

CHAPTER

# 9

# Statistics

## Statistics helping the environment

To classify and map the mosaic of plant or animal life in any ecological community requires the collection and analysis of large quantities of data. To organise raw data, a spreadsheet matrix can be developed using the columns for sample sites, rows for species and an 'abundance score' (i.e. frequency count) in each cell. Statistical cluster analysis bunches similar samples to create a tree-like graph. Multi-dimensional scaling creates 2D or 3D drawings of points showing samples grouped according to similarities or differences.

An important use of statistics is to determine the significance or importance of a set of results. For example, a pollution incident in a stream might result in some species, such as fish, shrimps or crayfish, slowly dying out. A scientist can measure the diversity of life in the stream with regular netting and counting the various species caught. Diversity versus time can be graphed and a trend line drawn through the scattered points. A downward sloping trend line suggests a decline in diversity. Statistical analysis is used to determine if the deviation of data



## 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.

## In this chapter

- 9A Collecting and using data
- 9B Review of statistical graphs (CONSOLIDATING)
- 9C Summary statistics
- 9D Box plots
- 9E Standard deviation (10A)
- 9F Time-series data
- 9G Bivariate data and scatter plots
- 9H Line of best fit by eye
- 9I Linear regression using technology (10A)

## 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)

(10A) Calculate and interpret the mean and standard deviation of data and use these to compare data sets. Investigate the effect of individual data values including outliers, on the standard deviation (VCMSP372)

(10A) Use digital technology to investigate bivariate numerical data sets. Where appropriate use a straight line to describe the relationship allowing for variation, make predictions based on this straight line and discuss limitations (VCMSP373)

points from the trend line is small enough for the trend to be significant. Using the procedure, a scientist has discovered that a species of fairy shrimp, *Branchinella latzi*, is now extinct in the pools on Uluru (Ayers Rock) due to human waste pollution.



## 9A Collecting and using data

### Learning intentions

- To understand how surveys work and the necessary considerations for their construction
- To understand the difference between a population and a sample
- To know how to describe types of data using the key words: categorical (nominal or ordinal) or numerical (discrete or continuous)
- To be able to decide if a survey sample is representative

There are many reports on television and radio 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.

Sometimes the results of these surveys are used to persuade people to change their behaviour. Sometimes they are used to pressure the government into changing the laws or to change the way the government spends public money.

Results of surveys and other statistics can sometimes be misused or displayed in a way to present a certain point of view.



Niche marketing is when a product or service is advertised to a specific group, such as people who train for obstacle competitions or dog-owners who use luxury dog groomers. Surveys provide valuable data for niche marketing and sales.

### LESSON STARTER Improving 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 way 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?
- Do you think Question 2 could be improved somehow?
- What will the answers to Question 3 look like? How could they be displayed?
- Do you think Question 3 could be improved somehow?

## KEY IDEAS

■ **Surveys** are used to collect statistical data.

- Survey questions need to be constructed carefully so that the person knows exactly what sort of answer to give. Survey questions 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:

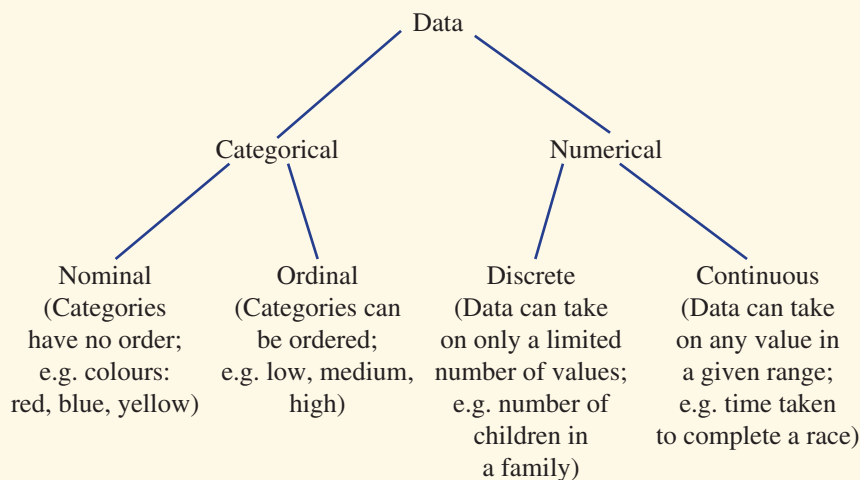
strongly agree	agree	unsure	disagree	strongly disagree
----------------	-------	--------	----------	-------------------

■ 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
- 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 Brisbane
- 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** can be divided into subgroups.



### BUILDING UNDERSTANDING

- 1** Match each word (a–e) with its definition (A–E).
- |                     |  |
|---------------------|--|
| <b>a</b> population | <b>A</b> a group chosen from a population                                |
| <b>b</b> census     | <b>B</b> a tool used to collect statistical data                         |
| <b>c</b> sample     | <b>C</b> all the people or objects in question                           |
| <b>d</b> survey     | <b>D</b> statistics collected from every member of the population        |
| <b>e</b> data       | <b>E</b> the factual information collected from a survey or other source |
- 2** Match each word (a–f) with its definition (A–F).
- |                      |   |
|----------------------|---|
| <b>a</b> numerical   | <b>A</b> categorical data that has no order                     |
| <b>b</b> continuous  | <b>B</b> data that are numbers                                  |
| <b>c</b> discrete    | <b>C</b> numerical data that take on a limited number of values |
| <b>d</b> categorical | <b>D</b> data that can be divided into categories               |
| <b>e</b> ordinal     | <b>E</b> numerical data that take any value in a given range    |
| <b>f</b> nominal     | <b>F</b> categorical data that can be ordered                   |
- 3** Classify each set of data as categorical or numerical.
- a** 4.7, 3.8, 1.6, 9.2, 4.8
- b** red, blue, yellow, green, blue, red
- c** low, medium, high, low, low, medium
- 4** 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 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?



### 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
- b** categorical and nominal

#### EXPLANATION

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

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 tall are the students in Year 10?
- b** What is your level of satisfaction (low, medium and high) with a meal at a restaurant?



## Example 2 Choosing a survey sample

A survey is carried out on the internet to determine Australia's favourite musical performer. Why will this sample not necessarily be representative of Australia's views?

### SOLUTION

An internet survey is restricted to people with a computer and internet access, ruling out some sections of the community from participating in the survey.

### EXPLANATION

The sample may not include some of the older members of the community or those in areas without access to the internet. Also, the survey would need to be set up so that people can do it only once so that 'fake' surveys are not completed.

### Now you try

A survey is carried out in a library to determine typical study habits of Year 12 students. Why will this sample not necessarily be representative of all Year 12 students?

## Exercise 9A

### FLUENCY

1-3

2, 3

2, 3

Example 1

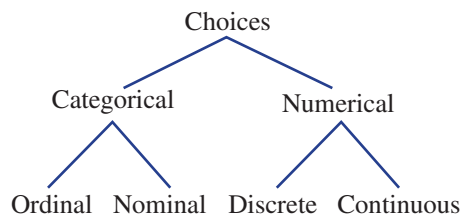
1 What type of data would the following survey questions generate?

- How many people are there in each office room?
- What was the time taken to complete the task?
- What colour are the jackets on a rack?
- How would you rate the movie: good, don't care, bad?

Example 1

2 Year 10 students were asked the following questions in a survey. Describe what type of data each question generates.

- How many people under the age of 18 years are there in your immediate family?
- How many letters are there in your first name?
- Which company is the carrier of your mobile telephone calls? Optus/Telstra/Vodafone/Virgin/Other (Please specify.)
- What is your height?
- How would you describe your level of application in Maths? (Choose from very high, high, medium or low.)



Example 2

3 Decide if the following surveys would be representative of the entire Australian population:

- a survey via social media to find out people's favourite news program
- a survey to find out the average number of pets in a household from people entering a pet store
- using census data to determine the average household income
- making 10000 random phone calls to find out who is likely to win the next federal election

## PROBLEM-SOLVING

4, 5

4–6

5–7

- 4 The principal decides to survey Year 10 students to determine their opinion of Mathematics.
- a 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.
- b Explain your choice of answer in part a. Describe what is wrong with the other three options.
- 5 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.



Is a train station survey of how people get to work representative?

- 6 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 people 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.

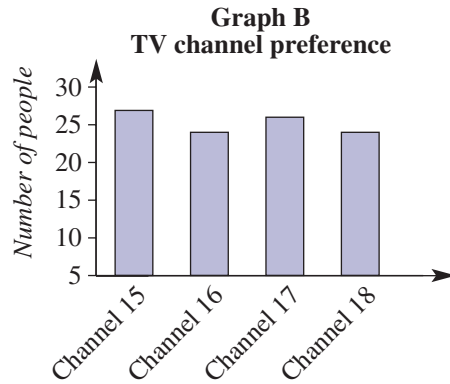
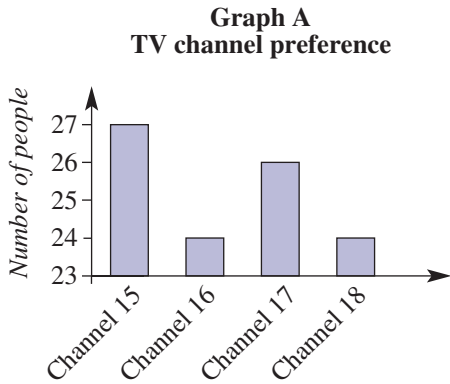
Question 3 must produce data that are numerical and discrete.

Question 4 must produce data that are numerical and continuous.

- 7 A television news reporter surveyed four companies and found that the profits of three of these companies had reduced over the past year. They report that this means the country is facing an economic downturn and that only one in four companies is making a profit.
- a What are some of the problems in this media report?
  - b How could the news reporter improve their sampling methods?
  - c Is it correct to say that only one in four companies is making a profit? Explain.

**REASONING** 8 8, 9 9, 10

8 Here are two column graphs, each showing the same results of a survey that asked people which TV channel they preferred.



- a Which graph could be titled ‘Channel 15 is clearly most popular’?
  - b Which graph could be titled ‘All TV channels have similar popularity’?
  - c What is the difference between the two graphs?
  - d Which graph is misleading and why?
- 9 Describe three ways that graphs or statistics could be used to mislead people and give a false impression about the data.
- 10 Search the internet or newspaper for ‘misleading graphs’ and ‘how to lie with statistics’. Explain why they are misleading.

**ENRICHMENT: The 2016 Australian Census** – – 11, 12

- 11 Research the 2016 Australian Census on the website of the Australian Bureau of Statistics. Find out something interesting from the results of the 2016 Australian Census and write a short news report.
- 12 It is often said that Australia has an ageing population. What does this mean? Search the internet for evidence showing that the ‘average’ Australian is getting older every year.



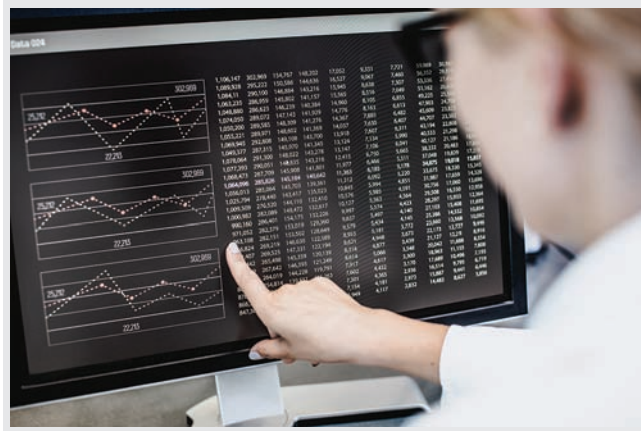


## 9B Review of statistical graphs CONSOLIDATING

### Learning intentions

- To review the types of graphs that can be used to display categorical data or numerical data
- To know how to construct a frequency table and histogram from numerical data using class intervals
- To know how to find the measures of centre, mean and median, of a set of data

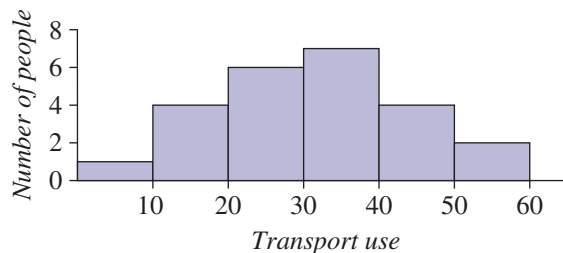
Statistical graphs are an essential element in the analysis and representation of data. Graphs can help to show the most frequent category, the range of values, the shape of the distribution and the centre of the data. By looking at statistical graphs the reader can quickly draw conclusions about the numbers or categories in the data set and interpret this within the context of the data.



People who specialise in medical biostatistics apply statistical techniques to analyse results from health-related research, such as in genetics, medicine and pharmacy. Data presentation includes using bar charts, line charts, histograms and scatter plots.

### LESSON STARTER Public transport

A survey was carried out to find out how many times people in the 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:

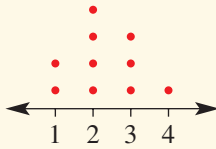
- How many people were surveyed?
- Is the data symmetrical or skewed?
- Is it possible to work out the exact mean? Why/why not?
- Do you think these people were selected from a group in your own community? Give reasons.

**KEY IDEAS**

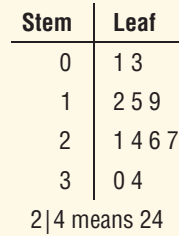
■ The types of **statistical data** that we saw in the previous section; i.e. categorical (nominal or ordinal) and numerical (discrete or continuous), can be displayed using different types of graphs to represent the different data.

■ Graphs for a single set of categorical or discrete data

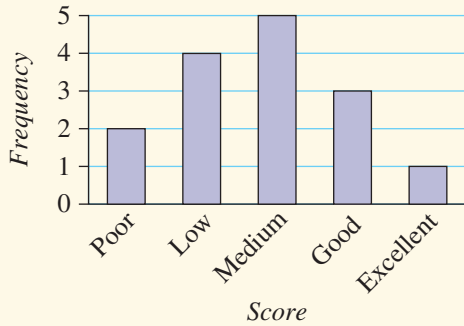
• **Dot plot**



• **Stem-and-leaf plot**

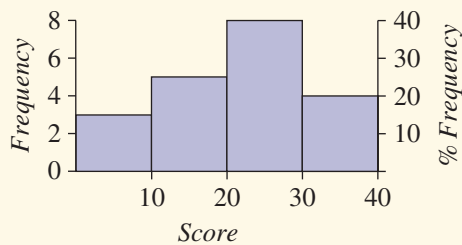


• **Column graph**



■ **Histograms** can be used for grouped discrete or continuous numerical data. The interval 10– includes all numbers from 10 (including 10) to fewer than 20.

Class interval	Frequency	Percentage frequency
0–	3	15
10–	5	25
20–	8	40
30–40	4	20

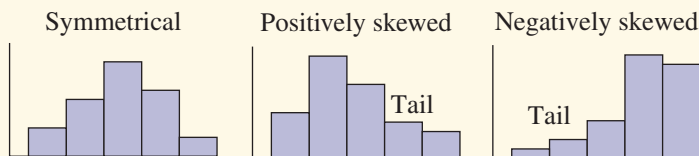


■ Measures of centre include:

- **mean** ( $\bar{x}$ )  $\bar{x} = \frac{\text{sum of all data values}}{\text{number of data values}}$
- **median** the middle value when data are placed in order

■ The **mode** of a data set is the most common value.

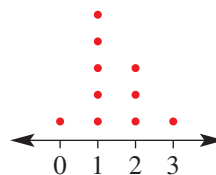
■ Data can be **symmetrical** or **skewed**.





## BUILDING UNDERSTANDING

- 1 A number of families were surveyed to find the number of children in each. The results are shown in this dot plot.
- How many families were surveyed?
  - Find the mean number of children in the families surveyed.
  - State the median number of children in the families surveyed.
  - State the mode for the number of children in the families surveyed.
  - What percentage of the families have, at most, two children?
- 2 State the missing values in this frequency table.



Class interval	Frequency	Percentage frequency
0–	2	
10–	1	
20–	5	
30–40	2	
<b>Total</b>		



## Example 3 Presenting and analysing data

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

- 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.
- Construct a stem-and-leaf plot for the data.
- Use your stem-and-leaf plot to find the median.

## SOLUTION

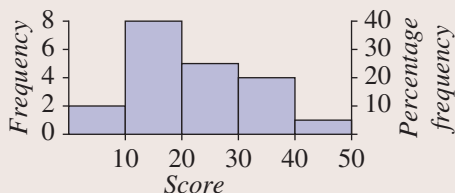
a

Class interval	Frequency	Percentage frequency
0–	2	10
10–	8	40
20–	5	25
30–	4	20
40–50	1	5
<b>Total</b>	<b>20</b>	<b>100</b>

## EXPLANATION

Calculate each percentage frequency by dividing the frequency by the total (i.e. 20) and multiplying by 100.

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



Transfer the data from the frequency table to the histogram. Axis 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**

Stem	Leaf
0	2 5
1	0 1 3 4 5 6 8 9
2	1 4 5 8 9
3	1 1 2 3
4	0

3|1 means 31

Order the data in each leaf and also show a key (e.g. 3|1 means 31).

**d** Median =  $\frac{19 + 21}{2}$   
= 20

After counting the scores in order from the lowest value (i.e. 2), the two middle values are 19 and 21, so the median is the mean of these two numbers.

### Now you try

Sixteen people were surveyed to find out how many phone texts they send in one day. The raw data are as follows.

10, 7, 2, 5, 22, 14, 7, 9, 11, 29, 32, 18, 5, 24, 12, 14

- 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.
- Construct a stem-and-leaf plot for the data.
- Use your stem-and-leaf plot to find the median.

## Exercise 9B

### FLUENCY

1-4

1, 3, 4

1, 3, 4

Example 3

- 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 and include a percentage frequency column.
  - Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
  - Construct a stem-and-leaf plot for the data.
  - Use your stem-and-leaf plot to find the median.

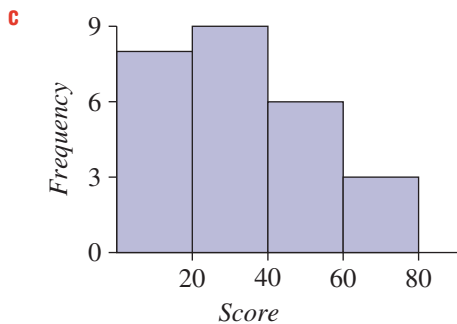
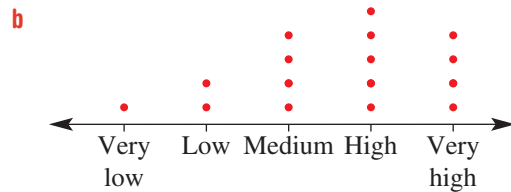
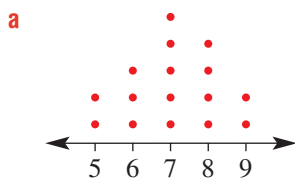


2 This frequency table displays the way in which 40 people travel to and from work.

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

3 Describe each graph as symmetrical, positively skewed or negatively skewed.



d

Stem	Leaf
4	1 6
5	0 5 4 8
6	1 8 9 9 9
7	2 7 8
8	3 8
4   6	means 46



- 4 For the data in these stem-and-leaf plots, find:
- i the mean (rounded to one decimal place)
  - ii the median
  - iii the mode

a

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

b

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

**PROBLEM-SOLVING** 5, 6 6, 7 7, 8

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

<b>Game</b>	1	2	3	4	5	6	7	8	9	10	11	12
<b>Nick</b>	0	2	2	0	3	1	2	1	2	3	0	1
<b>Jack</b>	0	0	4	1	0	5	0	3	1	0	4	0


- a Draw a dot plot to display Nick’s goal-scoring achievement.
  - b Draw a dot plot to display Jack’s goal-scoring achievement.
  - c How would you describe Nick’s scoring habits?
  - d How would you describe Jack’s scoring habits?
- 6 Three different electric sensors, A, B and C, are used to detect movement in Harvey’s backyard over a period of 3 weeks. An in-built device counts the number of times the sensor detects movement each night. The results are as follows.

<b>Day</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
<b>Sensor A</b>	0	0	1	0	0	1	1	0	0	2	0	0	0	0	0	1	1	0	0	1	0
<b>Sensor B</b>	0	15	1	2	18	20	2	1	3	25	0	0	1	15	8	9	0	0	2	23	2
<b>Sensor C</b>	4	6	8	3	5	5	5	4	8	2	3	3	1	2	2	1	5	4	0	4	9

- a Using class intervals of 3 and starting at 0, draw up a frequency table for each sensor.
- b Draw histograms for each sensor.
- c Given that it is known that stray cats consistently wander into Harvey’s backyard, how would you describe the performance of:
  - i sensor A?
  - ii sensor B?
  - iii sensor C?



Possoms could set off the sensors.

 7 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–	
30–35	





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

Section	Tally
String	
Woodwind	
Brass	
Percussion	

- Construct and complete a percentage frequency table for the data.
- What is the total number of students in the school orchestra?
- What percentage of students play in the string section?
- What percentage of students do not play in the string section?
- If the number of students in the string section increases by 3, what will be the percentage of students who play in the percussion section? Round your answer to one decimal place.
- What will be the percentage of students in the string section of the orchestra if the entire woodwind section is absent? Round your answer to one decimal place.



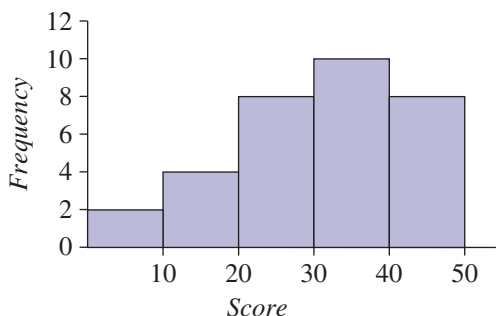
### REASONING

9

9, 10

10, 11

9 This histogram shows the distribution of test scores for a class. Explain why the percentage of scores in the 20–30 range is 25%.



- Explain why the exact value of the mean, median and mode cannot be determined directly from a histogram.
- State the possible values of  $a$ ,  $b$  and  $c$  in this ordered stem-and-leaf plot.

Stem	Leaf
3	2 3 $a$ 7
4	$b$ 4 8 9 9
5	0 1 4 9 $c$
6	2 6

## ENRICHMENT: Cumulative frequency curves and percentiles

12

- 12** Cumulative frequency is obtained by adding a frequency to the total of its predecessors. It is sometimes referred to as a 'running total'.

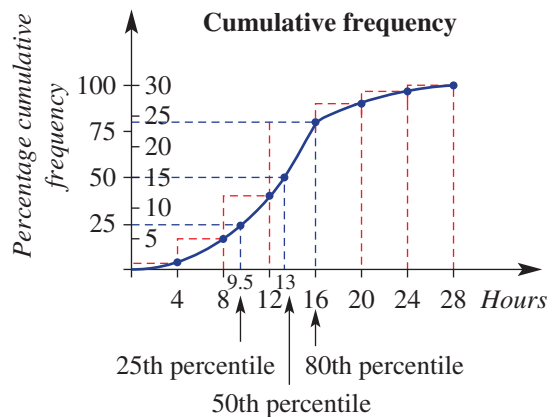
$$\text{Percentage cumulative frequency} = \frac{\text{cumulative frequency}}{\text{total number of data elements}} \times 100$$

A cumulative frequency graph is one in which the heights of the columns are proportional to the corresponding cumulative frequencies.

The points in the upper right-hand corners of these rectangles join to form a smooth curve called the cumulative frequency curve.

If a percentage scale is added to the vertical axis, the same graph can be used as a percentage cumulative frequency curve, which is convenient for the reading of percentiles.

Number of hours	Frequency	Cumulative frequency	Percentage cumulative frequency
0–	1	1	3.3
4–	4	5	16.7
8–	7	12	40.0
12–	12	24	80.0
16–	3	27	90.0
20–	2	29	96.7
24–28	1	30	100.0



The following information relates to the amount, in dollars, of winter gas bills for houses in a suburban street.

Amount (\$)	Frequency	Cumulative frequency	Percentage cumulative frequency
0–	2		
40–	1		
80–	12		
120–	18		
160–	3		
200–240	1		

- Copy and complete the table. Round the percentage cumulative frequency to one decimal place.
- Find the number of houses that have gas bills of less than \$120.
- Construct a cumulative frequency curve for the gas bills.
- Estimate the following percentiles.
  - 50th
  - 20th
  - 80th
- In this street, 95% of households pay less than what amount?
- What percentage of households pay less than \$100?



## Using calculators to graph grouped data

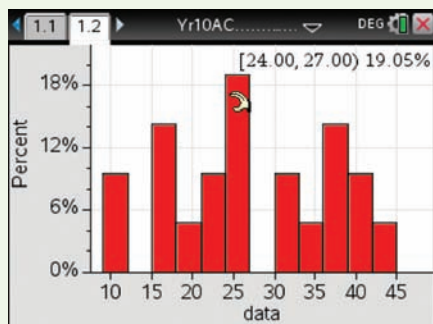
- 1 Enter the following data in a list called *data* and find the mean and median.  
21, 34, 37, 24, 19, 11, 15, 26, 43, 38, 25, 16, 9, 41, 36, 31, 24, 21, 30, 39, 17
- 2 Construct a histogram using intervals of 3 and percentage frequency for the data above.

### Using the TI-Nspire:

- 1 In a **Lists and spreadsheets** page type in the list name *data* and enter the values as shown. Use **menu** > **Statistics** > **Stat Calculations** > **One-Variable Statistics** and press **enter**. Scroll to view the statistics.

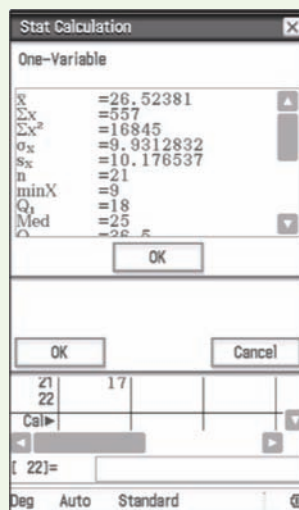
A	B	C	D
data			=OneVar(
1	21	Title	One-Va...
2	34	$\bar{x}$	26.5238
3	37	$\Sigma x$	557.
4	24	$\Sigma x^2$	16845.
5	19	$s_x := s_n \dots$	10.1765

- 2 Insert a **Data and Statistics** page and select the *data* variable for the horizontal axis. Use **menu** > **Plot Type** > **Histogram**. Then use **menu** > **Plot Properties** > **Histogram Properties** > **Bin Settings** > **Equal Bin Width**. Choose the **Width** to be 3 and **Alignment** to be 0. Use **menu** > **Window/Zoom** > **Zoom-Data** to auto rescale. Use **menu** > **Plot Properties** > **Histogram Properties** > **Histogram Scale** > **Percent** to show the percentage frequency.

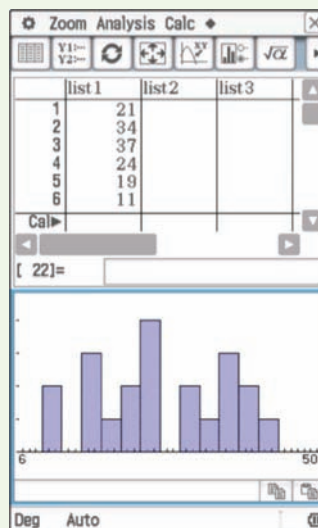


### Using the ClassPad:

- 1 In the **Statistics** application enter the data into list1. Tap **Calc**, **One-Variable** and then **OK**. Scroll to view the statistics.



- 2 Tap **SetGraph**, ensure StatGraph1 is ticked and then tap **Setting**. Change the **Type** to **Histogram**, set **XList** to **list1**, **Freq** to **1** and then tap on **Set**. Tap **W** and set **HStart** to 9 and **HStep** to 3.



## 9C Summary statistics

### Learning intentions

- To understand the concept of quartiles for a set of data
- To be able to find the five-figure summary for a set of data
- To understand how the range and interquartile range describe the spread of a data set
- To know how to determine the outliers of a set of data

In addition to the median of a single set of data, there are two related statistics called the upper and lower quartiles. When data are placed in order, then the lower quartile is central to the lower half of the data and 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 determine whether or not any data points are outliers.

### LESSON STARTER House prices

A real estate agent tells you that the median house price for a suburb in 2019 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.



Australians who rent have a wide spread of ages: roughly 27% are 15–25 years; 31% are 25–35 years; 20% are 35–45 years; 15% are 45–55 years; and 7% are older than 55. A five-figure summary and a box plot would more effectively show this age spread.

### KEY IDEAS

#### ■ Five-figure summary

- **Minimum value** (min): the minimum value
- **Lower quartile** ( $Q_1$ ): the number above 25% of the ordered data
- **Median** ( $Q_2$ ): the middle value above 50% of the ordered data
- **Upper quartile** ( $Q_3$ ): the number above 75% of the ordered data
- **Maximum value** (max): the maximum value

#### ■ Measures of spread

- **Range** = max value – min value
- **Interquartile range** (IQR)  
IQR = upper quartile – lower quartile  
=  $Q_3 - Q_1$

- **Odd number**

1 2 2 3) 5 (6 6 7 9  
 $\downarrow$     $\downarrow$     $\downarrow$

$Q_1(2)$   $Q_2(5)$   $Q_3(6.5)$

IQR (4.5)

- **Even number**

2 3 3 4 7 | 8 8 9 9 9  
 $\downarrow$     $\downarrow$     $\downarrow$

$Q_1(3)$   $Q_2(7.5)$   $Q_3(9)$

IQR (6)

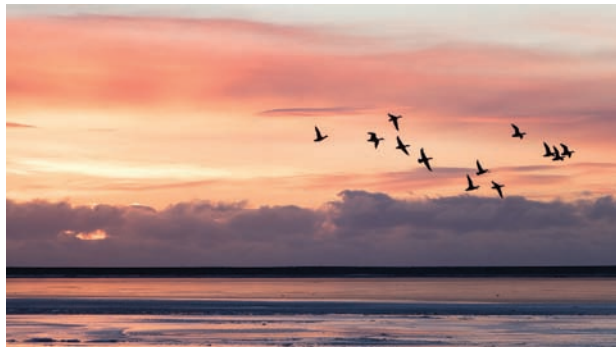
- The **standard deviation** is discussed in **Section 9E**.

■ **Outliers** are data elements outside the vicinity of the rest of the data. More formally, a data point is an outlier when it is below the **lower fence** (i.e. lower limit) or above the **upper fence** (i.e. upper limit).

- Lowerfence =  $Q_1 - 1.5 \times \text{IQR}$
- Upperfence =  $Q_3 + 1.5 \times \text{IQR}$
- An outlier does not significantly affect the median of a data set.
- An outlier does significantly affect the mean of a data set.

## BUILDING UNDERSTANDING

- 1
  - a State the types of values that must be calculated for a five-figure summary.
  - b Explain the difference between the range and the interquartile range.
  - c What is an *outlier*?
  - d How do you determine if a score in a single data set is an outlier?
- 2 This data set shows the number of cars in 13 families surveyed.  
0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 8
  - a Find the median (i.e. the middle value).
  - b By first removing the middle value, determine:
    - i the lower quartile  $Q_1$  (middle of lower half)
    - ii the upper quartile  $Q_3$  (middle of upper half).
  - c Determine the interquartile range (IQR).
  - d Calculate  $Q_1 - 1.5 \times \text{IQR}$  and  $Q_3 + 1.5 \times \text{IQR}$ .
  - e Are there any values that are outliers (numbers below  $Q_1 - 1.5 \times \text{IQR}$  or above  $Q_3 + 1.5 \times \text{IQR}$ )?
- 3 The number of ducks spotted in eight different flocks are given in this data set.  
2, 7, 8, 10, 11, 11, 13, 15
  - a
    - i Find the median (i.e. average of the middle two numbers).
    - ii Find the lower quartile (i.e. middle of the smallest four numbers).
    - iii Find the upper quartile (i.e. middle of the largest four numbers).
  - b Determine the IQR.
  - c Calculate  $Q_1 - 1.5 \times \text{IQR}$  and  $Q_3 + 1.5 \times \text{IQR}$ .
  - d Are there any outliers (i.e. numbers below  $Q_1 - 1.5 \times \text{IQR}$  or above  $Q_3 + 1.5 \times \text{IQR}$ )?







### Example 4 Finding the range and IQR

Determine the range and IQR for these data sets by finding the five-figure summary.

- a** 2, 2, 4, 5, 6, 8, 10, 13, 16, 20  
**b** 1.6, 1.7, 1.9, 2.0, 2.1, 2.4, 2.4, 2.7, 2.9

#### SOLUTION

**a** Range =  $20 - 2 = 18$

2	2	4	5	6	8	10	13	16	20
		↑		↑			↑		
		Q <sub>1</sub>		Q <sub>2</sub> (7)			Q <sub>3</sub>		

$Q_2 = 7$ , so  $Q_1 = 4$  and  $Q_3 = 13$ .

IQR =  $13 - 4 = 9$

**b** Range =  $2.9 - 1.6 = 1.3$

1.6	1.7	1.9	2.0	2.1	2.4	2.4	2.7	2.9
	↑		↑		↑			
	Q <sub>1</sub>		Q <sub>2</sub>		Q <sub>3</sub>			

$Q_1 = \frac{1.7 + 1.9}{2} = 1.8$

$Q_3 = \frac{2.4 + 2.7}{2} = 2.55$

IQR =  $2.55 - 1.8 = 0.75$

#### EXPLANATION

Range = max – min

First, split the ordered data in half to locate the median, which is  $\frac{6 + 8}{2} = 7$ .

$Q_1$  is the median of the lower half and  $Q_3$  is the median of the upper half.

IQR =  $Q_3 - Q_1$

Max = 2.9, min = 1.6

Leave the median out of the upper and lower halves when locating  $Q_1$  and  $Q_3$ .

Average the two middle values of the lower and upper halves to find  $Q_1$  and  $Q_3$ .

#### Now you try

Determine the range and IQR for these data sets by finding the five-figure summary.

- a** 3, 5, 5, 6, 7, 9, 10, 12  
**b** 3.8, 3.9, 4.0, 4.2, 4.5, 4.5, 4.7



### Example 5 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:
- i the minimum and maximum number of geese spotted
  - ii the median
  - iii the upper and lower quartiles
  - iv the IQR
  - v any outliers by determining the lower and upper fences.
- b** Can you give a possible reason for why the outlier occurred?

*Continued on next page*

**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 =  $\frac{2+2}{2}$   
 $= 2$   
Upper quartile =  $\frac{5+6}{2}$   
 $= 5.5$   
**iv** IQR =  $5.5 - 2$   
 $= 3.5$   
**v** Lower fence =  $Q_1 - 1.5 \times \text{IQR}$   
 $= 2 - 1.5 \times 3.5$   
 $= -3.25$   
Upper fence =  $Q_3 + 1.5 \times \text{IQR}$   
 $= 5.5 + 1.5 \times 3.5$   
 $= 10.75$   
 $\therefore$  The outlier is 18.  
**b** 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 upper halves when  $Q_1$  and  $Q_3$  are calculated.

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

A data point is an outlier when it is less than  $Q_1 - 1.5 \times \text{IQR}$  or greater than  $Q_3 + 1.5 \times \text{IQR}$ .

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

**Now you try**

The following data set represents the number of people on 11 buses in a local area.

36, 24, 15, 23, 26, 0, 19, 24, 26, 33, 19

- a** For the data, find:  
**i** the minimum and maximum number of people on the buses.  
**ii** the median  
**iii** the upper and lower quartiles  
**iv** the IQR  
**v** any outliers by determining the lower and upper fences.  
**b** Can you give a possible reason for why the outlier occurred?

**Exercise 9C****FLUENCY**

1–3

1(1/2), 2, 3

2–4

Example 4

- 1** Determine the range and IQR for these data sets by finding the five-figure summary.  
**a** 3, 4, 6, 8, 8, 10, 13  
**b** 10, 10, 11, 14, 14, 15, 16, 18  
**c** 1.2, 1.8, 1.9, 2.3, 2.4, 2.5, 2.9, 3.2, 3.4  
**d** 41, 49, 53, 58, 59, 62, 62, 65, 66, 68

- Example 5** 2 The following numbers of cars, travelling on a quiet suburban street, were counted on each day for 15 days.

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

For the given data, find:

- a the minimum and maximum number of cars counted
- b the median
- c the lower and upper quartiles
- d the IQR
- e any outliers by determining the lower and upper fences
- f a possible reason for the outlier.



- 3 Summarise the data sets below by finding:

- i the minimum and maximum values
- ii the median ( $Q_2$ )
- iii the lower and upper quartiles ( $Q_1$  and  $Q_3$ )
- iv the IQR
- v 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

-  4 The number 20 is an outlier in this data set 1, 2, 2, 3, 4, 20.

- a Calculate the mean if the outlier is:
  - i included
  - ii excluded
- b Calculate the median if the outlier is:
  - i included
  - ii excluded
- c By how much does including the outlier increase the:
  - i mean?
  - ii median?

### PROBLEM-SOLVING

5,  $6\frac{1}{2}$

$6\frac{1}{2}$ , 7

7, 8

- 5 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:
  - i the minimum and maximum number of buttons on the calculators
  - ii the median
  - iii the lower and upper quartiles
  - iv the IQR
  - v any outliers
  - vi the mean.
- b Which is a better measure of the centre of the data, the mean or the median? Explain.
- c Can you give a possible reason why the outlier has occurred?



- 6 Using the definition of an outlier, decide whether or not any outliers exist in the following sets of data. If so, list them.

- a** 3, 6, 1, 4, 2, 5, 9, 8, 6, 3, 6, 2, 1  
**b** 8, 13, 12, 16, 17, 14, 12, 2, 13, 19, 18, 12, 13  
**c** 123, 146, 132, 136, 139, 141, 103, 143, 182, 139, 127, 140  
**d** 2, 5, 5, 6, 5, 4, 5, 6, 7, 5, 8, 5, 5, 4

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

- a** the IQR  
**b** any outliers  
**c** the median if the number 37 is added to the list  
**d** the median if the number 22 is added to the list instead of 37.

Stem	Leaf
0	1
1	68
2	046
3	23

2 | 4 means 24

- 8 Three different numbers have median 2 and range 2. Find the three numbers.

### REASONING

9

9, 10

10–12

- 9 Explain what happens to the mean of a data set if all the values are:
- a** increased by 5                      **b** multiplied by 2                      **c** divided by 10.
- 10 Explain what happens to the IQR of a data set if all values are:
- a** increased by 5                      **b** multiplied by 2                      **c** divided by 10.
- 11 Give an example of a small data set that satisfies the following.
- a** median = mean                      **b** median = upper quartile                      **c** range = IQR
- 12 Explain why, in many situations, the median is preferred to the mean as a measure of centre.

### ENRICHMENT: 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:
- i** a stem-and-leaf plot                      **ii** a 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 partner.

## 9D Box plots

### Learning intentions

- To understand the features of a box plot in describing the spread of a set of data
- To know how to construct a box plot with outliers
- To be able to compare data sets using parallel box plots

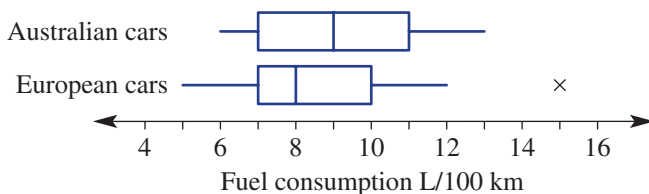
The five-figure summary (min,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , max) can be represented in graphical form as a box plot. Box plots are graphs that summarise single data sets. They clearly display the minimum and maximum values, the median, the quartiles and any outliers. Box plots also give a clear indication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



Medical researchers analyse data about the health of babies and mothers. Parallel box plots comparing birth weights of full-term babies born to smoking and non-smoking mothers show significantly lower weights for babies whose mothers smoke.

### LESSON STARTER Fuel consumption

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

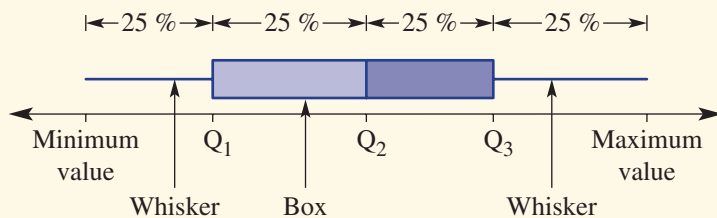


- What do the box plots say about how the fuel consumption compares between Australian-made and European-made cars?
- What does each part of the box plot represent?
- What do you think the cross (×) represents on the European cars box plot?

## KEY IDEAS

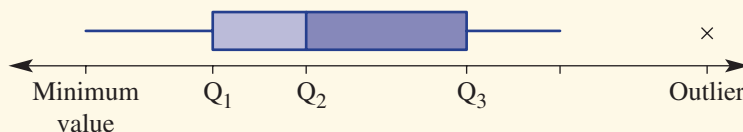
■ A **box plot** (also called a box-and-whisker plot) can be used to summarise a data set.

- The number of data values in each quarter (25%) are approximately equal.



■ An **outlier** is marked with a cross (×).

- An outlier is greater than  $Q_3 + 1.5 \times \text{IQR}$  or less than  $Q_1 - 1.5 \times \text{IQR}$ .
- The whiskers stretch to the lowest and highest data values that are not outliers.

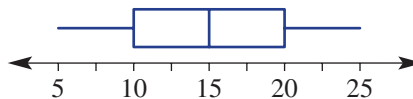


■ **Parallel box plots** are two or more box plots drawn on the same scale. They are used to compare data sets within the same context.

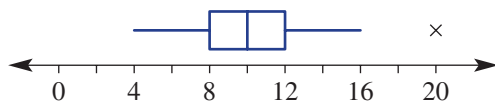
## BUILDING UNDERSTANDING

1 For this simple box plot, state:

- |                                  |                                |
|----------------------------------|--------------------------------|
| a the median ( $Q_2$ )           | b the minimum                  |
| c the maximum                    | d the range                    |
| e the lower quartile ( $Q_1$ )   | f the upper quartile ( $Q_3$ ) |
| g the interquartile range (IQR). |                                |



2 Complete the following for this box plot.



- Find the IQR.
- Calculate  $Q_1 - 1.5 \times \text{IQR}$ .
- Calculate  $Q_3 + 1.5 \times \text{IQR}$ .
- State the value of the outlier.
- Check that the outlier is greater than  $Q_3 + 1.5 \times \text{IQR}$ .



### Example 6 Constructing box plots

Consider the given data set:

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

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

#### SOLUTION

$$\begin{array}{cccccccccccc} \text{a} & 2 & 3 & 3 & 4 & 5 & 5 & 5 & 6 & 6 & 7 & 9 & 12 \\ & & & \uparrow & & \uparrow & & \uparrow & & & & & \\ & & & Q_1 & & Q_2 & & Q_3 & & & & & \end{array}$$

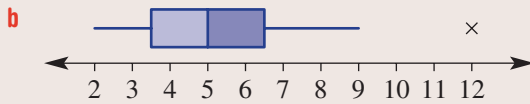
$$Q_1 = \frac{3+4}{2} = 3.5 \qquad Q_3 = \frac{6+7}{2} = 6.5$$

$$\therefore \text{IQR} = 6.5 - 3.5 = 3$$

$$Q_1 - 1.5 \times \text{IQR} = 3.5 - 1.5 \times 3 = -1$$

$$Q_3 + 1.5 \times \text{IQR} = 6.5 + 1.5 \times 3 = 11$$

$\therefore$  12 is an outlier.



#### EXPLANATION

Order the data to help find the quartiles.

Locate the median  $Q_2$  then split the data in half above and below this value.

$Q_1$  is the middle value of the lower half and  $Q_3$  the middle value of the upper half.

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

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

There are no data values 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:

12, 8, 19, 13, 22, 15, 1, 17, 24, 19

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

## Exercise 9D

### FLUENCY

1

 $1-2(1/2)$  $1-2(1/2)$ 

Example 6

- Consider the data sets below.

- Determine whether any outliers exist by first finding  $Q_1$  and  $Q_3$ .

- Draw a box plot to summarise the data, marking outliers if they exist.

**a** 4, 6, 5, 2, 3, 4, 4, 13, 8, 7, 6

**b** 1.8, 1.7, 1.8, 1.9, 1.6, 1.8, 2.0, 1.1, 1.4, 1.9, 2.2

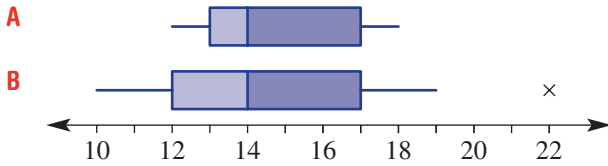
**c** 21, 23, 18, 11, 16, 19, 24, 21, 23, 22, 20, 31, 26, 22

**d** 0.04, 0.04, 0.03, 0.03, 0.05, 0.06, 0.07, 0.03, 0.05, 0.02





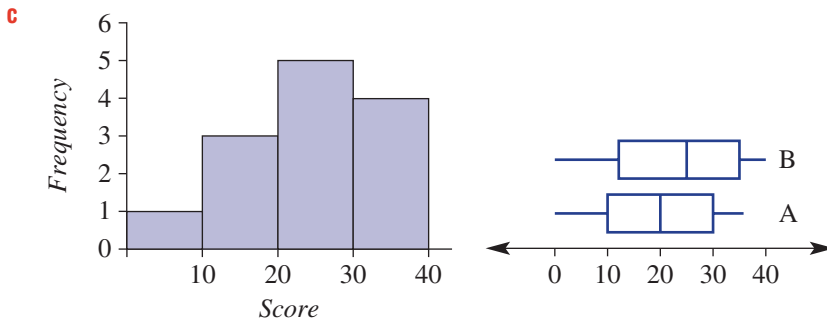
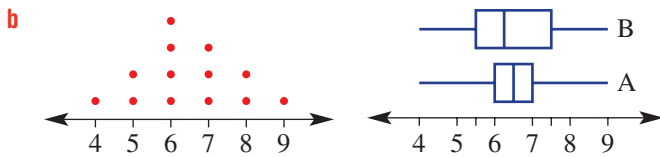
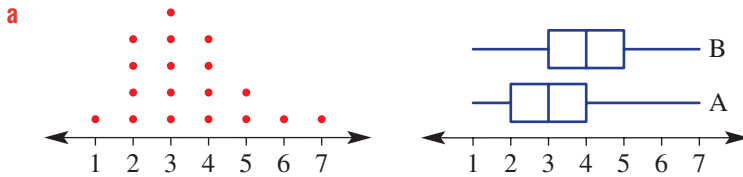
5 Two data sets can be compared using parallel box plots on the same scale, as shown below.



- a What statistical measures 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?

**REASONING** 6 6, 7 7, 8

6 Select the box plot (A or B) that best matches the given dot plot or histogram.



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

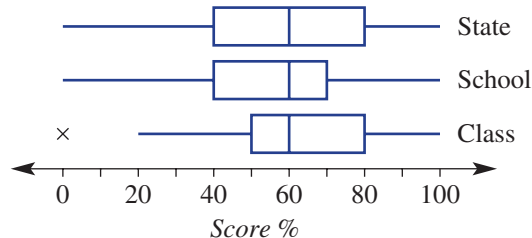
3, 2, 4, 6, 8, 4, 6, 7, 6, 1, 7, 12, 7, 3, 8

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.

- 8 The results for a Year 12 class are to be compared with the Year 12 results of the school and the State, using the parallel box plots shown.



- a Describe the main differences between the performance of:
- the class against the school
  - the class against the State
  - the school against the State.
- b Why is an outlier shown on the class box plot but not shown on the school box plot?

### ENRICHMENT: Creating your own parallel box plots

-

-

9

- 9 a Choose an area of study for which you can collect data easily, for example:
- heights or weights of students
  - maximum temperatures over a weekly period
  - amount of pocket money received each week for a group of students.
- b Collect at least two sets of data for your chosen area of study – perhaps from two or three different sources, including the internet.

Examples:

- Measure student heights in your class and from a second class in the same year level.
  - Record maximum temperatures for 1 week and repeat for a second week to obtain a second data set.
  - Use the internet to obtain the football scores of two teams for each match in the previous season.
- c Draw parallel box plots for your data.
- d Write a report on the characteristics of each data set and the similarities and differences between the data sets collected.



## Using calculators to draw box plots

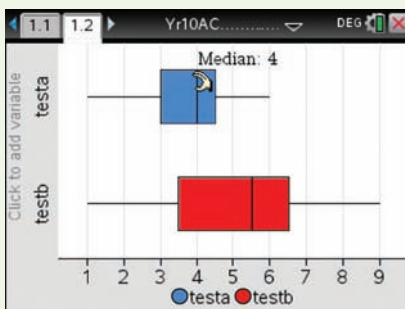
- Type these data into lists and define them as Test A and Test B.  
 Test A: 4, 6, 3, 4, 1, 3, 6, 4, 5, 3, 4, 3  
 Test B: 7, 3, 5, 6, 9, 3, 6, 7, 4, 1, 4, 6
- Draw parallel box plots for the data.

### Using the TI-Nspire:

- In a **Lists and spreadsheets** page type in the list names *testa* and *testb* and enter the values as shown.

	A testa	B testb	C	D
1	4	7		
2	6	3		
3	3	5		
4	4	6		
5	1	9		

- Insert a **Data and Statistics** page and select the *testa* variable for the horizontal axis. Change to a box plot using **Plot Type > Box Plot**. Trace (or hover over) to reveal the statistical measures. To show the box plot for *testb*, use **Plot Properties > Add X Variable** and select *testb*.

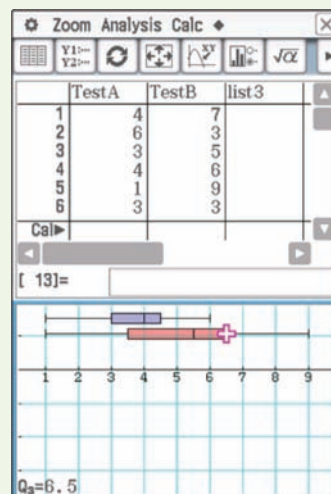


### Using the ClassPad:

- In the **Statistics** application enter the data into the lists. Give each column a title.

	TestA	TestB	list3
1	4	7	
2	6	3	
3	3	5	
4	4	6	
5	1	9	
6	3	3	
7	6	6	
8	4	7	
9	5	4	
10	3	1	
11	4	4	
12	3	6	
13			
14			
15			
16			
17			
18			

- Tap . For graph 1, set **Draw** to **On**, **Type** to **MedBox**, **XList** to **mainTestA** and **Freq** to **1**. For graph 2, set **Draw** to **On**, **Type** to **MedBox**, **XList** to **mainTestB** and **Freq** to **1**. Tap **Set**. Tap .





## 9E Standard deviation 10A

### Learning intentions

- To understand that standard deviation is a number that describes the spread of the data about the mean
- To know that a small standard deviation means data are concentrated about the mean
- To know how to calculate the standard deviation for a small set of data
- To be able to compare two sets of data referring to the mean and standard deviation

For a single data set we have already discussed the range and interquartile range to describe the spread of the data. Another statistic commonly used to describe spread is standard deviation. The standard deviation is a number that describes how far data values are from the mean. A data set with a relatively small standard deviation will have data values concentrated about the mean, and if a data set has a relatively large standard deviation then the data values will be more spread out from the mean.

The standard deviation can be calculated by hand but, given the tedious nature of the calculation, technology can be used for more complex data sets. In this section technology is not required but you will be able to find a function on your calculator (often denoted  $s$  or  $\sigma$ ) that can be used to find the standard deviation.

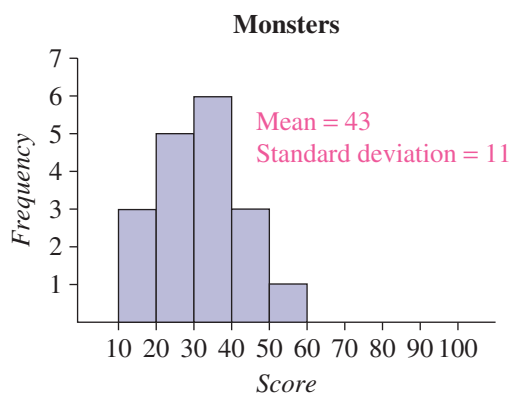
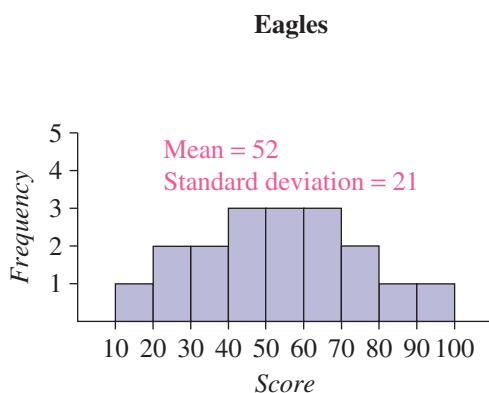


When selecting a sportsperson for a competition, the average and standard deviation of past results are useful. Two cricketers may have equal average runs per game, but the player with the smaller standard deviation is the more consistent batter.

### LESSON STARTER Which is the better team?

These histograms show the number of points scored by the Eagles and the Monsters basketball teams in an 18-round competition. The mean and standard deviation are given for each team.

- Which team has the higher mean? What does this say about the team's performance?
  - Which team has the smaller standard deviation? What does this say about the team's performance?
- Discuss.



## KEY IDEAS

- The **standard deviation** is a number that describes the spread of data about the mean.
  - The sample standard deviation is for a sample data set drawn from the population.
  - If every data value from a population is used, then we calculate the population standard deviation.

- To calculate the **sample standard deviation** ( $s$ ), follow these steps.

- 1 Find the mean ( $\bar{x}$ ).
- 2 Find the difference between each value and the mean (called the deviation).
- 3 Square each deviation.
- 4 Sum the squares of each deviation.
- 5 Divide by the number of data values less 1 (i.e.  $n - 1$ ).
- 6 Take the square root.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}}$$

- If the data represent the complete population, then divide by  $n$  instead of  $(n - 1)$ . This would give the **population standard deviation** ( $\sigma$ ). Dividing by  $(n - 1)$  for the sample standard deviation gives a better estimate of the population standard deviation.
- If data are concentrated about the mean, then the standard deviation is relatively small.
- If data are spread out from the mean, then the standard deviation is relatively large.
- In many common situations we can expect 95% of the data to be within two standard deviations of the mean.

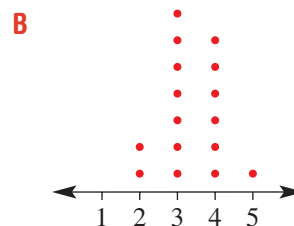
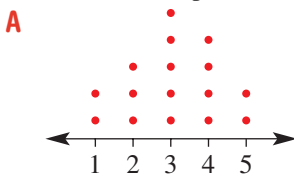
## BUILDING UNDERSTANDING

- 1 Use the word *smaller* or *larger* to complete each sentence.

a If data are more spread out from the mean, then the standard deviation is \_\_\_\_\_.

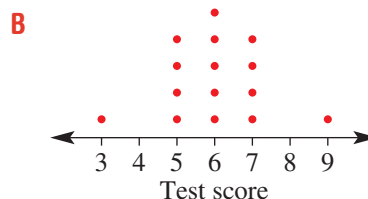
b If data are more concentrated about the mean, then the standard deviation is \_\_\_\_\_.

- 2 Here are two dot plots, A and B.



- a Which data set (A or B) would have the higher mean?
- b Which data set (A or B) would have the higher standard deviation?

- 3 These dot plots show the results for a class of 15 students who sat tests A and B. Both sets of results have the same mean and range.



Which data set (A or B) would have the higher standard deviation? Give a reason.

- 4 This back-to-back stem-and-leaf plot compares the number of trees or shrubs in the backyards of homes in the suburbs of Gum Heights and Oak Valley.
- a** Which suburb has the smaller mean number of trees or shrubs? Do not calculate the actual means.
- b** Without calculating the actual standard deviations, which suburb has the smaller standard deviation?

Gum Heights Leaf	Stem	Oak Valley Leaf
7 3 1	0	
8 6 4 0	1	0
9 8 7 2	2	0 2 3 6 8 8 9
9 6 4	3	4 6 8 9
	4	3 6
2   8 means 28		



### Example 7 Calculating the standard deviation

Calculate the mean and sample standard deviation for this small data set, correct to one decimal place.  
2, 4, 5, 8, 9

#### SOLUTION

$$\bar{x} = \frac{2 + 4 + 5 + 8 + 9}{5}$$

$$= 5.6$$

$$s = \sqrt{\frac{(2 - 5.6)^2 + (4 - 5.6)^2 + (5 - 5.6)^2 + (8 - 5.6)^2 + (9 - 5.6)^2}{5 - 1}}$$

$$= \sqrt{\frac{(-3.6)^2 + (-1.6)^2 + (-0.6)^2 + (2.4)^2 + (3.4)^2}{4}}$$

$$= 2.9 \text{ (to 1 d.p.)}$$

#### EXPLANATION

Sum all the data values and divide by the number of data values (i.e. 5) to find the mean.

Deviation 1 is  $2 - 5.6$  (the difference between the data value and the mean).

Sum the square of all the deviations, divide by  $(n - 1)$  (i.e. 4) and then take the square root.

#### Now you try

Calculate the mean and sample standard deviation for this small data set, correct to one decimal place.  
1, 2, 2, 4, 5



### Example 8 Interpreting the standard deviation

This back-to-back stem-and-leaf plot shows the distribution of distances that 17 people in Darwin and Sydney travel to work. The means and standard deviations are given.

Darwin Leaf	Stem	Sydney Leaf	
8 7 4 2	0	1 5	Sydney $\bar{x} = 27.9$
9 9 5 5 3	1	2 3 7	$s = 15.1$
8 7 4 3 0	2	0 5 5 6	
5 2 2	3	2 5 9 9	Darwin
	4	4 4 6	$\bar{x} = 19.0$
	5	2	$s = 10.1$

3 | 5 means 35 km

Consider the position and spread of the data and then answer the following.

- By looking at the stem-and-leaf plot, suggest why Darwin's mean is less than that of Sydney.
- Why is Sydney's standard deviation larger than that of Darwin?
- Give a practical reason for the difference in centre and spread for the data for Darwin and Sydney.

#### SOLUTION

- The maximum score for Darwin is 35. Sydney's mean is affected by several values larger than 35.
- The data for Sydney are more spread out from the mean. Darwin's scores are more closely clustered near its mean.
- Sydney is a larger city and more spread out, so people have to travel farther to get to work.

#### EXPLANATION

The mean depends on every value in the data set.

Sydney has more scores with a large distance from its mean. Darwin's scores are closer to the Darwin mean.

Higher populations often lead to larger cities and longer travel distances.

#### Now you try

This stem-and-leaf plot shows the distribution of hours of television watched by 20 students from each of Year 7 and Year 12 over a 1-month period. The means and standard deviations are given.

Year 7	Stem	Year 12	
9	0	4 7	Year 7 $\bar{x} = 30.1$
9 5 2	1	0 1 3 4 4 6 7 9	$s = 10.7$
9 8 8 4 1	2	1 2 2 4 5 7 8 9	
9 8 7 5 3 2 0	3	3 5	Year 12
6 3 2 2	4		$\bar{x} = 19.6$

2 | 4 means 24 hours  $s = 8.5$

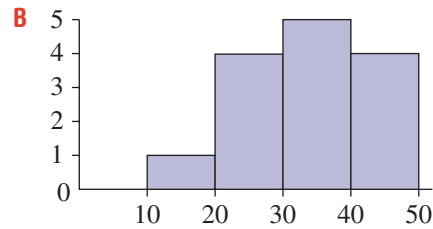
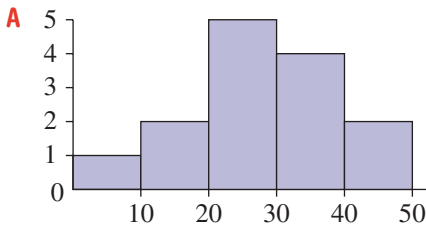
Consider the position and spread of the data and then answer the following.

- Why is the mean for Year 12 less than that for Year 7?
- Why is Year 7's standard deviation larger than that for Year 12?
- Give a practical reason for the difference in centre and spread for the Year 7 and Year 12 data.





5 Consider these two histograms, and then state whether the following are true or false.



- a The mean for set A is greater than the mean for set B.
- b The range for set A is greater than the range for set B.
- c The standard deviation for set A is greater than the standard deviation for set B.



6 Find the mean and sample standard deviation for the scores in these frequency tables. Round the standard deviations to one decimal place.

**a**

Score	Frequency
1	3
2	1
3	3

**b**

Score	Frequency
4	1
5	4
6	3

**REASONING** 7 7, 8 8, 9

7 Two simple data sets, A and B, are identical except for the maximum value, which is an outlier for set B.

A: 4, 5, 7, 9, 10

B: 4, 5, 7, 9, 20

- a Is the range for set A equal to the range for set B?
  - b Is the mean for each data set the same?
  - c Is the median for each data set the same?
  - d Would the standard deviation be affected by the outlier? Explain.
- 8 Data sets 1 and 2 have means  $\bar{x}_1$  and  $\bar{x}_2$ , and standard deviations  $s_1$  and  $s_2$ .
- a If  $\bar{x}_1 > \bar{x}_2$ , does this necessarily mean that  $s_1 > s_2$ ? Give a reason.
  - b If  $s_1 < s_2$  does this necessarily mean that  $\bar{x}_1 < \bar{x}_2$ ?
- 9 Data sets A and B each have 20 data values and are very similar except for an outlier in set A. Explain why the interquartile range might be a better measure of spread than the range or the standard deviation.

## ENRICHMENT: Study scores

–

–

10

- 10 The Mathematics study scores (out of 100) for 50 students in a school are as listed.

71, 85, 62, 54, 37, 49, 92, 85, 67, 89  
 96, 44, 67, 62, 75, 84, 71, 63, 69, 81  
 57, 43, 64, 61, 52, 59, 83, 46, 90, 32  
 94, 84, 66, 70, 78, 45, 50, 64, 68, 73  
 79, 89, 80, 62, 57, 83, 86, 94, 81, 65

The mean ( $\bar{x}$ ) is 69.16 and the sample standard deviation ( $s$ ) is 16.0.

- a** Calculate:

- i**  $\bar{x} + s$
- ii**  $\bar{x} - s$
- iii**  $\bar{x} + 2s$
- iv**  $\bar{x} - 2s$
- v**  $\bar{x} + 3s$
- vi**  $\bar{x} - 3s$

- b** Use your answers from part **a** to find the percentage of students with a score within:

- i** one standard deviation from the mean
- ii** two standard deviations from the mean
- iii** three standard deviations from the mean.

- c i** Research what it means when we say that the data are ‘normally distributed’. Give a brief explanation.

- ii** For data that are normally distributed, find out what percentage of data are within one, two and three standard deviations from the mean. Compare this with your results for part **b** above.



9A

- 1 What type of data would these survey questions generate?
- How many pets do you have?
  - What is your favourite ice-cream flavour?

9B

- 2 A Year 10 class records the length of time (in minutes) each student takes to travel from home to school. The results are listed here.

15	32	6	14	44	28	15	9	25	18
8	16	13	20	19	27	23	12	38	15

- 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.
- Construct a stem-and-leaf plot for the data.
- Use your stem-and-leaf plot to find the median.

9C



- 3 Determine the range and IQR for these data sets by finding the five-figure summary.
- 4, 9, 12, 15, 16, 18, 20, 23, 28, 32
  - 4.2, 4.3, 4.7, 5.1, 5.2, 5.6, 5.8, 6.4, 6.6

9C

- 4 The following numbers of parked cars were counted in the school car park and adjacent street each day at morning recess for 14 school days.

36, 38, 46, 30, 69, 31, 40, 37, 55, 34, 44, 33, 47, 42

- For the data, find:
  - the minimum and maximum number of cars
  - the median
  - the upper and lower quartiles
  - the IQR
  - any outliers.
- Can you give a possible reason for why the outlier occurred?

9D

- 5 The ages of a team of female gymnasts are given in this data set:

18, 23, 14, 28, 21, 19, 15, 32, 17, 18, 20, 13, 21

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

9E



- 6 Find the sample standard deviation for this small data set, correct to one decimal place.

2, 3, 5, 6, 9

Use the sample standard deviation formula.



## 9F Time-series data

### Learning intentions

- To understand that time-series data are data recorded at regular time intervals
- To know how to plot a time-series graph with time on the horizontal axis
- To be able to use a time-series plot to describe any trend in the data

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 and these can help to analyse the data, describe trends and make predictions about the future.

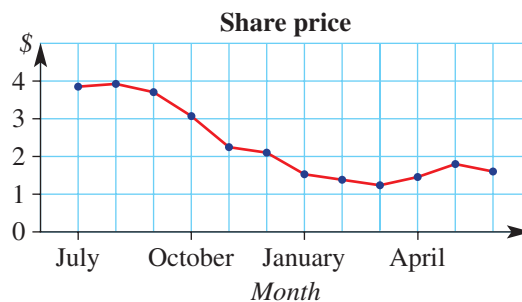


The BOM (Bureau of Meteorology) publishes time-series graphs of Australian annual and monthly mean temperature anomalies, i.e. deviations from the overall average. Over recent decades, these graphs show an upward trend of positive and increasing anomalies.

### LESSON STARTER Share price trends

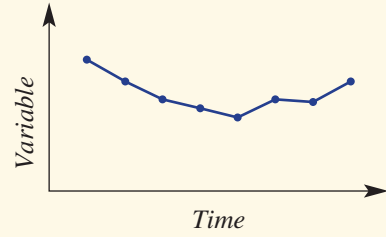
A company's share price is recorded at the end of each month of the financial year, as shown in this time-series graph.

- Describe the trend in the data at different times of the year.
- At what time of year do you think the company starts reporting bad profit results?
- Does it look like the company's share price will return to around \$4 in the next year? Why?

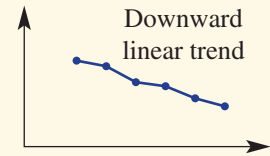


## KEY IDEAS

- **Time-series data** are recorded at regular time intervals.
- The graph or plot of a time series uses:
  - time on the horizontal axis as the **independent** variable
  - line segments connecting points on the graph.
  - the variable being considered on the vertical axis as the **dependent** variable

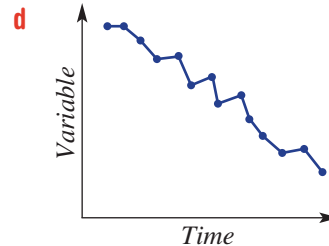
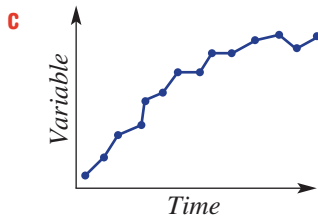
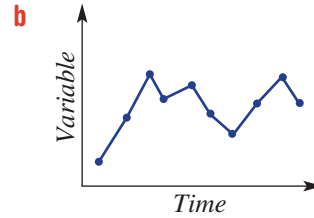
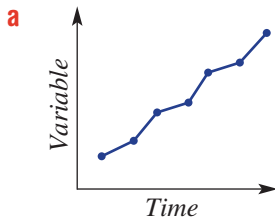


- If the time-series plot results in points being on or near a straight line, then we say that the trend is **linear**.



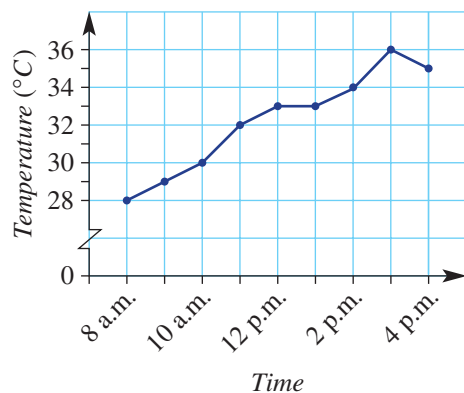
## BUILDING UNDERSTANDING

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



- 2 This time-series graph shows the temperature over the course of an 8-hour school day.

- a** State the temperature at:
- i** 8 a.m.
  - ii** 12 p.m.
  - iii** 1 p.m.
  - iv** 4 p.m.
- b** What was the maximum temperature?
- c** During what times did the temperature:
- i** stay the same?
  - ii** decrease?
- d** Describe the general trend in the temperature for the 8-hour school day.





### Example 9 Plotting and interpreting a time-series plot

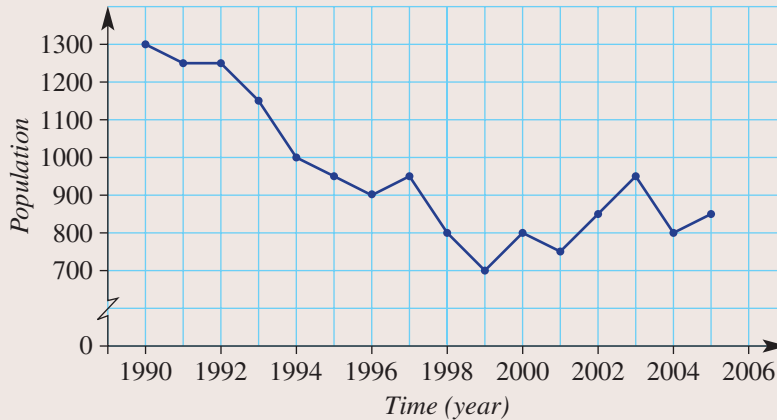
The approximate population of an outback town is recorded from 1990 to 2005.

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Population	1300	1250	1250	1150	1000	950	900	950	800	700	800	750	850	950	800	850

- Plot the time series.
- Describe the trend in the data over the 16 years.

#### SOLUTION

**a**



- 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.

#### EXPLANATION

Use time on the horizontal axis. Break the y-axis so as to not include 0–700. Join points with line segments.

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

#### Now you try

The average price of lambs at a market over 14 weeks is given in this table.

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Price (\$)	82	80	85	89	91	87	93	104	100	111	108	105	112	119

- Plot the time series.
- Describe the trend in the data over the 14 weeks.

## Exercise 9F

### FLUENCY

Example 9

- 1 The approximate population of a small village is 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

- Plot the time-series graph.
- Describe the general trend in the data over the 11 years.
- For the 11 years, what was the:
  - minimum population?
  - maximum population?

Example 9

- 2 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

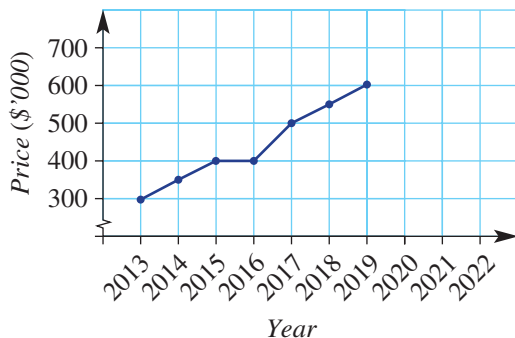
- Plot the time-series graph. Break the y-axis to exclude values from \$0 to \$1.20.
  - Describe the way in which the share price has changed over the 12 months.
  - What is the difference between the maximum and minimum share price in the 12 months?
- 3 The pass rate (%) for a particular examination is given in a table over 10 years.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Pass rate (%)	74	71	73	79	85	84	87	81	84	83

- Plot the time-series graph for the 10 years.
- Describe the way in which the pass rate for the examination has changed in the given time period.
- In what year was the pass rate a maximum?
- By how much had the pass rate improved from 1995 to 1999?

### PROBLEM-SOLVING

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



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

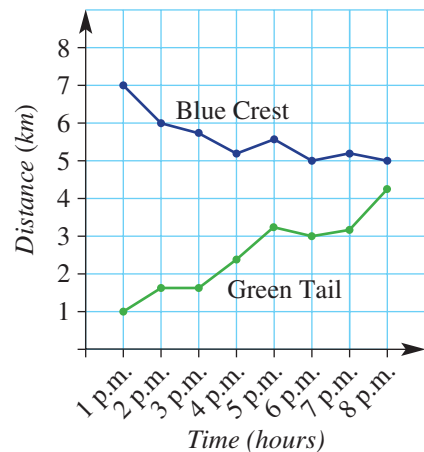
- 5 The two top-selling book stores for a company list their sales figures for the first 6 months of the year. Sales amounts are in thousands of dollars.

	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 was the difference in the sales volume for:
- August?
  - December?
- b In 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:
- City Central
  - 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?

- 6 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.

- a State the distance from the machine at 3 p.m. for:
- Blue Crest
  - Green Tail
- b Describe the trend in the distance from the recording machine for:
- Blue Crest
  - Green Tail
- c Assuming that the given trends continue, predict the time when the pigeons will be the same distance from the recording machine.



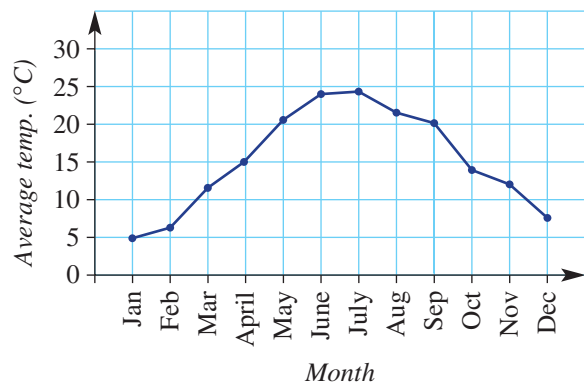
## REASONING

7

7, 8

8, 9

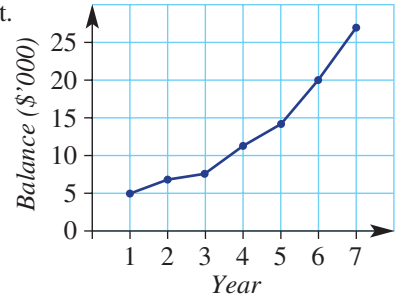
- 7 The average monthly maximum temperature for a city is illustrated in this graph.
- a Explain why the average maximum temperature for December is close to the average maximum temperature for January.
- b Do you think this graph is for an Australian city?
- c Do you think the data are for a city in the Northern Hemisphere or the Southern Hemisphere? Give a reason.





8 The balance of an investment account is shown in this time-series plot.

- a Describe the trend in the account balance over the 7 years.
- b Give a practical reason for the shape of the curve that models the trend in the graph.



9 A drink at room temperature is placed in a fridge that is at 4°C.

- a Sketch a time-series plot that might show the temperature of the drink after it has been placed in the fridge.
- b Would the temperature of the drink ever get to 3°C? Why?

**ENRICHMENT: Moving run average**      -      -      10

10 In this particular question, 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 batsman in cricket with the following runs scored from 10 completed innings.

<b>Innings</b>	1	2	3	4	5	6	7	8	9	10
<b>Score</b>	26	38	5	10	52	103	75	21	33	0
<b>Moving average</b>	26	32	23							

- a Complete the table by calculating the moving average for innings 4–10. Round to the nearest whole number where required.
- b Plot the score and moving averages for the batter on the same set of axes.
- 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.

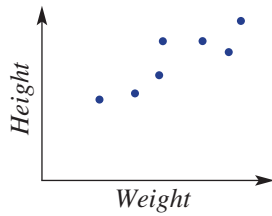


## 9G Bivariate data and scatter plots

### Learning intentions

- To understand that bivariate data involve data about two variables in a given context
- To know how to draw a scatter plot to compare data from two variables
- To be able to use a scatter plot to describe the correlation between the two variables using key terms

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 illustrate a relationship that may exist between the variables. Scatter plots make it very easy to see the strength of the association between the two variables.



Market research analysts find a positive correlation in scatter plots of advertising spending versus product sales. AI (artificial intelligence) algorithms use automated marketing to create highly effective digital advertising, specifically targeted to each person's online presence.

### 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 think 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



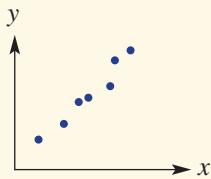
How might the size of a ship and its cargo capacity be related?

## KEY IDEAS

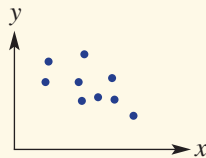
- **Bivariate data** include data for two variables.
  - The two variables are usually related; for example, height and weight.
  - The variable that is changed or controlled is the independent variable and is on the  $x$ -axis.
  - The variable being tested or measured is the dependent variable and is on the  $y$ -axis.
- A **scatter plot** is a graph on a number plane in which the axes variables correspond to the two variables from the bivariate data.
- The words *relationship*, *correlation* and *association* are used to describe the way in which variables are related.
- Types of correlation:

*Examples*

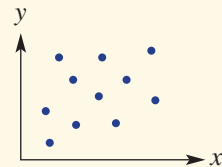
Strong positive correlation



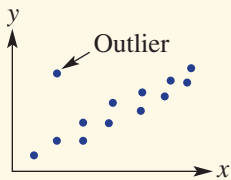
Weak negative correlation



No correlation



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



## BUILDING UNDERSTANDING

- 1 Decide if it is likely for there to be a strong correlation between these pairs of variables.
  - a Height of door and Thickness of door handle
  - b Weight of car and Fuel consumption
  - c Temperature and Length of phone calls
  - d Size of textbook and Number of textbook
  - e Diameter of flower and Number of bees
  - f Amount of rain and Size of vegetables in the vegetable garden
- 2 For each of the following sets of bivariate data with variables  $x$  and  $y$ , decide whether  $y$  generally increases or decreases as  $x$  increases.

a

$x$	1	2	3	4	5	6	7	8	9	10
$y$	3	2	4	4	5	8	7	9	11	12

b

$x$	0.1	0.3	0.5	0.9	1.0	1.1	1.2	1.6	1.8	2.0	2.5
$y$	10	8	8	6	7	7	7	6	4	3	1



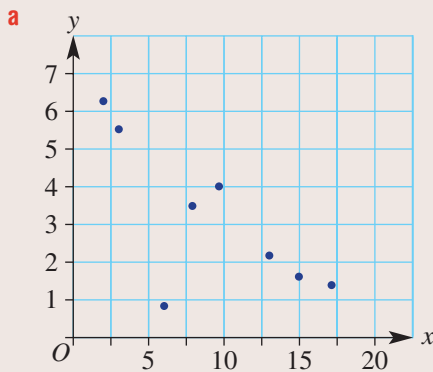
### Example 10 Constructing and interpreting scatter plots

Consider this simple bivariate data set.

<b>x</b>	13	9	2	17	3	6	8	15
<b>y</b>	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



- Negative correlation
- Strong correlation
- The outlier is  $(6, 0.9)$ .

#### EXPLANATION

Plot each point using a  $\bullet$  on graph paper.

- As  $x$  increases,  $y$  decreases.
- The downwards trend in the data is clearly defined.
- This point defies the trend.

#### Now you try

Consider this simple bivariate data set.

<b>x</b>	12	2	15	10	4	5	8	13	7
<b>y</b>	4.0	1.3	4.5	3.6	1.8	2.0	2.5	2.0	2.9

- 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.

## Exercise 9G

**FLUENCY** 1-4 2-4 2-4

Example 10

1 Consider this simple bivariate data set. (Use technology to assist if desired. See page 709.)

<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

- 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.

Example 10

2 Consider this simple bivariate data set. (Use technology to assist if desired. See page 709.)

<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

- 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.

3 By completing scatter plots (by hand or using technology) 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

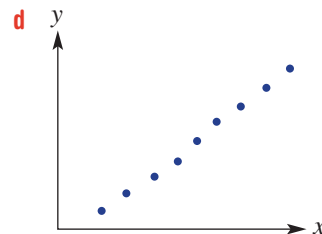
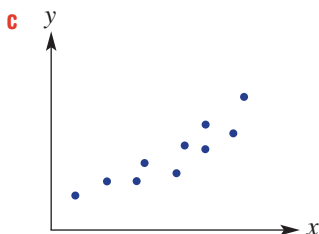
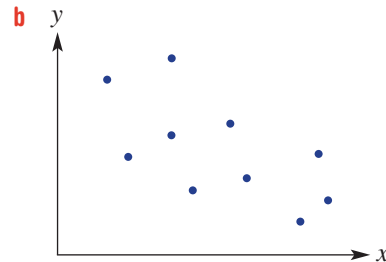
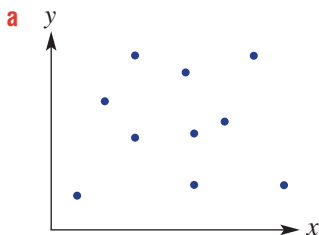
b

<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

4 For the following scatter plots, describe the correlation between  $x$  and  $y$ .





## PROBLEM-SOLVING

5, 6

6, 7

6, 8

- 5 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**?
  - Which car gives a fuel economy reading that does not support the general trend?
- 6 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 in saying that the amount of fertiliser fed to tomato plants does affect the size of the tomato produced?







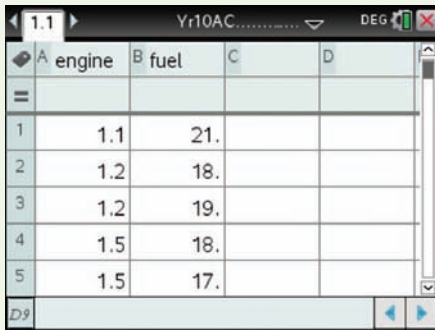
## Using calculators to draw scatter plots

Type the following data about car fuel economy into two lists and draw a scatter plot.

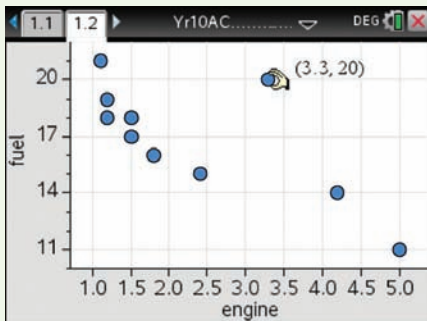
Car	A	B	C	D	E	F	G	H	I	J
Engine size	1.1	1.2	1.2	1.5	1.5	1.8	2.4	3.3	4.2	5.0
Fuel economy	21	18	19	18	17	16	15	20	14	11

### Using the TI-Nspire:

- 1 In a **Lists and spreadsheets** page type in the list names *engine* and *fuel* and enter the values as shown.

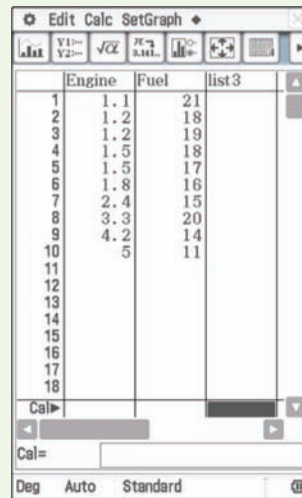


- 2 Insert a **Data and Statistics** page and select the *engine* variable for the horizontal axis and *fuel* for the vertical axis. Hover over points to reveal coordinates.

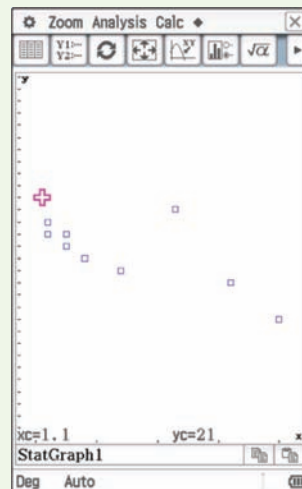


### Using the ClassPad:

- 1 In the **Statistics** application, assign a title to each column then enter the data into the lists.



- 2 Tap . For graph 1 set **Draw** to **On**, **Type** to **Scatter**, **XList** to **mainEngine**, **YList** to **mainFuel**, **Freq** to **1** and **Mark** to **square**. Tap **Set**. Tap . Tap **Analysis**, **Trace** to reveal coordinates.



## Applications and problem-solving

The following problems will investigate practical situations drawing upon knowledge and skills developed throughout the chapter. In attempting to solve these problems, aim to identify the key information, use diagrams, formulate ideas, apply strategies, make calculations and check and communicate your solutions.

### Twenty20

- 1 Two teams, the Auckland Aces and the Sunrisers Hyderabad, are part of an international 20/20 cricket tournament. They each play 10 round-robin matches and their batting totals are shown below.

<b>Aces</b>	148	172	186	179	194	132	112	154	142	177
<b>Sunrisers</b>	147	160	166	182	171	163	170	155	152	166

*You are to compare the statistics of the two cricket teams using box plots and discuss each team's performance in terms of the number of runs and the consistency of the run scoring across the season.*

- Draw parallel box plots for these two data sets.
- Compare the box plots of the two teams, commenting on which team appears capable of getting higher scores and which team appears more consistent.
- The Auckland Aces' lowest two scores were the result of rain delays and the restricted number of overs that they faced. If these two innings were increased by 40 runs each, what changes occur on the box plot?
- In their first final, the Sunrisers Hyderabad's batting total would be an outlier if included in their above set of scores. What possible scores did they get in this innings?

### Salaries and payrise

- 2 A small business has 20 employees with the following monthly salaries.

Salary (\$)	Number of employees
4500	5
5400	8
5800	5
6400	2

*The small business wishes to calculate measures of centre and spread for its salary data and then investigate the impact on these summary statistics given changes in some specific salaries.*

- Calculate the mean, median, range and standard deviation (to the nearest dollar) of these salaries.
  - The top two earning employees are given an increase of \$ $x$  per month. Describe the impact on the mean, median and range in terms of  $x$ .
  - Describe the impact on the standard deviation from part **ii**.



- b** Employees at another small business think they are paid less given their mean monthly salary is \$4800 with standard deviation \$800.
- In this company 95% of salaries lie within two standard deviations of the mean. What would employees who are in the top or bottom 2.5% of earners be earning?
  - If each employee in this business is given a pay rise of \$ $x$ , give the new mean and standard deviation of employee salaries in terms of  $x$  where appropriate.
  - The employees instead decide to give each person a percentage increase in their salary. If each person's salary is increased by a factor of  $k$ , give the new mean and standard deviation of the salaries in terms of  $k$ .



## Winter getaway

- 3** A family is planning to escape the winter cold and spend July in Noosa. They wish to be prepared for varying temperatures throughout the day and compare the daily maximum and minimum temperatures of a recent July as shown in the table below.

Min. temp (°C)	19	17	17	16	18	19	18	11	12	14	14	15	15	12	15	15
Max. temp (°C)	23	21	20	20	23	23	24	23	19	17	17	19	20	21	21	23

Min. temp (°C)	14	16	16	16	13	14	15	16	16	16	17	17	17	17	15
Max. temp (°C)	23	23	24	24	22	19	19	23	24	24	24	25	25	26	23

*The family is interested in the relationship between the maximum and minimum temperatures for the month of July and use this to make predictions for their upcoming holiday.*

- Prepare a scatter plot of these data with the minimum temperature on the horizontal axis.
- To make predictions the family use a straight line to model the data. If the line passes through the points shown in red, find the equation of this line by completing the following:  
max. temp = \_\_\_\_\_  $\times$  min. temp + \_\_\_\_\_.
- Use your equation in part **b** to find the likely:
  - max. temp on a day with a min. temp of 13°C, rounding to the nearest degree
  - min. temp on a day with a max temp of 28°C, rounding to the nearest degree.
- Which of your results in part **c** seem the most accurate? Why?
- Select two other points on the graph that a straight line modelling the data could reasonably pass through. Find the equation of this line and repeat part **c**. Comment on the similarities or differences in your results for part **c** using the two different equations.



## 9H Line of best fit by eye

### Learning intentions

- To understand that a line of best fit can be used as a model for the data when there is a strong linear association
- To know how to fit a line of best fit by eye
- To know how to find the equation of a line of best fit
- To be able to use the line of best fit and its equation to estimate data values within and outside the data range

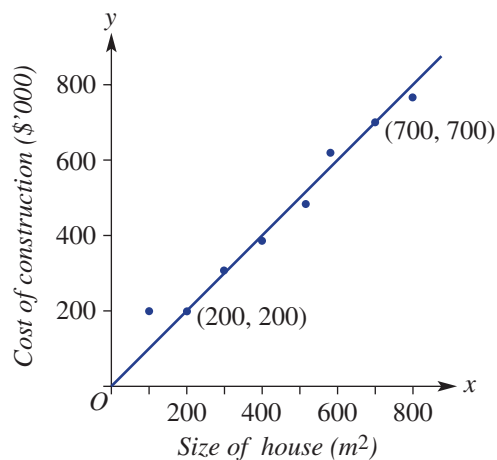
When bivariate data have a strong linear correlation, we can model the data with a straight line. This line is called a trend line or line of best fit. When we fit the line ‘by eye’, we try to balance the number of data points above the line with the number of points below the line. This trend line and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



A scatter plot of product price ( $y$ ) versus demand ( $x$ ) shows a negative correlation, with a downward sloping trend line. Businesses use demand equations to forecast sales and make informed decisions about future stock and staffing levels.

### LESSON STARTER Size versus cost

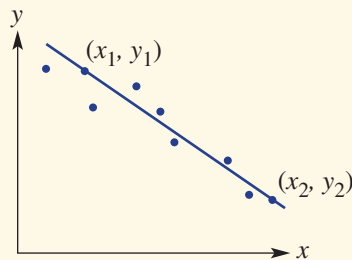
This scatter plot shows the estimated cost of building a house of a given size, as quoted by a building company. The given trend line passes through the points  $(200, 200)$  and  $(700, 700)$ .



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

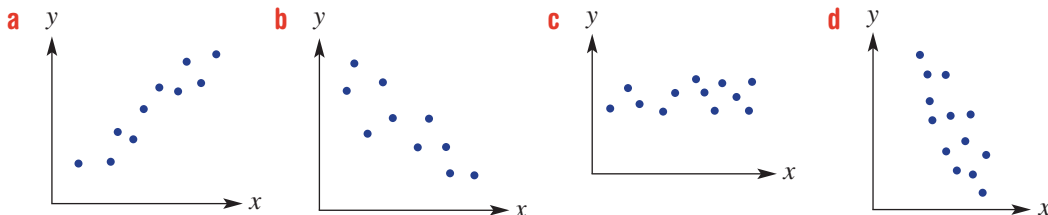
## KEY IDEAS

- 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 must be taken into account.
- The equation of the line of best fit can be found using two points that are on the line of best fit.
- For  $y = mx + c$ :
 
$$m = \frac{y_2 - y_1}{x_2 - x_1}$$
 and substitute a point to find the value of  $c$ .
  - Alternatively, use  $y - y_1 = m(x - x_1)$ .
- The line of best fit and its equation can be used for:
  - **interpolation**: constructing points within the given data range
  - **extrapolation**: constructing points outside the given data range.

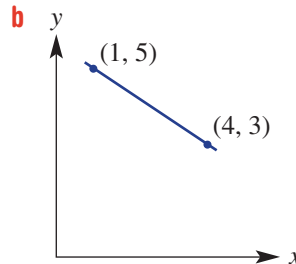
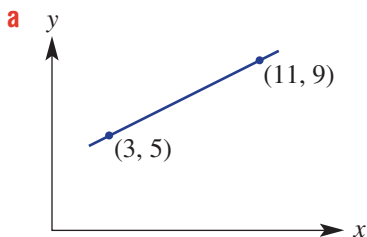


## BUILDING UNDERSTANDING

- 1 Practise fitting a line of best fit on these scatter plots by trying to balance the number of points above the line with the numbers of points below the line. (Use the side of a ruler if you don't want to draw a line.)



- 2 For each graph find the equation of the line in the form  $y = mx + c$ . First, find the gradient  $m = \frac{y_2 - y_1}{x_2 - x_1}$  and then substitute a point.



- 3 Using  $y = \frac{5}{4}x - 3$ , find:

a y when:

i  $x = 16$

ii  $x = 7$

b x when:

i  $y = 4$

ii  $y = \frac{1}{2}$



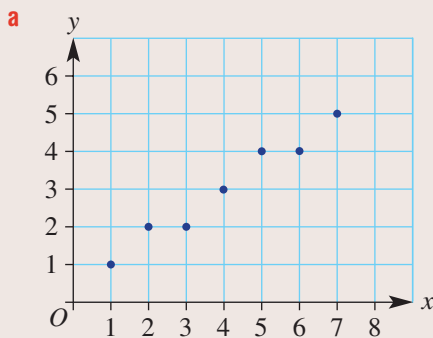
### Example 11 Fitting a line of best fit

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

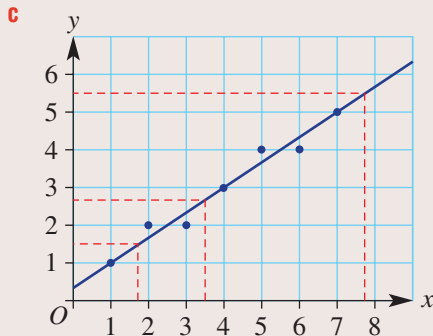
$x$	1	2	3	4	5	6	7
$y$	1	2	2	3	4	4	5

- 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



- b** Positive correlation



- d**
- $y \approx 2.7$
  - $y \approx 0.4$
  - $x \approx 1.7$
  - $x \approx 7.8$

#### EXPLANATION

Plot the points on graph paper.

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.)

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.

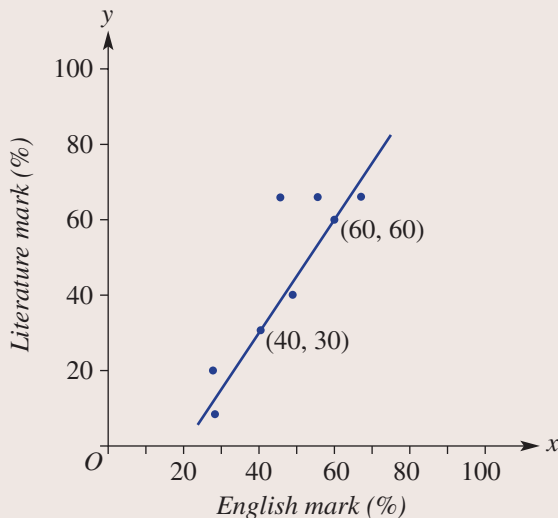
$x$	1	2	3	4	5	6
$y$	10	8	8	6	5	3

- 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  $x = 3.5$
  - ii**  $y$  when  $x = 0$
  - iii**  $x$  when  $y = 1.5$
  - iv**  $x$  when  $y = 5.5$

**Example 12 Finding the equation of a line of best fit**

This scatter plot shows a linear relationship between English marks and Literature marks in a small class of students. A trend line passes through  $(40, 30)$  and  $(60, 60)$ .

- a** Find the equation of the trend line.  
**b** Use your equation to estimate a Literature score if the English score is:
- i** 50
  - ii** 86
- c** Use your equation to estimate the English score if the Literature score is:
- i** 42
  - ii** 87

**SOLUTION**

**a**  $y = mx + c$   

$$m = \frac{60 - 30}{60 - 40} = \frac{30}{20} = \frac{3}{2}$$

$$\therefore y = \frac{3}{2}x + c$$

$$(40, 30): 30 = \frac{3}{2}(40) + c$$

$$30 = 60 + c$$

$$c = -30$$

$$\therefore y = \frac{3}{2}x - 30$$

**EXPLANATION**

Use  $m = \frac{y_2 - y_1}{x_2 - x_1}$  for the two given points.

Substitute either  $(40, 30)$  or  $(60, 60)$  to find  $c$ .

*Continued on next page*

$$\mathbf{b \ i} \quad y = \frac{3}{2}(50) - 30 = 45$$

$\therefore$  Literature score is 45.

$$\mathbf{ii} \quad y = \frac{3}{2}(86) - 30 = 99$$

$\therefore$  Literature score is 99.

$$\mathbf{c \ i} \quad 42 = \frac{3}{2}x - 30$$

$$72 = \frac{3}{2}x$$

$$x = 48$$

$\therefore$  English score is 48.

$$\mathbf{ii} \quad 87 = \frac{3}{2}x - 30$$

$$117 = \frac{3}{2}x$$

$$x = 78$$

$\therefore$  English score is 78.

Substitute  $x = 50$  and find the value of  $y$ .

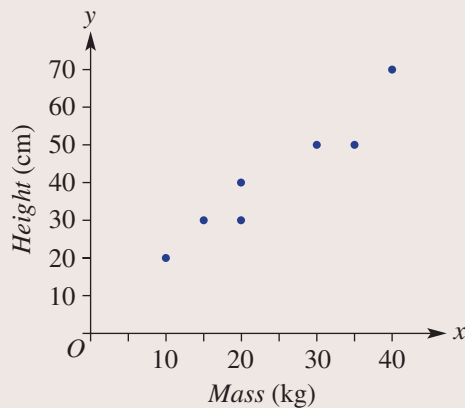
Repeat for  $x = 86$ .

Substitute  $y = 42$  and solve for  $x$ .

Repeat for  $y = 87$ .

### Now you try

This scatter plot shows a linear relationship between the mass and height of a small number of dogs. A trend line passes through (10, 20) and (40, 70).



**a** Find the equation of the trend line.

**b** Use your equation to estimate a dog height if its mass is:

**i** 25 kg

**ii** 52 kg

**c** Use your equation to estimate a dog mass if its height is:

**i** 60 cm

**ii** 80 cm





## PROBLEM-SOLVING

4

4, 5

4, 5

- 4 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.

Rainfall (mm)	450	620	560	830	680	650	720	540
Growth (cm)	25	45	25	85	50	55	50	20

- a Draw a scatter plot for the data, showing growth on the vertical axis.
- b Fit a line of best fit by eye.
- c Use your line of best fit to estimate the growth expected for the following rainfall readings.  
You do not need to find the equation of the line.
- i 500 mm                      ii 900 mm
- d Use your line of best fit to estimate the rainfall for a given year if the growth of the bamboo was:
- i 30 cm                      ii 60 cm



- 5 A line of best fit for a scatter plot, relating the weight (kg) and length (cm) of a group of dogs, passes through the points (15, 70) and (25, 120). Assume weight is on the  $x$ -axis.
- a Find the equation of the trend line.
- b Use your equation to estimate the length of an 18 kg dog.
- c Use your equation to estimate the weight of a dog that has a length of 100 cm.

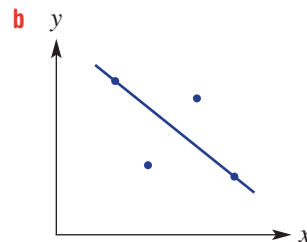
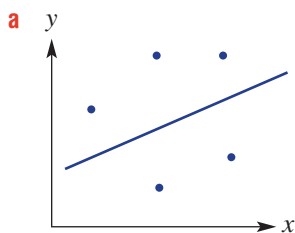
## REASONING

6

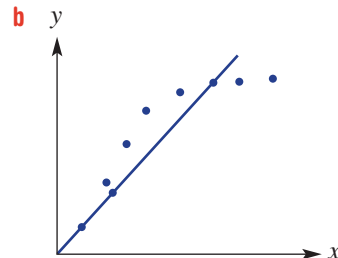
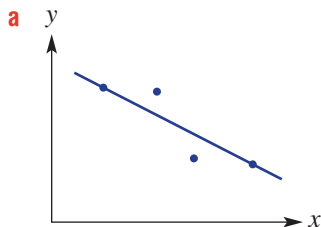
6, 7

7, 8

- 6 Describe the problem when using each trend line below for interpolation.



- 7 Describe the problem when using each trend line below for extrapolation.



- 8 A trend line relating the percentage scores for Music performance ( $y$ ) and Music theory ( $x$ ) is given by  $y = \frac{4}{5}x + 10$ .
- Find the value of  $x$  when:
    - $y = 50$
    - $y = 98$
  - What problem occurs in predicting Music theory scores when using high Music performance scores?

## ENRICHMENT: Heart rate and age

9

- 9 Two independent scientific experiments confirmed a correlation between *Maximum heart rate* (in beats per minute or b.p.m.) and *Age* (in years). The data for the two experiments are as follows.

Experiment 1													
Age (years)	15	18	22	25	30	34	35	40	40	52	60	65	71
Max. heart rate (b.p.m.)	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 (b.p.m.)	205	195	180	185	175	160	160	145	150	150	135	140	90

- Sketch separate scatter plots for experiment 1 and experiment 2.
- 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, 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 according to age and compare with the results given above.



Different watches are used by people to record heart rate.

# 9 | Linear regression using technology 10A

## Learning intentions

- To understand that there are different methods for fitting a straight line to bivariate data
- To know how to use technology to find the least squares regression line
- To be able to use the regression line equation as a model to make predictions

In **Section 9H** we used a line of best fit by eye to describe a general linear (i.e. straight line) trend for bivariate data. In this section we look at the more formal methods for fitting straight lines to bivariate data. This is called linear regression. There are many different methods used by statisticians to model bivariate data. One of the most common methods is called least squares regression. This is best handled with the use of technology.



Data scientists use machine learning algorithms, such as multiple linear regression, to extract relationships from big data. Predictive modelling applications include insurance premiums, financial services, healthcare, stock market trading and effects of climate change.

## LESSON STARTER What can my calculator or software do?

Explore the menus of your chosen technology to see what kind of regression tools are available. For CAS calculator users, refer to page 722.

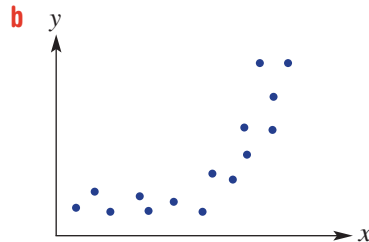
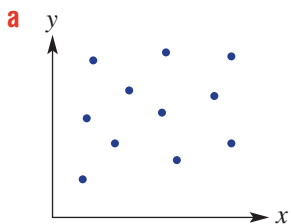
- Can you find the least squares regression tools?
- Use your technology to try **Example 13**.

## KEY IDEAS

- **Linear regression** involves using a method to fit a straight line to bivariate data.
  - The result is a straight line equation that can be used for interpolation and extrapolation.
- The **least squares** regression line minimises the sum of the square of the deviations of each point from the line.
  - Outliers have an effect on the least squares regression line because all deviations are included in the calculation of the equation of the line.

## BUILDING UNDERSTANDING

- 1 A regression line for a bivariate data set is given by  $y = 2.3x - 4.1$ . Use this equation to find:
- a the value of  $y$  when  $x$  is:
- i 7 ii 3.2
- b the value of  $x$  when  $y$  is:
- i 12 ii 0.5
- 2 Give a brief reason why a linear regression line is not very useful in the following scatter plots.



## Example 13 Finding and using regression lines

Consider the following data and use a graphics or CAS calculator or software to help answer the questions below. Round answers to two decimal places where necessary.

$x$	1	2	2	4	5	5	6	7	9	11
$y$	1.8	2	1.5	1.6	1.7	1.3	0.8	1.1	0.8	0.7

- a Construct a scatter plot for the data.
- b Find the equation of the least squares regression line.
- c Sketch the graph of the regression line onto the scatter plot.
- d Use the least squares regression line to estimate the value of  $y$  when  $x$  is:
- i 4.5 ii 15

## Now you try

Consider the following data and use a graphics or CAS calculator or software to help answer the questions below. Round answers to two decimal places where necessary.

$x$	1	3	3	4	6	7	9	10
$y$	0.5	2	1.6	3	5	6.5	5.6	8

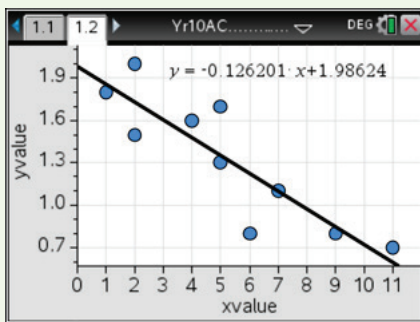
- a Construct a scatter plot for the data.
- b Find the equation of the least squares regression line.
- c Sketch the graph of the regression line onto the scatter plot.
- d Use the least squares regression line to estimate the value of  $y$  when  $x$  is:
- i 4.5 ii 15

## Using calculators to find equations of regression

## Using the TI-Nspire:

**a, b, c** In a **Lists & Spreadsheet** page enter the data in the lists named *xvalue* and *yvalue*. Insert a **Data & Statistics** page and select *xvalue* as the variable on the horizontal axis and *yvalue* as the variable on the vertical axis.

To show the linear regression line and equation use  $\left(\text{menu}\right) > \text{Analyze} > \text{Regression} > \text{Show Linear}(mx + b)$



Least squares:  $y = -0.126201x + 1.986245$

- d i**  $y \approx 1.42$   
**ii**  $y \approx 0.09$

## Using the ClassPad:

**a, b, c** In the **Statistics** application enter the data into the lists. Tap **Calc**, **Regression**, **Linear Reg** and set **XList** to *list1*, **YList** to *list2*, **Freq** to **1**, **Copy Formula** to **y1** and **Copy Residual** to **Off**. Tap **OK** to view the regression equation. Tap on **OK** again to view the regression line.

Tap **Analysis**, **Trace** and then scroll along the regression line.





## Exercise 9I

### FLUENCY

1, 2

1, 2

1, 2

Example 13



- 1 Consider the data in tables **A–C** and use a graphics or CAS calculator or software to help answer the following questions. Round answers to two decimal places where necessary.

**A**

<b>x</b>	1	2	3	4	5	6	7	8
<b>y</b>	3.2	5	5.6	5.4	6.8	6.9	7.1	7.6

**B**

<b>x</b>	3	6	7	10	14	17	21	26
<b>y</b>	3.8	3.7	3.9	3.6	3.1	2.5	2.9	2.1

**C**

<b>x</b>	0.1	0.2	0.5	0.8	0.9	1.2	1.6	1.7
<b>y</b>	8.2	5.9	6.1	4.3	4.2	1.9	2.5	2.1

- Construct a scatter plot for the data.
- Find the equation of the least squares regression line.
- Sketch the graphs of the regression lines onto the scatter plot.
- Use the least squares regression line to estimate the value of  $y$  when  $x$  is:
  - 7
  - 12



- 2 The values and ages of 14 cars are summarised in these tables.

<b>Age (years)</b>	5	2	4	9	10	8	7
<b>Price (\$'000)</b>	20	35	28	14	11	12	15

<b>Age (years)</b>	11	2	1	4	7	6	9
<b>Price (\$'000)</b>	5	39	46	26	19	17	14

- Using Age for the  $x$ -axis and rounding your coefficients to two decimal places, find the least squares regression line.
- Use your least squares regression line to estimate the value of a 3-year-old car, correct to the nearest dollar.
- Use your least squares regression line to estimate the age of a \$15000 car, correct to the nearest year.




## PROBLEM-SOLVING

3, 4

3, 4


4, 5

-  **3** A factory that produces denim jackets does not have air-conditioning. It was suggested that high temperatures inside the factory were having an effect on the number of jackets able to be produced, so a study was completed and data collected on 14 consecutive days.

<b>Max. daily temp. inside factory (<math>^{\circ}\text{C}</math>)</b>	28	32	36	27	24	25	29	31	34	38	41	40	38	31
<b>Number of jackets produced</b>	155	136	120	135	142	148	147	141	136	118	112	127	136	132

Use a graphics or CAS calculator to complete the following.


- Draw a scatter plot for the data.
- Find the equation of the least squares regression line, rounding coefficients to two decimal places.
- Graph the line onto the scatter plot.
- Use the regression line to estimate how many jackets, correct to the nearest whole number, would be able to be produced if the maximum daily temperature in the factory was:
  - $30^{\circ}\text{C}$
  - $35^{\circ}\text{C}$
  - $45^{\circ}\text{C}$

-  **4** A particular brand of electronic photocopier is considered for scrap once it has broken down more than 50 times or if it has produced more than 200000 copies. A study of one particular copier gave the following results.

<b>Number of copies (<math>\times 1000</math>)</b>	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150
<b>Total number of breakdowns</b>	0	0	1	2	2	5	7	9	12	14	16	21	26	28	33

- Sketch a scatter plot for the data.
- Find the equation of the least squares regression line.
- Graph the least squares regression line onto the scatter plot.
- Using your regression line, estimate the number of copies the photocopier will have produced at the point when you would expect 50 breakdowns.
- Would you expect this photocopier to be considered for scrap because of the number of breakdowns or the number of copies made?



-  **5** 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 and the most recent record was 118.2 m in 1996.

Further details are as follows.

<b>Year</b>	1967	1968	1969	1976	1978	1983	1987	1996
<b>New record (m)</b>	72.3	73.4	82.7	94.2	99.1	101.2	111.6	118.2

- Draw a scatter plot for the data.
- Find the equation of the least squares regression line.
- Use your regression equation to estimate the distance record for the hammer throw for:
  - 2000
  - 2020
- Would you say that it is realistic to use your regression equation to estimate distance records beyond 2020? Why?

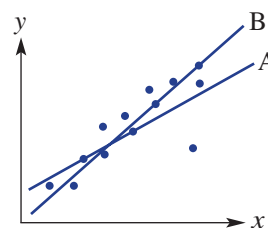
### REASONING

6

6, 7

6, 7

- Briefly explain why the least squares regression line is affected by outliers.
- This scatter plot shows both the least squares regression line and another type of regression line. Which line (i.e. A or B) do you think is the least squares line? Give a reason.



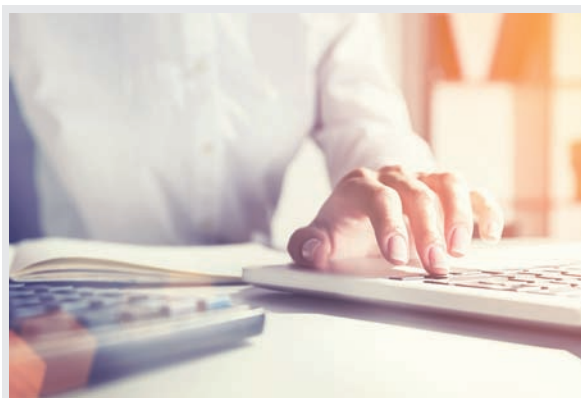
### ENRICHMENT: Correlation coefficient

-

-

8

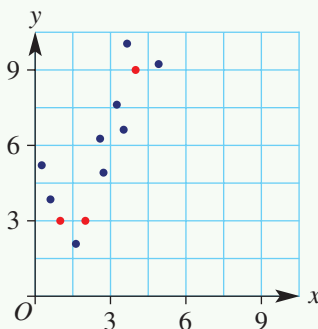
- Use the internet to find out about the Pearson correlation coefficient and then answer these questions.
  - What is the coefficient used for?
  - Do most calculators include the coefficient as part of their statistical functions?
  - What does a relatively large or small correlation coefficient mean?



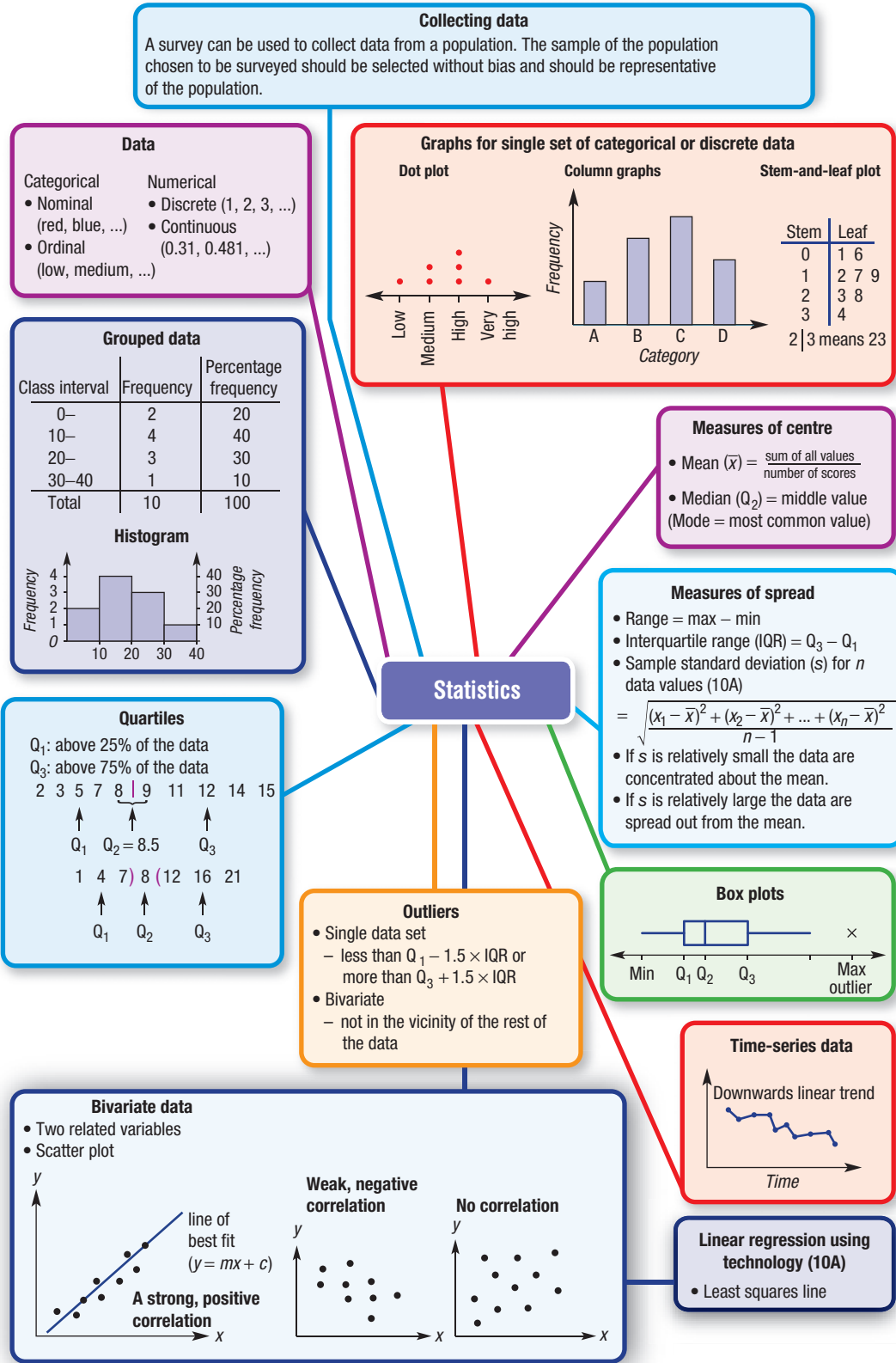
Statisticians work in many fields of industry, business, finance, research, government and social services. As computers are used to process the data, they can spend more time on higher-level skills, such as designing statistical investigations and data analysis.

- 1 The mean mass of six boys is 71 kg, and the mean mass of five girls is 60 kg. Find the average mass of all 11 people put 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 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?
- 4 Find the interquartile range for a set of data if 75% of the data are above 2.6 and 25% of the data are above 3.7.
- 5 A single data set has 3 added to every value. Describe the change in:
  - a the mean
  - b the median
  - c the range
  - d the interquartile range
  - e the standard deviation.
- 6 Three key points on a scatter plot have coordinates (1, 3), (2, 3) and (4, 9). Find a quadratic equation that fits these three points exactly.

Up for a challenge? If you get stuck on a question, check out the 'Working with unfamiliar problems' poster at the end of the book to help you.



- 7 Six numbers are written in ascending order: 1.4, 3, 4.7, 5.8,  $a$ , 11.  
Find all possible values of  $a$  if the number 11 is considered to be an outlier.
- 8 The class mean,  $\bar{x}$ , and standard deviation,  $s$ , for some Year 10 term tests are:  
Maths ( $\bar{x} = 70\%$ ,  $s = 9\%$ ); Physics ( $\bar{x} = 70\%$ ,  $s = 6\%$ ); Biology ( $\bar{x} = 80\%$ ,  $s = 6.5\%$ ).  
If Emily gained 80% in each of these subjects, which was her best and worst result? Give reasons for your answer.



## Chapter checklist: Success criteria

		✓										
9A	<p><b>1. I can describe types of data.</b> e.g. What type of data would the survey question 'How many pairs of shoes do you own?' generate?</p>											
9A	<p><b>2. I can choose a survey sample.</b> e.g. A survey is carried out by calling people listed in the phone book, to determine their voting preferences for a state election. Why will this sample not necessarily be representative of the state's views?</p>											
9B	<p><b>3. I can present data in a histogram.</b> e.g. 15 people were surveyed to find out how many hours they spend on the internet in a week. The data are: 7, 12, 14, 20, 2, 26, 8, 11, 17, 12, 21, 5, 6, 18, 14 Construct a histogram for the data using class intervals of 5, showing both the frequency and percentage frequency on the one graph.</p>											
9B	<p><b>4. I can analyse data in a statistical graph.</b> e.g. For the stem and leaf plot shown below, find the mean correct to one decimal place, the median and the mode.</p> <table style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th style="border-right: 1px solid black; padding: 5px;">Stem</th> <th style="padding: 5px;">Leaf</th> </tr> </thead> <tbody> <tr> <td style="border-right: 1px solid black; text-align: center;">0</td> <td style="text-align: center;">2 5 7</td> </tr> <tr> <td style="border-right: 1px solid black; text-align: center;">1</td> <td style="text-align: center;">1 1 4 6</td> </tr> <tr> <td style="border-right: 1px solid black; text-align: center;">2</td> <td style="text-align: center;">0 3 9</td> </tr> <tr> <td style="border-right: 1px solid black; text-align: center;">3</td> <td style="text-align: center;">2 5</td> </tr> </tbody> </table> <p style="text-align: center;">2 3 means 23</p>	Stem	Leaf	0	2 5 7	1	1 1 4 6	2	0 3 9	3	2 5	
Stem	Leaf											
0	2 5 7											
1	1 1 4 6											
2	0 3 9											
3	2 5											
9C	<p><b>5. I can find the five-figure summary and interquartile range.</b> e.g. For the data set below find the minimum, maximum, median, upper and lower quartiles and the range and IQR. 7, 10, 12, 12, 14, 18, 22, 25, 26, 30</p>											
9C	<p><b>6. I can find any outliers in a data set.</b> e.g. The following data represent the number of aces by a tennis player in 11 grand slam matches for the year: 15, 12, 22, 2, 10, 18, 16, 14, 15, 20, 16 For the data find the upper and lower quartiles and use these to help determine if there are any outliers.</p>											
9D	<p><b>7. I can construct a box plot.</b> e.g. For the data set: 5, 8, 2, 1, 6, 3, 3, 1, 4, 18, 2, 8, 5, draw a box plot to summarise the data, marking outliers if they exist.</p>											
9E	<p><b>8. I can calculate the standard deviation.</b> e.g. For the data set 10, 5, 4, 7, 2, calculate the mean and standard deviation correct to one decimal place.</p>	10A										







9E

**9. I can interpret a standard deviation value.** 10A  
 e.g. This back-to-back stem-and-leaf plot shows the average monthly maximum temperatures for a year in New York and Melbourne. The mean and standard deviation are given.

<b>New York Leaf</b>	<b>Stem</b>	<b>Melbourne Leaf</b>	Melbourne: $\bar{x} = 19.8, s = 4.6$ New York: $\bar{x} = 17.1, s = 9.4$
7 5 4	0		
8 7 2 0	1	3 4 5 7 7	
9 9 7 5 2	2	0 0 2 4 4 6 6	
	1   7 means 17°C		

By looking at the stem-and-leaf plot, suggest why New York's mean is less than that of Melbourne and why New York's standard deviation is larger than that of Melbourne.

9F

**10. I can plot and interpret a time-series plot.**  
 e.g. The approximate number of DVD rental stores in a city over a 10-year period is shown below.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Number of DVD stores	72	65	56	56	31	22	14	14	8	6

Plot the time series and describe the trend in the data over the 10 years.

9G

**11. I can construct and interpret a scatter plot.**  
 e.g. For the bivariate data set below, draw a scatter plot and describe the correlation between  $x$  and  $y$  as positive or negative and strong or weak.

$x$	5	8	12	4	6	15	11	3
$y$	4.4	6	11	4.7	5.3	11.6	10.3	2.4

9H

**12. I can fit a line of best fit by eye and use the line to make predictions.**  
 e.g. For the scatter plot from the data set above, fit a line of best fit by eye on the scatter plot and use it to estimate  $y$  when  $x = 10$  and  $x$  when  $y = 8$ .

9H

**13. I can find the equation of a line of best fit.**  
 e.g. A scatter plot shows a linear relationship between two variables  $x$  and  $y$ . If the trend line passes through (20, 15) and (40, 25), find the equation of the trend line and use it to estimate  $x$  when  $y = 50$ .

9I

**14. I can find and use a regression line using technology.** 10A  
 e.g. For the data set below, use technology to construct a scatter plot for the data and find the equation of the least squares regression line. Use the equation to estimate the value of  $y$  when  $x$  is 12.

$x$	2	3	5	6	7	8	8	9
$y$	10.8	10.6	9.2	4.7	7.3	5.6	6.2	4.1

## Short-answer questions

9B

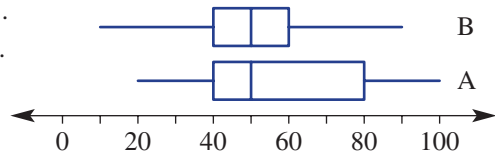
- 1 A group of 16 people was surveyed to find the number of hours of television they watch in a week. The raw data are listed:
- 6, 5, 11, 13, 24, 8, 1, 12  
7, 6, 14, 10, 9, 16, 8, 3
- Organise the data into a table with class intervals of 5 and include a percentage frequency column.
  - Construct a histogram for the data, showing both the frequency and percentage frequency on the graph.
  - Would you describe the data as symmetrical, positively skewed or negatively skewed?
  - Construct a stem-and-leaf plot for the data, using 10s as the stem.
  - Use your stem-and-leaf plot to find the median.

9D

- 2 For each set of data below, complete the following tasks.
- Find the range.
  - Find the lower quartile ( $Q_1$ ) and the upper quartile ( $Q_3$ ).
  - Find the interquartile range.
  - 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

9D

- 3 Compare these parallel box plots, A and B, and answer the following as true or false.
- The range for A is greater than the range for B.
  - The median for A is equal to the median for B.
  - The interquartile range is smaller for B.
  - 75% of the data for A sit below 80.



9G

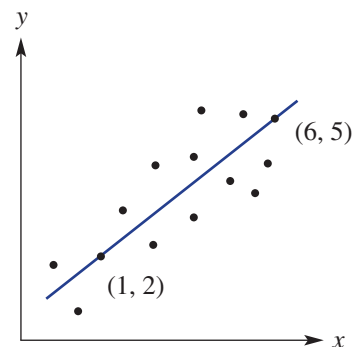
- 4 Consider the simple bivariate data set.

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

- 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.

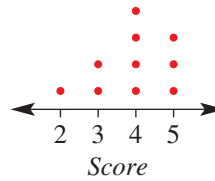
9H

- 5 The line of best fit passes through the two points labelled on this graph.
- Find the equation of the line of best fit.
  - Use your equation to estimate the value of  $y$  when:
    - $x = 4$
    - $x = 10$
  - Use your equation to estimate the value of  $x$  when:
    - $y = 3$
    - $y = 12$





Questions 2–4 refer to the dot plot shown at right.

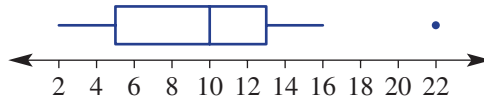


- 9B 2 The mean of the scores in the data is:
- A 3.5
  - B 3.9
  - C 3
  - D 4
  - E 5

- 9B 3 The mode for the data is:
- A 3.5
  - B 2
  - C 3
  - D 4
  - E 5

- 9B 4 The dot plot is:
- A symmetrical
  - B positively skewed
  - C negatively skewed
  - D bimodal
  - E correlated

Questions 5 and 6 refer to this box plot.



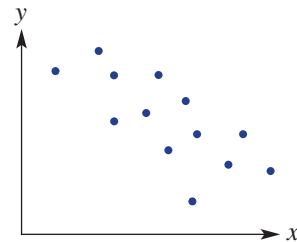
- 9D 5 The interquartile range is:
- A 8
  - B 5
  - C 3
  - D 20
  - E 14

- 9D 6 The range is:
- A 5
  - B 3
  - C 20
  - D 14
  - E 8

- 9E 7 The sample standard deviation for the small data set 1, 1, 2, 3, 3 is:
- A 0.8
  - B 2
  - C 1
  - D 0.9
  - E 2.5

10A

- 9G 8 The variables  $x$  and  $y$  in this scatter plot could be described as having:
- A no correlation
  - B a strong, positive correlation
  - C a strong, negative correlation
  - D a weak, negative correlation
  - E a weak, positive correlation



- 9H 9 The equation of the line of best fit for a set of bivariate data is given by  $y = 2.5x - 3$ . An estimate for the value of  $x$  when  $y = 7$  is:
- A -1.4
  - B 1.2
  - C 1.6
  - D 7
  - E 4

- 9H 10 The equation of the line of best fit connecting the points (1, 1) and (4, 6) is:
- A  $y = 5x + 3$
  - B  $y = \frac{5}{3}x - \frac{2}{3}$
  - C  $y = -\frac{5}{3}x + \frac{8}{3}$
  - D  $y = \frac{5}{3}x - \frac{8}{3}$
  - E  $y = \frac{3}{5}x - \frac{2}{3}$

## Extended-response questions

- 1 The number of flying foxes taking refuge in two different fig trees was recorded over a period of 14 days. The data collected are given here.

<b>Tree 1</b>	56	38	47	59	63	43	49	51	60	77	71	48	50	62
<b>Tree 2</b>	73	50	36	82	15	24	73	57	65	86	51	32	21	39

- a Find the IQR for:
- tree 1
  - tree 2
- b Identify any outliers for:
- tree 1
  - tree 2
- c Draw parallel box plots for the data.
- d By comparing your box plots, describe the difference in the ways the flying foxes use the two fig trees for taking refuge.



- 2 The approximate number of shoppers in an air-conditioned shopping plaza was recorded for 14 days, along with the corresponding maximum daily outside temperatures for those days.

<b>Day</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14
<b>Max. daily temp. (T) (°C)</b>	27	26	28	33	38	36	28	30	32	25	25	27	29	33
<b>No. of shoppers (N)</b>	1050	950	1200	1550	1750	1800	1200	1450	1350	900	850	700	950	1250

- a Draw a scatter plot for the number of shoppers versus the maximum daily temperatures, with the number of shoppers, correct to the nearest whole number, on the vertical axis, and describe the correlation between the variables as either positive, negative or none.
- b Use technology to determine the least squares regression line for the data, rounding coefficients to two decimal places.
- c Use your least squares regression equation to estimate:
- the number of shoppers on a day, correct to the nearest whole number, with a maximum daily temperature of 24°C
  - the maximum daily temperature, correct to one decimal place, if the number of shoppers at the plaza is 1500.