



# Chapter 5

## Statistics

### Essential mathematics: why skills with statistics are important

Vast quantities of data are collected from online records and surveys. To find meaning, statistics such as typical measures and graphs are used to show trends and allow comparisons.

Statistical calculations and interpretations provide essential information for decision making by governments, farmers, research scientists, technicians, and business and finance personnel.

- The Australian Government uses statistics to compare the health of Indigenous and non-Indigenous Australians to improve policy; for example, time-series graphs of average lifespan versus year.
- Farmers use digital sensors to record data which is then statistically analysed. For example, crop and soil moisture levels are used to automate irrigation and improve crop production.
- Stem-and-leaf plots can be used to record fuel economies and tailpipe emissions of carbon dioxide for various vehicle models.
- Medical researchers use parallel box plots that show significantly lower birth weights for babies with mothers who smoke compared to non-smoking mothers. Low birth weights can cause health problems.



## In this chapter

- 5A Collecting data
- 5B Frequency tables, column graphs and histograms (**Consolidating**)
- 5C Dot plots and stem-and-leaf plots (**Consolidating**)
- 5D Range and measures of centre
- 5E Quartiles and outliers
- 5F Box plots
- 5G Time-series data
- 5H Bivariate data and scatter plots
- 5I Line of best fit by eye ★

## Victorian Curriculum

### STATISTICS AND PROBABILITY

#### Data representation and interpretation

Determine quartiles and interquartile range and investigate the effect of individual data values, including outliers on the interquartile range (VCMSP349)

Construct and interpret box plots and use them to compare data sets (VCMSP350)

Compare shapes of box plots to corresponding histograms and dot plots and discuss the distribution of data (VCMSP351)

Use scatter plots to investigate and comment on relationships between two numerical variables (VCMSP352)

Investigate and describe bivariate numerical data, including where the independent variable is time (VCMSP353)

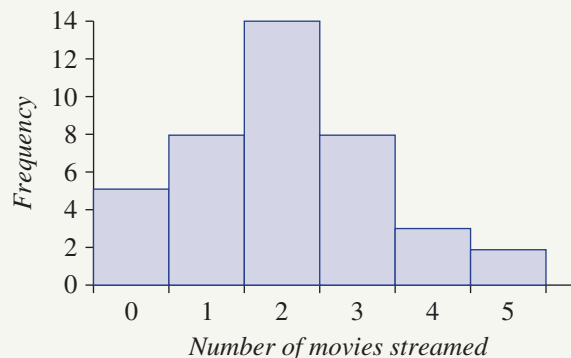
Evaluate statistical reports in the media and other places by linking claims to displays, statistics and representative data (VCMSP354)

© Victorian Curriculum and Assessment Authority (VCAA)

## Online resources

A host of additional online resources are included as part of your Interactive Textbook, including HOTmaths content, video demonstrations of all worked examples, auto-marked quizzes and much more.

- 1 The number of movies streamed in a month online by a number of people is shown in this graph.
- How many people streamed three movies?
  - How many people were surveyed?
  - How many movies were streamed during the survey?
  - How many people streamed fewer than two movies?



- 2 This table shows the frequency of scores in a test.

| Score  | Frequency |
|--------|-----------|
| 0–     | 2         |
| 20–    | 3         |
| 40–    | 6         |
| 60–    | 12        |
| 80–100 | 7         |

- How many scores were in the 40 to less than 60 range?
- How many scores were:
  - at least 60?
  - less than 80?
- How many scores were there in total?
- What percentage of scores were in the 20 to less than 40 range?

- 3 Calculate:

a  $\frac{6+10}{2}$

b  $\frac{8+9}{2}$

c  $\frac{2+4+5+9}{4}$

d  $\frac{3+5+8+10+14}{5}$



- 4 For each of these data sets, find:
- the mean (i.e. average)
  - the mode (i.e. most frequent)
  - the median (i.e. middle value of ordered data)
  - the range (i.e. difference between highest and lowest)
- 38, 41, 41, 47, 58
  - 2, 2, 2, 4, 6, 6, 7, 9, 10, 12

- 5 This stem-and-leaf plot shows the weight, in grams, of some small calculators.

- How many calculators are represented in the plot?
- What is the mode (i.e. most frequent)?
- What is the minimum calculator weight and maximum weight?
- Find the range (i.e. maximum value – minimum value).

| Stem | Leaf              |
|------|-------------------|
| 9    | 8                 |
| 10   | 2 6               |
| 11   | 1 1 4 9           |
| 12   | 3 6               |
| 13   | 8 9 9             |
| 14   | 0 2 5             |
| 13   | 6 means 136 grams |

## 5A Collecting data

### Learning intentions

- To understand what is required to construct a good survey
- To know the difference between a population and a sample
- To be able to categorise types of statistical data

**Key vocabulary:** survey, statistical data, categorical data, numerical data, nominal data, ordinal data, discrete data, continuous data, population, sample, census

A statistician is a person who is employed to design surveys. They also collect, analyse and interpret data. They assist the government, companies and other organisations to make decisions and plan for the future.

Statisticians:

- **Formulate** and **refine** questions for a survey.
- Choose some **subjects** (i.e. people) to complete the survey.
- **Collect** the data.
- **Organise and display** the data using the most appropriate graphs and tables.
- **Analyse** the data.
- **Interpret the data** and draw conclusions.

There are many reports in the media that begin with the words 'A recent study has found that...'. These are usually the result of a survey or investigation that a researcher has conducted to collect information about an important issue, such as unemployment, crime or obesity.

### → Lesson starter: Critiquing survey questions

Here is a short survey. It is not very well constructed.

Question 1: How old are you?

Question 2: How much time did you spend sitting in front of the television or a computer yesterday?

Question 3: Some people say that teenagers like you are lazy and spend too much time sitting around when you should be outside exercising. What do you think of that comment?

Have a class discussion about the following.

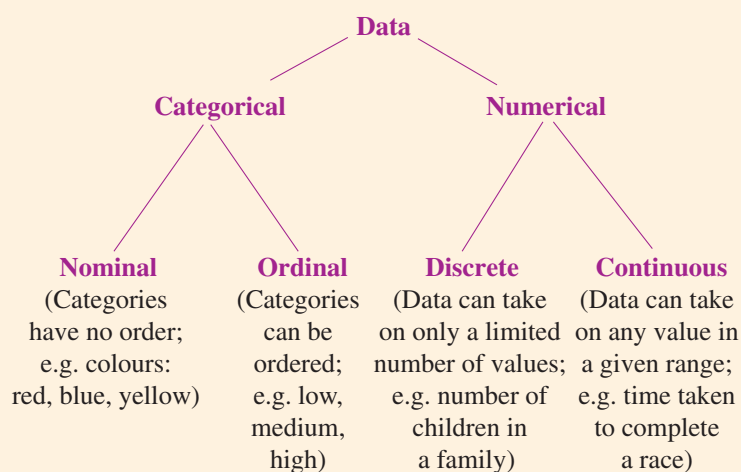
- What will the answers to Question 1 look like? How could they be displayed?
- What will the answers to Question 2 look like? How could they be displayed?
- Is Question 2 going to give a realistic picture of your normal daily activity?
- How could Question 2 be improved?
- What will the answers to Question 3 look like? How could they be displayed?
- How could Question 3 be improved?

### Key ideas

- **Surveys** are used to collect statistical data by asking randomly selected people questions.
  - Survey questions need to be constructed carefully so that the person knows exactly what sort of answer to give. They should use simple language and should not be ambiguous.
  - Survey questions should not be worded so that they deliberately try to provoke a certain kind of response.
  - If the question contains an option to be chosen from a list, the number of options should be an odd number, so that there is a 'neutral' choice. For example, the options could be:

|                   |          |        |       |                |
|-------------------|----------|--------|-------|----------------|
| strongly disagree | disagree | unsure | agree | strongly agree |
|-------------------|----------|--------|-------|----------------|

- A **population** is a group of people, animals or objects with something in common. Some examples of populations are:
  - all the people in Australia on Census night (a **census** is a set of statistics collected from the entire population)
  - all the students in your school
  - all the boys in your Maths class
  - all the tigers in the wild in Sumatra
  - all the cars in Sydney
  - all the wheat farms in NSW
- A **sample** is a group that has been chosen from a population. Sometimes information from a sample is used to describe the whole population, so it is important to choose the sample carefully.
- **Statistical data** is information gathered by observation, survey or measurement. It can be categorised as follows.



## Exercise 5A

### Understanding

1–3

2, 3

- 1 What are some of the considerations needed when constructing a survey?
- 2 Match each word (a–f) with its definition (A–F).
 

|   |   |
|---|---|
| <p><b>a</b> Population</p> <p><b>b</b> Census</p> <p><b>c</b> Sample</p> <p><b>d</b> Survey</p> <p><b>e</b> Data</p> <p><b>f</b> Statistics</p> | <p><b>A</b> A group chosen from a population</p> <p><b>B</b> A tool used to collect statistical data</p> <p><b>C</b> All the people or objects in question</p> <p><b>D</b> Statistics collected from an entire population</p> <p><b>E</b> The practice of collecting and analysing data</p> <p><b>F</b> The factual information collected from a survey or other source</p> |
|---|---|
- 3 Match each word (a–f) with its definition (A–F).
 

|   |   |
|---|---|
| <p><b>a</b> Numerical</p> <p><b>b</b> Continuous</p> <p><b>c</b> Discrete</p> <p><b>d</b> Categorical</p> <p><b>e</b> Ordinal</p> <p><b>f</b> Nominal</p> | <p><b>A</b> Categorical data that has no order</p> <p><b>B</b> Data that are numbers</p> <p><b>C</b> Numerical data that take on a limited number of values</p> <p><b>D</b> Data that can be divided into categories</p> <p><b>E</b> Numerical data that take any value in a given range</p> <p><b>F</b> Categorical data that can be ordered</p> |
|---|---|

## Fluency

4–6

4, 5, 7



## Example 1 Describing types of data

What type of data would the following survey questions generate?

- a How many televisions do you have in your home?
- b To what type of music do you most like to listen?

**Solution**

a Numerical and discrete

**Explanation**

The answer to the question is a number with a limited number of values; in this case, a whole number.

b Categorical and nominal

The answer is a type of music and these categories have no order.

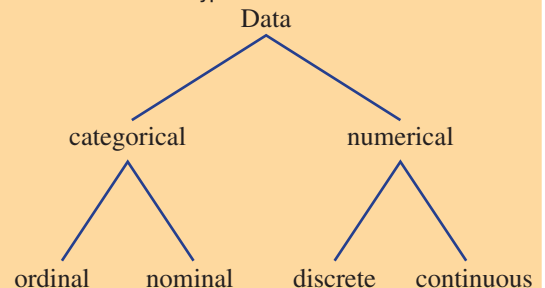
**Now you try**

What type of data would the following survey questions generate?

- a How would you rate your hotel stay (poor, satisfactory, good, excellent)?
- b How many times in a week do you exercise?

- 4 Which one of the following survey questions would generate numerical data?
- A What is your favourite colour?
  - B What type of car does your family own?
  - C How long does it take for you to travel to school?
  - D What type of dog do you own?
- 5 Which one of the following survey questions would generate categorical data?
- A How many times do you eat at your favourite fast-food place in a typical week?
  - B How much money do you usually spend buying your favourite fast food?
  - C How many items did you buy last time you went to your favourite fast-food place?
  - D Which is your favourite fast food?
- 6 Year 10 students were asked the following questions in a survey. Describe what type of data each question generates.
- a How many people under the age of 18 years are there in your immediate family?
  - b How many letters are there in your first name?
  - c Which company is the carrier of your mobile telephone calls?  
Optus/Telstra/Vodafone/Boost/Other (Please specify.)
  - d What is your height?
  - e How would you describe your level of application in Maths? (Choose from *very high*, *high*, *medium* or *low*.)

Hint: Recall the four types of data:



## 5A

- 7 Every student in Years 7 to 12 votes in the prefect elections. The election process is an example of:
- A a population
  - B continuous data
  - C a representative sample
  - D a census

## Problem-solving and reasoning

8, 9

9–11

- 8 A popular Australian 'current affairs' television show recently investigated the issue of spelling. They suspected that people in their twenties are not as good at spelling as people in their fifties, so they decided to conduct a statistical investigation. They chose a sample of 12 people aged 50–59 years and 12 people aged 20–29 years.
- Answer the following questions on paper, then discuss in a small group or as a whole class.
- a Do you think that the number of people surveyed is enough?
  - b Do you think it is fair and reasonable to compare the spelling ability of these two groups of people?
  - c How would you go about comparing the spelling ability of these two groups of people?
  - d Would you give the two groups the same set of words to spell?
  - e How could you give the younger people an unfair advantage?
  - f What sorts of words would you include in a spelling test for the survey?
  - g How and where would you choose the people to do the spelling test?
- 9 The principal decides to survey Year 10 students to determine their opinion of Mathematics. In order to increase the chance of choosing a representative sample, the principal should:
- A Give a survey form to the first 30 Year 10 students who arrive at school.
  - B Give a survey form to all the students studying the most advanced maths subject.
  - C Give a survey form to five students in every Maths class.
  - D Give a survey form to 20% of the students in every class.
- 10 Discuss some of the problems with the selection of a survey sample for each given topic.
- a A survey at the train station of how Australians get to work.
  - b An email survey on people's use of computers.
  - c Phoning people on the electoral roll to determine Australia's favourite sport.
- 11 Choose a topic in which you are especially interested, such as football, cricket, movies, music, cooking, food, computer games or social media.
- Make up a survey about your topic that you could give to the other students in your class. It must have *four* questions.
- Question 1 must produce data that are categorical and ordinal.
- Question 2 must produce data that are categorical and nominal.
- Question 3 must produce data that are numerical and discrete.
- Question 4 must produce data that are numerical and continuous.



## The Australian Census

—

12

- 12 The Australian Census is conducted by the Australian Bureau of Statistics every five years. Research either the 2011 or 2016 Australian Census on the website of the Australian Bureau of Statistics. Find out something interesting from the results of the Census.
- Write a short news report or record a 3-minute news report on your computer.

## 5B Frequency tables, column graphs and histograms

**CONSOLIDATING**

### Learning intentions

- To be able to construct a frequency table from a set of data
- To know the difference between a column graph and histogram when choosing to represent data from a frequency table
- To be able to construct and analyse column graphs and histograms
- To be able to describe data as symmetrical or skewed

**Key vocabulary:** frequency table, column graph, histogram, symmetrical, skewed, data

As a simple list, data can be difficult to interpret. Sorting the data into a frequency table allows us to make sense of it and draw conclusions from it.

Statistical graphs are an essential part of the analysis and representation of data. By looking at statistical graphs, we can draw conclusions about the numbers or categories in the data set.



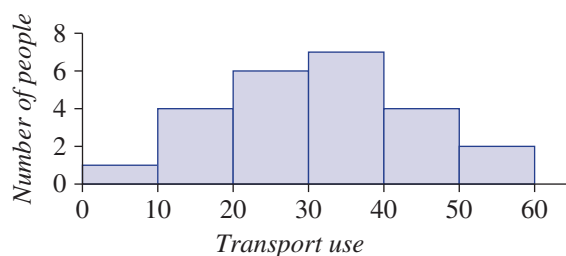
### → Lesson starter: Public transport analysis

A survey was carried out to find out how many times people in a particular group had used public transport in the past month. The results are shown in this histogram.

Discuss what the histogram tells you about this group of people and their use of public transport.

You may wish to include these points:

- How many people were surveyed?
- Is the data symmetrical or skewed?
- Is it possible to work out all the individual data values from this graph?
- Do you think these people were selected from a group in your own community? Give reasons.



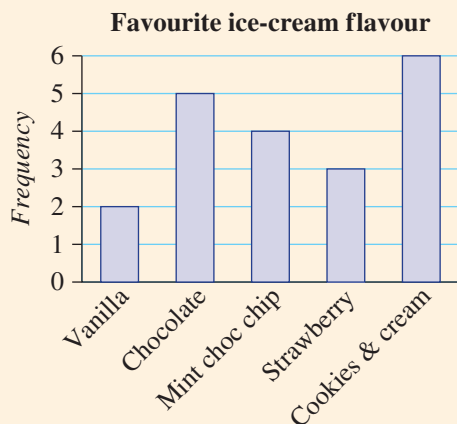
### Key ideas

- A **frequency table** displays data by showing the number of values within a set of categories or class intervals. It may include a tally column to help count the data.

| Favourite ice-cream flavour | Tally | Frequency |
|-----------------------------|-------|-----------|
| Vanilla                     |       | 2         |
| Chocolate                   |       | 5         |
| Mint choc chip              |       | 4         |
| Strawberry                  |       | 3         |
| Cookies and cream           |       | 6         |



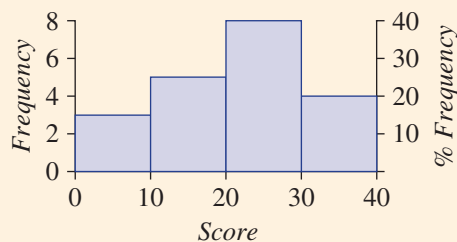
- A **column graph** can be used for a single set of categorical or discrete data to show the frequency.



- **Histograms** can be used for grouped discrete or continuous numerical data. The frequency of particular class intervals is recorded.

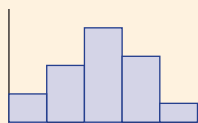
- The interval 10– (in the table below) includes all numbers from 10 (including 10) to less than 20.
- The percentage frequency is calculated as  $\% \text{Frequency} = \frac{\text{frequency}}{\text{total}} \times 100$ .

| Class interval | Frequency | Percentage frequency             |
|----------------|-----------|----------------------------------|
| 0–             | 3         | $\frac{3}{20} \times 100 = 15\%$ |
| 10–            | 5         | $\frac{5}{20} \times 100 = 25\%$ |
| 20–            | 8         | 40%                              |
| 30–40          | 4         | 20%                              |
| <b>Total</b>   | 20        | 100%                             |

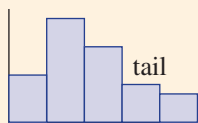


- Data can be **symmetrical** (same shape either side of the middle) or **skewed** (data weighted to the left or the right).

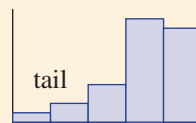
Symmetrical



Positively skewed



Negatively skewed



## Exercise 5B

### Understanding

1–3

3

- Decide whether a histogram or column graph would be best used to display the following types of data
  - heights of students in a class
  - favourite colour
  - rating of hotel service (low, medium, high)
  - hourly wage at a number of restaurants

2 Complete these frequency tables.

a

| Car colour   | Tally | Frequency |
|--------------|-------|-----------|
| Red          |       |           |
| White        |       |           |
| Green        |       |           |
| Silver       |       |           |
| <b>Total</b> |       |           |

Hint: In the tally, |||| is 5.



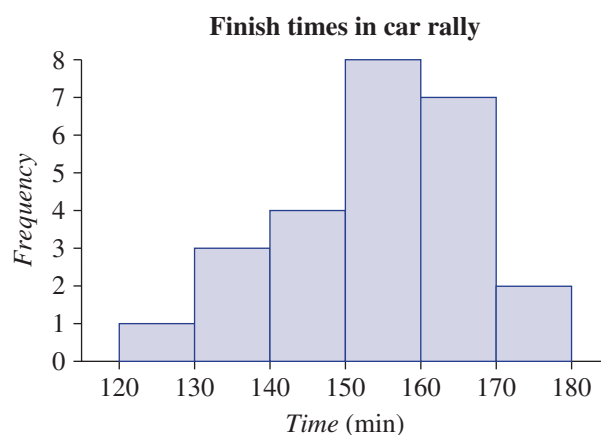
b

| Class interval | Frequency | Percentage frequency             |
|----------------|-----------|----------------------------------|
| 80–            | 8         | $\frac{8}{50} \times 100 = 16\%$ |
| 85–            | 23        |                                  |
| 90–            | 13        |                                  |
| 95–100         |           |                                  |
| <b>Total</b>   | <b>50</b> |                                  |

3 This frequency histogram shows how many competitors in a car rally race finished within a given time interval.

- a How many competitors finished in a time between 150 and 160 minutes?  
 b How many cars were there in the race?

Hint: There was one competitor in the 120–130 interval, three competitors in the 130–140 interval etc.



c Determine the following.

- i How many competitors finished in fewer than 150 minutes?  
 ii What percentage of competitors finished in fewer than 150 minutes?

Hint:  
 Percentage =  $\frac{\text{number} < 150}{\text{total}} \times 100$



## Fluency

4–6

4, 5, 7



### Example 2 Constructing a frequency table and column graph

Twenty people checking out at a hotel were surveyed on the level of service provided by the hotel staff. The results were:

|         |             |             |             |             |
|---------|-------------|-------------|-------------|-------------|
| Poor    | First class | Poor        | Average     | Good        |
| Good    | Average     | Good        | First class | First class |
| Good    | Good        | First class | Good        | Average     |
| Average | Good        | Poor        | First class | Good        |

- a Construct a frequency table to record the data, with headings Category, Tally and Frequency.  
 b Construct a column graph for the data.

*Continued on next page*

## Solution

**a**

| Category    | Tally | Frequency |
|-------------|-------|-----------|
| Poor        |       | 3         |
| Average     |       | 4         |
| Good        | <br>  | 8         |
| First class | <br>  | 5         |
| Total       | 20    | 20        |

## Explanation

Construct a table with the headings Category, Tally, Frequency.

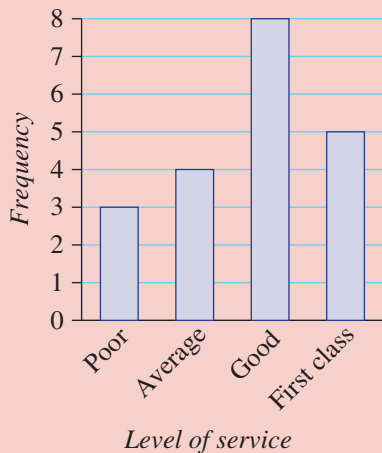
Fill in each category shown in the data. Work through the data in order, recording a tally mark (|) next to the category. It is a good idea to tick the data as you go, to keep track.

On the 5th occurrence of a category, place a diagonal line through the tally marks (||||). Then start again on the 6th. Do this every five values, as it makes the tally marks easy to count up.

Once all data is recorded, count the tally marks for the frequency.

Check that the frequency total adds up to the number of people surveyed (20).

**b** **Hotel service satisfaction**



Draw a set of axes with frequency going up to 8.

For each category, draw a column with height up to its frequency value.

Leave gaps between each column.

Give your graph an appropriate heading.

## Now you try

A class of 24 students was surveyed on their favourite genre of movie. The results were:

|        |         |         |         |        |         |
|--------|---------|---------|---------|--------|---------|
| Action | Comedy  | Comedy  | Romance | Action | Sci-Fi  |
| Horror | Sci-Fi  | Comedy  | Action  | Comedy | Romance |
| Comedy | Action  | Romance | Horror  | Action | Comedy  |
| Action | Romance | Horror  | Action  | Comedy | Action  |

**a** Construct a frequency table to record the data, with headings Category, Tally and Frequency.

**b** Construct a column graph for the data.

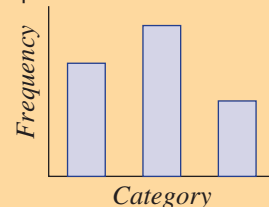
**4** For the data on the next page, which was obtained from surveys:

**i** Copy and complete this frequency table.

| Category | Tally | Frequency |
|----------|-------|-----------|
|          |       |           |
|          |       |           |
|          |       |           |
| ⋮        | ⋮     | ⋮         |

**ii** Construct a column graph for the data and include a heading.

Hint: In the column graph, leave spaces between each column.



- a The results from 10 subjects on a student's school report showing their level of application are:  
 Good      Low      Good      Good      Excellent  
 Very Low      Low      Good      Good      Low
- b The favourite sports of a class of students are:  
 Football      Tennis      Basketball      Tennis      Football  
 Netball      Football      Tennis      Football      Basketball  
 Basketball      Tennis      Netball      Football      Football  
 Football      Basketball      Football      Netball      Tennis



### Example 3 Constructing and analysing a histogram

Twenty people were surveyed to find out how many times they use the internet in a week. The raw data are listed.

21, 19, 5, 10, 15, 18, 31, 40, 32, 25  
 11, 28, 31, 29, 16, 2, 13, 33, 14, 24

- a Organise the data into a frequency table, using class intervals of 10. Include a percentage frequency column.
- b Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- c Which interval is the most frequent?
- d What percentage of people used the internet 20 times or more?

#### Solution

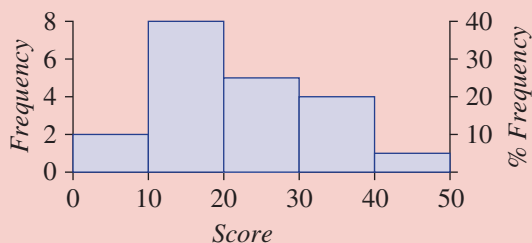
| Class interval | Tally | Frequency | Percentage frequency |
|----------------|-------|-----------|----------------------|
| 0–             |       | 2         | 10%                  |
| 10–            |       | 8         | 40%                  |
| 20–            |       | 5         | 25%                  |
| 30–            |       | 4         | 20%                  |
| 40–49          |       | 1         | 5%                   |
| Total          | 20    | 20        | 100%                 |

#### Explanation

The interval 10– includes all numbers from 10 (including 10) to less than 20, so 10 is in this interval but 20 is not.

Count the tally marks to record the frequency. Add the frequency column to ensure all 20 values have been recorded. Calculate each percentage frequency by dividing the frequency by the total (20) and multiplying by 100%; i.e.  $\frac{2}{20} \times 100 = 10$ .

#### b Number of times the internet is accessed



Transfer the data from the frequency table to the histogram. Axes scales are evenly spaced and the histogram bar is placed across the boundaries of the class interval. There is no space between the bars.

- c The 10– interval is the most frequent.
- d 50% of those surveyed used the internet 20 or more times.

The frequency (8) is highest for this interval. It is the highest bar on the histogram.

Sum the percentages for the class intervals from 20– and above.  
 $25 + 20 + 5 = 50$

*Continued on next page*

## 5B

## Now you try

The number of points scored by a basketball team in its 25-game season are shown below:

74 82 77 101 91  
 66 86 87 90 88  
 79 108 94 89 70  
 75 81 72 89 97  
 86 78 78 82 88

- Organise the data into a frequency table, using class intervals of 10. Include a percentage frequency column.
- Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- Which interval is the most frequent?
- What percentage of their games did they score 80 or more?

- 5 The Maths test results of a class of 25 students were recorded as:

74 65 54 77 85 68 93 59 75  
 71 82 57 98 73 66 88 76  
 92 70 77 65 68 81 79 80

- Organise the data into a frequency table, using class intervals of 10. Include a percentage frequency column.
- Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- Which interval is the most frequent?
- If an A is awarded for a mark of 80 or more, what percentage of the class received an A?

Hint: Construct a frequency table like this:

| Class interval | Tally | Frequency | Percentage frequency                           |
|----------------|-------|-----------|--|
| 50–            |       | 3         | $\frac{\text{freq.}}{\text{total}} \times 100$ |
| 60–            |       |           |  |
| 70–            |       |           |  |
| 80–            |       |           |  |
| 90–100         |       |           |  |
| Total          |       |           |  |



- 6 The number of wins scored this season is given for 20 hockey teams. Here are the raw data.  
 4, 8, 5, 12, 15, 9, 9, 7, 3, 7  
 10, 11, 1, 9, 13, 0, 6, 4, 12, 5

- Organise the data into a frequency table, using class intervals of 5. Start with 0–, then 5– etc. and include a percentage frequency column.
- Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- Which interval is the most frequent?
- What percentage of teams scored 5 or more wins?



- 7 This frequency table displays the way in which 40 people travel to and from work.

| Type of transport | Frequency | Percentage frequency |
|-------------------|-----------|----------------------|
| Car               | 16        |                      |
| Train             | 6         |                      |
| Tram              | 8         |                      |
| Walking           | 5         |                      |
| Bicycle           | 2         |                      |
| Bus               | 3         |                      |
| <b>Total</b>      | <b>40</b> |                      |



- a Copy and complete the table.  
 b Use the table to find the:  
 i frequency of people who travel by train  
 ii most popular form of transport  
 iii percentage of people who travel by car  
 iv percentage of people who walk or cycle to work  
 v percentage of people who travel by public transport, including trains, buses and trams

Hint: Percentage frequency:

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

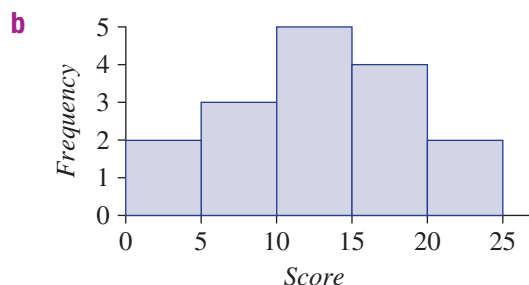
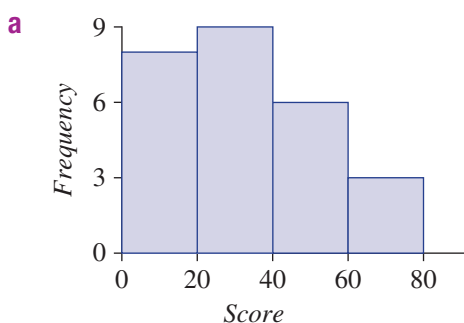


## Problem-solving and reasoning

8, 9

8, 10, 11

- 8 Which of these histograms shows a symmetrical data set and which one shows a skewed data set?



- 9 This tally records the number of mice that were weighed and categorised into particular mass intervals for a scientific experiment.

- a Construct a table using these column headings: Mass, Frequency and Percentage frequency.  
 b Find the total number of mice weighed in the experiment.  
 c State the percentage of mice that were in the 20– gram interval.  
 d Which was the most common weight interval?  
 e What percentage of mice were in the most common mass interval?  
 f What percentage of mice had a mass of 15 grams or more?

| Mass (grams) | Tally |
|--------------|-------|
| 10–          |       |
| 15–          |       |
| 20–          |       |
| 25–          |       |
| 30–34        |       |

- 10 A school orchestra contains four musical sections: string, woodwind, brass and percussion. The number of students playing in each section is summarised in this tally.

- a Construct and complete a percentage frequency table for the data.  
 b What is the total number of students in the school orchestra?  
 c What percentage of students play in the string section?  
 d What percentage of students do not play in the string section?  
 e If the number of students in the string section increased by three, what would be the percentage of students who play in the percussion section? (Round your answer to one decimal place.)

| Section    | Tally |
|------------|-------|
| String     |       |
| Woodwind   |       |
| Brass      |       |
| Percussion |       |

- 11 Describe the information that is lost when displaying data using a histogram.

## 5B



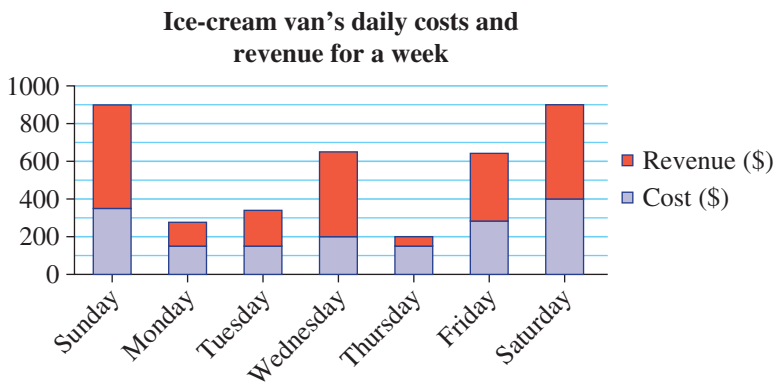
## Interpreting other graphical displays

12, 13

- 12 The graph shown compares the life expectancy of males and females in 10 different countries. Use the graph to answer the following questions.



- Which country has the biggest difference in life expectancy for males and females? Approximately how many years is this difference?
  - Which country appears to have the smallest difference in life expectancy between males and females?
  - From the information in the graph, write a statement comparing the life expectancy of males and females.
  - South Africa is clearly below the other countries. Provide some reasons why you think this may be the case.
- 13 This graph shows the amount spent (Cost) on the purchase and storage of ice-cream each day by an ice-cream vendor, and the amount of money made from the daily sales of ice-cream (Revenue) over the course of a week.



- On which days was the cost highest for the purchase and storage of ice-cream? Why do you think the vendor chose these days to spend the most?
- Wednesday had the greatest revenue for any weekday. What factors may have led to this?
- Daily profit is determined by the difference in revenue and cost. Identify:
  - on which day the largest profit was made and state this profit (in dollars)
  - on which day the vendor suffered the biggest financial loss
- Describe some problems associated with this type of graph.

### Using technology 5B: Using calculators to graph grouped data

This activity is available on the companion website as a printable PDF.

# 5C Dot plots and stem-and-leaf plots

## CONSOLIDATING

### Learning intentions

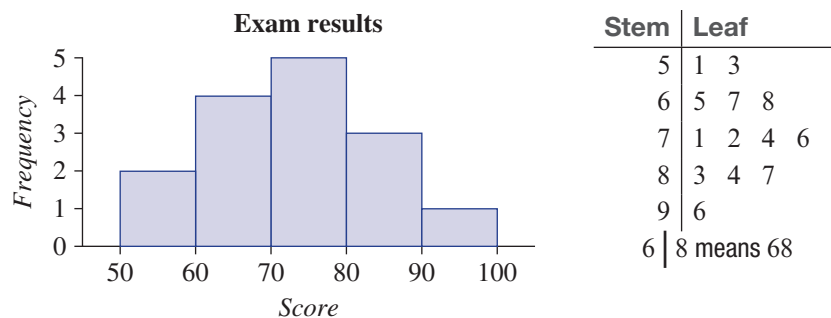
- To be able to construct and interpret a dot plot
- To be able to construct and interpret a stem-and-leaf plot and back-to-back stem-and-leaf-plots
- To know when it is appropriate to use a dot plot or stem-and-leaf plot to represent a set of data
- To be able to interpret the shape of these graphs to describe the distribution of the data as symmetrical or skewed

**Key vocabulary:** dot plot, stem-and-leaf plot, symmetrical data, skewed data

In addition to column graphs, dot plots and stem-and-leaf plots can be used to display categorical or discrete data. They can also display two related sets for comparison. Like a histogram, they help to show how the data are distributed. A stem-and-leaf plot has the advantage of still displaying all the individual data items.

### → Lesson starter: Alternative representations

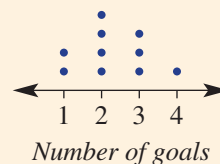
The histogram and stem-and-leaf plot below represent the same set of data. They show the scores achieved by a class in an exam.



- Describe the similarities in what the two graphs display.
- What does the stem-and-leaf provide that the histogram does not? What is the advantage of this?
- Which graph do you prefer?
- Discuss any other types of graphs that could be used to present the data.

### Key ideas

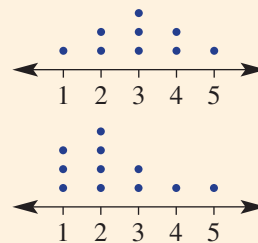
- A **dot plot** records the frequency of each category or discrete value in a data set.
  - Each occurrence of the value is marked with a dot.
- A **stem-and-leaf plot** displays each value in the data set using a stem number and a leaf number.
  - The data are displayed in two parts: a stem and a leaf.
  - The 'key' tells you how to interpret the stem and leaf parts.
  - The graph is similar to a histogram with class intervals, but the original data values are not lost.



| Stem  | Leaf     |
|-------|----------|
| 1     | 0 1 1 5  |
| 2     | 3 7      |
| 3     | 4 4 6    |
| 4     | 2 9      |
| 2   3 | means 23 |
| key   |          |



- The stem-and-leaf plot is ordered to allow for further statistical calculations.
- Back-to-back stem and leaf plots, with leaves either side of the stem, can be used to compare two related sets of data.
- The shape of each of these graphs gives information about the distribution of the data.
  - A graph that is even either side of the centre is symmetrical.
  - A graph that is bunched to one side of the centre is skewed.



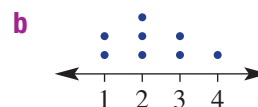
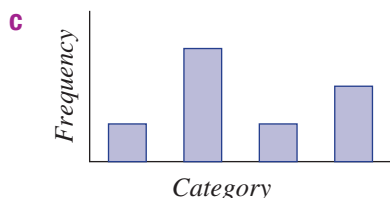
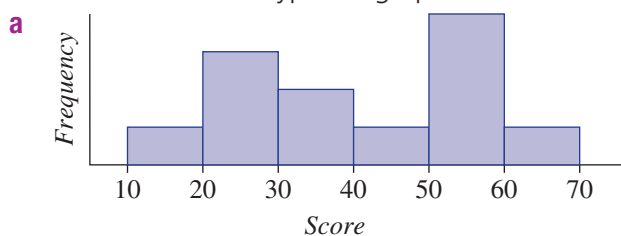
## Exercise 5C

### Understanding

1-3

3

- 1 Name each of these types of graphs.



**d**

| Stem | Leaf      |
|------|-----------|
| 0    | 1 1 3     |
| 1    | 2 4 7     |
| 2    | 0 2 2 5 8 |
| 3    | 1 3       |

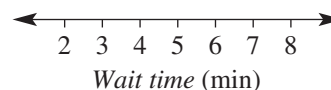
2 | 5 means 25

- 2 A student records the following wait times, in minutes, for his school bus over 4 school weeks.

5 4 2 8 4 2 7 5 3 3

5 4 2 5 4 5 8 7 2 6

Copy and complete this dot plot of the data by placing a dot for each occurrence of a value.



- 3 List all the data shown in these stem-and-leaf plots (e.g. 32, 35, ...).

**a**

| Stem | Leaf       |
|------|------------|
| 3    | 2 5        |
| 4    | 1 3 7      |
| 5    | 4 4 6      |
| 6    | 0 2        |
| 7    | 1 1        |
| 4    | 1 means 41 |

**b**

| Stem | Leaf        |
|------|-------------|
| 0    | 2 3 7       |
| 1    | 4 4 8 9     |
| 2    | 3 6 6       |
| 3    | 0 5         |
| 2    | 3 means 2.3 |

Hint: Look at the key '4 | 1 means 41' to see how the stems and leaves go together.



## Fluency

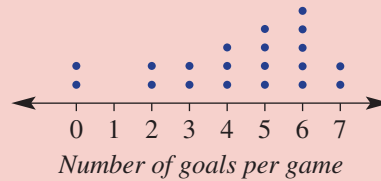
4, 5, 6–7(½)

4, 5, 6–7(½)



## Example 4 Interpreting a dot plot

This dot plot shows the number of goals per game scored by a team during the soccer season.



- How many games were played?
- What was the most common number of goals per game?
- How many goals were scored for the season?
- Describe the data in the dot plot.

## Solution

## Explanation

- a** There were 20 matches played.

Each dot represents a match. Count the number of dots.

- b** 6 goals in a game occurred most often.

The most common number of goals has the most dots.

$$\begin{aligned} \mathbf{c} \quad & 2 \times 0 + 2 \times 2 + 2 \times 3 + 3 \times 4 + 4 \times 5 \\ & + 5 \times 6 + 2 \times 7 \\ & = 0 + 4 + 6 + 12 + 20 + 30 + 14 \\ & = 86 \text{ goals} \end{aligned}$$

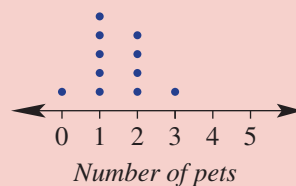
Count the number of games (dots) for each number of goals and multiply by the number of goals. Add these together.

- d** Two games resulted in no goals but the data were generally skewed towards a higher number of goals.

Consider the shape of the graph; it is bunched towards the 6 end of the goal scale.

## Now you try

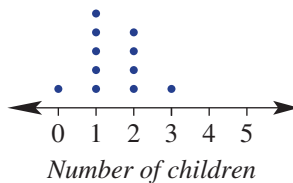
This dot plot shows the number of pets owned by a number of households in a street.



- How many households were surveyed?
- What was the most common number of pets?
- How many pets are there in the street?
- Describe the data in the dot plot.

5C

- 4 A number of families were surveyed to find the number of children in each. The results are shown in this dot plot.

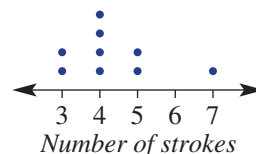


Hint: 4 families had 2 children (4 dots), so that represents 8 children from these families.



- How many families were surveyed?
- What was the most common number of children in a family?
- How many children were there in total?
- Describe the data in the dot plot.

- 5 This dot plot shows the number of strokes a golfer played, each hole, in his round of golf.



- How many holes did he play?
- How many strokes did he play in the round?
- Describe his round of golf.



### Example 5 Constructing a stem-and-leaf plot

Consider this data set.

22 62 53 44 35 47 51 64 72  
32 43 57 64 70 33 51 68 59

- Organise the data into an ordered stem-and-leaf plot.
- Describe the distribution of the data as symmetrical or skewed.

#### Solution

| Stem | Leaf       |
|------|------------|
| 2    | 2          |
| 3    | 2 3 5      |
| 4    | 3 4 7      |
| 5    | 1 1 3 7 9  |
| 6    | 2 4 4 8    |
| 7    | 0 2        |
| 5    | 1 means 51 |

#### Explanation

For two-digit numbers, select the tens value as the stem and the units as the leaves.

The data ranges from 22 to 72, so the graph will need stems 2 to 7.

Work through the data and record the leaves in the order of the data.

| Stem | Leaf       |
|------|------------|
| 2    | 2          |
| 3    | 5 2 3      |
| 4    | 4 7 3      |
| 5    | 3 1 7 1 9  |
| 6    | 2 4 4 8    |
| 7    | 2 0        |
| 5    | 1 means 51 |

51 occurs twice, so the leaf 1 is recorded twice in the 5 stem row. Once data are recorded, redraw and order the leaves from smallest to largest.

Include a key to explain how the stem and leaf go together; i.e. 5 | 1 means 51.

- The distribution of the data is symmetrical.
- The shape of the graph is symmetrical (i.e. evenly spread) either side of the centre.

#### Now you try

Consider this data set.

15 11 18 22 31 24 24 26  
14 63 54 40 44 32 28 10

- Organise the data into an ordered stem-and-leaf plot.
- Describe the distribution of the data as symmetrical or skewed.

- 6 Consider the following sets of data.
- Organise the data into an ordered stem-and-leaf plot.
  - Describe the distribution of the data as symmetrical or skewed.
- a 46 22 37 15 26 38 52 24  
31 20 15 37 21 25 26
- b 35 16 23 55 38 44 12 48 21 42  
53 36 35 25 40 51 27 31 40 36 32
- c 153 121 124 117 125 118 135 137 162  
145 147 119 127 149 116 133 160 158
- d 4.9 3.7 4.5 5.8 3.8 4.3 5.2 7.0 4.7  
4.4 5.5 6.5 6.1 3.3 5.4 2.0 6.3 4.8

Hint: Remember to include a key such as '4 | 6 means 46'.



Hint:

| Symmetrical |         | Skewed |         |
|-------------|---------|--------|---------|
| Stem        | Leaf    | Stem   | Leaf    |
| 1           | 1 2     | 1      | 2 5 7 8 |
| 2           | 1 2 3   | 2      | 3 4 7 6 |
| 3           | 1 2 3 4 | 3      | 1 2     |
| 4           | 1 2 7   | 4      | 5       |
| 5           | 3       |        |         |



### Example 6 Constructing back-to-back stem-and-leaf plots

Two television sales representatives sell the following number of televisions each week over a 15-week period.

*Employee 1*

23 38 35 21 45 27 43 36  
19 35 49 20 39 58 18

*Employee 2*

28 32 37 20 30 45 48 17  
32 37 29 17 49 40 46

- Construct an ordered back-to-back stem-and-leaf plot.
- Describe the distribution of each employee's sales.

#### Solution

| Employee 1     |      | Employee 2 |  |
|----------------|------|------------|--|
| Leaf           | Stem | Leaf       |  |
| 9 8            | 1    | 7 7        |  |
| 7 3 1 0        | 2    | 0 8 9      |  |
| 9 8 6 5 5      | 3    | 0 2 2 7 7  |  |
| 9 5 3          | 4    | 0 5 6 8 9  |  |
| 8              | 5    |            |  |
| 3   7 means 37 |      |            |  |

#### Explanation

Construct an ordered stem-and-leaf plot with the sales by employee 1 on the left-hand side and the sales by employee 2 on the right-hand side. Include a key.

- Sales by employee 1 are symmetrical, whereas sales by employee 2 are skewed. Observe the shape of each employee's graph. If appropriate, use the words symmetrical (i.e. spread evenly around the centre) or skewed (i.e. bunched to one side of the centre).

#### Now you try

Two friends spent the following amounts in dollars on take-away food over a 12-week period.

*Friend 1*

54 44 30 32 46 62  
22 66 41 36 57 48

*Friend 2*

16 24 30 44 29 19  
22 28 18 52 32 41

- Construct an ordered back-to-back stem-and-leaf plot.
- Describe the distribution of each friend's spending.

## 5C

- 7 Consider the following sets of data.
- Draw a back-to-back stem-and-leaf plot.
  - Comment on the distribution of the two data sets.

**a** Set 1: 61 38 40 53 48 57 64  
39 42 59 46 42 53 43

Set 2: 41 55 64 47 35 63 61  
52 60 52 56 47 67 32

**b** Set 1: 176 164 180 168 185 187 195 166 201  
199 171 188 175 192 181 172 187 208

Set 2: 190 174 160 170 186 163 182 171  
167 187 171 165 194 182 163 178

Hint: For part **a** use a key like '3 | 7 means 37' and for part **b** use a key like '15 | 6 means 156'.



### Problem-solving and reasoning

8, 9

8, 10, 11

- 8 Two football players, Logan and Max, compare their personal tallies of the number of goals scored for their team over a 12-match season. Their tallies are as follows.

| Game  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-------|---|---|---|---|---|---|---|---|---|----|----|----|
| Logan | 0 | 2 | 2 | 0 | 3 | 1 | 2 | 1 | 2 | 3  | 0  | 1  |
| Max   | 0 | 0 | 4 | 1 | 0 | 5 | 0 | 3 | 1 | 0  | 4  | 0  |

- Draw a dot plot to display Logan's goal-scoring achievement.
  - Draw a dot plot to display Max's goal-scoring achievement.
  - How would you describe Logan's scoring habits?
  - How would you describe Max's scoring habits?
- 9 This stem-and-leaf plot shows the times, in minutes, that Chris has achieved in the past 14 fun runs she competed in.
- What is the difference between her slowest and fastest times?
  - Just by looking at the stem-and-leaf plot, what would you estimate to be Chris' average time?
  - If Chris records another time of 24.9 minutes, how would this affect your answer to part **b**?

| Stem   | Leaf       |
|--------|------------|
| 20     | 5 7        |
| 21     | 1 2 6      |
| 22     | 2 4 6 8    |
| 23     | 4 5 6      |
| 24     | 3 6        |
| 22   4 | means 22.4 |

- 10 The data below show the distances travelled (in km) by students at an inner-city and an outer-suburb school.

Inner city: 3 10 9 14 21 6  
1 12 24 1 19 4

Outer suburb: 12 21 18 9 34 19  
24 3 23 41 18 4

- Draw a back-to-back stem-and-leaf plot for the data.
- Comment on the distribution of distances for each school.
- Give a practical reason for the distribution of the data.

11 Determine the possible values of the pronumerals in the following ordered stem-and-leaf plots.

**a**

| Stem     | Leaf           |
|----------|----------------|
| 1        | 2 4            |
| 2        | 3 6 9 <i>b</i> |
| <i>a</i> | 1 4            |
| 4        | 7 <i>c</i> 8   |

2 | 3 means 2.3

**b**

| Stem | Leaf           |
|------|----------------|
| 20   | <i>a</i> 1 4   |
| 21   | 2 2 9          |
| 22   | 0 <i>b</i> 5 7 |
| 23   | 1 4            |

22 | 7 means 227

Hint: The stems and leaves are ordered from smallest to largest. A leaf can appear more than once.



### Splitting stems

—

12

12 The back-to-back stem-and-leaf plot below shows the maximum daily temperature for two cities over a 2-week period.

| Maximum temperature |      |             |
|---------------------|------|-------------|
| City A              |      | City B      |
| leaf                | Stem | leaf        |
|                     | 0    |             |
| 9 8 8               | 0*   |             |
| 4 3 3 1 1 1         | 1    |             |
| 8 8 6 6 5           | 1*   | 7 9         |
|                     | 2    | 0 2 2 3 4 4 |
|                     | 2*   | 5 6 7 7 8   |
|                     | 3    | 1           |

1 | 4 means 14  
1\* | 5 means 15

- Describe the difference between the stems 1 and 1\*.
- To which stem would these numbers be allocated?
  - 12°C
  - 5°C
- Why might you use this process of splitting stems, like that used for 1 and 1\*?
- Compare and comment on the differences in temperatures between the two cities.
- What might be a reason for these different temperatures?



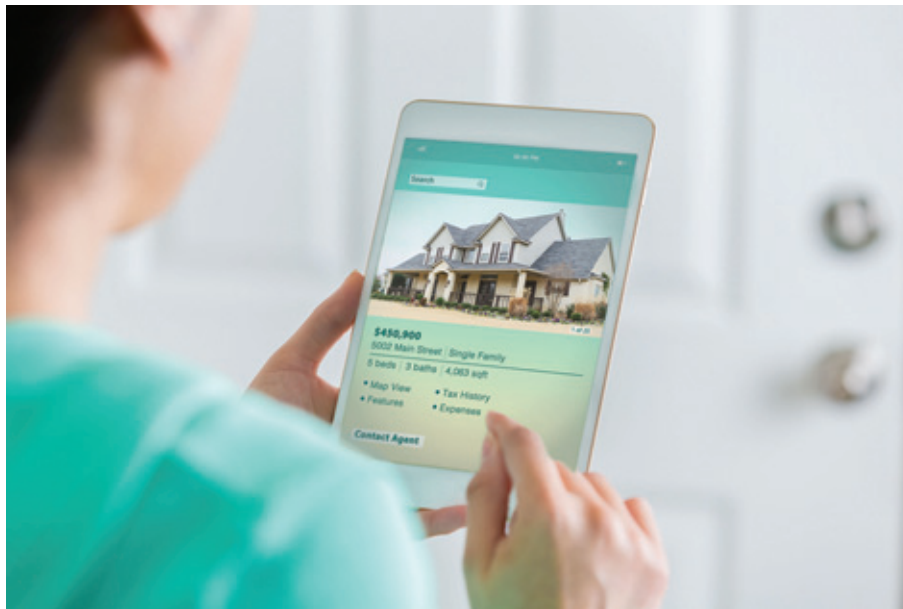
## 5D Range and measures of centre

### Learning intentions

- To know some measures of centre and spread used to summarise a data set.
- To know how the median is found differently for data sets with an odd or even number of values
- To be able to find the mean, median, mode and range of a set of data in list or graphical display form
- To understand the effect of extreme values on the values of the summary statistics

**Key vocabulary:** mode, bimodal, mean, range, median, stem-and-leaf plot, dot plot

In the previous sections you have seen how to summarise data in the form of a frequency table and to display data using graphs. Key summary statistics also allow us to describe the data using a single numerical value. The mean (i.e. average), for example, may be used to describe a student's performance over a series of tests. The median (i.e. middle value when data are ordered) is often used when describing the house prices in a suburb. These are termed *measures of centre*. Providing information about the spread of the data is the range, which measures the difference between the maximum and minimum values.



### → Lesson starter: Mean, median or mode?

The following data represent the number of goals scored by Ellie in each game of a 9-game netball season.

24 18 25 16 3 23 27 19 25

It is known that the figures below represent, in some order, the mean, median and mode.

25 20 23

- Without doing any calculations, can you suggest which statistic is which? Explain.
- From the data, what gives an indication that the mean (i.e. average) will be less than the median (i.e. middle value)?
- Describe how you would calculate the mean, median and mode from the data values.

## Key ideas

- The **mean** (or average) is calculated by summing all the data values and dividing by the total number of values.

$$\text{Mean } (\bar{x}) = \frac{\text{sum of all data values}}{\text{number of data values}}$$

- The mean is affected by extreme values in the data.
- The **mode** is the most commonly occurring value in the data set.
  - A data set can have two modes (called **bimodal**) or no unique mode at all.
- The **median** is the middle value of a data set when the data are arranged in order.
  - When the data set has an even number of values, the median is the average of the two middle values.

For example,

2 3 **6** 8 12

Median = 6

4 7 **8** **10** 13 17

$$\begin{aligned} \text{Median} &= \frac{8 + 10}{2} \\ &= 9 \end{aligned}$$

- The median is not significantly affected by extreme values in the data.
- The **range** is a measure of how spread out the data is.
  - Range = maximum value – minimum value

## Exercise 5D

### Understanding

1–3

1, 3

- 1 Use the words from the list below to fill in the missing word in these sentences.

*mean, median, mode, bimodal, range*

- The \_\_\_\_\_ is the most frequently occurring value in a data set.
- Dividing the sum of all the data values by the total number of values gives the \_\_\_\_\_.
- The middle value of a data set ordered from smallest to largest is the \_\_\_\_\_.
- A data set with two most common values is \_\_\_\_\_.
- A data set has a maximum value of 7 and a minimum value of 2. The \_\_\_\_\_ is 5.

- 2 Circle the middle value(s) of these ordered data sets.

**a** 2 4 6 7 8 10 11

**b** 6 9 10 14 17 20

Hint: Recall that an even number of data values will have two middle values.



- 3 Aaron drinks the following number of cups of coffee each day in a week.

4 5 3 6 4 3 3

- How many cups of coffee does he drink in the week (sum of the data values)?
- How many days are in the week (total number of data values)?
- What is the mean number of cups of coffee Aaron drinks each day (part **a** ÷ part **b**)?



## 5D

## Fluency

4–5(½)

4–5(½), 6



## Example 7 Finding the mean, mode and range

For the following data sets, find:

- i the mean
- ii the mode
- iii the range

**a** 2, 4, 5, 8, 8

**b** 3, 15, 12, 9, 12, 15, 6, 8

**Solution****Explanation**

**a i** Mean =  $\frac{2+4+5+8+8}{5}$   
 $= \frac{27}{5}$   
 $= 5.4$

Mean =  $\frac{\text{sum of all data values}}{\text{number of data values}}$

Add all the data and divide by the number of values (5).

**ii** The mode is 8.

The mode is the most common value in the data.

**iii** Range =  $8 - 2$   
 $= 6$

Range = maximum value – minimum value

**b i** Mean =  $\frac{3+15+12+9+12+15+6+8}{8}$   
 $= \frac{80}{8}$   
 $= 10$

Mean =  $\frac{\text{sum of all data values}}{\text{number of data values}}$

Add all the data and divide by the number of values (8).

**ii** There are two modes: 12 and 15.

The data set is bimodal: 12 and 15 are the most common data values.

**iii** Range =  $15 - 3$   
 $= 12$

Range = maximum value – minimum value


**Now you try**

For the following data sets, find:

- i the mean
- ii the mode
- iii the range

**a** 8, 10, 2, 6, 12, 3, 10, 1

**b** 22, 25, 18, 22, 3, 18

-  4 For each of the following data sets, find:
- i the mean
  - ii the mode
  - iii the range
- a 2 4 5 8 8  
 b 5 8 10 15 20 12 10 50  
 c 55 70 75 50 90 85 50 65 90  
 d 27 30 28 29 24 12  
 e 2.0 1.9 2.7 2.9 2.6 1.9 2.7 1.9  
 f 1.7 1.2 1.4 1.6 2.4 1.3

Hint:  
 $\text{Mean} = \frac{\text{sum of data values}}{\text{number of data values}}$   
 Mode is the most common value.  
 Range = maximum – minimum



### Example 8 Finding the median

Find the median of each data set.

- a 4, 7, 12, 2, 9, 15, 1      b 16, 20, 8, 5, 21, 14

#### Solution

- a 1 2 4 7 9 12 15  
 Median = 7

#### Explanation

The data must first be ordered from smallest to largest.  
 The median is the middle value.  
 For an odd number of data values, there will be one middle value.

- b 5 8 14 16 20 21  
 $\text{Median} = \frac{14 + 16}{2}$   
 = 15

Order the data from smallest to largest.  
 For an even number of data values, there will be two middle values.  
 The median is the average of these two values  
 (i.e. the value halfway between the two middle numbers).

#### Now you try

Find the median of each data set.

- a 10, 16, 2, 7, 1, 18, 5, 10, 14      b 10, 35, 18, 24, 12, 28, 16, 31

- 5 Find the median of each data set.
- a 1 4 7 8 12
  - b 1 2 2 4 4 7 9
  - c 11 13 6 10 14 13 11
  - d 62 77 56 78 64 73 79 75 77
  - e 2 4 4 5 6 8 8 10 12 22
  - f 1 2 2 3 7 12 12 18
  - g 30 36 31 38 27 40
  - h 2.4 2.0 3.2 2.8 3.5 3.1 3.7 3.9
- 6 Nine people watch the following number of hours of television on a weekend.
- 4 4 6 6 6 8 9 9 11
- a Find the mean number of hours of television watched.
  - b Find the median number of hours of television watched.
  - c Find the range of the television hours watched.
  - d What is the mode number of hours of television watched?

Hint: First make sure that the data values are in order. For two middle values, find their average.



## 5D

## Problem-solving and reasoning

7–9

9–11

7 Eight students compare the amount of pocket money they receive. The data are as follows.

\$12 \$15 \$12 \$24 \$20 \$8 \$50 \$25

- Find the range of pocket money received.
- Find the median amount of pocket money.
- Find the mean amount of pocket money.
- Why is the mean larger than the median?



### Example 9 Calculating summary statistics from a stem-and-leaf plot

For the data in this stem-and-leaf plot, find the:

- range
- mode
- mean
- median

| Stem | Leaf       |
|------|------------|
| 2    | 5 8        |
| 3    | 1 2 2 2 6  |
| 4    | 0 3 3      |
| 5    | 2 6        |
| 5    | 2 means 52 |

#### Solution

- Minimum value = 25  
Maximum value = 56  
Range =  $56 - 25$   
= 31

- Mode = 32

- Mean =  $\frac{25 + 28 + 31 + 32 + 32 + 32 + 36 + 40 + 43 + 43 + 52 + 56}{12}$   
=  $\frac{450}{12}$   
= 37.5

- Median =  $\frac{32 + 36}{2}$   
= 34

#### Explanation

In an ordered stem-and-leaf plot, the first data item is the minimum and the last is the maximum. Use the key '5 | 2 means 52' to see how to put the stem and leaf together.  
Range = maximum value – minimum value

The mode is the most common value. The leaf 2 appears three times with the stem 3.

Form each data value from the graph and add them all together. Then divide by the number of data values in the stem-and-leaf plot.

There is an even number of data values: 12. The median will be the average of the middle two values (i.e. the 6th and 7th data values).

#### Now you try

For the data in this stem-and-leaf plot, find the:

- range
- mode
- mean
- median

| Stem | Leaf       |
|------|------------|
| 2    | 1 3 7      |
| 3    | 2 8 9 9    |
| 4    | 4 6 8      |
| 3    | 2 means 32 |



8 For the data in these stem-and-leaf plots, find the:

i range

ii mode

iii mean (rounded to one decimal place)

iv median

Hint: Use the key to see how the stem and leaf go together.

**a**

| Stem | Leaf       |
|------|------------|
| 0    | 4 4        |
| 1    | 0 2 5 9    |
| 2    | 1 7 8      |
| 3    | 2          |
| 2    | 7 means 27 |

**b**

| Stem | Leaf        |
|------|-------------|
| 10   | 1 2 4       |
| 11   | 2 6         |
| 12   | 5           |
| 11   | 6 means 116 |

**c**

| Stem | Leaf        |
|------|-------------|
| 3    | 0 0 5       |
| 4    | 2 7         |
| 5    | 1 3 3       |
| 6    | 0 2         |
| 3    | 2 means 3.2 |



9 This back-to-back stem-and-leaf plot shows the results of two students, Hugh and Mark, on their end-of-year examination in each subject.

**a** For each student, find:

i the mean

ii the median

iii the range

**b** Compare the performance of the two students using your answers to part **a**.

| Hugh leaf | Stem | Mark leaf   |
|-----------|------|-------------|
| 8 8 5     | 6    | 4           |
| 7 3       | 7    | 4 7         |
| 5 4 2 1 1 | 8    | 2 4 6 8     |
|           | 9    | 2 4 5       |
|           | 7    | 4 means 74% |

10 A real estate agent recorded the following amounts for the sale of five properties:

\$120 000 \$210 000 \$280 000 \$370 000 \$1 700 000

The mean is \$536 000 and the median is \$280 000.

Which is a better measure of the centre of the five property prices: the mean or the median?

Give a reason.

11 This dot plot shows the number of wins recorded by a school sports team in the past 10 eight-game seasons.

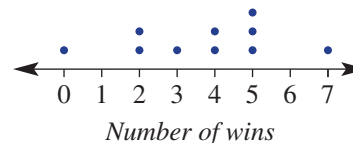
**a** What was the median number of wins?

**b** What was the mean number of wins?

**c** The following season, the team records 3 wins. What effect will this have (i.e. increase/decrease/no change) on the:

i median?

ii mean?



## Moving run average

12

12 A moving average is determined by calculating the average of all data values up to a particular time or place in the data set.

Consider a batter in cricket with the following runs scored from 10 completed innings.

| Innings        | 1  | 2  | 3 | 4  | 5  | 6   | 7  | 8  | 9  | 10 |
|----------------|----|----|---|----|----|-----|----|----|----|----|
| Score          | 26 | 38 | 5 | 10 | 52 | 103 | 75 | 21 | 33 | 0  |
| Moving average | 26 | 32 |   |    |    |     |    |    |    |    |

In the table, 26 is the average after 1 innings and 32 is the average after 2 innings.

**a** Complete the table by calculating the moving average for innings 3–10. Round your answer to the nearest whole number where required.

**b** Plot the score and moving averages for the batter on the same set of axes, with the innings number on the horizontal axis. Join the points to form two line graphs.

**c** Describe the behaviour of the:

i score graph

ii moving average graph

**d** Describe the main difference in the behaviour of the two graphs. Give reasons.

## 5E Quartiles and outliers

### Learning intentions

- To know which statistics make up the five-figure summary
- To be able to calculate the quartiles of a data set
- To be able to find the interquartile range and know what it represents
- To know what is meant by an outlier and be able to find any outliers in a data set
- To understand the impact of outliers on various statistics

**Key vocabulary:** five-figure summary, upper quartile, lower quartile, median, interquartile range, outlier, upper fence, lower fence

In addition to the median of a single set of data, there are two related statistics called the upper and lower quartiles. If data are placed in order, then the lower quartile is central to the lower half of the data. The upper quartile is central to the upper half of the data. These quartiles are used to calculate the interquartile range, which helps to describe the spread of the data, and show whether or not any data points do not fit the rest of the data (outliers).



### → Lesson starter: House prices

A real estate agent tells you that the median house price for a suburb in 2020 was \$753 000 and the mean was \$948 000.

- Is it possible for the median and the mean to differ by so much?
- Under what circumstances could this occur? Discuss.

### Key ideas

■ The **five-figure summary** uses the following statistical measures to summarise a set of data:

- |                                   |  |
|-----------------------------------|--|
| • Minimum value (min)             | the lowest value                               |
| • <b>Lower quartile</b> ( $Q_1$ ) | the number above 25% of the ordered data       |
| • Median ( $Q_2$ )                | the middle value above 50% of the ordered data |
| • <b>Upper quartile</b> ( $Q_3$ ) | the number above 75% of the ordered data       |
| • Maximum value (max)             | the highest value                              |

Odd number of data values

$$1 \quad \underline{2 \quad 2} \quad 3 \quad ) \textcircled{5} ( \quad \underline{6 \quad 6} \quad \underline{7 \quad 9}$$

$$Q_1 = \frac{2+2}{2} = 2 \quad Q_2 = 5 \quad Q_3 = \frac{6+7}{2} = 6.5$$

Even number of data values

$$2 \quad 3 \quad \textcircled{3} \quad 4 \quad \underline{7 \quad 8} \quad 8 \quad \textcircled{9} \quad 9 \quad 9$$

$$Q_1 = 3 \quad Q_2 = 7.5 \quad Q_3 = 9$$

■ Another measure of the spread of the data is the **interquartile range (IQR)**.

IQR = upper quartile – lower quartile

$$= Q_3 - Q_1$$

■ **Outliers** are data elements outside the vicinity of the rest of the data.

A data point is an outlier if it is either:

- less than the **lower fence**, where lower fence =  $Q_1 - 1.5 \times \text{IQR}$  or
- greater than the **upper fence**, where upper fence =  $Q_3 + 1.5 \times \text{IQR}$

■ Outliers significantly affect the range of a data set but have limited to no effect on the IQR.

# Exercise 5E

## Understanding

1-3

3

- 1
  - a State the five values that need to be calculated for a five-figure summary.
  - b Explain the difference between the range and the interquartile range.
  - c What is an *outlier*?
  
- 2 The data show, in order, the numbers of cars owned by 10 families surveyed.

0, 1, 1, 1, 1, 2, 2, 2, 3, 3

- a Find the median (the middle value).
- b By splitting the data in half, determine the:
  - i lower quartile  $Q_1$  (middle of lower half)
  - ii upper quartile  $Q_3$  (middle of upper half)

Hint: For an even number of data values, average the two middle values for the median.



- 3 For the data set with  $Q_1 = 3$  and  $Q_3 = 8$ :
  - a Find  $IQR = Q_3 - Q_1$ .
  - b Calculate  $Q_1 - 1.5 \times IQR$  (lower fence) and  $Q_3 + 1.5 \times IQR$  (upper fence).
  - c Identify the name that would be given to the value 18 in the data set.

## Fluency

4-6

4-6, 7(½)

### Example 10 Finding quartiles and IQR for an even number of data values

Consider this data set:

2, 2, 4, 5, 6, 8, 10, 13, 16, 20

- a Find the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ ).
- b Determine the IQR.

#### Solution

$$\begin{array}{cccccccccccc}
 \text{a} & 2 & 2 & \textcircled{4} & 5 & 6 & | & 8 & 10 & \textcircled{13} & 16 & 20 \\
 & & & \uparrow & & & & \uparrow & & & & \\
 & & & Q_1 & & & & Q_3 & & & & \\
 & & & & & & & Q_2 = \frac{6+8}{2} & & & & \\
 & & & & & & & = 7 & & & & \\
 & & & Q_1 = 4 & \text{ and } & Q_3 = 13 & & & & & & 
 \end{array}$$

#### Explanation

The data are already ordered. Since there is an even number of values, split the data in half to locate the median.

$Q_1$  is the middle value of the lower half:

2 2  $\textcircled{4}$  5 6

$Q_3$  is the middle value of the upper half:

8 10  $\textcircled{13}$  16 20

b  $IQR = 13 - 4$   
 $= 9$

$IQR = Q_3 - Q_1$

#### Now you try

Consider this data set:

3, 6, 6, 10, 12, 13, 15, 19, 24, 27, 28, 31, 33, 33

- a Find the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ ).
- b Determine the IQR.

## 5E

- 4 For these data sets, find the:
- upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ )
  - IQR
- 3, 4, 6, 8, 8, 10
  - 10, 10, 11, 14, 14, 15, 16, 18, 20, 21
  - 41, 49, 53, 58, 59, 62, 62, 65
  - 1.2, 1.7, 1.9, 2.2, 2.4, 2.5, 2.9, 3.2

Hint: For an even number of data values, split the ordered data in half:

2 4 7 | 8 10 12



$Q_1$



$Q_3$

$$\text{IQR} = Q_3 - Q_1$$



### Example 11 Finding quartiles and IQR for an odd number of data values

Consider this data set:

2.2, 1.6, 3.0, 2.7, 1.8, 3.6, 3.9, 2.8, 3.8

- Find the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ ).
- Determine the IQR.

#### Solution

a

|     |     |  |     |     |       |     |     |  |     |     |
|-----|-----|--|-----|-----|-------|-----|-----|--|-----|-----|
| 1.6 | 1.8 |  | 2.2 | 2.7 | 2.8   | 3.0 | 3.6 |  | 3.8 | 3.9 |
|     |     |  |     |     | ↑     |     |     |  |     |     |
|     |     |  |     |     | $Q_2$ |     |     |  |     |     |

$$Q_1 = \frac{1.8 + 2.2}{2}$$

$$= \frac{4.0}{2}$$

$$= 2.0$$

$$Q_3 = \frac{3.6 + 3.8}{2}$$

$$= \frac{7.4}{2}$$

$$= 3.7$$

b

$$\text{IQR} = 3.7 - 2.0$$

$$= 1.7$$

#### Explanation

First order the data and locate the median ( $Q_2$ ).

Split the data in half; i.e. either side of the median.

$Q_1$  is the middle value of the lower half; for two middle values, average the two numbers.

$Q_3$  is the middle value of the upper half.

$$\text{IQR} = Q_3 - Q_1$$

#### Now you try

Consider this data set:

14.4, 15.2, 16.0, 13.7, 18.2, 21.4, 19.7, 19.9, 12.8, 20.6, 21.4

- Find the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ ).
- Determine the IQR.

- 5 For these data sets, find:
- the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ )
  - the IQR
- 1, 2, 4, 8, 10, 11, 14
  - 10, 7, 14, 2, 5, 8, 3, 9, 2, 12, 1
  - 0.9, 1.3, 1.1, 1.2, 1.7, 1.5, 1.9, 1.1, 0.8
  - 21, 7, 15, 9, 18, 16, 24, 33, 4, 12, 13, 18, 24

Hint: For an odd number of data values, split ordered data in half, leaving out the middle value.

0 (2) 4) 7 (9) 14) 16



$Q_1$



$Q_3$





### Example 12 Finding the five-figure summary and outliers

The following data set represents the number of flying geese spotted on each day of a 13-day tour of England.

5, 1, 2, 6, 3, 3, 18, 4, 4, 1, 7, 2, 4

- a** For the data, find the:
- minimum and maximum number of geese spotted
  - median
  - upper and lower quartiles
  - IQR
- b** Find any outliers.
- c** Can you give a possible reason for why the outlier occurred?
- d** If the outlier's value is corrected to 8, determine the range and IQR of the data set. Explain the similarities or differences compared to the original data set.

#### Solution

- a i** Min = 1, max = 18
- ii** 1, 1, 2, 2, 3, 3, (4), 4, 4, 5, 6, 7, 18  
 $\therefore$  Median = 4
- iii** Lower quartile ( $Q_1$ ) =  $\frac{2+2}{2}$   
 $= 2$
- Upper quartile ( $Q_3$ ) =  $\frac{5+6}{2}$   
 $= 5.5$
- iv** IQR =  $5.5 - 2$   
 $= 3.5$

- b** Lower fence =  $Q_1 - 1.5 \times \text{IQR}$   
 $= 2 - 1.5 \times 3.5$   
 $= 2 - 5.25$   
 $= -3.25$

$$\begin{aligned} \text{Upper fence} &= Q_3 + 1.5 \times \text{IQR} \\ &= 5.5 + 1.5 \times 3.5 \\ &= 5.5 + 5.25 \\ &= 10.75 \end{aligned}$$

$\therefore$  The outlier is 18.

- c** Perhaps a flock of geese was spotted that day.

#### Explanation

Look for the largest and smallest numbers and order the data:

$$1 \ 1 \ 2 \ | \ 2 \ 3 \ 3 \ ) \ 4 \ ( \ 4 \ 4 \ 5 \ | \ 6 \ 7 \ 18$$

$\uparrow$                      $\uparrow$                      $\uparrow$   
 $Q_1$                      $Q_2$                      $Q_3$

Since  $Q_2$  falls on a data value, it is not included in the lower or higher halves when  $Q_1$  and  $Q_3$  are calculated.

$$\text{IQR} = Q_3 - Q_1$$

A data point is an outlier if it is less than the lower fence =  $Q_1 - 1.5 \times \text{IQR}$  or greater than the upper fence =  $Q_3 + 1.5 \times \text{IQR}$ .

There are no numbers less than  $-3.25$  but 18 is greater than 10.75.

*Continued on next page*



## 5E

- d 1, 1, 2, 2, 3, 4, 4, 4, 5, 6, 7, 8

$$\text{Range} = 8 - 1$$

$$= 7$$

$$\text{IQR} = 5.5 - 2$$

$$= 3.5$$

The range is significantly reduced by correcting the outlier. The IQR here is unchanged. Extreme values do not significantly affect the IQR.

18 is replaced in the data set with 8 and the range and IQR calculated.

The outlier affected the range as the maximum value is changed but the IQR is not affected by the value of the maximum.

### Now you try

The following data set represents the number of hours a ball girl worked at the Australian Open over the 14-day event.

4, 6, 6, 5, 8, 7, 8, 14, 5, 4, 2, 3, 5, 6

- a For the data, find the:
- minimum and maximum number of hours worked
  - median
  - upper and lower quartiles
  - IQR
- b Find any outliers.
- c Can you give a possible reason for why the outlier occurred?
- d If the outlier's value is corrected to 9, determine the range and IQR of the data set. Explain the similarities or differences compared to the original data set.

- 6 The following numbers of cars were counted on each day for 15 days, travelling on a quiet suburban street.

10, 9, 15, 14, 10, 17, 15, 0, 12, 14, 8, 15, 15, 11, 13

- a For the given data, find the:
- minimum and maximum number of cars counted
  - median
  - lower and upper quartiles ( $Q_1$  and  $Q_3$ )
  - IQR
- b Find any outliers.
- c Give a possible reason for the outlier.
- d The outlier value is changed to 8. Find the new range and IQR and comment on any similarities or differences compared to the original data set.

Hint:  
Outliers are more than  $Q_3 + 1.5 \times \text{IQR}$   
or  
less than  $Q_1 - 1.5 \times \text{IQR}$



- 7 Summarise the data sets below by finding:
- the minimum and maximum values
  - the median ( $Q_2$ )
  - the lower and upper quartiles ( $Q_1$  and  $Q_3$ )
  - the IQR
  - any outliers
- a 4, 5, 10, 7, 5, 14, 8, 5, 9, 9  
 b 24, 21, 23, 18, 25, 29, 31, 16, 26, 25, 27  
 c 10, 13, 2, 11, 10, 8, 24, 12, 13, 15, 12  
 d 3, 6, 10, 11, 17, 4, 4, 1, 8, 4, 10, 8

## Problem-solving and reasoning

8, 9

8, 10–12

- 8 Twelve different calculators had the following numbers of buttons:

36, 48, 52, 43, 46, 53, 25, 60, 128, 32, 52, 40

- a For the given data, find:
- the minimum and maximum number of buttons on the calculators
  - the median
  - the lower and upper quartiles ( $Q_1$  and  $Q_3$ )
  - the IQR
  - any outliers
  - the mean
- b Can you give a possible reason why the outlier has occurred?
- c Which is a better measure of the centre of the data: the mean or the median? Explain.
- d Which is a better measure of the spread of the data: the range or the IQR? Explain.



- 9 At an airport, Paul checks the weight of 20 luggage items. If the weight of a piece of luggage is an outlier, then the contents undergo a further check. The weights in kilograms are:

1 4 5 5 6 7 7 7 8 8  
 10 10 10 13 15 16 17 19 32 33

How many luggage items will undergo a further check?



- 10 The prices of nine fridges are displayed in a sale catalogue. They are:  
 \$350 \$1000 \$850 \$900 \$1100 \$1200  
 \$1100 \$1000 \$1700  
 How many of the fridge prices could be considered outliers?



5E

11 For the data in this stem-and-leaf plot, find:

- a the IQR
- b any outliers
- c any outliers if the number 32 was added to the list

| Stem | Leaf       |
|------|------------|
| 0    | 1          |
| 1    | 6 8        |
| 2    | 0 4 6 8    |
| 3    | 0          |
| 2    | 4 means 24 |

Hint: Split the data in half to find  $Q_2$ , then find  $Q_1$  and  $Q_3$ .



12 For the data in this stem-and-leaf plot the value 42 was incorrectly recorded. What are the possible values it should have been if the:

- a median is not changed
- b IQR is not changed

| Stem | Leaf       |
|------|------------|
| 1    | 7 8        |
| 2    | 1 4 6      |
| 3    | 5 5        |
| 4    | 2 9        |
| 5    | 0 3        |
| 4    | 2 means 42 |



### Some research

13

13 Use the internet to search for data about a topic that interests you. Try to choose a single set of data that includes between 15 and 50 values.

- a Organise the data using a:
  - i stem-and-leaf plot
  - ii frequency table and histogram
- b Find the mean and the median.
- c Find the range and the interquartile range.
- d Write a brief report describing the centre and spread of the data, referring to parts a to c above.
- e Present your findings to your class or a classmate.



## 5F Box plots

### Learning intentions

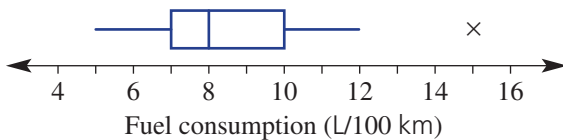
- To know that a box plot can be used to represent the five-figure summary of a data set
- To understand that the sections of a box plot each represent approximately 25% of the data and show the spread of the data
- To be able to construct box plots both with and without outliers
- To know that parallel box plots can be used to compare two or more sets of data in the same context

**Key vocabulary:** box plot, parallel box plot, outlier, five-figure summary, upper quartile, lower quartile, median, interquartile range, upper fence, lower fence

The five-figure summary (min,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , max) can be represented in graphical form as a box plot. Box plots are graphs that summarise single data sets. They clearly display the minimum and maximum values, the median, the quartiles and any outliers.  $Q_1$ ,  $Q_2$  and  $Q_3$  divide the data into quarters. Box plots also give a clear indication of how data are spread, as the IQR (interquartile range) is shown by the width of the central box.

### → Lesson starter: Fuel consumption

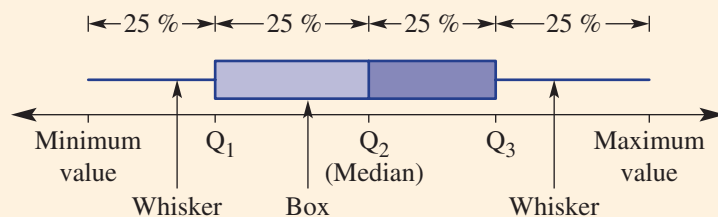
This box plot summarises the average fuel consumption (litres per 100 km) for a group of European-made cars.



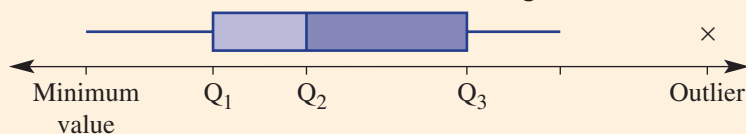
- What does each part of the box plot represent in terms of the five-figure summary?
- What do you think the cross (x) represents?
- Describe how you can use the box plot to find the IQR.
- For the top 25% of cars, what would you expect the fuel consumption to be above?

### Key ideas

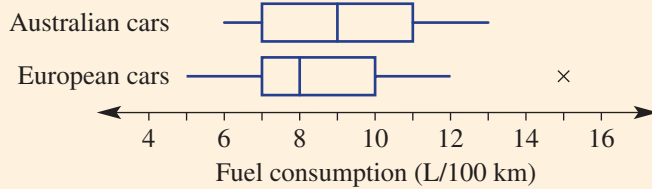
- A **box plot** (also called a box-and-whisker plot) can be used to summarise a data set and show the spread of the data. It displays the five-figure summary (min,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , max), as shown.



- An outlier is marked with a cross (x).
  - An outlier is greater than  $Q_3 + 1.5 \times \text{IQR}$  (upper fence) or less than  $Q_1 - 1.5 \times \text{IQR}$  (lower fence).
  - $\text{IQR} = Q_3 - Q_1$
  - The whiskers stretch to the lowest and highest data values that are not outliers.



- **Parallel box plots** are box plots drawn on the same scale. They are used to compare data sets within the same context.



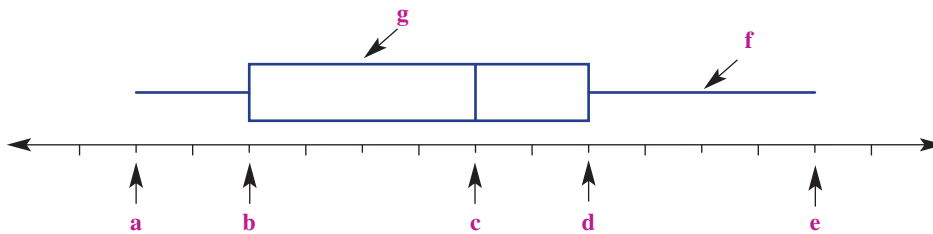
## Exercise 5F

### Understanding

1-4

4

- 1 Label the parts **a–g** of the box plot below.



- 2 For this simple box plot, find the:

**a** median ( $Q_2$ )

**b** minimum

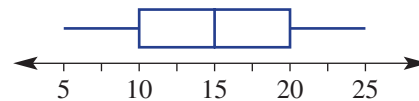
**c** maximum

**d** range

**e** lower quartile ( $Q_1$ )

**f** upper quartile ( $Q_3$ )

**g** interquartile range (IQR)



- 3 Construct a box plot showing these features.

**a** Min = 1,  $Q_1 = 3$ ,  $Q_2 = 4$ ,  $Q_3 = 7$ , max = 8

**b** Outlier = 5, minimum above outlier = 10,  
 $Q_1 = 12$ ,  $Q_2 = 14$ ,  $Q_3 = 15$ , max = 17

- 4 Select from the list below to fill in the blanks.

*minimum,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , maximum.*

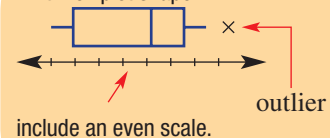
**a** The top 25% of data are above \_\_\_\_\_.

**b** The middle 50% of data are between \_\_\_\_\_ and \_\_\_\_\_.

**c** The lowest or first 25% of data are between the \_\_\_\_\_ and \_\_\_\_\_.

**d** The highest or last 25% of data are between \_\_\_\_\_ and the \_\_\_\_\_.

Hint: Box plot shape





## 5F



### Example 14 Constructing box plots with outliers

Consider the given data set.

5, 9, 4, 3, 5, 6, 6, 5, 7, 12, 2, 3, 5

- Determine the quartiles  $Q_1$ ,  $Q_2$  and  $Q_3$ .
- Determine whether any outliers exist.
- Draw a box plot to summarise the data, marking outliers if they exist.

#### Solution

**a**

2 3 3 4 5 5) 5 (5 6 6 7 9 12

↑                    ↑                    ↑

$Q_1$                      $Q_2$                      $Q_3$

$$Q_1 = \frac{3+4}{2} = 3.5$$

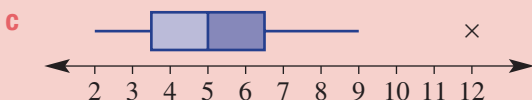
$$Q_3 = \frac{6+7}{2} = 6.5$$

**b**  $IQR = 6.5 - 3.5 = 3$

$$Q_1 - 1.5 \times IQR = 3.5 - 1.5 \times 3 = -1$$

$$Q_3 + 1.5 \times IQR = 6.5 + 1.5 \times 3 = 11$$

$\therefore 12$  is an outlier.



#### Explanation

Order the data to help find the quartiles.

Locate the median  $Q_2$  (i.e. the middle value), then split the data in half above and below this value.

$Q_1$  is the middle value of the lower half and  $Q_3$  is the middle value of the upper half. Average the two middle values to find the median.

Determine  $IQR = Q_3 - Q_1$ .

Check for any outliers; i.e. numbers below  $Q_1 - 1.5 \times IQR$  or above  $Q_3 + 1.5 \times IQR$ .

There are no data below  $-1$  but  $12 > 11$ .

Draw a line and mark in a uniform scale, reaching from 2 to 12. Sketch the box plot by marking the minimum 2 and the outlier 12, and  $Q_1$ ,  $Q_2$  and  $Q_3$ . The end of the five-point summary is the nearest value below 11; i.e. 9.

#### Now you try

Consider the given data set.

10, 8, 11, 9, 8, 22, 15, 10, 12

- Determine the quartiles  $Q_1$ ,  $Q_2$  and  $Q_3$ .
- Determine whether any outliers exist.
- Draw a box plot to summarise the data, marking outliers if they exist.

- 6** Consider the data sets below.
- Determine the quartiles  $Q_1$ ,  $Q_2$  and  $Q_3$ .
  - Determine whether any outliers exist.
  - Draw a box plot to summarise the data, marking outliers if they exist.
- 4, 6, 5, 2, 3, 4, 4, 13, 8, 7, 6
  - 1.8, 1.7, 1.8, 1.9, 1.6, 1.8, 2.0, 1.1, 1.4, 1.9, 2.2
  - 21, 23, 18, 11, 16, 19, 24, 21, 23, 22, 20, 31, 26, 22
  - 37, 48, 52, 51, 51, 42, 48, 47, 39, 41, 65

**Hint:**  
Outliers are more than  $Q_3 + 1.5 \times IQR$  or less than  $Q_1 - 1.5 \times IQR$ .  
Mark with a X.  
The next value above or below an outlier is used as the new end of the whisker.



## Problem-solving and reasoning

7, 8

7–9

- 7 A butcher records the weight (in kilograms) of a dozen parcels of sausages sold on one morning.

1.6 1.9 2.0 2.0 2.1 2.2  
2.2 2.4 2.5 2.7 3.8 3.9

- a Write down the value of:

i the minimum                      ii  $Q_1$   
iii  $Q_2$                                   iv  $Q_3$   
v the maximum                      vi IQR

- b Find any outliers.

- c Draw a box plot for the weight of the parcels of sausages.



- 8 Joel the gardener records the number of days that it takes for 11 special bulbs to germinate. The results are:

8 14 15 15 16 16 16 17 19 19 24

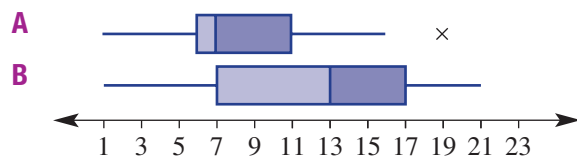
- a Write down the value of:

i the minimum                      ii  $Q_1$                                   iii  $Q_2$   
iv  $Q_3$                                   v the maximum                      vi IQR

- b Are there any outliers? If so, what are they?

- c Draw a box plot for the number of days it takes for the bulbs to germinate.

- 9 Consider these parallel box plots, A and B.



- a What statistical measure do these box plots have in common?

- b Which data set (A or B) has a wider range of values?

- c Find the IQR for:

i data set A                              ii data set B

- d How would you describe the main difference between the two sets of data from which the parallel box plots have been drawn?

Hint: Parallel box plots are two box plots that can be compared using the same scale.

Compare the box plots at each point of the five-figure summary.



## Creating parallel box plots

—

10

- 10 Fifteen essays are marked for spelling errors by a particular examiner and the following numbers of spelling errors are counted.

3, 2, 4, 6, 8, 4, 6, 7, 6, 1, 7, 12, 7, 3, 8

The same 15 essays are marked for spelling errors by a second examiner and the following numbers of spelling errors are counted.

12, 7, 9, 11, 15, 5, 14, 16, 9, 11, 8, 13, 14, 15, 13

- a Draw parallel box plots for the data.

- b Do you believe there is a major difference in the way the essays were marked by the two examiners? If yes, describe this difference.

## Using technology 5F: Using calculators to draw box plots

This activity can be found in the interactive textbook in the form of a printable PDF.



5A

- 1 Name the type of data that would be generated by the following survey questions.
- What is your favourite sport?
  - How many times did you exercise in the past week?

5B

- 2 Twenty-five people were surveyed as to the number of hours of sleep they had the previous night. The data are listed below.

6 5 8 9 7 10 5 8 8  
 11 7 9 4 2 9 8 7  
 10 8 7 5 7 10 6 9

- Organise the data into a frequency table using class intervals of 3, starting with 0–, then 3– etc. Include a percentage frequency column.
- Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- What percentage of people surveyed had fewer than 6 hours sleep?
- Which interval is the most frequent?

5C/D

- 3 This dot plot shows the number of days a group of students did homework in a week.

- How many students were surveyed?
- What was the mode number of days of homework?
- What was the median number of days of homework?
- What was the mean number of days of homework?
- What was the range of the number of days of homework?



5C/D

- 4 Consider the given data set.

7 11 46 42 34 40 24 45 15 22  
 21 16 49 25 33 30 47 30 3 48

- Organise the data into an ordered stem-and-leaf plot.
- Describe the distribution of the data as symmetrical or skewed.
- Use the stem-and-leaf plot to find:
  - the mode
  - the median

5D



- 5 For the following data set, find the:  
 8, 15, 23, 12, 3, 19, 42, 33

- mean
- median
- range

5E

- 6 Consider the given data set.

10 13 8 15 24 18 11 20

- Find the upper quartile ( $Q_3$ ) and lower quartile ( $Q_1$ ).
- Determine the IQR.

5F



- 7 The data below show the number of cars that travel down a particular road between 5:30 p.m. and 6 p.m. each day for a week.

0 82 75 49 102 110 97

- Determine the quartiles  $Q_1$ ,  $Q_2$  and  $Q_3$ .
- Determine whether any outliers exist.
- Draw a box plot to summarise the data.



# 5G Time-series data

## Learning intentions

- To understand that time-series data is data recorded at regular time intervals
- To be able to plot a time-series graph
- To be able to describe any trends in the data of a time-series graph

**Key vocabulary:** time-series data, linear, trend

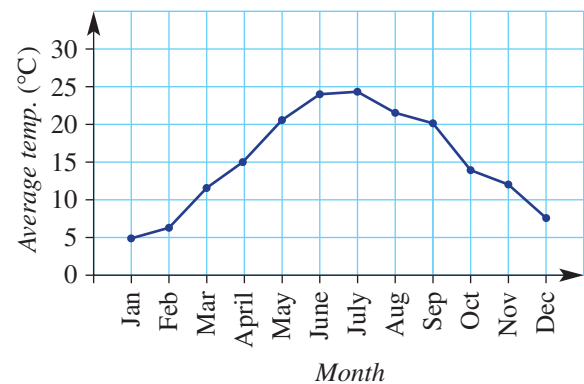
A time series is a sequence of data values that are recorded at regular time intervals. Examples include temperature recorded on the hour, speed recorded every second, population recorded every year and profit recorded every month. A line graph can be used to represent time-series data. This can help to analyse the data, describe trends and make predictions about the future.



## → Lesson starter: Changing temperatures

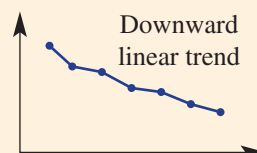
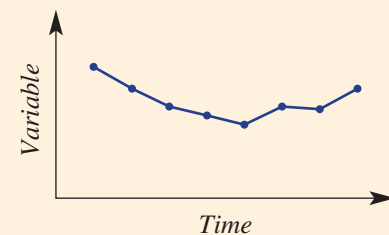
The average monthly maximum temperature for a city is illustrated by this graph.

- Describe the trend in the data at different times of the year.
- Explain why the average maximum temperature for December is close to the average maximum temperature for January.
- Do you think this graph is for an Australian city? Explain.
- If another year of temperatures was included on this graph, what would you expect the shape of the graph to look like?
- Do you think this city is in the Northern Hemisphere or the Southern Hemisphere? Give a reason.



## Key ideas

- **Time-series data** are recorded at regular time intervals.
- The graph or plot of a time series uses:
  - time on the horizontal axis
  - line segments connecting points on the graph
- If the time-series plot results in points being on or near a straight line, then we say that the **trend is linear**.



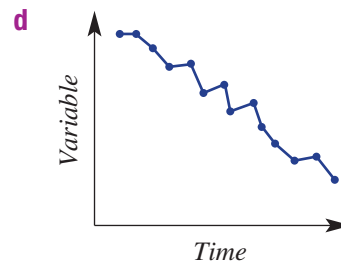
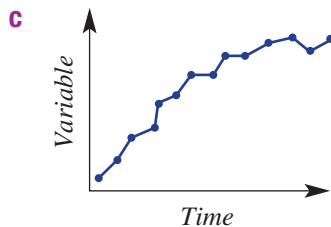
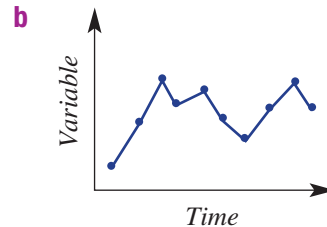
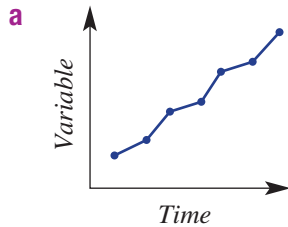
## Exercise 5G

### Understanding

1, 2

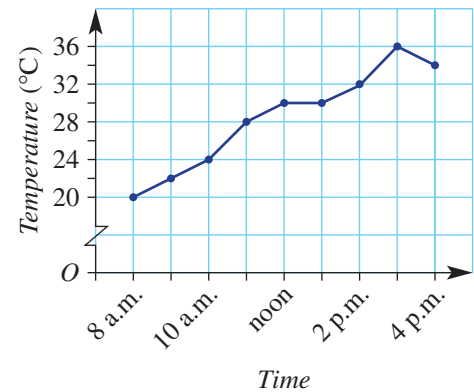
2

- 1 Describe the following time-series plots as having a linear (i.e. straight line) trend, non-linear trend (i.e. a smooth curve) or no trend.



- 2 This time-series graph shows the temperature over the course of 8 hours of a day.

- a** State the temperature at:
- 8 a.m.
  - noon
- b** What was the maximum temperature?
- c** During what times did the temperature:
- stay the same?
  - decrease?
- d** Describe the general trend in the temperature for the 8 hours.



### Fluency

3, 4

3, 4



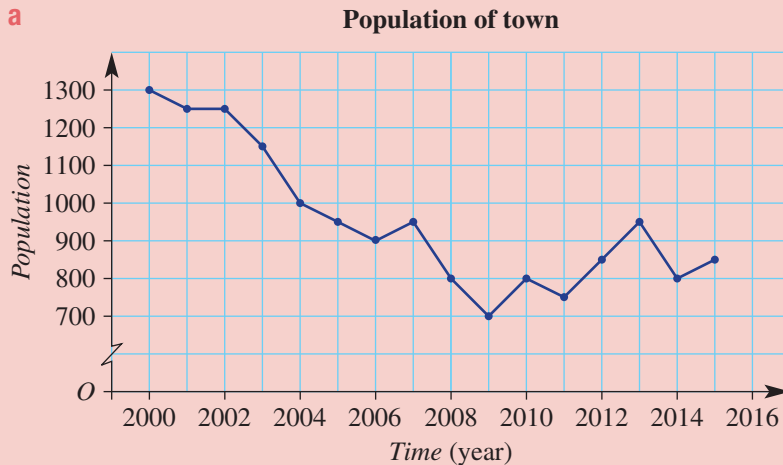
#### Example 15 Plotting and interpreting a time-series plot

The approximate population of a small town was recorded from 2000 to 2015.

| Year       | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Population | 1300 | 1250 | 1250 | 1150 | 1000 | 950  | 900  | 950  | 800  | 700  | 800  | 750  | 850  | 950  | 800  | 850  |

- a** Plot the time-series graph.
- b** Describe the trend in the data over the 16 years.

*Continued on next page*

**Solution****a****Explanation**

Use time on the horizontal axis. Break the  $y$ -axis so as to not include 0–700. Label an even scale on each axis. Join points with line segments.

- b** The population declines steadily for the first 10 years. The population rises and falls in the last 6 years, resulting in a slight upwards trend.

Interpret the overall rise and fall of the lines on the graph.

**Now you try**

The approximate population of a small town was recorded from 2005 to 2015.

| Year       | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|------------|------|------|------|------|------|------|------|------|------|------|------|
| Population | 550  | 500  | 550  | 600  | 700  | 650  | 750  | 750  | 850  | 950  | 900  |

- a** Plot the time-series graph. Break the  $y$ -axis so it does not include 0–500.  
**b** Describe the general trend in the data over the 11 years.

- 3** A company's share price over 12 months is recorded in this table.

| Month      | J    | F    | M    | A    | M    | J    | J    | A    | S    | O    | N    | D    |
|------------|------|------|------|------|------|------|------|------|------|------|------|------|
| Price (\$) | 1.30 | 1.32 | 1.35 | 1.34 | 1.40 | 1.43 | 1.40 | 1.38 | 1.30 | 1.25 | 1.22 | 1.23 |

- a** Plot the time-series graph. Break the  $y$ -axis to exclude values from \$0 to \$1.20.  
**b** Describe the way in which the share price has changed over the 12 months.  
**c** What is the difference between the maximum and minimum share price in the 12 months?

Hint: The scale on the vertical axis will need to include from \$1.20 to \$1.43. Choose an appropriate scale. Month will be on the horizontal axis.



- 4** The pass rate (%) for a particular examination is given in this table.

| Year          | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---------------|------|------|------|------|------|------|------|------|------|------|
| Pass rate (%) | 74   | 71   | 73   | 79   | 85   | 84   | 87   | 81   | 84   | 83   |

- a** Plot the time-series graph for the 10 years.  
**b** Describe the way in which the pass rate for the examination has changed in the given time period.  
**c** In what year is the pass rate a maximum?  
**d** By how much has the pass rate improved from 2006 to 2010?

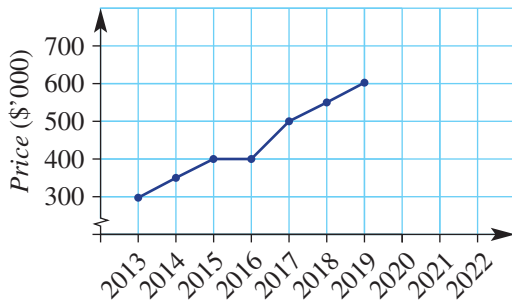
## 5G

## Problem-solving and reasoning

5, 6

5, 7, 8

- 5 This time-series plot shows the upwards trend of house prices in an Adelaide suburb over 7 years from 2013 to 2019.



Hint: Recall that a linear trend has the points on or near a straight line.



- a Would you say that the general trend in house prices is linear or non-linear?  
 b Assuming that the trend in house prices continues for this suburb, what would you expect the house price to be in:  
 i 2020?  
 ii 2022?

- 6 The following data show the monthly sales of strawberries (\$'000s) for a particular year.

| Month           | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Sales (\$'000s) | 22  | 14  | 9   | 11  | 12  | 9   | 7   | 9   | 8   | 10  | 18  | 25  |

Hint: \$'000s means 22 represents \$22 000.



- a Plot the time-series graph for the year.  
 b Describe any trends in the data over the year.  
 c Give a reason why you think the trends you observed may have occurred.



- 7 The two top-selling book stores for a company list their sales figures for the first 6 months of the year. Sales amounts are in thousands of dollars.

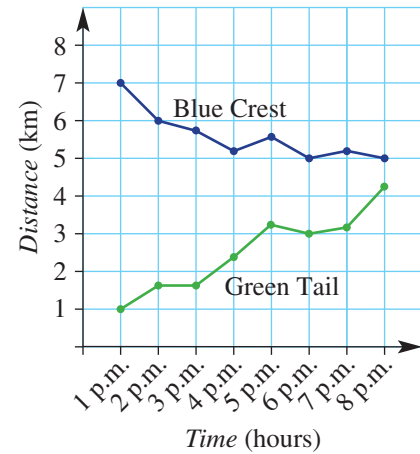
|                       | July | August | September | October | November | December |
|-----------------------|------|--------|-----------|---------|----------|----------|
| City Central (\$'000) | 12   | 13     | 12        | 10      | 11       | 13       |
| Southbank (\$'000)    | 17   | 19     | 16        | 12      | 13       | 9        |

- a What is the difference in the sales volume for:  
 i August?  
 ii December?  
 b How many months did the City Central store sell more books than the Southbank store?  
 c Construct a time-series plot for both stores on the same set of axes.  
 d Describe the trend of sales for the 6 months for:  
 i City Central  
 ii Southbank  
 e Based on the trend for the sales for the Southbank store, what would you expect the approximate sales volume to be in January?

Hint: Use different colours for the two line graphs.



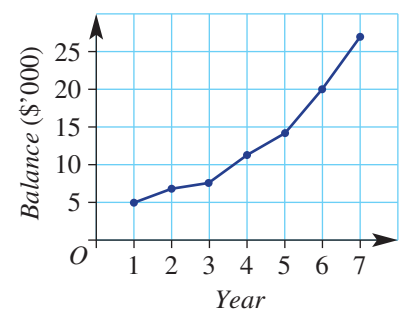
- 8 Two pigeons (Green Tail and Blue Crest) each have a beacon that communicates with a recording machine. The distance of each pigeon from the machine is recorded every hour for 8 hours.
- State the distance from the machine at 3 p.m. of:
    - Blue Crest
    - Green Tail
  - Describe the trend in the distance from the recording machine for:
    - Blue Crest
    - Green Tail
  - Assuming that the given trends continue, predict the time when the pigeons will be the same distance from the recording machine.



### Non-linear trends

9, 10

- 9 The balance of an investment account is shown in this time-series plot.
- Describe the trend in the account balance over the 7 years.
  - Give a practical reason for the shape of the curve that models the trend in the graph.
- 10 A drink at room temperature is placed in a fridge that is at  $4^{\circ}\text{C}$ .
- Sketch a time-series plot that might show the temperature of the drink after it has been placed in the fridge.
  - Would the temperature of the drink ever get to  $3^{\circ}\text{C}$ ? Why?
  - Record the temperature at regular intervals of a drink at room temperature that is placed in a fridge. Plot your results and compare them to your answer in part a.



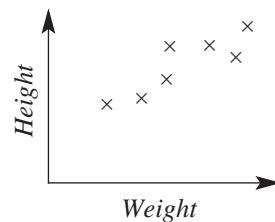
## 5H Bivariate data and scatter plots

### Learning intentions

- To know that bivariate data is data that involves two variables
- To be able to construct a scatterplot for bivariate data
- To be able to describe the correlation of the data in a scatter plot

**Key vocabulary:** bivariate data, scatter plot, correlation, association, outlier

When we collect information about two variables in a given context we are collecting bivariate data. As there are two variables involved in bivariate data, we use a number plane to graph the data. These graphs are called scatter plots and are used to show a relationship that may exist between the variables. Scatter plots make it very easy to see the strength of the relationship between the two variables.

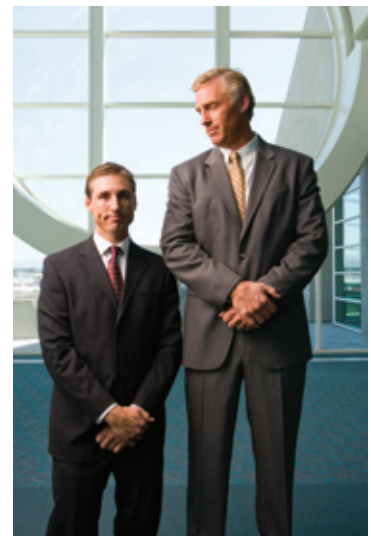


### → Lesson starter: A relationship or not?

Consider the two variables in each part below.

- Would you expect there to be some relationship between the two variables in each of these cases?
- If you feel that a relationship exists, would you expect the second-listed variable to increase or to decrease as the first variable increases?

- Height of person and Weight of person*
- Temperature and Life of milk*
- Length of hair and IQ*
- Depth of topsoil and Brand of motorcycle*
- Years of education and Income*
- Spring rainfall and Crop yield*
- Size of ship and Cargo capacity*
- Fuel economy and CD track number*
- Amount of traffic and Travel time*
- Cost of 2 litres of milk and Ability to swim*
- Background noise and Amount of work completed*



### Key ideas

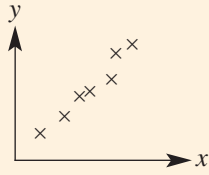
- **Bivariate data** is data that involves two variables.
  - The two variables are usually related; for example, height and weight.
- A **scatter plot** is a graph on a number plane in which the axes variables correspond to the two variables from the bivariate data. Points are marked with a cross.
- The words *relationship*, **correlation** and **association** are used to describe the way in which the variables are related.

■ Types of correlation:

- The correlation is positive if the  $y$  variable generally increases as the  $x$  variable increases.
- The correlation is negative if the  $y$  variable generally decreases as the  $x$  variable increases.

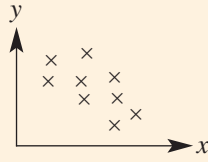
Examples:

Strong positive correlation



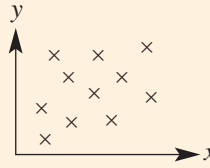
As  $x$  increases,  $y$  clearly increases.

Weak negative correlation



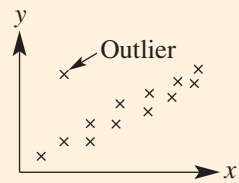
As  $x$  increases,  $y$  generally decreases.

No correlation



As  $x$  increases, there is no particular effect on  $y$ .

- An outlier can be clearly identified as a data point that is isolated from the rest of the data.



## Exercise 5H

### Understanding

1–3

3

- 1 Decide whether it is likely or unlikely that there will be a strong relationship between these pairs of variables.

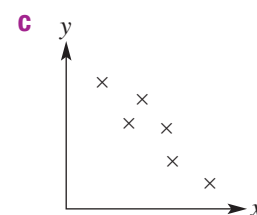
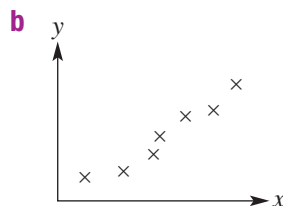
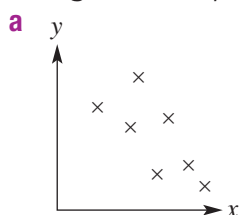
- Height of door and Width of door
- Weight of car and Fuel consumption
- Temperature and Length of phone calls
- Colour of flower and Strength of perfume
- Amount of rain and Size of vegetables in the vegetable garden

- 2 Complete the following using the word *increases* or *decreases*.

- For a positive correlation the  $y$  variable generally \_\_\_\_\_ as the  $x$  variable increases.
- For a negative correlation the  $y$  variable generally \_\_\_\_\_ as the  $x$  variable increases.

- 3 For these scatter plots, choose two words from those listed below to best describe the correlation between the two variables.

*strong weak positive negative*





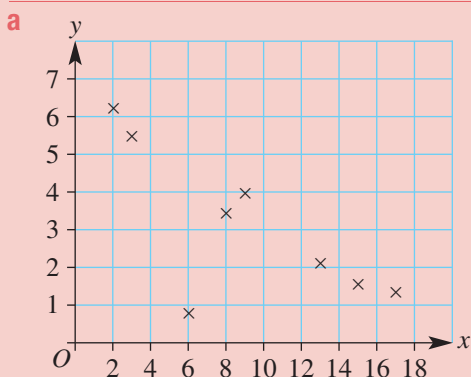
### Example 16 Constructing and interpreting scatter plots

Consider this simple bivariate data set.

|     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $x$ | 13  | 9   | 2   | 17  | 3   | 6   | 8   | 15  |
| $y$ | 2.1 | 4.0 | 6.2 | 1.3 | 5.5 | 0.9 | 3.5 | 1.6 |

- Draw a scatter plot for the data.
- Describe the correlation between  $x$  and  $y$  as positive or negative.
- Describe the correlation between  $x$  and  $y$  as strong or weak.
- Identify any outliers.

#### Solution



#### Explanation

Draw an appropriate scale on each axis by looking at the data:

- $x$  is up to 17
- $y$  is up to 6.2

The scale must be spread evenly on each axis. Plot each point using a cross symbol on graph paper.

- Negative correlation  
Looking at the scatter plot, as  $x$  increases  $y$  decreases.
- Strong correlation  
The downwards trend in the data is clearly defined.
- The outlier is (6, 0.9).  
This point defies the trend.

#### Now you try

Consider this simple bivariate data set.

|     |    |   |   |   |   |   |   |    |
|-----|----|---|---|---|---|---|---|----|
| $x$ | 9  | 6 | 4 | 3 | 8 | 2 | 1 | 10 |
| $y$ | 11 | 8 | 4 | 4 | 3 | 2 | 3 | 12 |

- Draw a scatter plot for the data.
- Describe the correlation between  $x$  and  $y$  as positive or negative.
- Describe the correlation between  $x$  and  $y$  as strong or weak.
- Identify any outliers.

- 4 Consider this simple bivariate data set.

|          |     |     |     |     |     |     |     |     |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|
| <b>x</b> | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
| <b>y</b> | 1.0 | 1.1 | 1.3 | 1.3 | 1.4 | 1.6 | 1.8 | 1.0 |

- Draw a scatter plot for the data.
- Describe the correlation between  $x$  and  $y$  as positive or negative.
- Describe the correlation between  $x$  and  $y$  as strong or weak.
- Identify any outliers.

- 5 Consider this simple bivariate data set.

|          |    |     |     |     |     |     |   |   |    |
|----------|----|-----|-----|-----|-----|-----|---|---|----|
| <b>x</b> | 14 | 8   | 7   | 10  | 11  | 15  | 6 | 9 | 10 |
| <b>y</b> | 4  | 2.5 | 2.5 | 1.5 | 1.5 | 0.5 | 3 | 2 | 2  |

- Draw a scatter plot for the data.
- Describe the correlation between  $x$  and  $y$  as positive or negative.
- Describe the correlation between  $x$  and  $y$  as strong or weak.
- Identify any outliers.

- 6 By completing scatter plots for each of the following data sets, describe the correlation between  $x$  and  $y$  as *positive*, *negative* or *none*.

**a**

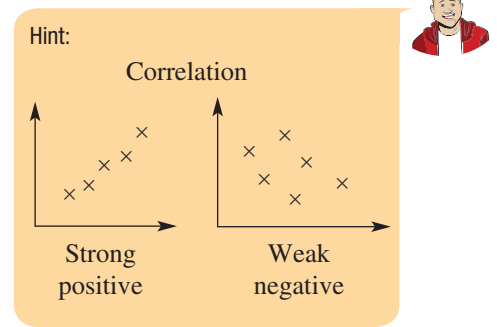
|          |     |     |     |     |     |     |     |     |     |     |     |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| <b>x</b> | 1.1 | 1.8 | 1.2 | 1.3 | 1.7 | 1.9 | 1.6 | 1.6 | 1.4 | 1.0 | 1.5 |
| <b>y</b> | 22  | 12  | 19  | 15  | 10  | 9   | 14  | 13  | 16  | 23  | 16  |

**b**

|          |     |     |     |     |     |     |     |     |     |     |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| <b>x</b> | 4   | 3   | 1   | 7   | 8   | 10  | 6   | 9   | 5   | 5   |
| <b>y</b> | 115 | 105 | 105 | 135 | 145 | 145 | 125 | 140 | 120 | 130 |

**c**

|          |    |    |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|----|----|
| <b>x</b> | 28 | 32 | 16 | 19 | 21 | 24 | 27 | 25 | 30 | 18 |
| <b>y</b> | 13 | 25 | 22 | 21 | 16 | 9  | 19 | 25 | 15 | 12 |



## Problem-solving and reasoning

7, 8

7, 9, 10

- 7 A tomato grower experiments with a new organic fertiliser and sets up five separate garden beds: A, B, C, D and E. The grower applies different amounts of fertiliser to each bed and records the diameter of each tomato picked.

The average diameter of a tomato from each garden bed and the corresponding amount of fertiliser are recorded below.

| Bed                         | A   | B   | C   | D   | E   |
|-----------------------------|-----|-----|-----|-----|-----|
| Fertiliser (grams per week) | 20  | 25  | 30  | 35  | 40  |
| Average diameter (cm)       | 6.8 | 7.4 | 7.6 | 6.2 | 8.5 |

- Draw a scatter plot for the data with 'Diameter' on the vertical axis and 'Fertiliser' on the horizontal axis. Label the points A, B, C, D and E.
- Which garden bed appears to go against the trend?
- According to the given results, would you be confident saying that the amount of fertiliser fed to tomato plants does affect the size of the tomato produced?



## 5H

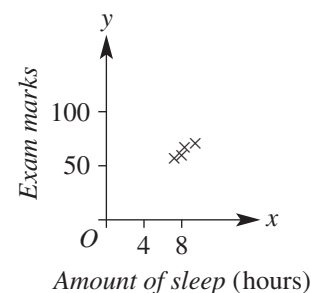
- 8 For common motor vehicles, consider the two variables *Engine size* (cylinder volume) and *Fuel economy* (number of kilometres travelled for every litre of petrol).
- Do you expect there to be some relationship between these two variables?
  - As the engine size increases, would you expect the fuel economy to increase or decrease?
  - The following data were collected for 10 vehicles.

| Car          | A   | B   | C   | D   | E   | F   | G   | H   | I   | J   |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Engine size  | 1.1 | 1.2 | 1.2 | 1.5 | 1.5 | 1.8 | 2.4 | 3.3 | 4.2 | 5.0 |
| Fuel economy | 21  | 18  | 19  | 18  | 17  | 16  | 15  | 20  | 14  | 11  |

- Do the data generally support your answers to parts **a** and **b** above?
  - Which car gives a fuel economy reading that does not support the general trend?
- 9 On 14 consecutive days a local council measures the volume of sound heard from a freeway at various points in a local suburb. The volume ( $V$ ) of sound, in decibels, is recorded against the distance ( $d$ ), in metres, between the freeway and the point in the suburb.

|     |     |     |     |     |      |     |     |     |     |     |     |      |     |      |
|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|------|-----|------|
| $d$ | 200 | 350 | 500 | 150 | 1000 | 850 | 200 | 450 | 750 | 250 | 300 | 1500 | 700 | 1250 |
| $V$ | 4.3 | 3.7 | 2.9 | 4.5 | 2.1  | 2.3 | 4.4 | 3.3 | 2.8 | 4.1 | 3.6 | 1.7  | 3.0 | 2.2  |

- Draw a scatter plot of  $V$  against  $d$ , plotting  $V$  on the vertical axis and  $d$  on the horizontal axis.
  - Describe the correlation between  $d$  and  $V$  as positive, negative or none.
  - Generally, as  $d$  increases, does  $V$  increase or decrease?
- 10 A person presents you with this scatter plot and suggests to you that there is a strong correlation between the amount of sleep and exam marks. What do you suggest is the problem with the person's graph and conclusions?



### Crime rates and police

—

11

- 11 A government department is interested in convincing the electorate that a large number of police on patrol leads to lower crime rates. Two separate surveys are completed over a 1-week period and the results are listed in the table below.

|          | Area               | A  | B  | C  | D  | E  | F  | G  |
|----------|--------------------|----|----|----|----|----|----|----|
| Survey 1 | Number of police   | 15 | 21 | 8  | 14 | 19 | 31 | 17 |
|          | Incidence of crime | 28 | 16 | 36 | 24 | 24 | 19 | 21 |
| Survey 2 | Number of police   | 12 | 18 | 9  | 12 | 14 | 26 | 21 |
|          | Incidence of crime | 26 | 25 | 20 | 24 | 22 | 23 | 19 |

- Using scatter plots, determine whether or not there is a relationship between the number of police on patrol and the incidence of crime, using the data in:
  - survey 1
  - survey 2
- Which survey results do you think the government will use to make its point? Why?

Hint: *Number of police* will be on the horizontal axis.



#### Using technology 5H: Using calculators to draw scatter plots

This activity is available on the companion website as a printable PDF.

## 51 Line of best fit by eye

### Learning intentions

- To know that a straight line can be fitted by eye to bivariate data with a positive or negative linear correlation
- To know how to fit a line of best fit by eye
- To be able use a line of best fit to find unknown points using both interpolation and extrapolation

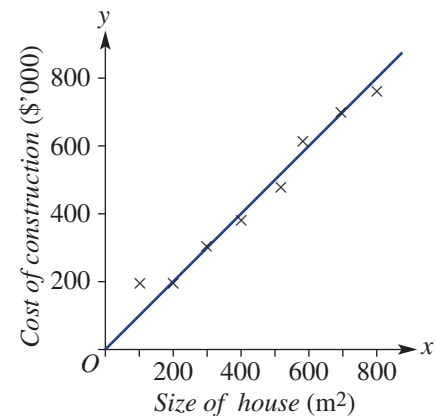
**Key vocabulary:** bivariate data, line of best fit (trend line), interpolation, extrapolation, linear, correlation

When bivariate data have a strong linear correlation, we can model the data with a straight line. This line is called a trend line or line of best fit. When we fit the line 'by eye', we try to balance the number of data points above the line with the number of points below the line. This trend line can then be used to construct other data points inside and outside the existing data points.

### Lesson starter: Size versus cost

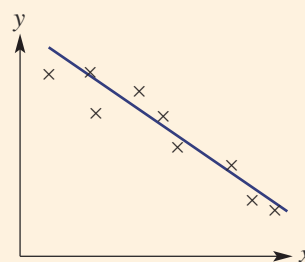
This scatter plot shows the estimated cost of building a house of a given size by a building company. A trend line has been added to the scatter plot.

- Why is it appropriate to fit a trend line to this data?
- Do you think the trend line is a good fit to the points on the scatter plot? Why?
- How can you predict the cost of a house of  $1000 \text{ m}^2$  with this building company?



### Key ideas

- For bivariate data showing a clearly defined positive or negative correlation, a straight line can be fitted by eye.
- A **line of best fit** (or trend line) is positioned by eye by balancing the number of points above the line with the number of points below the line.
  - The distance of each point from the trend line also needs to be taken into account.
  - Outliers should be ignored.
- The line of best fit can be used for:
  - **interpolation:** finding unknown points within the given data range
  - **extrapolation:** finding unknown points outside the given data range



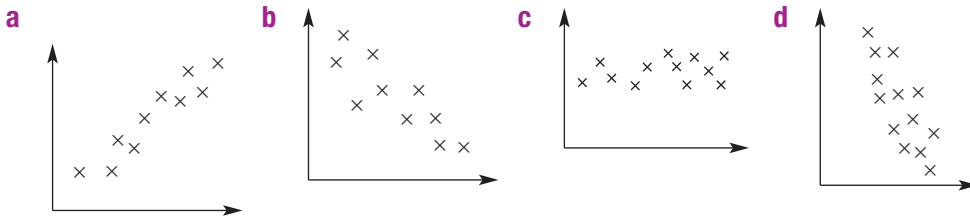
## Exercise 5I

### Understanding

1–3

2(½), 3

- When is it suitable to add a line of best fit to a scatterplot?
  - Describe the general guideline for placing a line of best fit.
- Practise fitting a line of best fit on these scatter plots by trying to balance the number of points above the line with the number of points below the line. (Using a pencil might help.)



- For the graph, with the given line of best fit shown, choose from *interpolation* or *extrapolation* to complete the following.

- Estimating the  $y$  value when  $x = 6$  is an example of \_\_\_\_\_.
- Estimating the  $y$  value when  $x = 10$  is an example of \_\_\_\_\_.



### Fluency

4–6

4, 6, 7



#### Example 17 Fitting a line of best fit

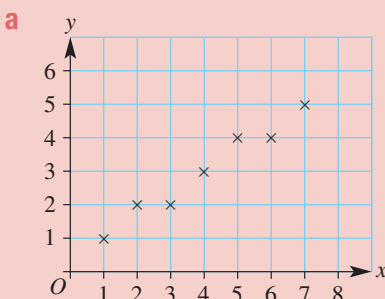
Consider the variables  $x$  and  $y$  and the corresponding bivariate data.

|     |   |   |   |   |   |   |   |
|-----|---|---|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $y$ | 1 | 2 | 2 | 3 | 4 | 4 | 5 |

- Draw a scatter plot for the data.
- Is there positive, negative or no correlation between  $x$  and  $y$ ?
- Fit a line of best fit by eye to the data on the scatter plot.
- Use your line of best fit to estimate:
  - $y$  when  $x = 3.5$
  - $y$  when  $x = 0$
  - $x$  when  $y = 1.5$
  - $x$  when  $y = 5.5$

#### Solution

#### Explanation

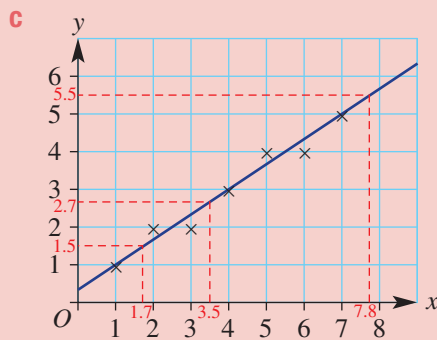


Plot the points on graph paper.

*Continued on next page*

**b** Positive correlation

As  $x$  increases,  $y$  increases.



Since a relationship exists, draw a line on the plot, keeping as many points above as below the line. (There are no outliers in this case.)

**d i**  $y \approx 2.7$

**ii**  $y \approx 0.4$

**iii**  $x \approx 1.7$

**iv**  $x \approx 7.8$

Start at  $x = 3.5$ . Draw a vertical line to the line of best fit, then draw a horizontal line to the  $y$ -axis and read off your solution.

Extend vertical and horizontal lines from the values given and read off your solution.

As they are approximations, we use the  $\approx$  sign and not the  $=$  sign.

**Now you try**

Consider the variables  $x$  and  $y$  and the corresponding bivariate data.

|          |    |   |   |   |   |   |
|----------|----|---|---|---|---|---|
| <b>x</b> | 1  | 2 | 3 | 4 | 5 | 6 |
| <b>y</b> | 10 | 7 | 6 | 6 | 5 | 3 |

**a** Draw a scatter plot for the data.

**b** Is there positive, negative or no correlation between  $x$  and  $y$ ?

**c** Fit a line of best fit by eye to the data on the scatter plot.

**d** Use your line of best fit to estimate:

**i**  $y$  when  $x = 4.5$

**ii**  $y$  when  $x = 0$

**iii**  $x$  when  $y = 7.5$

**iv**  $x$  when  $y = 1.5$

**4** Consider the variables  $x$  and  $y$  and the corresponding bivariate data.

|          |   |   |   |   |   |   |   |
|----------|---|---|---|---|---|---|---|
| <b>x</b> | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| <b>y</b> | 2 | 2 | 3 | 4 | 4 | 5 | 5 |

**a** Draw a scatter plot for the data.

**b** Is there positive, negative or no correlation between  $x$  and  $y$ ?

**c** Fit a line of best fit by eye to the data on the scatter plot.

**d** Use your line of best fit to estimate:

**i**  $y$  when  $x = 3.5$    **ii**  $y$  when  $x = 0$    **iii**  $x$  when  $y = 2$    **iv**  $x$  when  $y = 5.5$

**5** Consider the variables  $x$  and  $y$  and the corresponding data below.

|          |    |    |    |    |    |    |    |    |
|----------|----|----|----|----|----|----|----|----|
| <b>x</b> | 1  | 2  | 4  | 5  | 7  | 8  | 10 | 12 |
| <b>y</b> | 20 | 16 | 17 | 16 | 14 | 13 | 9  | 10 |

**a** Draw a scatter plot for the data.

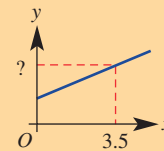
**b** Is there a positive, negative or no correlation between  $x$  and  $y$ ?

**c** Fit a line of best fit by eye to the data on the scatter plot.

**d** Use your line of best fit to estimate:

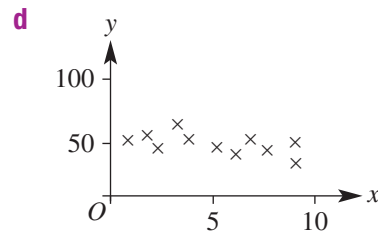
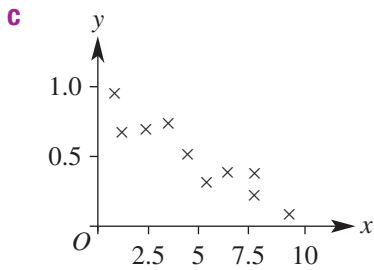
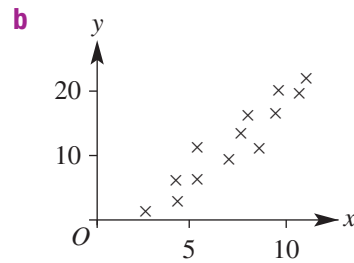
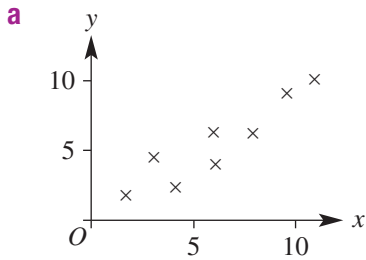
**i**  $y$  when  $x = 7.5$    **ii**  $y$  when  $x = 0$    **iii**  $x$  when  $y = 12$    **iv**  $x$  when  $y = 15$

Hint: Locate  $x = 3.5$  and read off the  $y$ -value for part **di**.



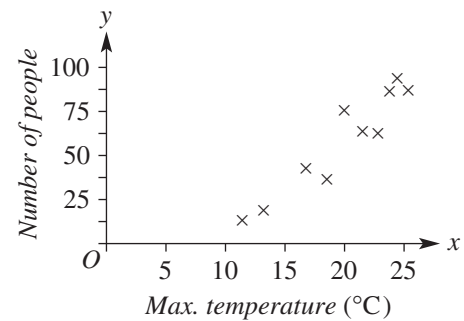
51

- 6 For the following scatter plots, pencil in a line of best fit by eye, and then use your line to estimate the value of  $y$  when  $x = 5$ .



- 7 This chart shows data for the *number of people* entering a suburban park and the corresponding *maximum temperature* for 10 spring days.

- a** Generally, as the maximum daily temperature increases, does the number of people who enter the park increase or decrease?
- b** Draw a line of best fit by eye on the given chart.
- c** Use your line of best fit to estimate:
- the number of people expected to enter the park if the maximum daily temperature is  $20^{\circ}\text{C}$
  - the maximum daily temperature when the total number of people who visit the park on a particular day is 25



## Problem-solving and reasoning

8, 9

8, 10

- 8 A small bookshop records its profit and number of customers for the past 8 days.

|                            |     |     |     |     |     |     |     |
|----------------------------|-----|-----|-----|-----|-----|-----|-----|
| <b>Number of customers</b> | 6   | 12  | 15  | 9   | 8   | 5   | 8   |
| <b>Profit (\$)</b>         | 200 | 450 | 550 | 300 | 350 | 250 | 300 |

- Draw a scatter plot for the data, using profit on the vertical axis.
- Fit a line of best fit by eye.
- Use your line of best fit to predict the profit for 17 customers.
- Use your line of best fit to predict the number of customers for a \$100 profit.

Hint: For 17 customers, you will need to extend your line beyond the data. This is called extrapolation.



- 9 Over eight consecutive years, a city nursery has measured the growth of an outdoor bamboo species for that year. The annual rainfall in the area where the bamboo is growing was also recorded. The data are listed in the table.

|                      |     |     |     |     |     |     |     |     |
|----------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| <b>Rainfall (mm)</b> | 450 | 620 | 560 | 830 | 680 | 650 | 720 | 540 |
| <b>Growth (cm)</b>   | 25  | 45  | 25  | 85  | 50  | 55  | 50  | 20  |

- Draw a scatter plot for the data, showing growth on the vertical axis.
- Fit a line of best fit by eye.
- Use your line of best fit to estimate the growth expected for the following rainfall readings.
  - 500 mm
  - 900 mm
- Use your line of best fit to estimate the rainfall for a given year if the growth of the bamboo is:
  - 30 cm
  - 60 cm







- 10 At a suburban sports club, the distance record for the hammer throw has increased over time. The first recorded value was 72.3 m in 1967. The most recent record was 118.2 m in 1996. Further details are in this table.

| Year           | 1967 | 1968 | 1969 | 1976 | 1978 | 1983  | 1987  | 1996  |
|----------------|------|------|------|------|------|-------|-------|-------|
| New record (m) | 72.3 | 73.4 | 82.7 | 94.2 | 99.1 | 101.2 | 111.6 | 118.2 |

- Draw a scatter plot for the data.
- Fit a line of best fit by eye.
- Use your line of best fit to estimate the distance record for the hammer throw for:
  - 2000
  - 2015
- Would you say that it is realistic to use your line of best fit to estimate distance records beyond 2015? Why?



### Heart rate and age

11

- 11 Two independent scientific experiments confirmed a correlation between *Maximum heart rate* and *Age*. The data for the two experiments are in this table.

| Experiment 1          |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Age (years)           | 15  | 18  | 22  | 25  | 30  | 34  | 35  | 40  | 40  | 52  | 60  | 65  | 71  |
| Max. heart rate (bpm) | 190 | 200 | 195 | 195 | 180 | 185 | 170 | 165 | 165 | 150 | 125 | 128 | 105 |
| Experiment 2          |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Age (years)           | 20  | 20  | 21  | 26  | 27  | 32  | 35  | 41  | 43  | 49  | 50  | 58  | 82  |
| Max. heart rate (bpm) | 205 | 195 | 180 | 185 | 175 | 160 | 160 | 145 | 150 | 150 | 135 | 140 | 90  |

- Sketch separate scatter plots for experiment 1 and experiment 2, with age on the horizontal axis.
- By fitting a line of best fit by eye to your scatter plots, estimate the maximum heart rate for a person aged 55 years, using the results from:
  - experiment 1
  - experiment 2
- Estimate the age of a person who has a maximum heart rate of 190 bpm, using the results from:
  - experiment 1
  - experiment 2
- For a person aged 25 years, which experiment estimates a lower maximum heart rate?
- Research the average maximum heart rate of people by age and compare with the results given above.



# Maths@Work: Project manager on a building site

Project managers can be responsible for coordinating a building site. They need to be organised, must be able to read and interpret plans, manage staff, coordinate the supply and costs of labour and equipment, as well as solve problems when unexpected situations arise on the worksite.

Project managers also need to evaluate their staff. Performance reviews are often used to evaluate staff performance. In some companies, bonuses and other incentives are linked to these reviews.

The performance review below is from a worksite, and used by the project manager to review their staff.



Use the performance review data in the table on the next page to answer the following questions that a project manager may face in their day-to-day job.

- 1 Consider trainee A.
  - a Calculate their mean score on this review.
  - b What is their median score?
  - c Write down the five-figure summary from the data given in this review.
- 2 Consider trainee B.
  - a Calculate their mean score on this review.
  - b What is their median score?
  - c Write down the five-figure summary from the data given in this review.
- 3 Display the data for all three trainees in parallel box plots on the same scale.
- 4 Compare the competency of trainees A, B and C by listing the following for *each* trainee. Then decide who is the most competent trainee, giving reasons.
  - a mean
  - b median
  - c lowest score
  - d highest score
  - e interquartile range
  - f range

Hint: Don't forget to order the data before finding the median.



Hint: The five-figure summary includes min,  $Q_1$ ,  $Q_2$ ,  $Q_3$  and max.



## Performance review

| Competency - please score 0 to 10   | Trainee A | Trainee B | Trainee C |
|---|-----------|-----------|-----------|
| 1. Please score the individual's punctuality when arriving to work.   | 9         | 8         | 8         |
| 2. Please score the individual's attitude when performing required tasks.   | 8.5       | 9         | 6.5       |
| 3. Once the individual has completed a task, do they ask for additional work?   | 8         | 8         | 5         |
| 4. How would you score the individual's willingness to learn?   | 9.5       | 8         | 6         |
| 5. How would you score the individual's application to competencies or requirements, as taught by the site team?  | 8         | 7         | 6         |
| 6. Please score the individual's accuracy when performing required tasks.   | 8.5       | 7.5       | 6.5       |
| 7. Please score the individual's skills in email, faxes and letter writing.   | 7         | 8         | 8.5       |
| 8. Please score the individual's skills in drafting site instructions.  | 7.5       | 7         | 5         |
| 9. Please score the individual's skills in drafting or issuing RFIs.  | 8         | 6.5       | 5         |
| 10. Please score the individual's respect earned within the site/department team for their required role.   | 8         | 7.5       | 5.5       |
| 11. Please score the individual's respect earned with subcontractors for their required role.   | 7.5       | 7         | 5.5       |
| 12. Please score the individual's respect earned with consultants for their required role.  | 8         | 7         | 5         |
| 13. How would you score the individual's application/focus to the required tasks you have assigned?   | 8.5       | 8         | 5         |
| 14. Do you consider that the individual has integrated well with the site/department team?  | 9         | 8         | 4.5       |
| 15. Do you find that the individual has improved their skills while under your supervision?   | 8         | 7.5       | 5.5       |
| 16. Would you take the individual to your next project to continue their competency training in other aspects of construction?<br>Yes - score 10    Maybe - score 5    No - score 0 | 10        | 5         | 0         |
| 17. How would you score the individual's progression through their required tasks?  | 9         | 8.5       | 3         |

## Using technology

5 The following data gives the competency scores (C) for trainee D.

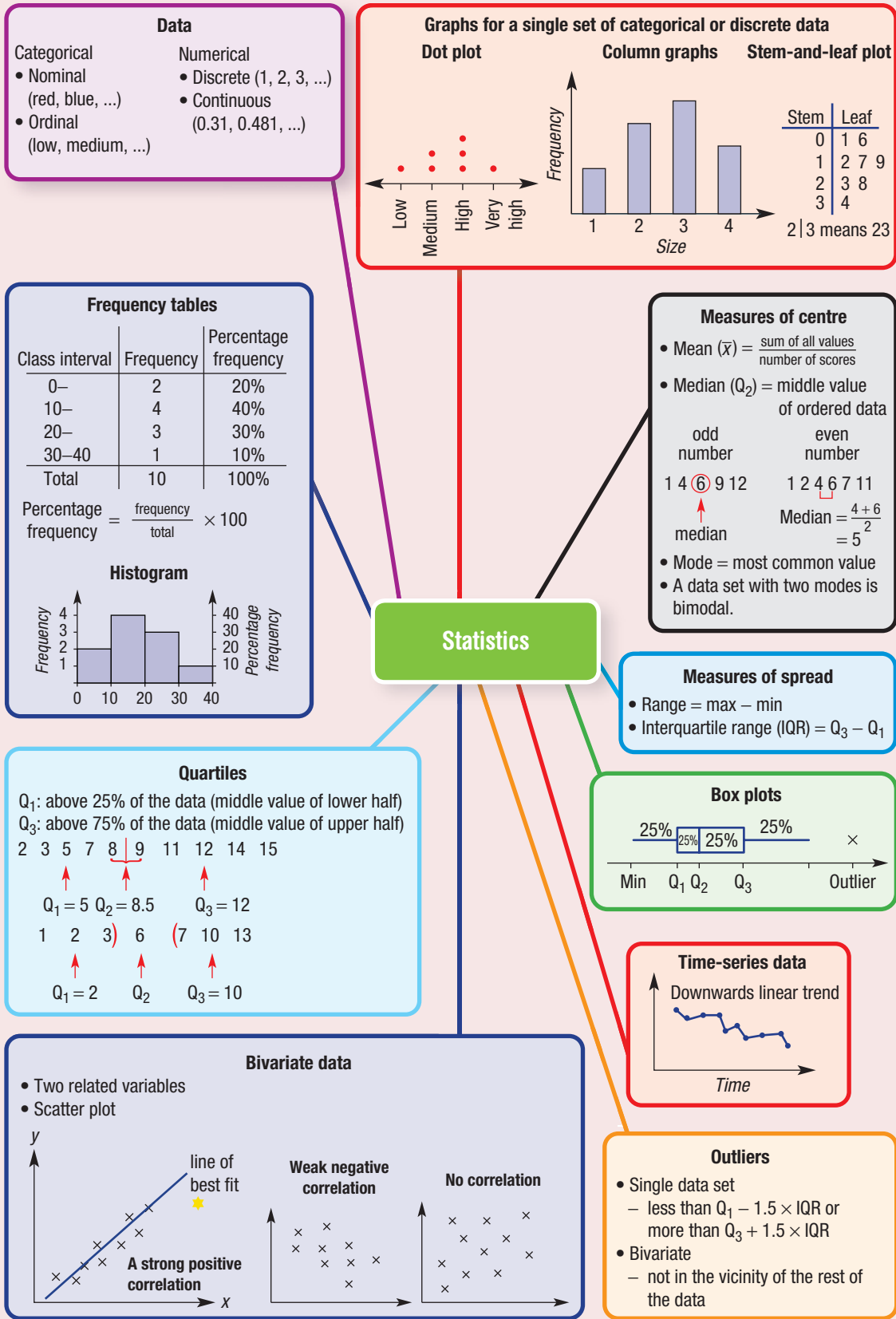
|      |      |      |      |      |      |      |      |     |
|------|------|------|------|------|------|------|------|-----|
| C 1  | C 2  | C 3  | C 4  | C 5  | C 6  | C 7  | C 8  | C 9 |
| 9    | 7    | 7    | 7    | 6    | 7    | 7    | 6    | 6   |
| C 10 | C 11 | C 12 | C 13 | C 14 | C 15 | C 16 | C 17 |     |
| 7    | 8    | 6    | 7    | 9    | 8    | 5    | 8    |     |

- Using a CAS calculator, enter the data for trainees B and D and draw parallel box plots. Note that an activity ('Using calculators to draw box plots') showing how to enter data and draw box plots for both the TI-Nspire and Class Pad is available in the interactive textbook.
- For trainees B and D, write down the five-figure summaries from the box plots.
- Compare the IQRs for trainee B and D. Which IQR shows more consistent competency? Give reasons for your answer.

- 1 The mean mass of six boys is 71 kg. The mean mass of five girls is 60 kg. Find the mean mass of all 11 children together.



- 2 Sean has a current four-topic average of 78% for Mathematics. What score does he need in the fifth topic to have an overall average of 80%?
- 3 I am a data set made up of five whole number values. My mode is 2 and both my mean and median are 5. What is my biggest possible range?
- 4 A single data set has 3 added to every value. Describe the change in the:
- a mean
  - b median
  - c range
- 5 Find the interquartile range for a set of data if 75% of the data is above 2.6 and 25% of the data is above 3.7.
- 6 I am a data set with four whole number values.
- I have a range of 8.
  - I have a mode of 3.
  - I have a median of 6.
- What are my four values?
- 7 A single-ordered data set includes the following data.  
2, 4, 5, 6, 8, 10,  $x$   
What is the largest possible value of  $x$  if it is not an outlier?
- 8 Describe what happens to the mean, median and mode of a data set if each value in the set is:
- a increased by 10
  - b multiplied by 10
- 9 At the end of 2019, the average rainfall in a town for the ten-year period just ended was 546 mm. A year later, the ten-year average was 562 mm and 654 mm had fallen in 2020. What was the rainfall in mm in 2010?



# Chapter checklist

A version of this checklist that you can print out and complete can be downloaded from your Interactive Textbook.

5A

## 1 I can describe the type of data generated by a survey question.

e.g. What type of data is generated by the survey questions, 'What is your favourite fruit?' and 'How many mobile phones are there in your household?'

5B

## 2 I can construct a frequency table and column graph.

e.g. Twenty students were surveyed on their favourite school subject. The results were:

|         |         |         |          |          |
|---------|---------|---------|----------|----------|
| Maths   | PE      | Science | Maths    | Language |
| Maths   | History | English | PE       | History  |
| Science | English | History | Language | Science  |
| Maths   | PE      | Maths   | History  | Maths    |

Construct a frequency table including the tally and frequency and construct a column graph for the data.

5B

## 3 I can work with a histogram.

e.g. A sample of 25 people leaving a movie are asked their age. The data are listed:

|    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|
| 16 | 19 | 22 | 23 | 42 | 44 | 36 | 41 | 28 |
| 26 | 28 | 34 | 52 | 55 | 46 | 20 | 22 |    |
| 34 | 35 | 48 | 46 | 50 | 29 | 26 | 35 |    |

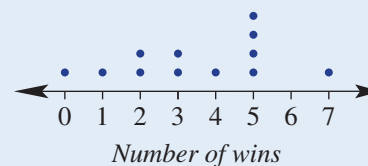
Organise the data into a frequency table, using class intervals of 10 and include a percentage frequency column. Construct a histogram showing both the frequency and percentage frequency and determine the percentage of people surveyed who were 30 years of age or older.

5C

## 4 I can interpret a dot plot.

e.g. This dot plot shows the number of wins by a tennis player in each of the grand slams they entered over 3 years.

- How many grand slams did they play in?
- What was the most common number of wins they had and the total number of wins in the 3 years?
- Describe the data in the dot plot.



5C

## 5 I can construct a stem-and-leaf plot or back-to-back stem-and-leaf plot.

e.g. Two sausage sizzles at two hardware stores sell the following number of sausages each week for a 10-week period.

| Store 1 |    |    |    |    | Store 2 |    |    |    |    |
|---------|----|----|----|----|---------|----|----|----|----|
| 32      | 44 | 28 | 36 | 52 | 27      | 35 | 32 | 43 | 52 |
| 56      | 31 | 45 | 42 | 47 | 34      | 29 | 24 | 38 | 22 |

Construct an ordered back-to-back stem-and-leaf plot and describe the distribution of each store's sausage sales.

5D

## 6 I can find the mean, mode and range.

e.g. For the data set: 12, 7, 11, 14, 3, 7, find the:  
**a** mode                      **b** mean                      **c** range

5D

## 7 I can find the median.

e.g. Find the median for the following data sets.

- 18, 14, 25, 28, 7, 11, 15
- 14, 19, 25, 16, 8, 1, 10, 30



5D

**8 I can calculate summary statistics from a graphical display.**

e.g. For the data in this stem-and-leaf plot, find the range, mode, median and mean.

| Stem  | Leaf     |
|-------|----------|
| 1     | 2 5      |
| 2     | 1 3 4 7  |
| 3     | 0 0 5    |
| 4     | 2        |
| 2   3 | means 23 |

5E

**9 I can find the upper and lower quartiles and the interquartile range of a data set.**

e.g. For the data sets listed below, find the upper quartile ( $Q_3$ ) and the lower quartile ( $Q_1$ ) and hence the IQR.

**a** 4, 8, 12, 18, 16, 20, 24, 19**b** 10.2, 11.4, 12.6, 9.8, 10.0, 15.6, 18

5E

**10 I can find the five-figure summary and outliers.**

e.g. The following data set represents the price of 10 different refrigerators in a department store.

|       |        |       |       |       |
|-------|--------|-------|-------|-------|
| \$620 | \$540  | \$642 | \$457 | \$585 |
| \$877 | \$1599 | \$918 | \$724 | \$840 |

For the data set, find:

**a** the five-figure summary    **b** the IQR**c** any outliers and give a possible reason why they are outliers

5F

**11 I can construct a box plot with and without outliers.**

e.g. For the data sets below, find the five-figure summary and whether any outliers exist. Draw a box plot to summarise the data, including outliers if they exist.

**a** 15 20 21 16 15 22 18 23 25 20**b** 2.9 3.5 2.6 4.1 3.2 2.4 1.8 3.1 4.4 9.2 5.6 3.0 2.8

5G

**12 I can plot and interpret a time-series graph.**

e.g. The approximate rainfall (in mm) for the 12 months of a year was recorded for Brisbane.

| Month         | J   | F   | M   | A   | M  | J  | J  | A  | S  | O  | N  | D   |
|---------------|-----|-----|-----|-----|----|----|----|----|----|----|----|-----|
| Rainfall (mm) | 110 | 125 | 115 | 130 | 95 | 70 | 60 | 40 | 30 | 80 | 95 | 105 |

Plot the time-series graph and describe the trend in the data over the 12 months.

5H

**13 I can construct and interpret a scatter plot.**

e.g. For the bivariate data set below, draw a scatter plot of the data, identify any outliers and describe the correlation between  $x$  and  $y$  choosing from the words: *positive*, *negative*, *weak* or *strong*.

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $x$ | 0.5 | 1.1 | 0.9 | 1.6 | 1.2 | 1.8 | 0.1 | 2.5 | 2.0 |
| $y$ | 8   | 7   | 6   | 6   | 7   | 9   | 10  | 1   | 3   |

5I

**14 I can fit a line of best fit by eye.**

e.g. Consider the bivariate data in the table.

|     |   |   |   |   |    |    |    |    |
|-----|---|---|---|---|----|----|----|----|
| $x$ | 1 | 2 | 4 | 5 | 8  | 10 | 12 | 15 |
| $y$ | 2 | 4 | 4 | 6 | 10 | 9  | 12 | 16 |

Draw a scatter plot for the data and fit a line of best fit by eye. Use the line of best fit to estimate:

**a**  $y$  when  $x = 6$ **b**  $x$  when  $y = 18$

## Short-answer questions

- 5B** 1 A group of 16 people was surveyed to find the number of hours of television they watch in a week. The raw data are listed:  
6, 5, 11, 13, 24, 8, 1, 12, 7, 6, 14, 10, 9, 16, 8, 3
- Organise the data into a table with class intervals of 5. Start at 0–, 5– etc. Include a tally, frequency and percentage frequency column.
  - Construct a histogram for the data, showing both the frequency and percentage frequency on the graph.
  - Would you describe the data as symmetrical or skewed?
- 5C** 2 A basketball team scores the following points per match for a season.  
20, 19, 24, 37, 42, 34, 38, 49, 28, 15, 38, 32, 50, 29
- Construct an ordered stem-and-leaf plot for the data.
  - Describe the distribution of scores.
- 5D** 3 For the following sets of data, determine:
- |  |                   |                     |                       |
|--|-------------------|---------------------|-----------------------|
|  | <b>i</b> the mean | <b>ii</b> the range | <b>iii</b> the median |
|--|-------------------|---------------------|-----------------------|
- a** 2, 7, 4, 8, 3, 6, 5  
**b** 10, 55, 67, 24, 11, 16  
**c** 1.7, 1.2, 1.4, 1.6, 2.4, 1.3
- 5D** 4 Thirteen adults compare their ages at a party. They are:  
40, 41, 37, 32, 48, 43, 32, 76, 29, 33, 26, 38, 87
- Find the mean age of the adults, to one decimal place.
  - Find the median age of the adults.
  - Why do you think the mean age is larger than the median age?

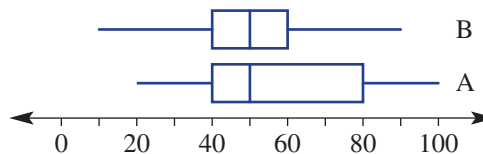


- 5E** 5 Determine  $Q_1$ ,  $Q_2$  and  $Q_3$  for these sets of data.
- 4, 5, 8, 10, 10, 11, 12, 14, 15, 17, 21
  - 14, 6, 2, 23, 11, 6, 15, 14, 12, 18, 16, 10
- 5E** 6 For the data set: 3, 7, 2, 10, 6, 21, 5, 9, 6, 2, 8, 10.
- Find the range and IQR of the data set
  - Find any outlier in the data set
  - Remove the outlier from the data set and find the new range and IQR. What do you notice?
- 5F** 7 For each set of data below, complete the following tasks.
- Find the lower quartile ( $Q_1$ ) and the upper quartile ( $Q_3$ ).
  - Find the interquartile range ( $IQR = Q_3 - Q_1$ ).
  - Locate any outliers.
  - Draw a box plot.
- 2, 2, 3, 3, 3, 4, 5, 6, 12
  - 11, 12, 15, 15, 17, 18, 20, 21, 24, 27, 28
  - 2.4, 0.7, 2.1, 2.8, 2.3, 2.6, 2.6, 1.9, 3.1, 2.2



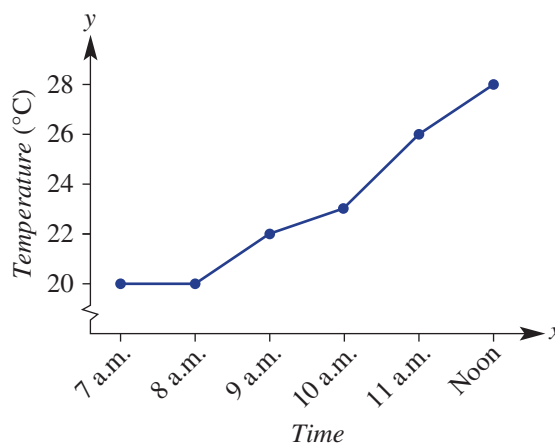
**5F** **8** Compare these parallel box plots, A and B, and answer the following as true or false.

- a** The range for A is greater than the range for B.
- b** The median for A is equal to the median for B.
- c** The interquartile range is smaller for B.
- d** 75% of the data for A sits below 80.

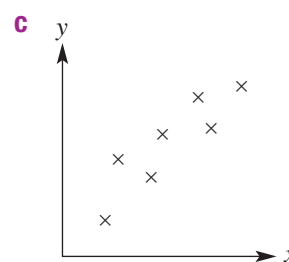
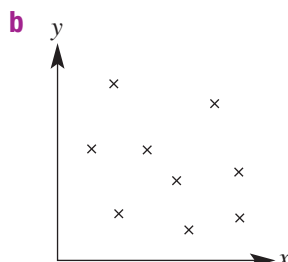
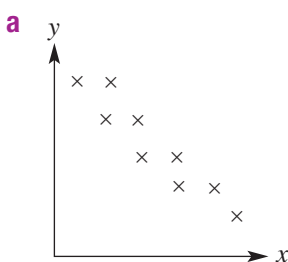


**5G** **9** This time series plot shows six temperature readings taken over 5 hours.

- a** Would you describe the trend as linear or non-linear?
- b** During which hour does the largest temperature increase occur?



**5H** **10** For the scatter plots below, describe the correlation between  $x$  and  $y$  as positive, negative or none.



**5H** **11** Consider the simple bivariate data set.

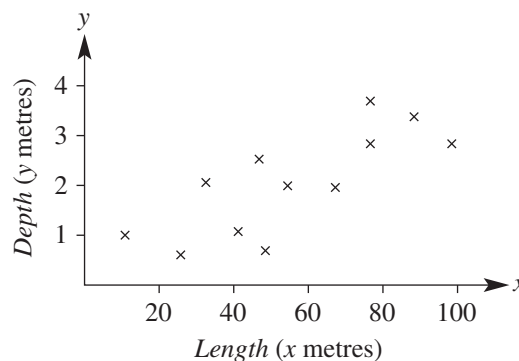
|     |    |    |    |    |    |    |   |    |   |   |
|-----|----|----|----|----|----|----|---|----|---|---|
| $x$ | 1  | 4  | 3  | 2  | 1  | 4  | 3 | 2  | 5 | 5 |
| $y$ | 24 | 15 | 16 | 20 | 22 | 11 | 5 | 17 | 6 | 8 |

- a** Draw a scatter plot for the data.
- b** Describe the correlation between  $x$  and  $y$  as positive or negative.
- c** Describe the correlation between  $x$  and  $y$  as strong or weak.
- d** Identify any outliers.

**5I** **12** The given scatter plot shows the maximum length ( $x$  metres) and depth ( $y$  metres) of 11 public pools around town.



- a** Draw a line of best fit by eye.
- b** Use your line to estimate the maximum depth of a pool that is 50 m in length.



## Multiple-choice questions

- 5A **1** The number of pets in 10 households are recorded after a survey is conducted. The type of data recorded could be described as:
- A** categorical and discrete                      **B** categorical and ordinal  
**C** categorical and nominal                      **D** numerical and discrete  
**E** numerical and continuous

Questions **2** and **3** refer to the stem-and-leaf plot below, at right.

- 5C **2** The minimum score in the data is:
- A** 4  
**B** 0  
**C** 24  
**D** 38  
**E** 54

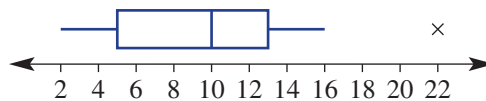
| Stem | Leaf       |
|------|------------|
| 2    | 4 9        |
| 3    | 1 1 7 8    |
| 4    | 2 4 6      |
| 5    | 0 4        |
| 4    | 2 means 42 |

- 5C **3** The mode is:
- A** 3  
**B** 31  
**C** 4  
**D** 38  
**E** 30

- 5D **4** The range and mean of 2, 4, 3, 5, 10 and 6 are:
- A** range = 8, mean = 5  
**B** range = 4, mean = 5  
**C** range = 8, mean = 4  
**D** range = 2 – 10, mean = 6  
**E** range = 8, mean = 6

- 5D **5** The median of 29, 12, 18, 26, 15 and 22 is:
- A** 18                      **B** 22                      **C** 20                      **D** 17                      **E** 26

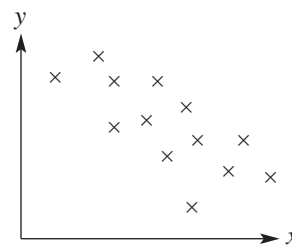
Questions **6–8** refer to the box plot below.



- 5E/F **6** The interquartile range (IQR) is:
- A** 8                      **B** 5                      **C** 3  
**D** 20                      **E** 14
- 5F **7** The outlier is:
- A** 2                      **B** 0                      **C** 20                      **D** 16                      **E** 22
- 5F **8** The median is:
- A** 2                      **B** 3                      **C** 10                      **D** 13                      **E** 16

5H 9 The variables  $x$  and  $y$  in this scatter plot could be described as having:

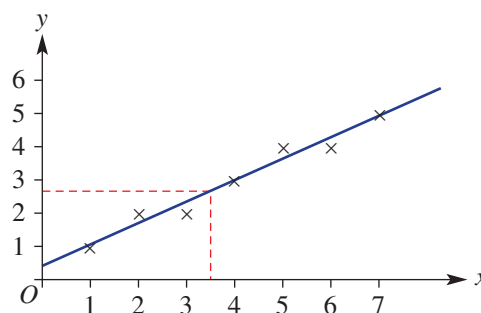
- A no correlation
- B strong positive correlation
- C strong negative correlation
- D weak negative correlation
- E weak positive correlation



5I 10 According to this scatter plot, when  $x$  is 3.5,  $y$  is approximately:



- A 4.4
- B 2.7
- C 2.5
- D 3.5
- E 5



### Extended-response questions

1 The number of flying foxes taking refuge in a fig tree is recorded over a period of 14 days. The data collected is given here.

|                               |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|-------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| <b>Number of flying foxes</b> | 73 | 50 | 36 | 82 | 15 | 24 | 73 | 57 | 65 | 86 | 51 | 32 | 21 | 39 |
|-------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|

- i Find the IQR.
- ii Identify any outliers.
- iii Draw a box plot for the data.



2 A newsagency records the *number of customers* and *profit* for 14 working days.

|                            |     |    |     |     |     |    |     |     |     |    |    |     |     |     |
|----------------------------|-----|----|-----|-----|-----|----|-----|-----|-----|----|----|-----|-----|-----|
| <b>Number of customers</b> | 18  | 13 | 15  | 24  | 29  | 12 | 18  | 16  | 15  | 11 | 4  | 32  | 26  | 21  |
| <b>Profit (\$)</b>         | 150 | 70 | 100 | 210 | 240 | 90 | 130 | 110 | 120 | 80 | 30 | 240 | 200 | 190 |

- a Draw a scatter plot for the data and draw a line of best fit by eye. Place *Number of customers* on the horizontal axis.
- b Use your line of best fit to predict the profit for:
  - i 10 customers
  - ii 20 customers
  - iii 30 customers
- c Use your line of best fit to predict the number of customers for a:
  - i \$50 profit
  - ii \$105 profit
  - iii \$220 profit