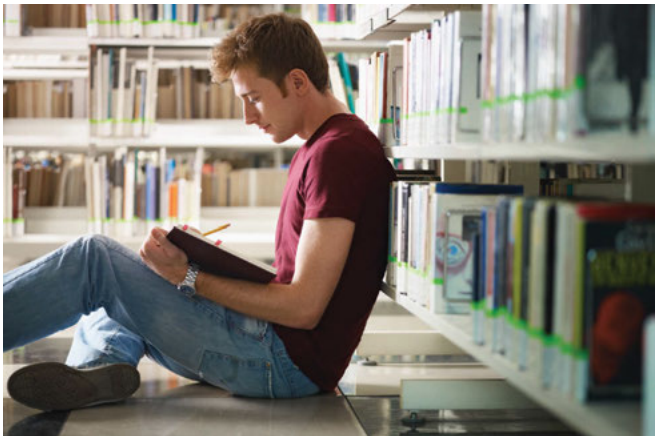
c What proportion of Alternative rock songs have lengths greater than the Indie median?

ISBN 978-1-107-65599-7

Enrichment

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

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



Females not reading

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

Females reading

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

Males not reading

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

Males reading

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

a What could you randomise in this experiment?

b Give an experimental condition that should stay the same.

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

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

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

ISBN 978-1-107-65599-7

1-2 Histograms

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

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

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

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

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

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

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

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

ISBN 978-1-107-65599-7

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

Bin: Interval of a histogram

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

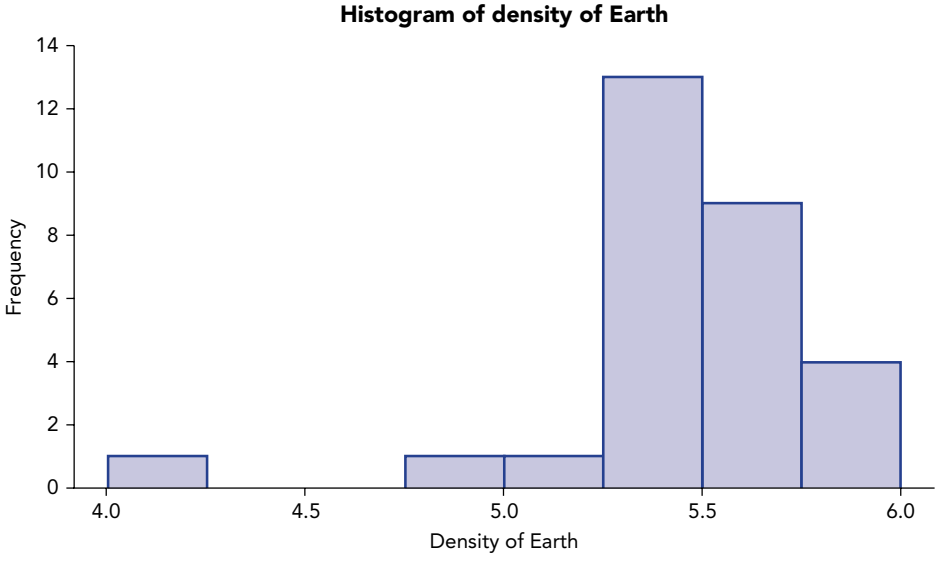
LET'S START A histogram of the famous Cavendish data

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

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

Question: How do we draw a histogram?

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

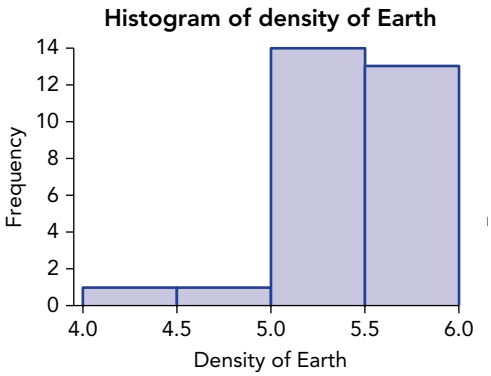


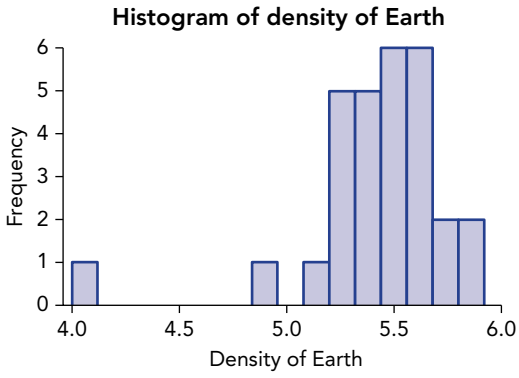
ISBN 978-1-107-65599-7

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

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

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





Key ideas

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

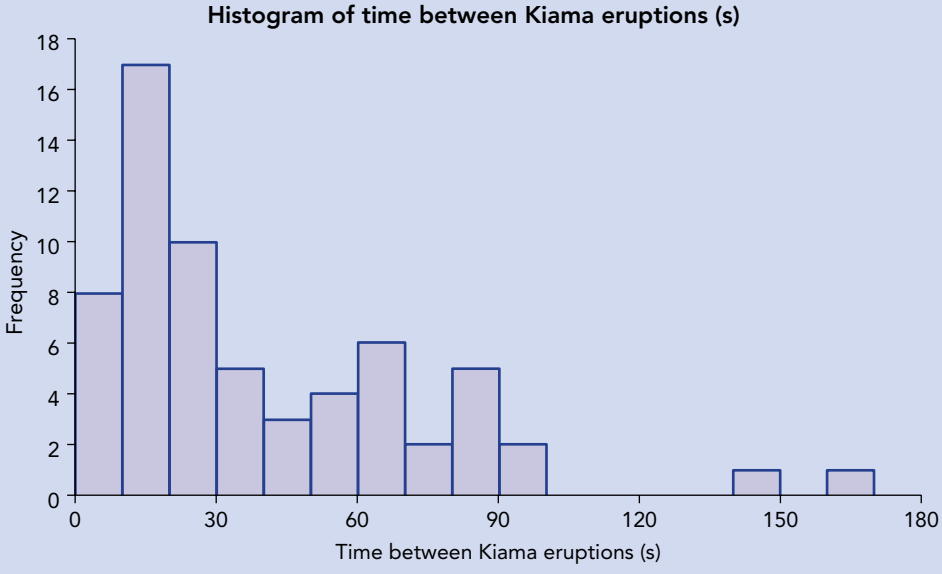
ISBN 978-1-107-65599-7

Example 2: How variable is the Kiama blowhole?

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

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

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



ISBN 978-1-107-65599-7

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

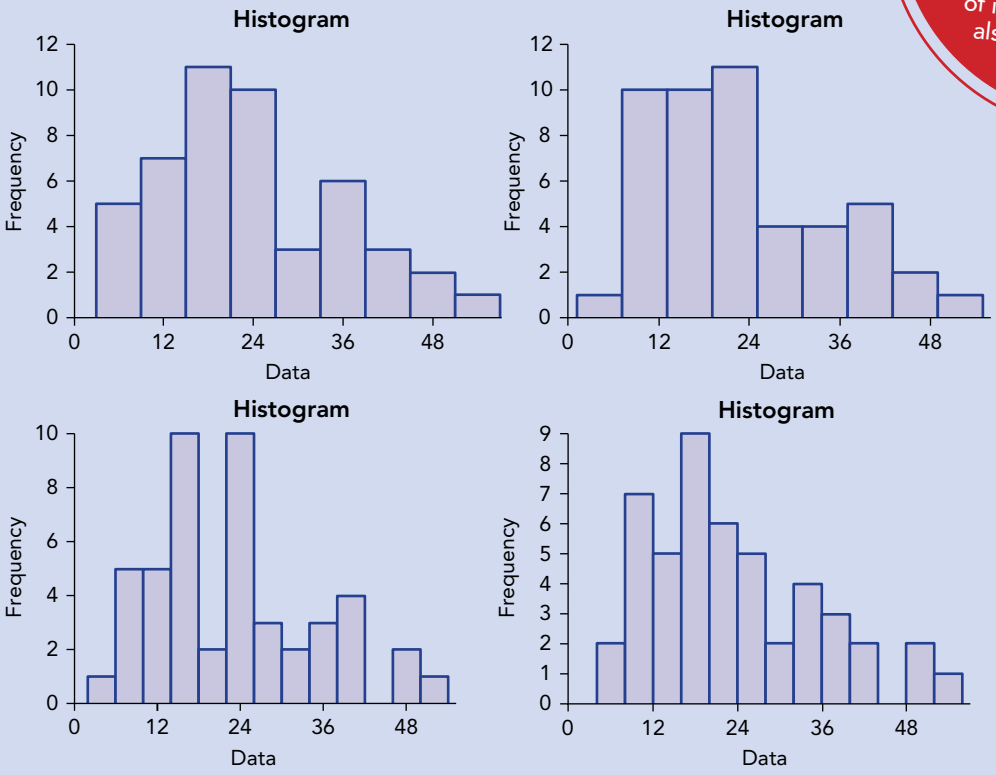
For the same dataset, the appearance of a histogram depends on where we start and on the number of bins. For very large datasets, different histograms will not look too different but for smaller datasets, the pictures provided by histograms can look very different. Below is a simple dataset of 48 observations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

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

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

Question: How different can histograms of these data look?

Below are four histograms of these data.



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

ISBN 978-1-107-65599-7

Exercise 1B

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

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

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

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

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

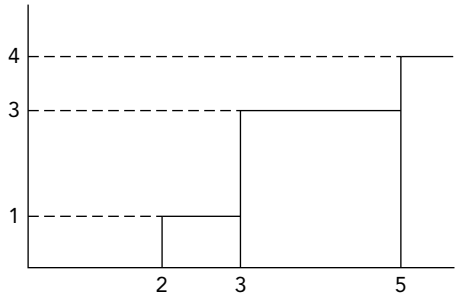
28 24 32 30 34 30 36 32 48 32 32 38 32 40
28 48 34 28 40 18 20 26 36 40 20 16 38 20
34 30

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

ISBN 978-1-107-65599-7

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

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



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

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

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

Enrichment

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

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

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

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

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

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

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

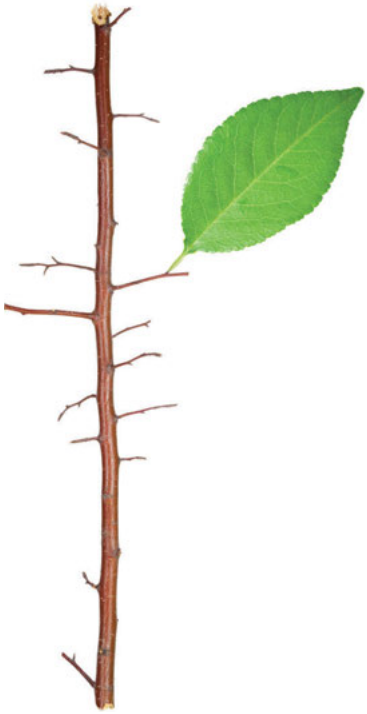
ISBN 978-1-107-65599-7

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

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

ISBN 978-1-107-65599-7

1-3 Comparing plots



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

The three types of plots

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

Same scale

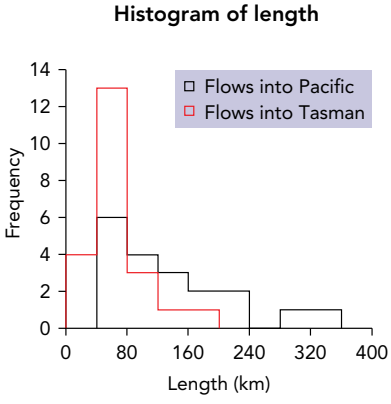
To compare across groups using any of these plots, the same scale *must* be used. Dotplots are readily placed above each other for comparisons, but stem-and-leaf plots need to be placed beside each other. Back-to-back stem-and-leaf plots can be used to compare two groups (e.g. males and females) but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

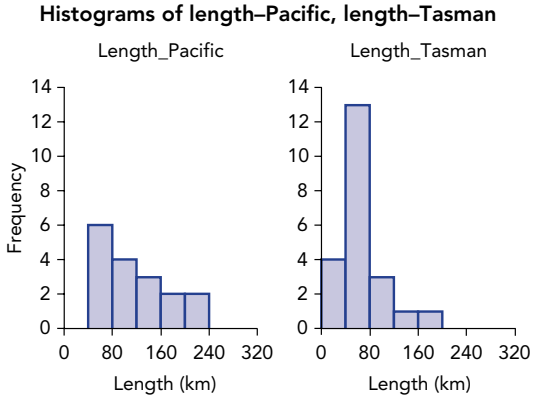
LET'S START Combining histograms

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

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

ISBN 978-1-107-65599-7

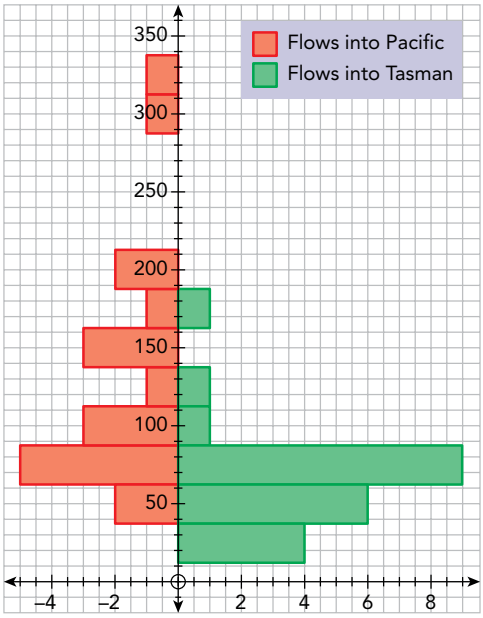




Histograms can also be placed side by side like back-to-back stem-and-leaf plots.

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

Which plot do you think shows the data best?



ISBN 978-1-107-65599-7

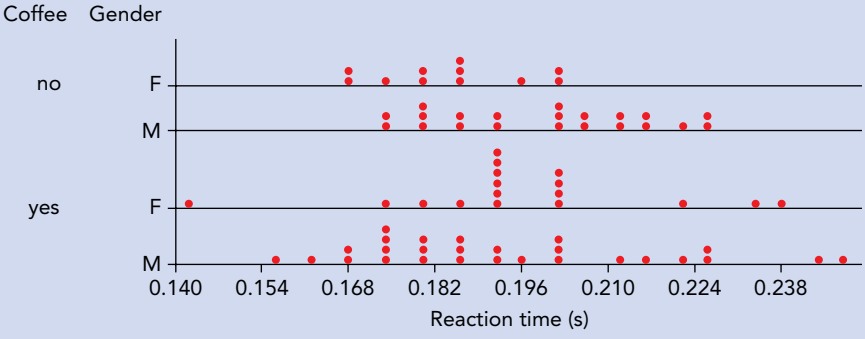
Key ideas

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

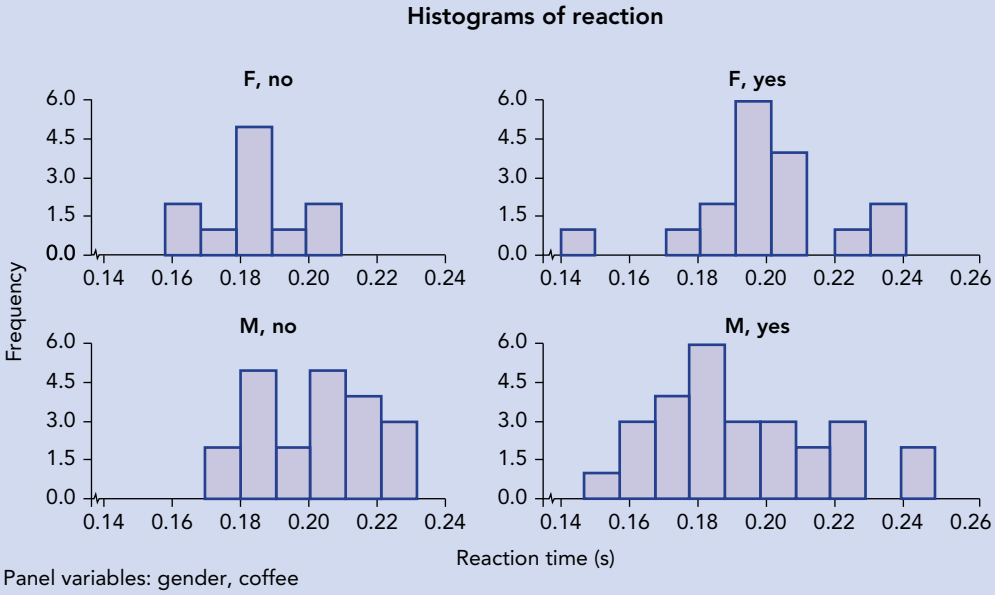
Example 4: Do males react faster than females? Does coffee affect reactions?

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

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



ISBN 978-1-107-65599-7



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

Exercise 1C

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

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

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

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

ISBN 978-1-107-65599-7

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

Speeds amber lights

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

Speeds green lights

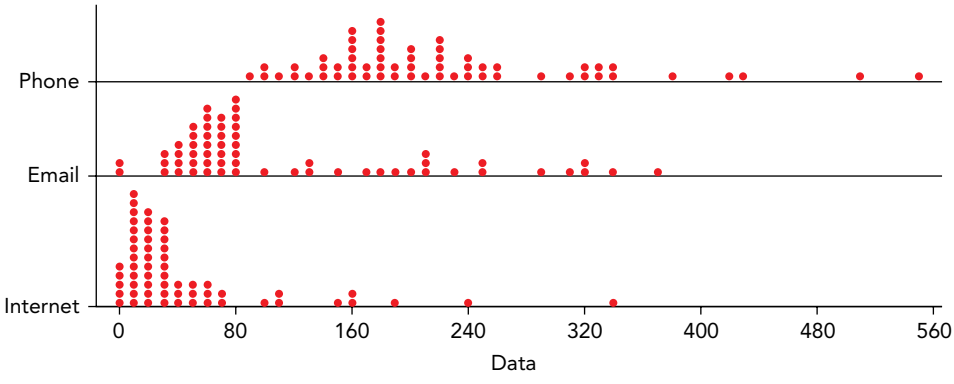
34.8	51.1	43.9	46.2	43.0	48.0	37.5	38.8	38.7	49.3	42.4	49.6
51.4	46.2	44.8	54.5	39.8	54.4	38.0	33.9	41.7	40.8	44.0	41.7
40.0	46.2	48.5	58.1	48.4	51.1	39.0	37.7	46.0	44.7	88.7	46.5
47.7	48.3	43.5	31.6	43.6	40.4	32.7	40.0	26.4	37.9	56.6	49.9
53.6	34.0										

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

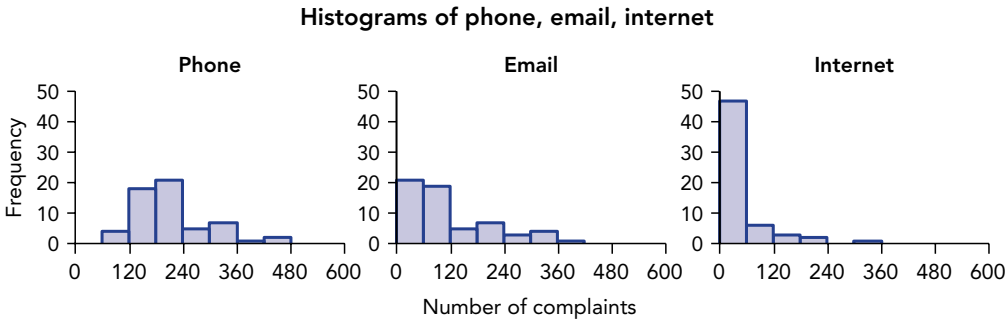
b The speed limit on this street was 60 km/h. Comment on the speeds.

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

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



ISBN 978-1-107-65599-7



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

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

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

Enrichment

How well do people estimate time?

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

ISBN 978-1-107-65599-7

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

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

i 5 s **ii** 10 s.

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

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

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

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

ISBN 978-1-107-65599-7

Chapter summary

Investigating quantitative data across groups

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

Histograms

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

Comparing plots

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

Multiple-choice questions

Questions 1–7 refer to the following situation.

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

1 Bar charts or column graphs can be used to plot

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

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

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

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

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

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

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

ISBN 978-1-107-65599-7

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

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

6 The rectangles of the histograms in question 5 above

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

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

A 2
B 3
C 5
D 6

Short-answer questions

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

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

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

a What would need care in carrying out this experiment?

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

c What should we plot to investigate the research question?

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

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

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

ISBN 978-1-107-65599-7

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

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

c What do the plots indicate?

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

No school holidays

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

Includes school holidays

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

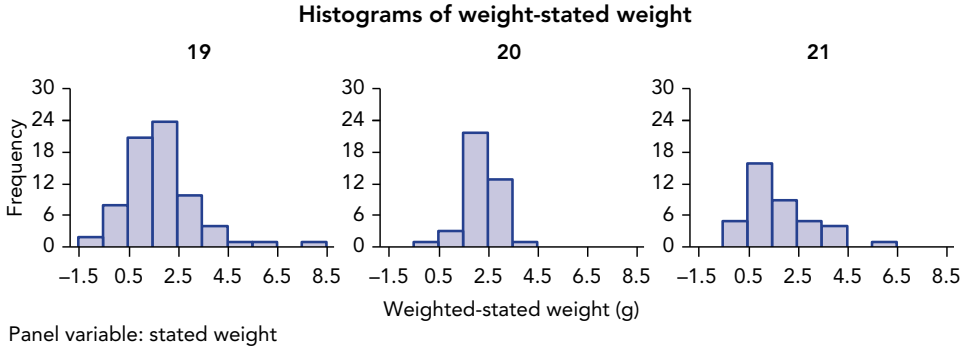
a Draw a back-to-back stem-and-leaf plot to compare the numbers of monthly phone complaints in months that do, or do not, include school holidays.

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

c What do the plots indicate?

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

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



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

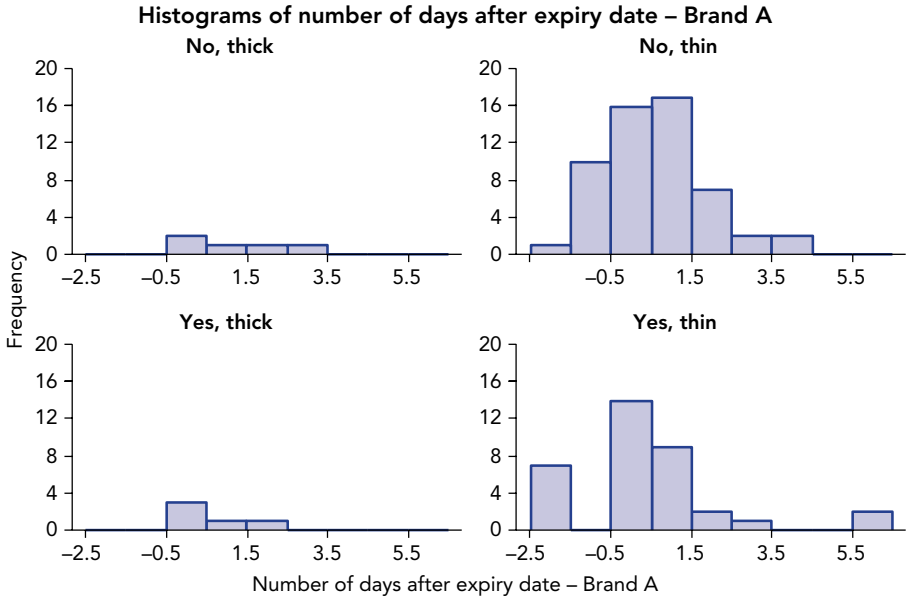
b Comment on the plots.

ISBN 978-1-107-65599-7

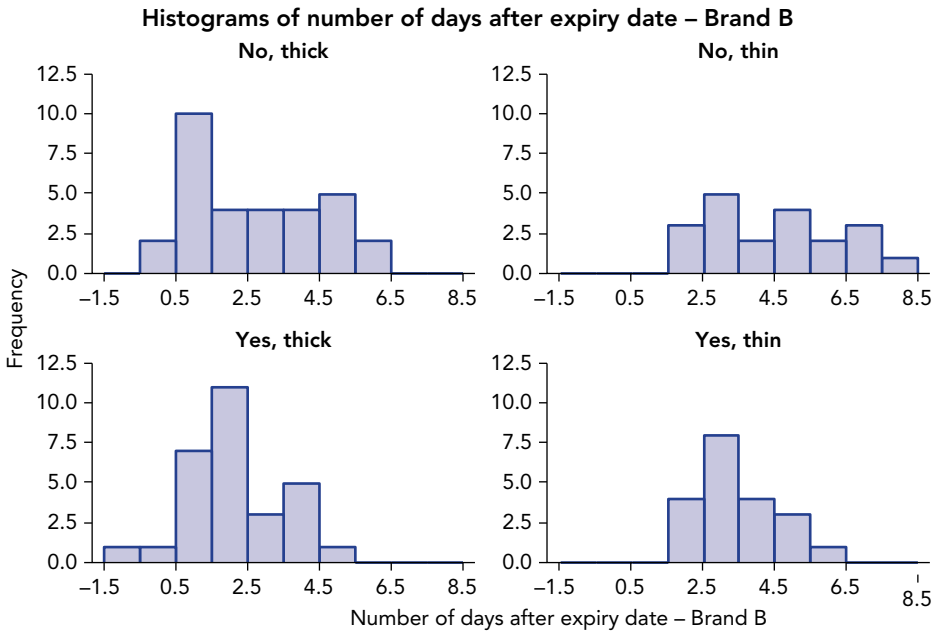
Extended-response question

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

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



ISBN 978-1-107-65599-7



Panel variables: Sunlight_Brand B, thick or thin_Brand B

a What do the negative values mean?

b Comment on the plots for brand A.

c Comment on the plots for brand B.

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

e Compare brands A and B.

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

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

ISBN 978-1-107-65599-7