

Chapter

9

Statistics

What you will learn

- 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 with technology (10A)

Australian curriculum

STATISTICS AND PROBABILITY

Data representation and interpretation

Determine quartiles and interquartile range.

Construct and interpret box plots and use them to compare data sets. Compare shapes of box plots with corresponding histograms and dot plots.

Use scatter plots to investigate and comment on relationships between two continuous variables.

Investigate and describe bivariate numerical data for which the independent variable is time.

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

(10A) Calculate and interpret the mean and standard deviation data and use these to compare data sets.

(10A) Use information technologies to investigate bivariate numerical data sets. Where appropriate, use a straight line to describe the relationship, allowing for variation.



Online resources

- Chapter pre-test
- Videos of all worked examples
- Interactive widgets
- Interactive walkthroughs
- Downloadable HOTsheets
- Access to HOTmaths Australian Curriculum courses

Closing the gap

According to the Australian Bureau of Statistics the population of Indigenous Australians has increased from 351 000 in 1991 to 548 000 in 2011. This represents an annual increase of 2.3% compared with a total Australian annual population increase of 1.2%.

The life expectancy for Indigenous Australians, however, is not as high as for non-Indigenous Australians. For males it is about 69 years (compared to about 80 years Australia-wide) and about

74 years for females (compared to about 83 years Australia-wide). This represents a life expectancy gap for Indigenous Australians of about 11 years for males and 9 years for females.

The Australian government works with departments such as the Australian Bureau of Statistics to collect accurate data, like the examples above, so it can best determine how to allocate resources in assisting Indigenous communities.

9A Collecting and using data



Interactive



Widgets



HOTSheets



Walkthroughs

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.



Let's start: 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?

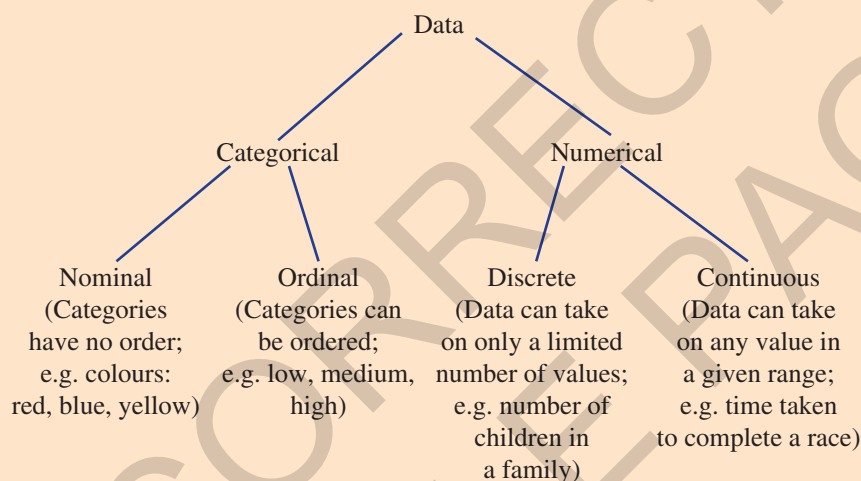
Key ideas

■ **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.



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.



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.

Exercise 9A

1-4

3-4

—

UNDERSTANDING

- 1 Match each word (a–e) with its definition (A–E).

<p>a population</p> <p>b census</p> <p>c sample</p> <p>d survey</p> <p>e data</p>	<p>A a group chosen from a population</p> <p>B a tool used to collect statistical data</p> <p>C all the people or objects in question</p> <p>D statistics collected from every member of the population</p> <p>E the factual information collected from a survey or other source</p>
---	--
- 2 Match each word (a–f) with its definition (A–F).

<p>a numerical</p> <p>b continuous</p> <p>c discrete</p> <p>d categorical</p> <p>e ordinal</p> <p>f nominal</p>	<p>A categorical data that has no order</p> <p>B data that are numbers</p> <p>C numerical data that take on a limited number of values</p> <p>D data that can be divided into categories</p> <p>E numerical data that take any value in a given range</p> <p>F categorical data that can be ordered</p>
---	---
- 3 Classify each set of data as categorical or numerical.

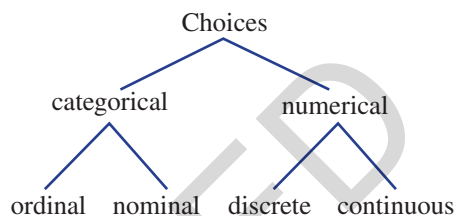
<p>a 4.7, 3.8, 1.6, 9.2, 4.8</p> <p>b red, blue, yellow, green, blue, red</p> <p>c low, medium, high, low, low, medium</p> <p>d 3 g, 7 g, 8 g, 7 g, 4 g, 1 g, 10 g</p>	
--	--
- 4 Which one of the following survey questions would generate categorical data?

<p>A How many times do you eat at your favourite fast-food place in a typical week?</p> <p>B How much do you usually spend buying your favourite fast food?</p> <p>C How many items did you buy last time you went to your favourite fast-food place?</p> <p>D Which is your favourite fast food?</p>	
---	--

Example 1

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

- a** How many people under the age of 18 years are there in your immediate family?
- b** How many letters are there in your first name?
- c** Which company is the carrier of your mobile telephone calls? Optus/Telstra/Vodafone/3/Virgin/Other (Please specify.)
- d** What is your height?
- e** How would you describe your level of application in Maths? (Choose from very high, high, medium or low.)



FLUENCY

Example 2

6 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.

7 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.

8 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.

PROBLEM-SOLVING

9A

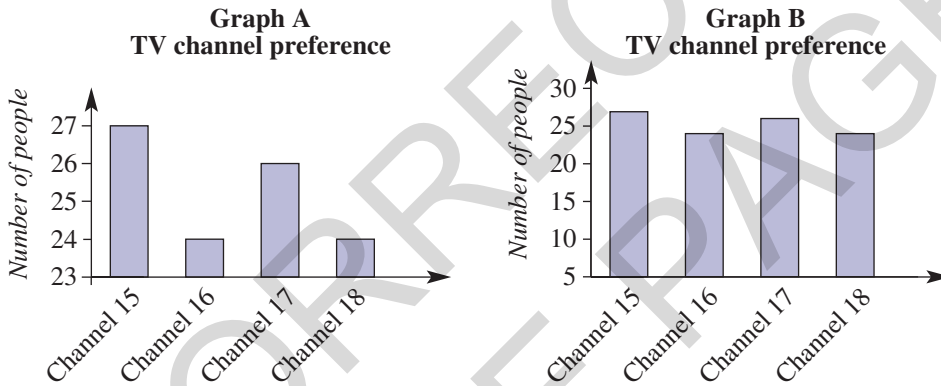
- 9 A television news reporter surveyed four companies and found that the profits of three of these companies had reduced over the past year. They report that this means the country is facing an economic downturn and that only one in four companies is making a profit.
- What are some of the problems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

10

10, 11

11, 12

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



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

The 2011 Australian Census

—

—

13, 14

- 13 Research the 2011 Australian Census on the website of the Australian Bureau of Statistics. Find out something interesting from the results of the 2011 Australian Census and write a short news report.
- 14 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



Interactive



Widgets



HOTSheets



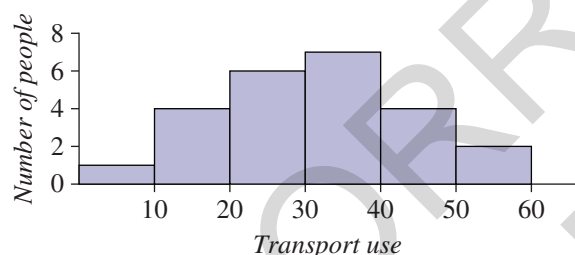
Walkthroughs

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.



Let's start: Public transport analysis

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.

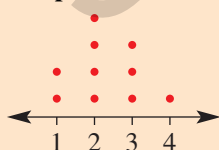
You may wish to include these points:

- 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.

■ 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

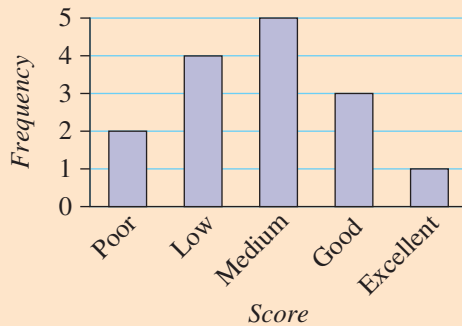
Stem	Leaf
0	1 3
1	2 5 9
2	1 4 6 7
3	0 4

2 | 4 means 24

Key
ideas

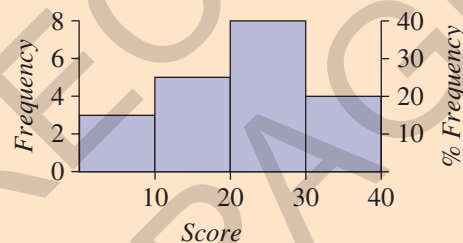
Key
ideas

- **Column graph**

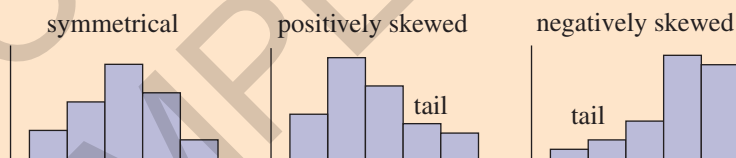


- **Histograms** can be used for grouped discrete or continuous numerical data. The interval 10– includes all numbers from 10 (including 10) to fewer than 20.

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**.



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

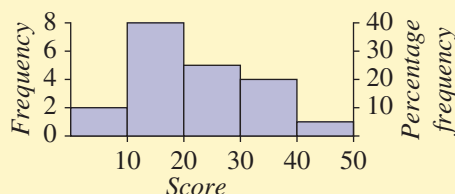
- Organise the data into a frequency table using class intervals of 10. Include a percentage frequency column.
- Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- Construct a stem-and-leaf plot for the data.
- Use your stem-and-leaf plot to find 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
Total	20	100

b Number of times the internet is accessed



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

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

EXPLANATION

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

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.

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

After counting the scores 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.

Exercise 9B

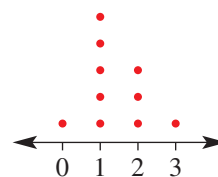
1–2

2(a)

—

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

- a** How many families were surveyed?
- b** Find the mean number of children in the families surveyed.
- c** Find the median number of children in the families surveyed.
- d** Find the mode for the number of children in the families surveyed.
- e** What percentage of the families have, at most, two children?



UNDERSTANDING

9B

2 Complete these frequency tables.

a

Class interval	Frequency	Percentage frequency
0–	2	
10–	1	
20–	5	
30–40	2	
Total		

b

Class interval	Frequency	Percentage frequency
80–	8	
85–	23	
90–	13	
95–100		
Total	50	

UNDERSTANDING

Example 3

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.

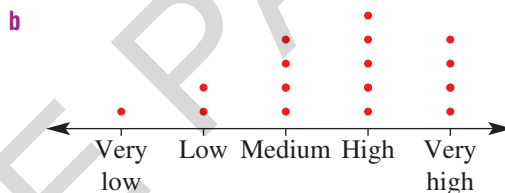
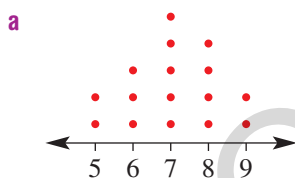


FLUENCY

- 4 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	
Total	40	

- a Copy and complete the table.
- b Use the table to find:
- the frequency of people who travel by train
 - the most popular form of transport
 - the percentage of people who travel by car
 - the percentage of people who walk or cycle to work
 - the percentage of people who travel by public transport, including trains, buses and trams.
- 5 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

- 6 For the data in these stem-and-leaf plots, find:
- the mean (rounded to one decimal place)
 - the median
 - 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

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

Game	1	2	3	4	5	6	7	8	9	10	11	12
Nick	0	2	2	0	3	1	2	1	2	3	0	1
Jayden	0	0	4	1	0	5	0	3	1	0	4	0

- Draw a dot plot to display Nick's goal-scoring achievement.
 - Draw a dot plot to display Jayden's goal-scoring achievement.
 - How would you describe Nick's scoring habits?
 - How would you describe Jayden's scoring habits?
- 8 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.

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

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



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

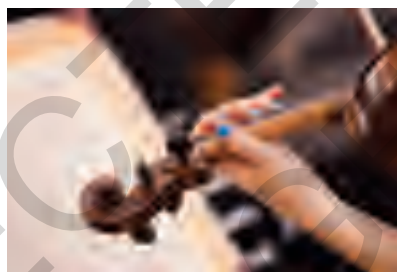
- Construct a table using these column headings: Mass, Frequency and Percentage frequency.
- Find the total number of mice weighed in the experiment.
- State the percentage of mice that were in the 20– gram interval.
- Which was the most common weight interval?
- What percentage of mice were in the most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

Mass (grams)	Tally
10–	
15–	
20–	
25–	
30–35	

- 10** 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.

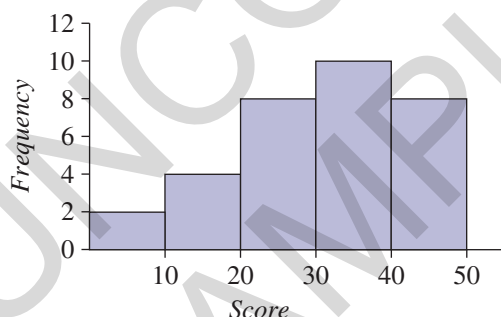


11

11, 12

12, 13

- 11** 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



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

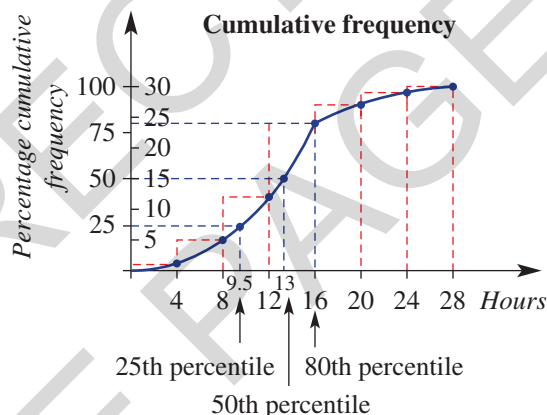
$$\text{Percentage cumulative frequency} = \frac{\text{cumulative frequency}}{\text{total number of data elements}} \times 100$$

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		

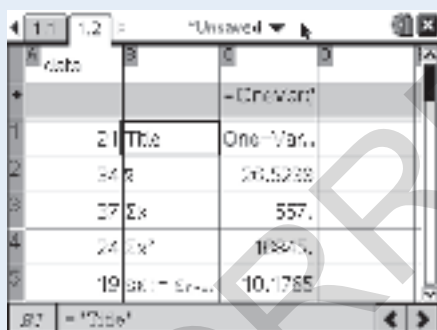
- Copy and complete the table. Round the percentage cumulative frequency to one decimal place.
- Find the number of houses that have gas bills of less than \$120.
- Construct a cumulative frequency curve for the gas bills.
- Estimate the following percentiles.
 - 50th
 - 20th
 - 80th
- In this street,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

Using calculators to graph grouped data

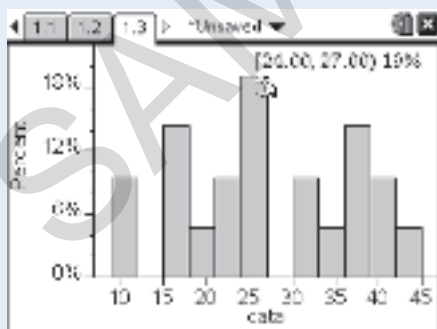
- 1 Enter the following data in a list called *data* and find the mean and median.
21, 34, 37, 24, 19, 11, 15, 26, 43, 38, 25, 16, 9, 41, 36, 31, 24, 21, 30, 39, 17
- 2 Construct a histogram using intervals of 3 and percentage frequency for the data above.

Using the TI-Nspire:

- 1 Go to a **Lists and spreadsheets** page and enter the data into list A. Select **menu, Statistics, Stat Calculations, One-Variable Statistics**. Press **enter** to finish and scroll to view the statistics.



- 2 Go to a **Data and Statistics** page and select the *data* variable for the horizontal axis. Select **menu, Plot Type, Histogram**. Then select **menu, Plot Properties, Histogram Properties, Bin Settings**. Choose the **Width** to be 3 and **Alignment** to be 0. Drag the scale to suit. Select **menu, Plot Properties, Histogram Properties, Histogram Scale** to receive the percentage frequency.

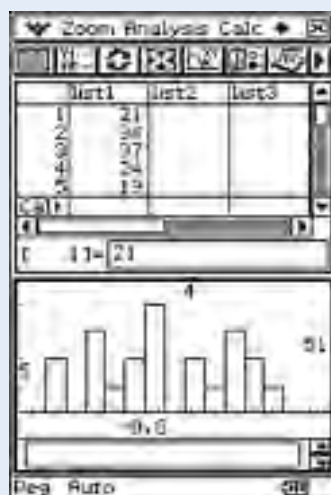


Using the ClassPad:

- 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 **Bin** and set **HStart** to 9 and **HStep** to 3.



9C Summary statistics



Interactive



Widgets



HOTSheets



Walkthroughs

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.



Let's start: House prices

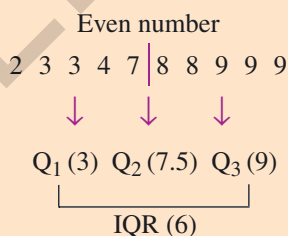
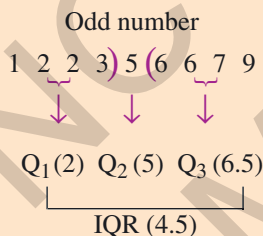
A real estate agent tells you that the median house price for a suburb in 2015 was \$753 000 and the mean was \$948 000.

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

Key ideas

■ 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
= $Q_3 - Q_1$
- 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).

- Lower fence = $Q_1 - 1.5 \times \text{IQR}$ or
- Upper fence = $Q_3 + 1.5 \times \text{IQR}$



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_1 Q_2 (7) Q_3

$Q_2 = 7$, 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

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

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

$$\text{IQR} = 2.55 - 1.8 = 0.75$$

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.

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

Max = 2.9, min = 1.6

1.6 1.7 | 1.9 2.0 2.1 2.4 2.4 | 2.7 2.9
 ↑ ↑ ↑
 Q_1 Q_2 Q_3

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



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?

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 higher 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.

Exercise 9C

1–3

3

—

- 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.
 1, 4, 2, 2, 3, 8, 1, 2, 2, 0, 3, 1, 2
- a** Arrange the data in ascending order.
 - b** Find the median (i.e. the middle value).
 - c** 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).
 - d** Determine the interquartile range (IQR).
 - e** Calculate $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$.
 - f** 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
- Find the median (i.e. average of the middle two numbers).
 - Find the lower quartile (i.e. middle of the smallest four numbers).
 - 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 $Q_3 + 1.5 \times \text{IQR}$)?

4–5($\frac{1}{2}$), 64–5($\frac{1}{2}$), 6, 74–5($\frac{1}{2}$), 6, 7

Example 4

- 4** 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

- 5** Determine the median and mean of the following sets of data. Round to one decimal place where necessary.

a 6, 7, 8, 9, 10**b** 2, 3, 4, 5, 5, 6**c** 4, 6, 3, 7, 3, 2, 5

Example 5

- 6** 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:

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



9C

7 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

8, 9

9, 10

10, 11

8 Twelve different calculators had the following numbers of buttons.

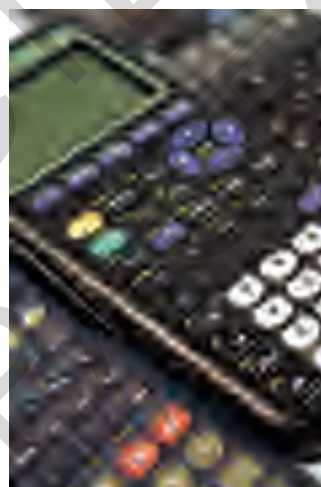
36, 48, 52, 43, 46, 53, 25, 60, 128, 32, 52, 40

a For the given data, find:

- 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?



9 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

10 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	6 8
2	0 4 6
3	2 3

2 | 4 means 24

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

12

12, 13

13–15

9C

REASONING

- 12** 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.
- 13** 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.
- 14** Give an example of a small data set that satisfies the following.
- a** median = mean **b** median = upper quartile **c** range = IQR
- 15** Explain why, in many situations, the median is preferred to the mean as a measure of centre.

Some research

16

ENRICHMENT

- 16** 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



Interactive



Widgets



HOTSheets

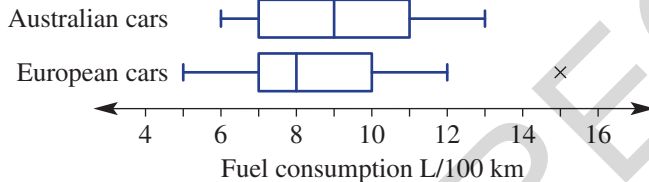


Walkthroughs

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.

Let's start: 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.

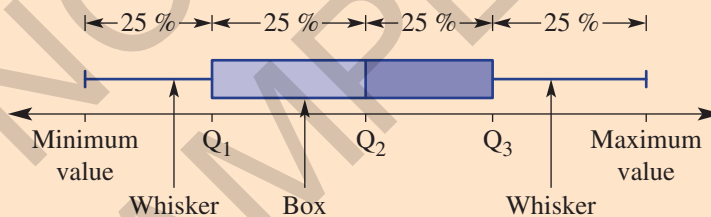


- What do the box plots say about how the fuel consumption compares between Australian-made and European-made cars?
- What does each part of the box plot represent?
- What do you think the cross (x) represents on the European cars box plot?

Key ideas

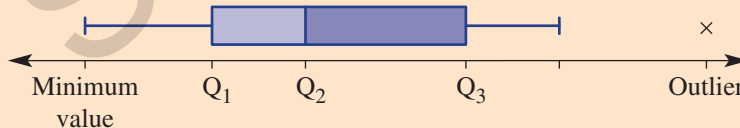
- A **box plot** (also called a box-and-whisker plot) can be used to summarise a data set.

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



- An **outlier** is marked with a cross (x).

- 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.



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

a

2	3	3	4	5	5	5	6	6	7	9	12
			↑			↑			↑		
			Q_1			Q_2			Q_3		

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

$$= 3.5$$

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

$$= 6.5$$

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

$$= 3$$

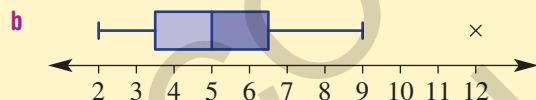
$$Q_1 - 1.5 \times \text{IQR} = 3.5 - 1.5 \times 3$$

$$= -1$$

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

$$= 11$$

$\therefore 12$ 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.

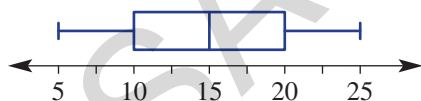
Exercise 9D

1–3

2

—

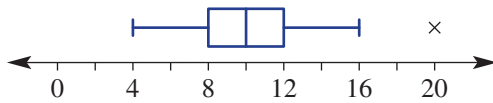
- 1 For this simple box plot, find:



- | | |
|----------------------------------|--------------------------------|
| 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). | |

UNDERSTANDING

- 2 Complete the following for this box plot.



- a Find the IQR. b Calculate $Q_1 - 1.5 \times \text{IQR}$.
 c Calculate $Q_3 + 1.5 \times \text{IQR}$. d State the value of the outlier.
 e Check that the outlier is greater than $Q_3 + 1.5 \times \text{IQR}$.
- 3 Construct a box plot that shows these features.
- a $\text{min} = 1$, $Q_1 = 3$, $Q_2 = 4$, $Q_3 = 7$, $\text{max} = 8$
 b outlier = 5, minimum above outlier = 10, $Q_1 = 12$, $Q_2 = 14$, $Q_3 = 15$, $\text{max} = 17$

4–5($\frac{1}{2}$)4–5($\frac{1}{2}$)4–5($\frac{1}{2}$)

Example 6

- 4 Consider the data sets below.

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

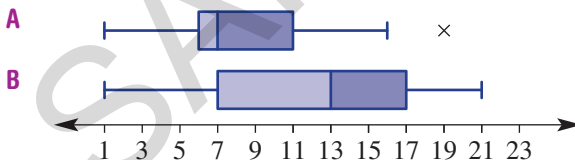
- a 4, 6, 5, 2, 3, 4, 4, 13, 8, 7, 6
 b 1.8, 1.7, 1.8, 1.9, 1.6, 1.8, 2.0, 1.1, 1.4, 1.9, 2.2
 c 21, 23, 18, 11, 16, 19, 24, 21, 23, 22, 20, 31, 26, 22
 d 0.04, 0.04, 0.03, 0.03, 0.05, 0.06, 0.07, 0.03, 0.05, 0.02
- 5 First, find Q_1 , Q_2 and Q_3 and then draw box plots for the given data sets. Remember to find outliers and mark them on your box plot if they exist.
- a 11, 15, 18, 17, 1, 2, 8, 12, 19, 15
 b 37, 48, 52, 51, 51, 42, 48, 47, 39, 41, 65
 c 0, 1, 5, 4, 4, 4, 2, 3, 3, 1, 4, 3
 d 124, 118, 73, 119, 117, 120, 120, 121, 118, 122

6, 7

7, 8

7, 8

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



- a What statistical measure do these box plots have in common?
 b Which data set (A or B) has a wider range of values?
 c Find the IQR for:
 i data set A ii data set B.
 d How would you describe the main difference between the two sets of data from which the parallel box plots have been drawn?

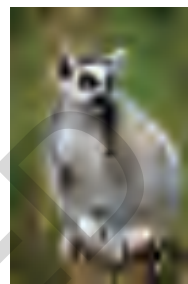
- 7 The following masses, in kilograms, of 15 Madagascan lemurs are recorded as part of a conservation project.

14.4, 15.5, 17.3, 14.6, 14.7

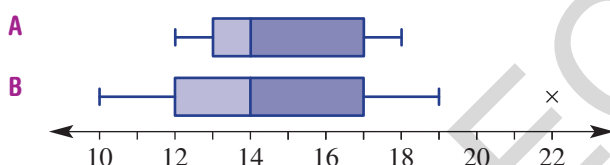
15.0, 15.8, 16.2, 19.7, 15.3

13.8, 14.6, 15.4, 15.7, 14.9

- Find Q_1 , Q_2 and Q_3 .
- Which masses, if any, would be considered outliers?
- Draw a box plot to summarise the lemurs' masses.



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



- What statistical measures do these box plots have in common?
- Which data set (A or B) has a wider range of values?
- Find the IQR for:
 - data set A
 - data set B.
- How would you describe the main difference between the two sets of data from which the parallel box plots have been drawn?

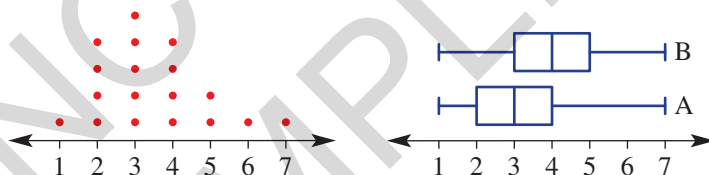
9

9, 10

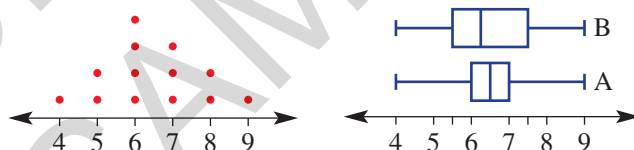
10, 11

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

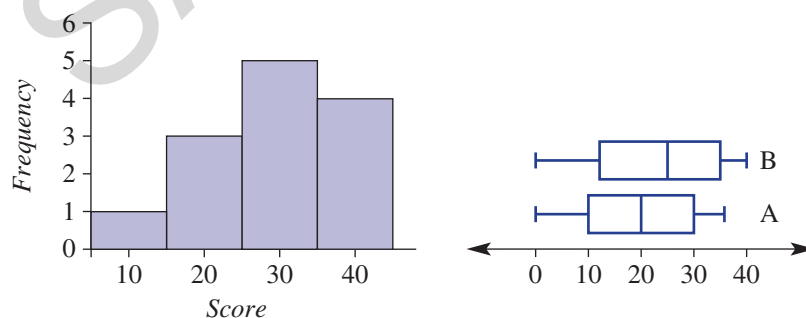
a



b



c



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

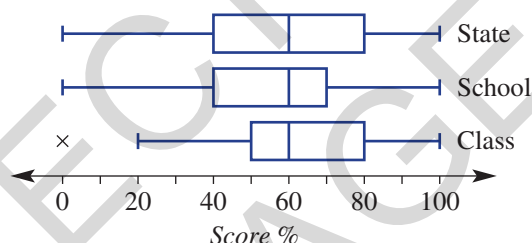
3, 2, 4, 6, 8, 4, 6, 7, 6, 1, 7, 12, 7, 3, 8

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

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

- a** Draw parallel box plots for the data.
b Do you believe there is a major difference in the way the essays were marked by the two examiners? If yes, describe this difference.

- 11** 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?

Creating your own parallel box plots

12

- 12 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

- 1 Type these data into lists and define them as Test A and Test B.

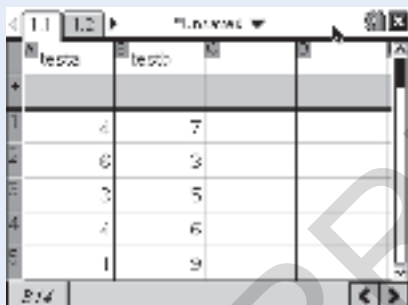
Test A: 4, 6, 3, 4, 1, 3, 6, 4, 5, 3, 4, 3

Test B: 7, 3, 5, 6, 9, 3, 6, 7, 4, 1, 4, 6

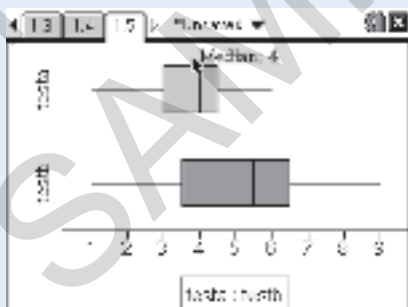
- 2 Draw parallel box plots for the data.

Using the TI-Nspire:

- 1 Go to a new **Lists and spreadsheets** page and enter the data into the lists. Give each column a title.

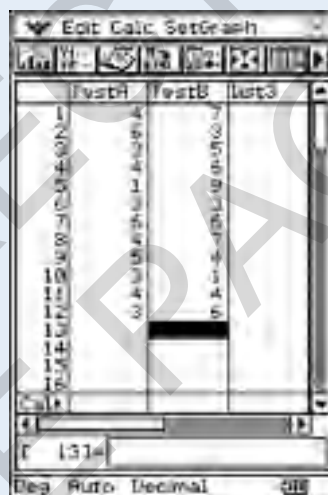


- 2 Go to a new **Data and Statistics** page and select the **testA** variable for the horizontal axis. Select **menu, Plot Type, Box Plot**. Trace to reveal the statistical measures. To show the box plot for **testB**, go to **menu, Plot Properties, Add X Variable** and click on **testB**.

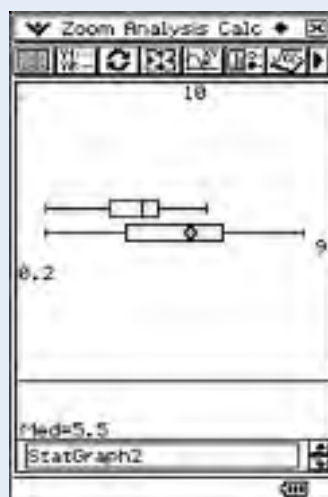


Using the ClassPad:

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



- 2 Tap . For graph 1, set **Draw** to **On**, **Type** to **Med-Box**, **XList** to **main\TestA** and **Freq** to **1**. For graph 2, set **Draw** to **On**, **Type** to **MedBox**, **XList** to **main\TestB** and **Freq** to **1**. Tap **Set**. Tap .



9E

Standard deviation

10A



Interactive



Widgets



HOTSheets



Walkthroughs

For a single data set we have already used 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 σ) that can be used to find the standard deviation.

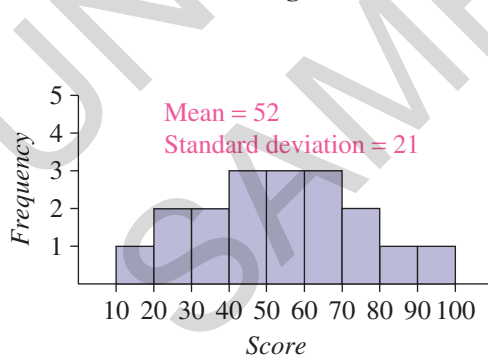
Let's start: Which is the better team?

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.

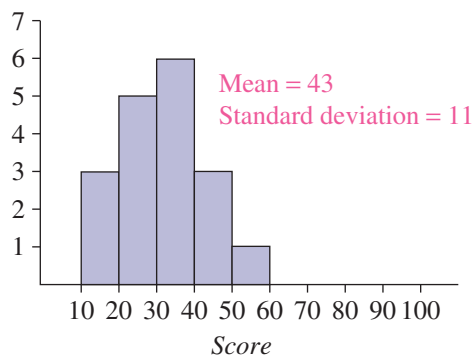
- Which team has the higher mean? What does this say about the team's performance?
- Which team has the smaller standard deviation? What does this say about the team's performance? Discuss.



Eagles



Monsters



- The **standard deviation** is a number that describes the spread of data about the mean.
 - The symbol used for standard deviation is σ (sigma).
 - 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**, 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.
- If the data represent the complete population, then divide by n instead of $(n - 1)$. This would give the **population standard deviation**. 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.

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



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

$$\begin{aligned}\bar{x} &= \frac{2 + 4 + 5 + 8 + 9}{5} \\ &= 5.6\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{(2 - 5.6)^2 + (4 - 5.6)^2 + (5 - 5.6)^2 + (8 - 5.6)^2 + (9 - 5.6)^2}{5 - 1}} \\ &= \sqrt{\frac{(-3.6)^2 + (-1.6)^2 + (-0.6)^2 + (2.4)^2 + (3.4)^2}{4}} \\ &= 2.9 \text{ (to 1 d.p.)}\end{aligned}$$

EXPLANATION

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

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

Deviation 1 is $2 - 5.6 = -3.6$

Deviation 2 is $4 - 5.6 = -1.6$ etc.



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

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

EXPLANATION

- | | |
|--|---|
| <p>a The maximum score for Darwin is 35. Sydney's mean is affected by several values larger than 35.</p> | <p>The mean depends on every value in the data set.</p> |
| <p>b The data for Sydney are more spread out from the mean. Darwin's scores are more closely clustered near its mean.</p> | <p>Sydney has more scores with a large distance from its mean. Darwin's scores are closer to the Darwin mean.</p> |
| <p>c Sydney is a larger city and more spread out, so people have to travel farther to get to work.</p> | <p>Higher populations often lead to larger cities and longer travel distances.</p> |



Exercise 9E

1–4

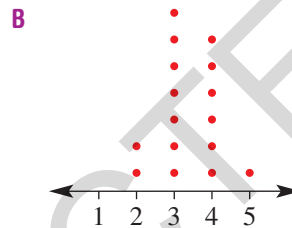
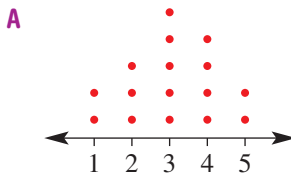
3, 4

UNDERSTANDING

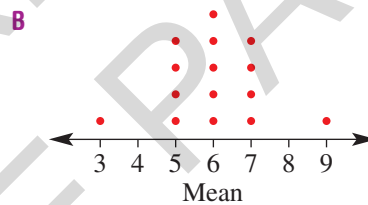
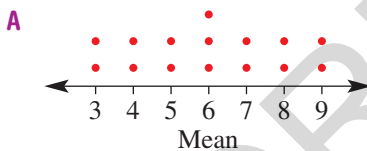
1 Insert the word *smaller* or *larger* into each sentence.

- 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.



- a Which data set (A or B) would have the higher standard deviation?
- b Give a reason for your answer in part a.



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 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

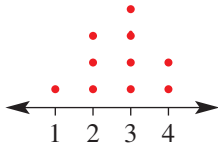
2 | 8 means 28

9E

Example 7

- 5 Calculate the mean and sample standard deviation for these small data sets. Round the standard deviation to one decimal place where necessary.
- a 3, 5, 6, 7, 9 b 1, 1, 4, 5, 7
c 2, 5, 6, 9, 10, 11, 13 d 28, 29, 32, 33, 36, 37
- 6 Calculate the mean and sample standard deviation for the data in these graphs, correct to one decimal place.

a



b

Stem	Leaf
0	4
1	1 3 7
2	0 2

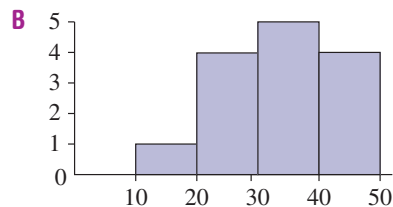
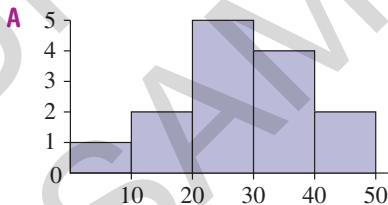
1 | 7 means 17

Example 8

- 7 This back-to-back stem-and-leaf plot shows the distribution of distances travelled by students at an inner-city and an outer-suburb school. The means and standard deviations are given.

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

- a Why is the mean for the outer-suburb school larger than that for the inner-city school?
b Why is the mean for the outer-suburb school larger than that for the inner-city school?
c Why is the standard deviation for the inner-city school smaller than that for the outer-suburb school?
d Give a practical reason for the difference in centre and spread for the two schools.
- 8 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.
- 9 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

FLUENCY

PROBLEM-SOLVING



10

10, 11

11, 12

9E

REASONING

- 10** 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

- Is the range for set A equal to the range for set B?
 - Is the mean for each data sets the same?
 - Is the median for each data set the same?
 - Would the standard deviation be affected by the outlier? Explain.
- 11** Data sets 1 and 2 have means \bar{x}_1 and \bar{x}_2 , and standard deviations σ_1 and σ_2 .
- If $\bar{x}_1 > \bar{x}_2$, does this necessarily mean that $\sigma_1 > \sigma_2$? Give a reason.
 - If $\sigma_1 < \sigma_2$ does this necessarily mean that $\bar{x}_1 < \bar{x}_2$?
- 12** 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.

Study scores

—

—

13

ENRICHMENT

- 13** 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 population standard deviation (σ) is 16.0.

- a** Calculate:

i $\bar{x} + \sigma$

ii $\bar{x} - \sigma$

iii $\bar{x} + 2\sigma$

iv $\bar{x} - 2\sigma$

v $\bar{x} + 3\sigma$

vi $\bar{x} - 3\sigma$

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

- one standard deviation from the mean
- two standard deviations from the mean
- three standard deviations from the mean.

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



Progress quiz

9A 1 What type of data would these survey questions generate?

- a** How many pets do you have?
- b** 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	33	20	19	27	23	12	38	15

- a** Organise the data into a frequency table, using class intervals of 10. Include a percentage frequency column.
- b** Construct a histogram for the data, showing both the frequency and percentage frequency on the one graph.
- c** Construct a stem-and-leaf plot for the data.
- d** 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.



- a** 4, 9, 12, 15, 16, 18, 20, 23, 28, 32
- b** 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

- a** For the data, find:
 - i** the minimum and maximum number of cars
 - ii** the median
 - iii** the upper and lower quartiles
 - iv** the IQR
 - v** any outliers.
- b** 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

- a** Determine whether any outliers exist by first finding Q_1 and Q_3 .
- b** 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

10A



9F Time-series data



Interactive



Widgets



HOTSheets



Walkthroughs

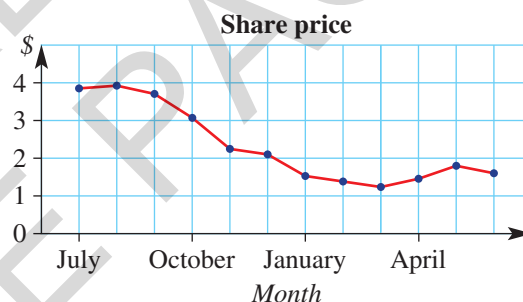
A time series is a sequence of data values that are recorded at regular time intervals. Examples include temperature recorded on the hour, speed recorded every second, population recorded every year and profit recorded every month. A line graph can be used to represent time-series data and these can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



Let's start: 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?

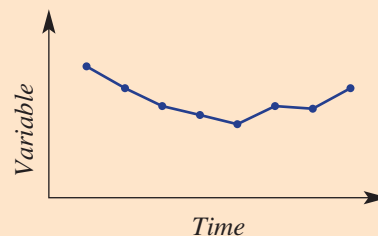


■ **Time-series data** are recorded at regular time intervals.

■ The graph or plot of a time series uses:

- time on the horizontal axis
- line segments connecting points on the graph.

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



Key ideas



Example 9 Plotting and interpreting a time-series plot

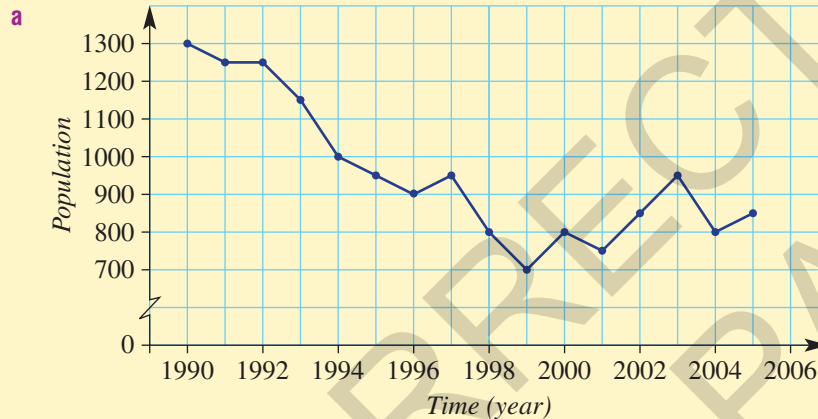
The approximate population of an outback town is recorded from 1990 to 2005.

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

a Plot the time series.

b Describe the trend in the data over the 16 years.

SOLUTION



EXPLANATION

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

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

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

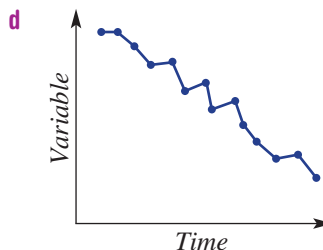
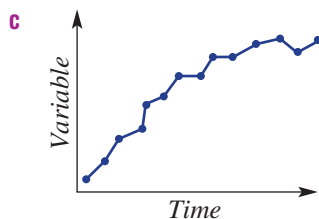
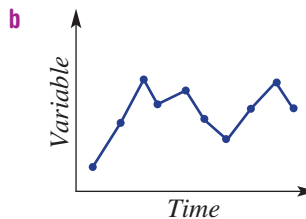
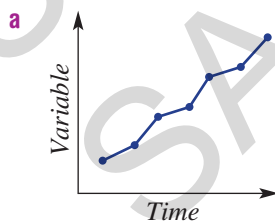
Exercise 9F

1, 2

2

—

1 Describe the following time-series plots as having a linear (i.e. straight-line trend), non-linear trend (i.e. a 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:

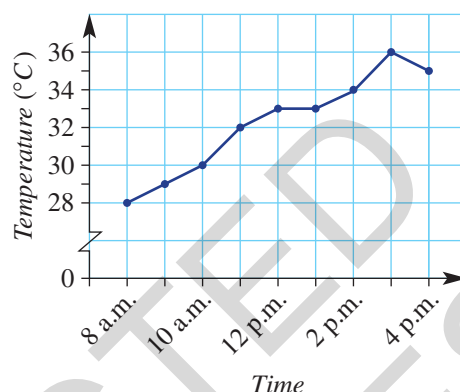
- i 8 a.m. ii 12 p.m.
iii 1 p.m. iv 4 p.m.

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.



UNDERSTANDING

3, 4

3-5

3, 5

Example 9

- 3 The approximate population of a small village is recorded from 2000 to 2010.

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

a Plot the time-series graph.

b Describe the general trend in the data over the 11 years.

c For the 11 years, what was the:

- i minimum population?
ii maximum population?



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

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

a Plot the time-series graph. Break the y-axis to exclude values from \$0 to \$1.20.

b Describe the way in which the share price has changed over the 12 months.

c What is the difference between the maximum and minimum share price in the 12 months?

- 5 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

a Plot the time-series graph for the 10 years.

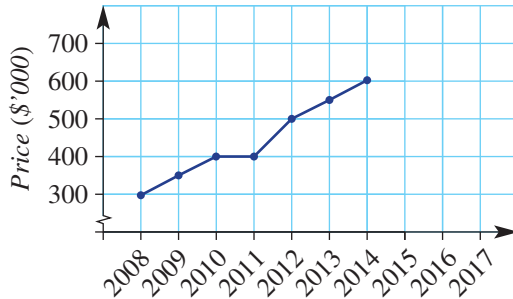
b Describe the way in which the pass rate for the examination has changed in the given time period.

c In what year was the pass rate a maximum?

d By how much had the pass rate improved from 1995 to 1999?

FLUENCY

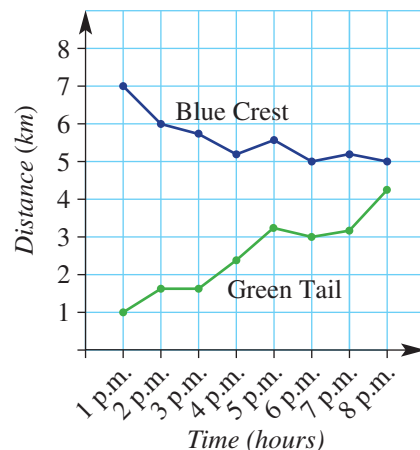
- 6** This time-series plot shows the upwards trend of house prices in an Adelaide suburb over 7 years from 2008 to 2014.



- a** Would you say that the general trend in house prices is linear or non-linear?
- b** Assuming the trend in house prices continues for this suburb, what would you expect the house price to be in:
- i** 2015? **ii** 2017?
- 7** The two top-selling book stores for a company list their sales figures for the first 6 months of the year. Sales amounts are in thousands of dollars.

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

- a** What was the difference in the sales volume for:
- i** August? **ii** 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:
- i** City Central **ii** Southbank
- e** Based on the trend for the sales for the Southbank store, what would you expect the approximate sales volume to be in January?
- 8** Two pigeons (Green Tail and Blue Crest) each have a beacon that communicates with a recording machine. The distance of each pigeon from the machine is recorded every hour for 8 hours.
- a** State the distance from the machine at 3 p.m. for:
- i** Blue Crest **ii** Green Tail
- b** Describe the trend in the distance from the recording machine for:
- i** Blue Crest **ii** Green Tail
- c** Assuming that the given trends continue, predict the time when the pigeons will be the same distance from the recording machine.



9

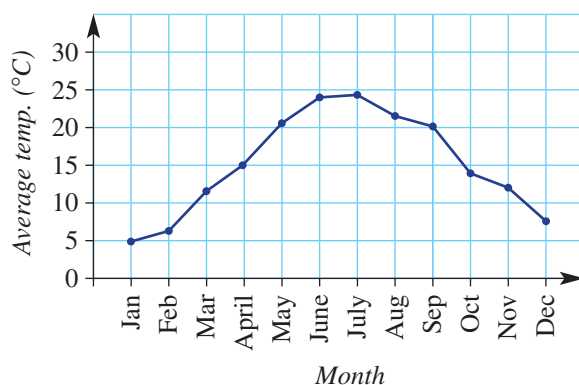
9, 10

10, 11

9F

REASONING

- 9 The average monthly maximum temperature for a city is illustrated in this graph.

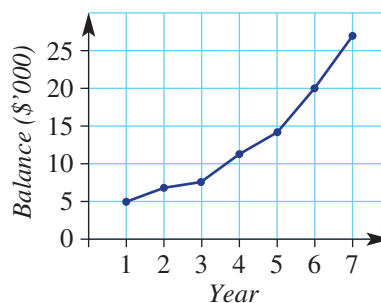


- Explain why the average maximum temperature for December is close to the average maximum temperature for January.
- Do you think this graph is for an Australian city?
- Do you think the data are for a city in the Northern Hemisphere or the Southern Hemisphere? Give a reason.



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

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



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

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

- 12** 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.

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

- Complete the table by calculating the moving average for innings 4–10. Round to the nearest whole number where required.
- Plot the score and moving averages for the batter on the same set of axes.
- Describe the behaviour of the:
 - score graph
 - moving average graph.
- Describe the main difference in the behaviour of the two graphs. Give reasons.



9G Bivariate data and scatter plots



Interactive



Widgets

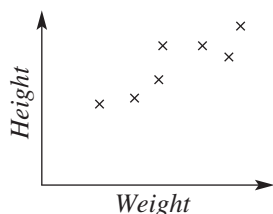


HOTSheets



Walkthroughs

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.



Let's start: 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?
- a Height of person and Weight of person
 - b Temperature and Life of milk
 - c Length of hair and IQ
 - d Depth of topsoil and Brand of motorcycle
 - e Years of education and Income
 - f Spring rainfall and Crop yield
 - g Size of ship and Cargo capacity
 - h Fuel economy and CD track number
 - i Amount of traffic and Travel time
 - j Cost of 2 litres of milk and Ability to swim
 - k Background noise and Amount of work completed

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

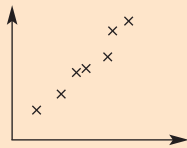
Key
ideas

Key
ideas

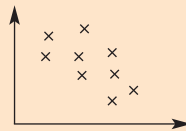
Types of correlation:

Examples

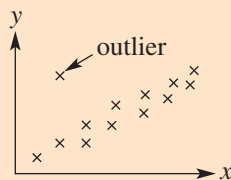
Strong positive correlation



Weak negative correlation



No correlation

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

Example 10 Constructing and interpreting scatter plots

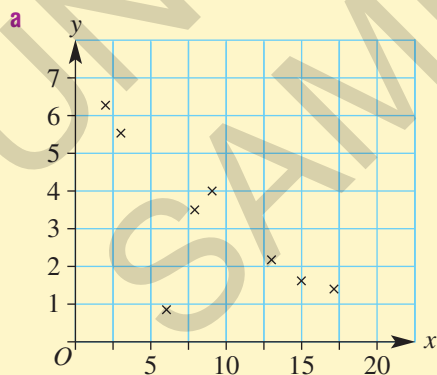
Consider this simple bivariate data set.

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

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

SOLUTION

EXPLANATION

Plot each point using a \times symbol on graph paper.

- negative correlation
- strong correlation
- The outlier is (6, 0.9).

As x increases, y decreases.

The downwards trend in the data is clearly defined.

This point defies the trend.

Exercise 9G

1, 2(a)

2(b)

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 :

- i Draw a scatter plot by hand.
- ii 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

3, 4, 5(a), 6

3, 5(b), 6

3, 5(c), 6

Example 10

3 Consider this simple bivariate data set. (Use technology to assist if desired. See page 648.)

x	1	2	3	4	5	6	7	8
y	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.

4 Consider this simple bivariate data set. (Use technology to assist if desired. See page 648.)

x	14	8	7	10	11	15	6	9	10
y	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.

FLUENCY

- 5 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

x	1.1	1.8	1.2	1.3	1.7	1.9	1.6	1.6	1.4	1.0	1.5
y	22	12	19	15	10	9	14	13	16	23	16

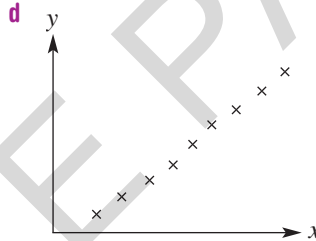
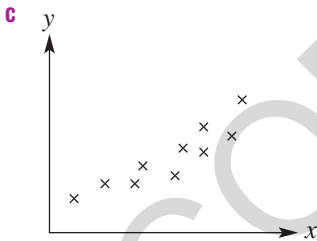
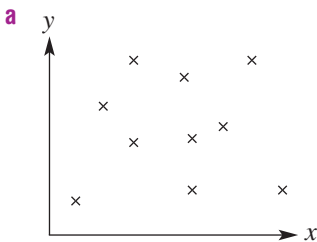
b

x	4	3	1	7	8	10	6	9	5	5
y	115	105	105	135	145	145	125	140	120	130

c

x	28	32	16	19	21	24	27	25	30	18
y	13	25	22	21	16	9	19	25	15	12

- 6 For the following scatter plots, describe the correlation between x and y .



7, 8

8, 9

8, 10

- 7 For common motor vehicles, consider the two variables *Engine size* (cylinder volume) and *Fuel economy* (number of kilometres travelled for every litre of petrol).

- a** Do you expect there to be some relationship between these two variables?
b As the engine size increases, would you expect the fuel economy to increase or decrease?
c 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

- i Do the data generally support your answers to parts **a** and **b**?
- ii Which car gives a fuel economy reading that does not support the general trend?



- 8** 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

- a** 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.
- b** Which garden bed appears to go against the trend?
- c** 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?



- 9** In a newspaper, the number of photos and number of words were counted for 15 different pages. Here are the results.

Page	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of photos	3	2	1	2	6	4	5	7	4	5	2	3	1	0	1
Number of words	852	1432	1897	1621	912	1023	817	436	1132	1201	1936	1628	1403	2174	1829

- a** Sketch a scatter plot using ‘Number of photos’ on the horizontal axis and ‘Number of words’ on the vertical axis.
- b** From your scatter plot, describe the general relationship between the number of photos and the number of words per page. Use the words positive, negative, strong correlation or weak correlation.
- 10** On 14 consecutive days, a local council measures the volume of sound heard from a freeway at various points in a local suburb. The volume of sound, in decibels, is recorded against the distance (in metres) between the freeway and the point in the suburb.

Distance (m)	200	350	500	150	1000	850	200	450	750	250	300	1500	700	1250
Volume (dB)	4.3	3.7	2.9	4.5	2.1	2.3	4.4	3.3	2.8	4.1	3.6	1.7	3.0	2.2

- Draw a scatter plot of *Volume* against *Distance*, plotting *Volume* on the vertical axis and *Distance* on the horizontal axis.
- Describe the correlation between *Distance* and *Volume* as positive, negative or none.
- Generally, as *Distance* increases does *Volume* increase or decrease?

11

11. 12

12, 13

- 11** A government department is interested in convincing the electorate that a larger number of police on patrol leads to a lower crime rate. Two separate surveys are completed over a one-week period and the results are listed in this table.

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

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

-

14

- ## ENRICHMENT

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Hours of TV	11	15	8	9	9	12	20	6	0	15	9	13	15	17	8	11	10	15	21	3
Test score	30	4	13	35	26	31	48	11	50	33	31	28	27	6	39	40	36	21	45	48

- a Which two students performed best on the general knowledge test, having watched TV for the following numbers of hours?
 - i fewer than 10
 - ii more than 4
- b Which two students performed worst on the general knowledge test, having watched TV for the following numbers of hours?
 - i fewer than 10
 - ii more than 4
- c Which four students best support the argument that the more hours of TV watched, the better your general knowledge will be?
- d Which four students best support the argument that the more hours of TV watched, the worse your general knowledge will be?
- e From the given data, would you say that the graduate should conclude that a student's general knowledge is definitely linked to the number of hours of TV watched per week?

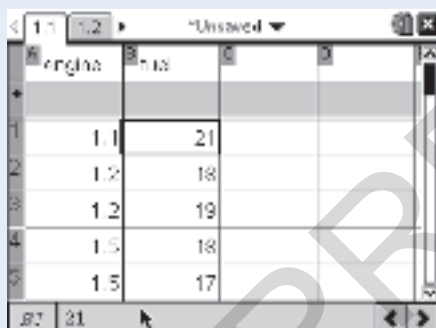
Using calculators to draw scatter plots

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

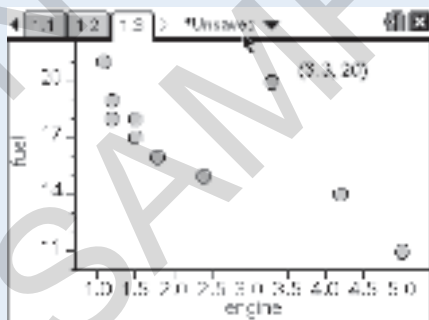
Car	A	B	C	D	E	F	G	H	I	J
Engine size	1.1	1.2	1.2	1.5	1.5	1.8	2.4	3.3	4.2	5.0
Fuel economy	21	18	19	18	17	16	15	20	14	11

Using the TI-Nspire:

- 1 Go to a new **Lists and spreadsheets** page and enter the data into the lists. Title each column.



- 2 Go to a new **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.

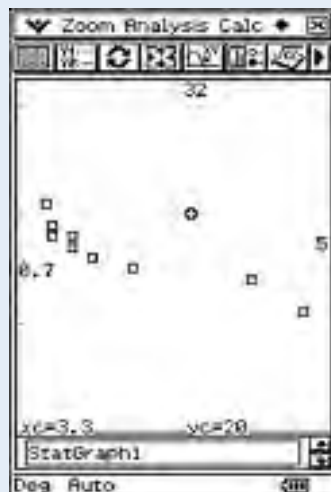


Using the ClassPad:

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



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



9H Line of best fit by eye



Interactive



Widgets



HOTSheets

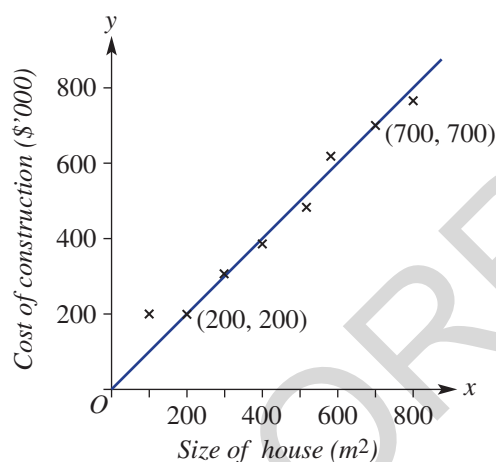


Walkthroughs

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.

Let's start: 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 m² with this building company?

■ 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.

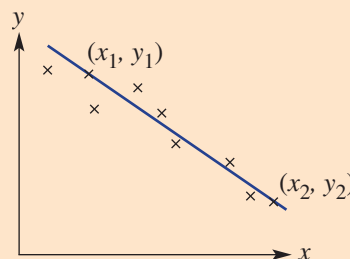
■ For $y = mx + c$:

$$m = \frac{y_2 - y_1}{x_2 - x_1} \text{ and substitute a point to find the value of } c.$$

- 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.



Key
ideas



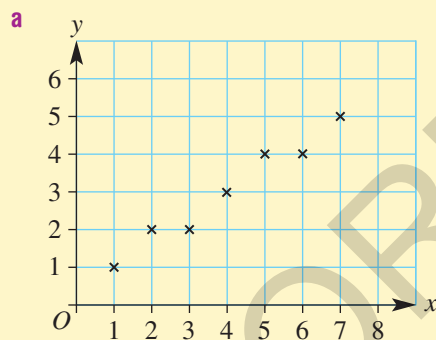
Example 11 Fitting a line of best fit

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

- 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 $y = 5.5$

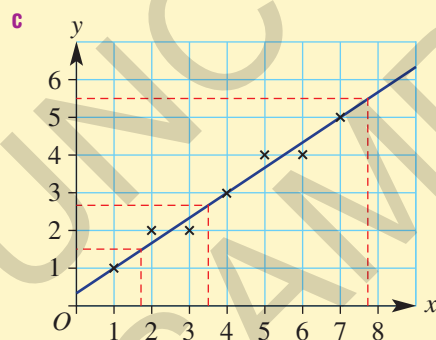
SOLUTION



Plot the points on graph paper.

- b Positive correlation

As x increases, y increases.



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

- d
- i $y \approx 2.7$
 - ii $y \approx 0.4$
 - iii $x \approx 1.7$
 - iv $x \approx 7.8$

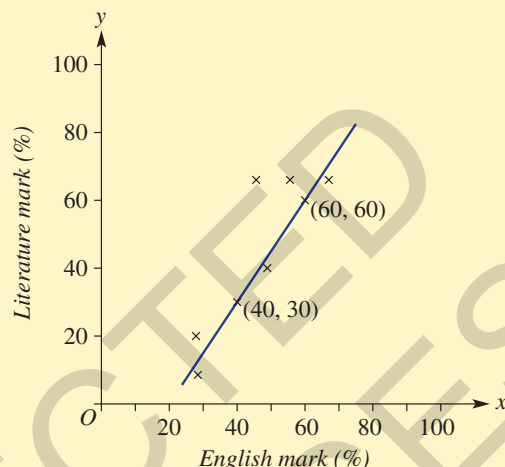
Extend vertical and horizontal lines from the values given and read off your solution. As they are approximations, we use the \approx sign and not the $=$ sign.



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

EXPLANATION

a $y = mx + c$

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

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

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

$$30 = 60 + c$$

$$c = -30$$

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

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

\therefore Literature score is 45.

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

\therefore Literature score is 99.

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

$$72 = \frac{3}{2}x, \text{ so } x = 48.$$

\therefore English score is 48.

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

$$117 = \frac{3}{2}x, \text{ so } x = 78.$$

\therefore English score is 78.

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 .

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

Repeat for $x = 86$.

Substitute $y = 42$ and solve for x .

Repeat for $y = 87$.

Exercise 9H

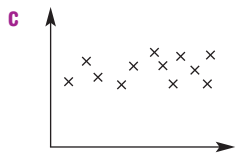
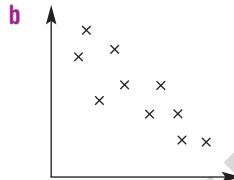
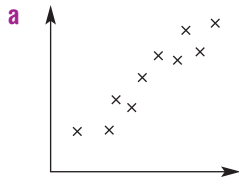
1–3

3

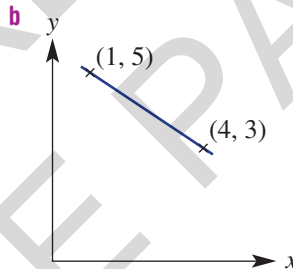
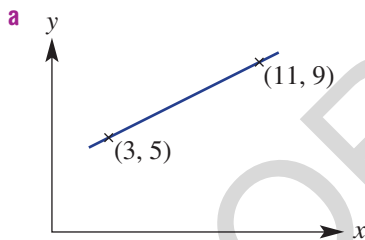
—

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. (Using a pencil might help.)



- 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 $x = 16$

b x when:

i $y = 4$

ii $x = 7$

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

4–6

4–6

4, 6

Example 11

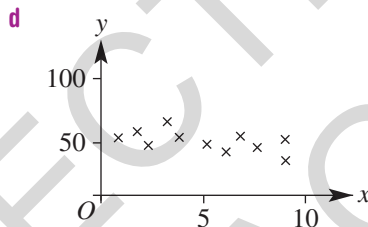
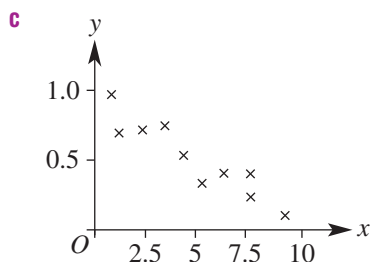
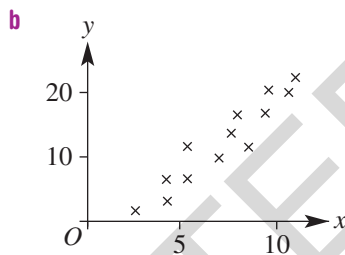
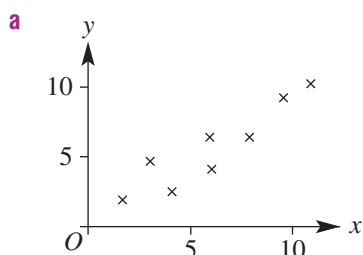
- 4 Consider the variables x and y and the corresponding bivariate data.

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

- a** Draw a scatter plot for the data.
b Is there positive, negative or no correlation between x and y ?
c Fit a line of best fit by eye to the data on the scatter plot.
d Use your line of best fit to estimate:
- i** y when $x = 3.5$ **ii** y when $x = 0$
iii x when $y = 2$ **iv** x when $y = 5.5$

FLUENCY

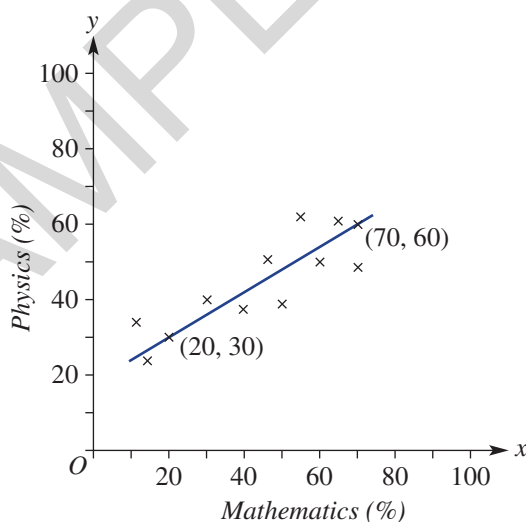
- 5** For the following scatter plots, pencil in a line of best fit by eye, and then use your line to estimate the value of y when $x = 5$.



Example 12

- 6** This scatter plot shows a linear relationship between Mathematics marks and Physics marks in a small class of students. A trend line passes through $(20, 30)$ and $(70, 60)$.

- a** Find the equation of the trend line.
b Use your equation to find the Physics score if the Mathematics score is:
i 40 **ii** 90
c Use your equation to find the Mathematics score if the Physics score is:
i 36 **ii** 78



9H

7

7, 8

7, 8

- 7 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

- Draw a scatter plot for the data, showing growth on the vertical axis.
 - Fit a line of best fit by eye.
 - Use your line of best fit to estimate the growth expected for the following rainfall readings. You do not need to find the equation of the line.
 - 500 mm
 - 900 mm
 - Use your line of best fit to estimate the rainfall for a given year if the growth of the bamboo was:
 - 30 cm
 - 60 cm
- 8 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.
- Find the equation of the trend line.
 - Use your equation to estimate the length of an 18 kg dog.
 - Use your equation to estimate the weight of a dog that has a length of 100 cm.

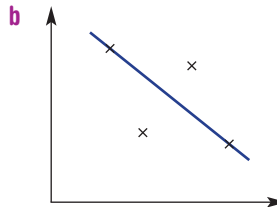
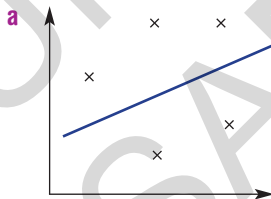
PROBLEM-SOLVING

9

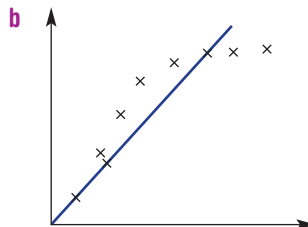
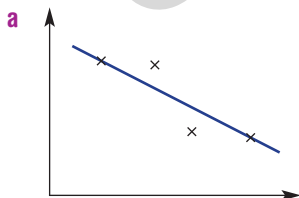
9, 10

10, 11

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



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



REASONING

- 11 A trend line relating the percentage scores for Music performance (y) and Music theory (x) is given by

$$y = \frac{4}{5}x + 10.$$

- a Find the value of x when:
- $y = 50$
 - $y = 98$
- b What problem occurs in predicting Music theory scores when using high Music performance scores?



Heart rate and age

12

- 12 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	

- a Sketch separate scatter plots for experiment 1 and experiment 2.
- b 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
- c Estimate the age of a person who has a maximum heart rate of 190, using the results from:
- experiment 1
 - experiment 2
- d For a person aged 25 years, which experiment estimates a lower maximum heart rate?
- e Research the average maximum heart rate of people according to age and compare with the results given above.



9I Linear regression with technology

10A



Interactive



Widgets



HOTSheets



Walkthroughs

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.

Two common methods are least squares regression and median–median regression. These methods are best handled with the use of technology.

Let's start: 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 667.

- Can you find the least squares regression and median–median regression tools?
- Use your technology to try Example 13 below.

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.
- The **median–median** regression line uses the three medians from the lower, middle and upper groups within the data set.
 - Outliers do not have an effect on the median–median line as they do not significantly alter the median values.



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.

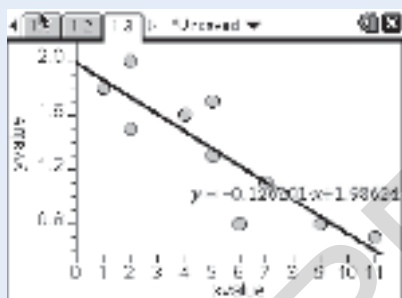
x	1	2	2	4	5	5	6	7	9	11
y	1.8	2	1.5	1.6	1.7	1.3	0.8	1.1	0.8	0.7

- Construct a scatter plot for the data.
- Find the equation of the least squares regression line.
- Find the equation of the median–median regression line.
- Sketch the graph of the regression line onto the scatter plot.
- Use the least squares regression line to estimate the value of y when x is:
 - 4.5
 - 15
- Use the median–median regression line to estimate the value of y when x is:
 - 4.5
 - 15

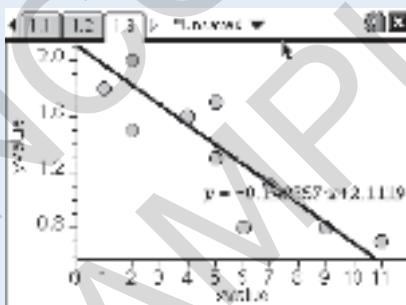
Using calculators to find equations of regression

Using the TI-Nspire:

a, b, d Go to a new **Lists and spreadsheets** page and enter the data into the lists. Title each column **xvalue** and **yvalue**. Go to a new **Data and Statistics** page and select **xvalue** variable for the horizontal axis and **yvalue** for the vertical axis. To produce the least squares regression line, go to menu, **Analyze, Regression, Show Linear (mx+b)**.



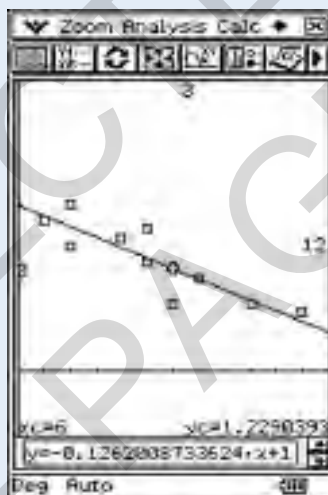
c, d To produce the median–median regression line, go to menu, **Analyze, Regression, Show Median-Median**.



- e i** $y \approx 1.39$
ii $y \approx 0.06$
f i $y \approx 1.47$
ii $y \approx 0.03$

Using the ClassPad:

a, b, d In the **Statistics** application enter the data into the lists. Tap **Calc, 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.



c, d To produce the median–median regression line, tap **Calc, MedMed Line** and apply the settings stated previously.



Exercise 9I

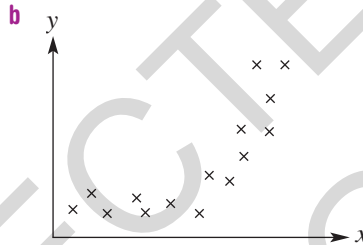
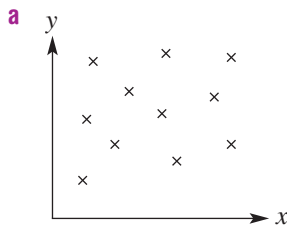
1, 2

2

—

UNDERSTANDING

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



3(A, B), 4

3(A, C), 4

3(B), 4

FLUENCY

Example 13



- 3 Consider the data in tables A–C and use a graphics or CAS calculator or software to help answer the following questions.

A

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

B

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

C

x	0.1	0.2	0.5	0.8	0.9	1.2	1.6	1.7
y	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.
- Find the equation of the median–median regression line.
- Sketch the graph of the regression line onto the scatter plot.
- Use the least squares regression line to estimate the value of y when x is:
 - 7
 - 12
- Use the median–median regression line to estimate the value of y when x is:
 - 7
 - 12

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

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

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

- a Using Age for the x -axis, find:
- i the least squares regression line ii the median–median regression line.
- b Use your least squares regression line to estimate the value of a 3-year-old car.
- c Use your median–median regression line to estimate the value of a 12-year-old car.
- d Use your least squares regression line to estimate the age of a \$15 000 car.
- e Use your median–median regression line to estimate the age of an \$8000 car.

5, 6

5, 6

6, 7

- 5 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.

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

Use a graphics or CAS calculator to complete the following.

- a Draw a scatter plot for the data.
- b Find the equation of the least squares regression line.
- c Graph the line onto the scatter plot.
- d Use the regression line to estimate how many jackets would be able to be produced if the maximum daily temperature in the factory was:
- i 30°C ii 35°C iii 45°C

- 6 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 200 000 copies. A study of one particular copier gave the following results.

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

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



- 7 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.

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

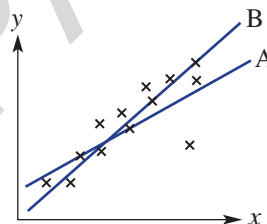
- Draw a scatter plot for the data.
- Find the equation of the median–median regression line.
- Use your regression equation to estimate the distance record for the hammer throw for:
 - 2000
 - 2020
- Would you say that it is realistic to use your regression equation to estimate distance records beyond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

8

8, 9

8, 9

- Briefly explain why the least squares regression line is affected by outliers.
 - Briefly explain why the median–median regression line is not strongly affected by outliers.
- This scatter plot shows both the least squares regression line and the median–median regression line. Which line (i.e. A or B) do you think is the least squares line? Give a reason.



Correlation coefficient

—

—

10

- 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.



Investigation

Indigenous population comparison

The following data was collected by the Australian Bureau of Statistics during the 2011 National Census. It shows the population of Indigenous and non-Indigenous people in Australia and uses class intervals of 5 years.

Cat. No. 2068.0 – 2006 Census Tables
2006 Census of Population and Housing
Australia (Australia)
INDIGENOUS STATUS BY AGE
Count of persons
Based on place of usual residence
Commonwealth of Australia 2007



Age	Total	Indigenous	Non-Indigenous	Indigenous status not stated
Total all ages	21 507 719	548 371	19 900 765	1 058 583
0–4 years	1 421 048	67 416	1 282 738	70 894
5–9 years	1 351 921	64 935	1 222 111	64 875
10–14 years	1 371 055	64 734	1 241 794	64 527
15–19 years	1 405 798	59 200	1 282 018	64 580
20–24 years	1 460 675	46 455	1 333 622	80 598
25–29 years	1 513 237	38 803	1 387 922	86 512
30–34 years	1 453 777	33 003	1 345 763	75 011
35–39 years	1 520 138	34 072	1 414 171	71 895
40–44 years	1 542 879	33 606	1 438 346	70 927
45–49 years	1 504 142	28 818	1 407 494	67 830
50–54 years	1 447 403	24 326	1 357 677	65 400
55–59 years	1 297 247	18 640	1 220 529	58 078
60–64 years	1 206 117	13 592	1 138 395	54 130
65 years and over	3 012 282	20 771	2 828 185	163 326

Indigenous histogram

- Use the given data to construct a histogram for the population of Indigenous people in Australia in 2011.
- Which age group contained the most Indigenous people?
- Describe the shape of the histogram. Is it symmetrical or skewed?

Non-Indigenous histogram

- Use the given data to construct a histogram for the population of non-Indigenous people in Australia in 2011.
Try to construct this histogram so it is roughly the same width and height as the histogram for the Indigenous population. You will need to rescale the y-axis.
- Which age group contains the most number of non-Indigenous people?

- Describe the shape of the histogram. Is it symmetrical or skewed?

Comparisons

- a Explain the main differences in the shapes of the two histograms.
- b What do the histograms tell you about the age of Indigenous and non-Indigenous people in Australia in 2011?
- c What do the graphs tell you about the difference in life expectancy for Indigenous and non-Indigenous people?

Antarctic ice

According to many different studies, the Antarctic ice mass is decreasing over time. The following data show the approximate change in ice mass in gigatonnes (Gt; 10^9 tonnes) from 2002 to 2009.

Year	2002	2003	2004	2005	2006	2007	2008	2009
Change (Gt)	450	300	400	200	-100	-400	-200	-700

A change of 300, for example, means that the ice mass has increased by 300 Gt in that year.

Data interpretation

- a By how much did the Antarctic ice mass increase in:
 - i 2002?
 - ii 2005?
- b By how much did the Antarctic ice mass decrease in:
 - i 2006?
 - ii 2009?
- c What was the overall change in ice mass from the beginning of 2005 to the end of 2007?



Time-series plot

- a Construct a time-series plot for the given data.
- b Describe the general trend in the change in ice mass over the 8 years.

Line of best fit

- a Fit a line of best fit by eye to your time-series plot.
- b Find an equation for your line of best fit.
- c Use your equation to estimate the change in ice mass for:
 - i 2010
 - ii 2020



Regression

- a Use technology to find either the least squares regression line or the median–median regression line for your time-series plot.
- b How does the equation of these lines compare to your line of best fit found above?

Problems and challenges



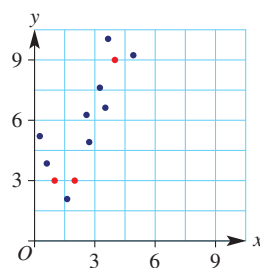
Check out the 'Working with unfamiliar problems' poster on the inside cover of your book to help you answer these questions.



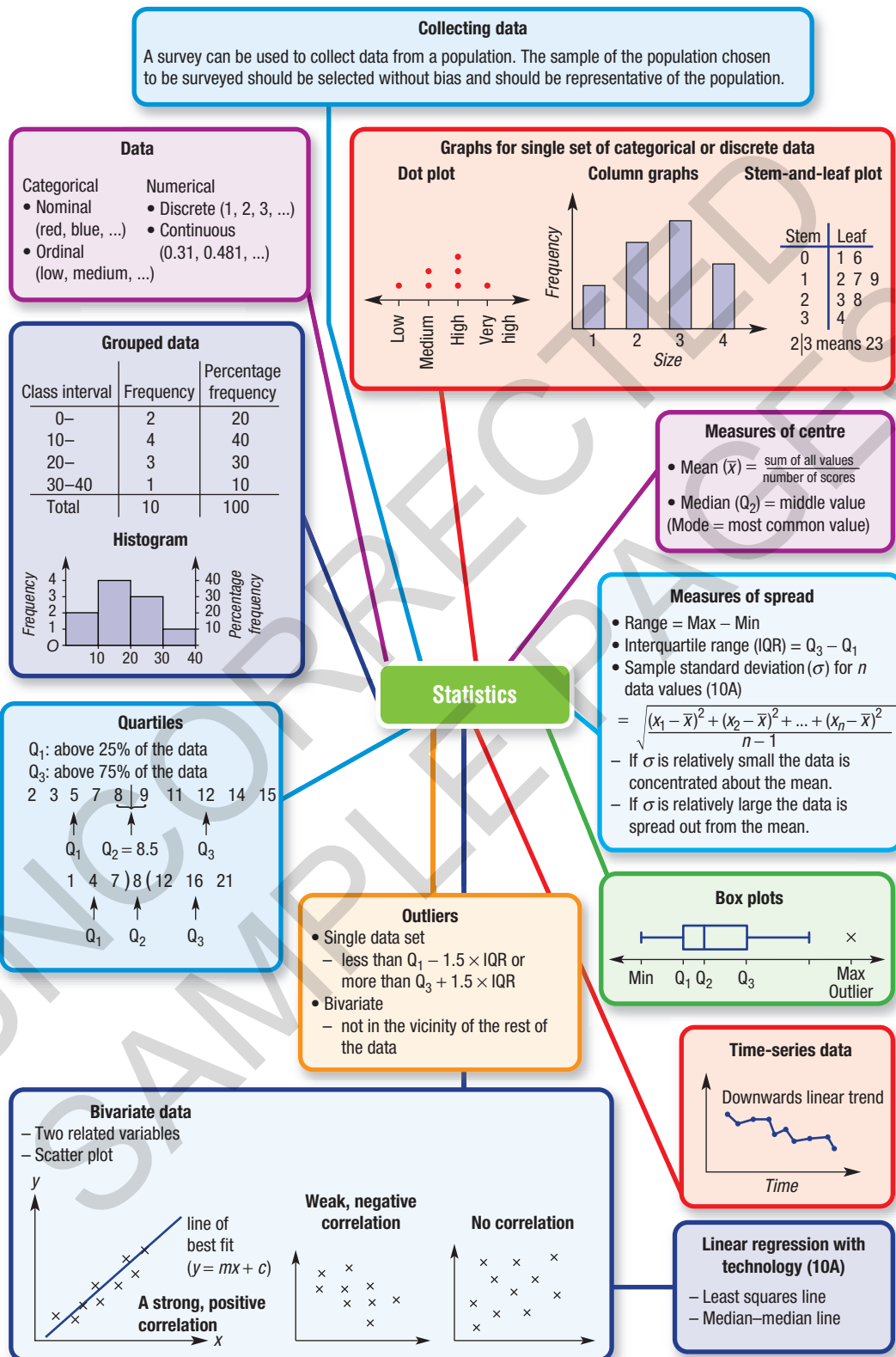
- 1 The mean mass of six boys is 71 kg, and the mean mass of five girls is 60 kg. Find the average mass of all 11 people put together.



- 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.



- 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, σ , for some Year 10 term tests are:
 Maths ($\bar{x} = 70\%$, $\sigma = 9\%$); Physics ($\bar{x} = 70\%$, $\sigma = 6\%$); Biology ($\bar{x} = 80\%$, $\sigma = 6.5\%$).
 If Emily gained 80% in each of these subjects, which was her best and worst result? Give reasons for your answer.



Multiple-choice questions

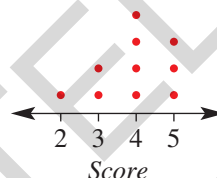
9A 1 The type of data generated by the survey question *What is your favourite food?* is:

- A numerical and discrete
B numerical and continuous
C a sample
D categorical and nominal
E categorical and ordinal

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



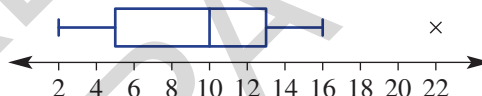
9B 3 The mode for the data is:

- 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:

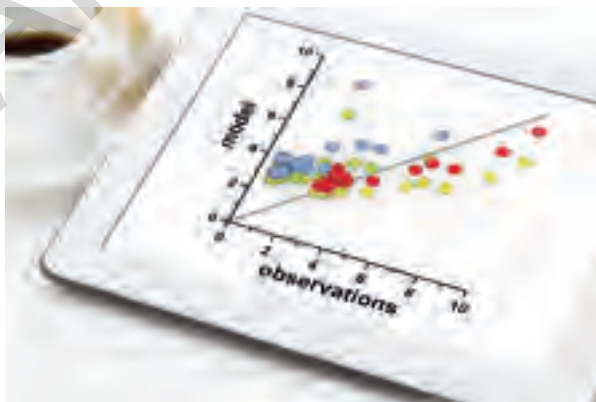
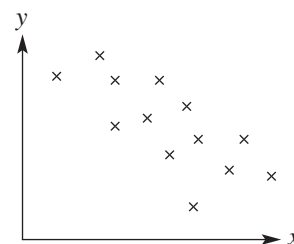
- A 8
B 5
C 3
D 20
E 14

9D 6 The range is:

- A 5
B 3
C 20
D 14
E 8

9G 7 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** 8 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

- 9E** 9 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



- 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** $y = \frac{3}{5}x - \frac{2}{3}$

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

- a** Organise the data into a table with class intervals of 5 and include a percentage frequency column.
b Construct a histogram for the data, showing both the frequency and percentage frequency on the graph.
c Would you describe the data as symmetrical, positively skewed or negatively skewed?
d Construct a stem-and-leaf plot for the data, using 10s as the stem.
e Use your stem-and-leaf plot to find the median.

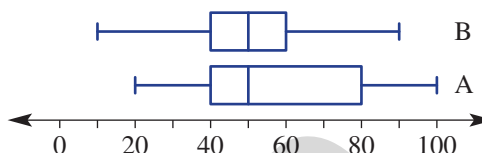


- 9D** 2 For each set of data below, complete the following tasks.

- i** Find the range.
ii Find the lower quartile (Q_1) and the upper quartile (Q_3).
iii Find the interquartile range.
iv Locate any outliers.
v Draw a box plot.
a 2, 2, 3, 3, 3, 4, 5, 6, 12
b 11, 12, 15, 15, 17, 18, 20, 21, 24, 27, 28
c 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.

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



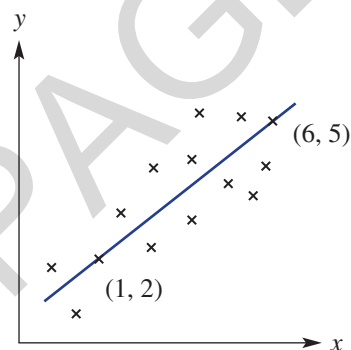
9G **4** Consider the simple bivariate data set.

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

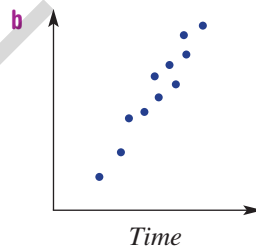
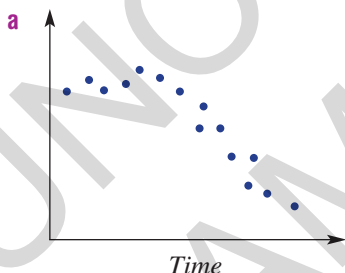
- a** Draw a scatter plot for the data.
- b** Describe the correlation between x and y as positive or negative.
- c** Describe the correlation between x and y as strong or weak.
- d** Identify any outliers.

9H **5** The line of best fit passes through the two points labelled on this graph.

- a** Find the equation of the line of best fit.
- b** Use your equation to estimate the value of y when:
 - i** $x = 4$
 - ii** $x = 10$
- c** Use your equation to estimate the value of x when:
 - i** $y = 3$
 - ii** $y = 12$



9F **6** Describe the trend in these time-series plots as linear, non-linear or no trend.



9E **7** Calculate the mean and sample standard deviation for these small data sets. Round the standard deviation to one decimal place.

a 4, 5, 7, 9, 10

b 1, 1, 3, 5, 5, 9

10A



- 9E** **8** The Cats and The Vipers basketball teams compare their number of points per match for a season. The data are presented in this back-to-back stem-and-leaf plot.

The Cats Leaf	Stem	The Vipers Leaf
	0	9
2	1	9
8 3	2	0 4 8 9
7 4	3	2 4 7 8 9
9 7 4 1 0	4	2 8
7 6 2	5	0

2 | 4 means 24



State which team has:

- a** the higher range
- b** the higher mean
- c** the higher median
- 10A** **d** the higher standard deviation.

- 9I** **9** For the simple bivariate data set in Question 4, use technology to find the equation of the:

- 10A** **a** least squares regression line
- b** median–median regression line.


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.

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

- a** Find the IQR for:
 - i** tree 1
 - ii** tree 2
- b** Identify any outliers for:
 - i** tree 1
 - ii** 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.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Max. daily temp. (T) ($^{\circ}\text{C}$)	27	26	28	33	38	36	28	30	32	25	25	27	29	33
No. of shoppers (N)	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 on the vertical axis, and describe the correlation between the variables as either positive, negative or none.
- b** Find the equation of the median–median regression line for the data, using technology.
- c** Use your regression equation to estimate:
- i** the number of shoppers on a day with a maximum daily temperature of 24°C
 - ii** the maximum daily temperature if the number of shoppers is 1500.
- d** Use technology to determine the least squares regression line for the data.
- e** Use your least squares regression equation to estimate:
- i** the number of shoppers on a day with a maximum daily temperature of 24°C
 - ii** the maximum daily temperature if the number of shoppers at the plaza is 1500.

