TOPIC 12

Univariate data

12.1 Overview

Why learn this?

According to the novelist Mark Twain, 'There are three kinds of lies: lies, damned lies and statistics.' There is so much information in our lives, increasingly so with the World Wide Web, smart phones and social media tracking our every move and accumulating vast amounts of data about us. The data are used to gather information about our likes and dislikes, our buying habits, our voting preferences and so on. Statistics can easily be used to manipulate people unless they have an understanding of the basic concepts involved.

What do you know?

assessON

1 **THINK** List what you know about data.
Use a thinking tool such as a concept map to show your list.
2 **PAIR** Share what you know with a partner and then with a small group.
3 **SHARE** As a class, create a thinking tool such as a large concept map that shows your class's knowledge of data.

Learning sequence

WATCH THIS VIDEO
The story of mathematics: Koby's bid to make the Olympic athletics team

SEARCHLIGHT ID: eles-1852

12.2 Measures of central tendency

Univariate data

- In this chapter you will learn how to measure and analyse **univariate data**. Univariate data are data with one variable; for example, the heights of Year 10 students.
- **Measures of central tendency** are summary statistics that measure the middle (or centre) of the data. These are known as the mean, median and mode.
  - The **mean** is the average of all observations in a set of data.
  - The **median** is the middle observation in an ordered set of data.
  - The **mode** is the most frequent observation in a data set.

Ungrouped data

Mean, median and mode of ungrouped data

Mean

- To obtain the mean of a set of ungrouped data, all numbers (scores) in the set are added together and then the total is divided by the number of scores in that set.

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

- Symbolically this is written $\bar{x} = \frac{\Sigma x}{n}$.

Median

- The median is the middle value of any set of data arranged in numerical order. In the set of n numbers, the median is located at the $\frac{n+1}{2}$th score. The median is:
  - the middle score for an odd number of scores arranged in numerical order
  - the average of the two middle scores for an even number of scores arranged in numerical order.

Mode

- The mode is the score that occurs most often in a set of data.
- A set of data may contain:
  1. no mode; that is, each score occurs once only
  2. one mode
  3. more than one mode.

WORKED EXAMPLE 1 TI CASIO

For the data 6, 2, 4, 3, 4, 5, 4, 5, find the:
a mean **b median** **c mode.**

THINK

a 1 Calculate the sum of the scores; that is, Σx.

2 Count the number of scores; that is, n.

3 Write the rule for the mean.

4 Substitute the known values into the rule.

WRITE

a $\Sigma x = 6 + 2 + 4 + 3 + 4 + 5 + 4 + 5$
$= 33$
$n = 8$
$\bar{x} = \frac{\Sigma x}{n}$
$= \frac{33}{8}$

5 Evaluate.
$= 4.125$

6 Answer the question.
The mean is 4.125.

b 1 Write the median scores in ascending numerical order.

b 2 3 4 4 4 5 5 6

2 Locate the position of the median using the rule $\frac{n+1}{2}$, where $n = 8$. This places the median as the 4.5th score; that is, between the 4th and 5th score.

$$\text{Median} = \frac{n+1}{2}\text{th score}$$
$$= \frac{8+1}{2}\text{th score}$$
$$= 4.5\text{th score}$$

2 3 4 **4** **4** 5 5 6

3 Obtain the average of the two middle scores.

$$\text{Median} = \frac{4+4}{2}$$
$$= \frac{8}{2}$$
$$= 4$$

4 Answer the question.
The median is 4.

c 1 Systematically work through the set and make note of any repeated values (scores).

c 2 3 4 4 4 5 5 6
↑ ↑ ↑

2 Answer the question.
The mode is 4.

Calculating mean, median and mode from a frequency distribution table

- If data are presented in a frequency distribution table, the formula used to calculate the mean is $\bar{x} = \frac{\Sigma(f \times x)}{n}$.
- Here, each value (score) in the table is multiplied by its corresponding frequency; then all the $f \times x$ products are added together and the total sum is divided by the number of observations in the set.
- To find the median, find the position of each score from the cumulative frequency column.
- The mode is the score with the highest **frequency**.

WORKED EXAMPLE 2 TI CASIO

For the table at right, find the:
a mean
b median
c mode.

Score (x)	Frequency (f)
4	1
5	2
6	5
7	4
8	3
Total	15

THINK

1. Rule up a table with four columns titled Score (x), Frequency (f), Frequency × score ($f \times x$) and Cumulative frequency (cf).
2. Enter the data and complete both the $f \times x$ and cumulative frequency columns.

WRITE

Score (x)	Frequency (f)	Frequency × score ($f \times x$)	Cumulative frequency (cf)
4	1	4	1
5	2	10	1 + 2 = 3
6	5	30	3 + 5 = 8
7	4	28	8 + 4 = 12
8	3	24	12 + 3 = 15
	$n = 15$	$\Sigma(f \times x) = 96$	

a 1. Write the rule for the mean.

2. Substitute the known values into the rule and evaluate.

3. Answer the question.

a $\bar{x} = \dfrac{\Sigma(f \times x)}{n}$

$\bar{x} = \dfrac{96}{15}$

$= 6.4$

The mean of the data set is 6.4.

b 1. Locate the position of the median using the rule $\dfrac{n+1}{2}$, where $n = 15$.

This places the median as the 8th score.

2. Use the cumulative frequency column to find the 8th score and answer the question.

b The median is the $\dfrac{15+1}{2}$th or 8th score.

The median of the data set is 6.

c 1. The mode is the score with the highest frequency.

2. Answer the question.

c The score with the highest frequency is 6.

The mode of the data set is 6.

Mean, median and mode of grouped data

Mean

- When the data are grouped into class intervals, the actual values (or data) are lost. In such cases we have to approximate the real values with the midpoints of the intervals into which these values fall. For example, when measuring heights of students in a class, if we found that 4 students had a height between 180 and 185 cm, we have to assume that each of those 4 students is 182.5 cm tall. The formula for calculating the mean:

$$\bar{x} = \frac{\sum(f \times x)}{n}$$

Here x represents the midpoint (or class centre) of each class interval, f is the corresponding frequency and n is the total number of observations in a set.

Median

- The median is found by drawing a **cumulative frequency** curve (ogive) of the data and estimating the median from the 50th percentile.

Modal class

- The modal class is the class interval that has the highest frequency.

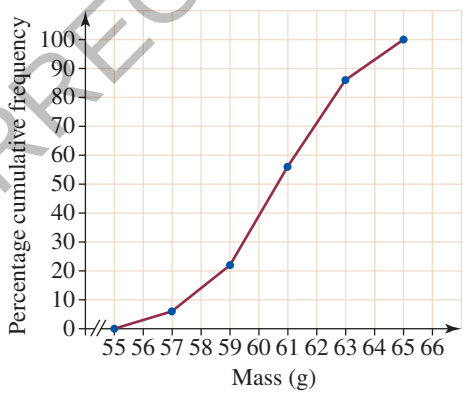
Cumulative frequency curves (ogives)

Ogives

- Data from a cumulative frequency table can be plotted to form a **cumulative frequency curve** (sometimes referred to as cumulative frequency polgons), which is also called an **ogive** (pronounced '*oh-jive*').
- To plot an ogive for data that is in class intervals, the maximum value for the class interval is used as the value against which the cumulative frequency is plotted.

For example, the following table and graph show the mass of cartons of eggs ranging from 55 g to 65 g.

Mass (g)	Frequency (f)	Cumulative frequency (cf)	Percentage cumulative frequency ($\%cf$)
55–<57	2	2	6%
57–<59	6	2 + 6 = 8	22%
59–<61	12	8 + 12 = 20	56%
61–<63	11	20 + 11 = 31	86%
63–<65	5	31 + 5 = 36	100%



Quantiles

- An ogive can be used to divide the data into any given number of equal parts called **quantiles**.
- Quantiles are named after the number of parts that the data are divided into.
  - **Percentiles** divide the data into 100 equal-sized parts.
  - **Quartiles** divide the data into 4 equal-sized parts. For example, 25% of the data values lie at or below the first quartile.

Percentile	Quartile and symbol	Common name
25th percentile	First quartile, Q_1	Lower quartile
50th percentile	Second quartile, Q_2	Median
75th percentile	Third quartile, Q_3	Upper quartile
100th percentile	Fourth quartile, Q_4	Maximum

- A percentile is named after the percentage of data that lies at or below that value. For example, 60% of the data values lie at or below the 60th percentile.
- Percentiles can be read off a percentage cumulative frequency curve.
- A percentage cumulative frequency curve is created by:
  - writing the cumulative frequencies as a percentage of the total number of data values
  - plotting the percentage cumulative frequencies against the maximum value for each interval.

WORKED EXAMPLE 3

For the given data:
a estimate the mean
b estimate the median
c find the modal class.

Class interval	Frequency
60–<70	5
70–<80	7
80–<90	10
90–<100	12
100–<110	8
110–<120	3
Total	45

THINK

1. Draw up a table with 5 columns headed Class interval, Class centre (x), Frequency (f), Frequency × class centre ($f \times x$) and Cumulative frequency (cf).
2. Complete the x, $f \times x$ and cf columns.

WRITE

Class interval	Class centre (x)	Freq. (f)	Frequency × class centre ($f \times x$)	Cumulative frequency (cf)
60–<70	65	5	325	5
70–<80	75	7	525	12
80–<90	85	10	850	22
90–<100	95	12	1140	34
100–<110	105	8	840	42
110–<120	115	3	345	45
		$n = 45$	$\Sigma(f \times x) = 4025$	

a **1** Write the rule for the mean.

a $\bar{x} = \frac{\Sigma(f \times x)}{n}$

2 Substitute the known values into the rule and evaluate.

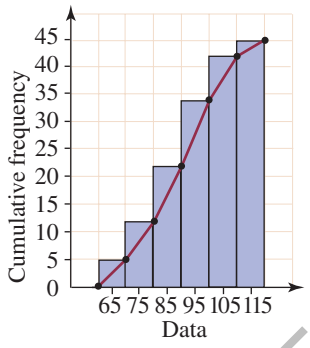
$\bar{x} = \frac{4025}{45}$
$\simeq 89.4$

3 Answer the question.

The mean for the given data is approximately 89.4.

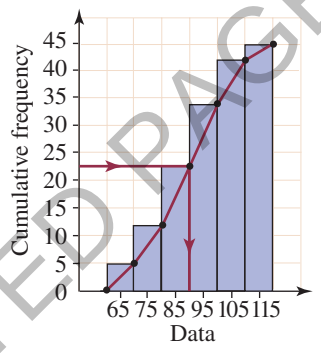
b **1** Draw a combined cumulative frequency histogram and ogive, labelling class centres on the horizontal axis and cumulative frequency on the vertical axis. Join the end-points of each class interval with a straight line to form the ogive.

b



2 Locate the middle of the cumulative frequency axis, which is 22.5.

3 Draw a horizontal line from this point to the ogive and a vertical line to the horizontal axis.



4 Read off the value of the median from the x-axis and answer the question.

The median for the given data is approximately 90.

c **1** The modal class is the class interval with the highest frequency.

c The class internal 90–100 occurs twelve times, which is the highest frequency.

2 Answer the question.

The modal class is the 90–100 class interval.

Exercise 12.2 Measures of central tendency

assessON

INDIVIDUAL PATHWAYS

PRACTISE	CONSOLIDATE	MASTER
Questions: 1–10, 12, 15, 16, 18	Questions: 1–10, 12, 14, 16, 18	Questions: 1–19

Individual pathway interactivity int-4621 eBookplus

REFLECTION
Under what circumstances might the median be a more reliable measure of centre than the mean?

eBookplus

Digital docs
SkillSHEET
Finding the mean of a small data set
doc-5299
SkillSHEET
Finding the median of a small data set
doc-5300
SkillSHEET
Finding the mode of a small data set
doc-5301
SkillSHEET
Finding the mean, median and mode from a stem-and-leaf plot
doc-5302
SkillSHEET
Presenting data in a frequency distribution table
doc-5303
SkillSHEET
Drawing statistical graphs
doc-5304

FLUENCY

1 **WE1** For each of the following sets of data find the:

i mean **ii** median **iii** mode.

a 3, 5, 6, 8, 8, 9, 10

b 4, 6, 7, 4, 8, 9, 7, 10

c 17, 15, 48, 23, 41, 56, 61, 52

d 4.5, 4.7, 4.8, 4.8, 4.9, 5.0, 5.3

e $7\frac{1}{2}$, $10\frac{1}{4}$, 12, $12\frac{1}{4}$, 13, $13\frac{1}{2}$, $13\frac{1}{2}$, 14

2 The back-to-back stem-and-leaf plot below shows the test results of 25 Year 10 students in Mathematics and Science. Find the mean, median and mode for each of the two subjects.

Key: 3 | 2 = 32

Leaf: Science	*Stem*	*Leaf: Mathematics*
8 7 3	3	2 9
9 6 2 2 1	4	0 6 8
8 7 6 1 1 0	5	1 3 5
9 7 4 3 2	6	2 6 7 9
8 5 1 0	7	3 6 7 8
7 3	8	0 4 4 6 8 9
	9	2 5 8

3 **WE2** For the data shown in each of the following frequency distribution tables, find the:

i mean **ii** median **iii** mode.

a

Score (x)	Frequency (f)
4	3
5	6
6	9
7	4
8	2
Total	24

b

Score (x)	Frequency (f)
12	4
13	5
14	10
15	12
16	9
Total	40

4 The following data show the number of bedrooms in each of the 10 houses in a particular neighbourhood: 2, 1, 3, 4, 2, 3, 2, 2, 3, 3.

a Calculate the mean and median number of bedrooms.

b A local motel contains 20 rooms. Add this observation to the set of data and recalculate the values of the mean and median.

c Compare the answers obtained in parts **a** and **b** and complete the following statement: When the set of data contains an unusually large value(s), called an outlier, the _____ (mean/median) is the better measure of central tendency, as it is less affected by this extreme value.

5 WE3 For the given data:

a estimate the mean **b** estimate the median **c** find the modal class.

Class interval	Frequency
40–<50	2
50–<60	4
60–<70	6
70–<80	9
80–<90	5
90–<100	4
Total	30

6 Calculate the mean of the grouped data shown in the table below.

Class interval	Frequency
100–109	3
110–119	7
120–129	10
130–139	6
140–149	4
Total	30

7 Find the modal class of the data shown in the table below.

Class interval	Frequency
50–<55	1
55–<60	3
60–<65	4
65–<70	5
70–<75	3
75–<80	2
Total	18

8 MC The number of textbooks sold by various bookshops during the second week of December was recorded. The results are summarised in the table below.

Number of books sold	Frequency
220–229	2
230–239	2
240–249	3
250–259	5
260–269	4
270–279	4
Total	20

a The modal class of the data is given by the class interval(s):
A 220–229 and 230–239 **B** 250–259
C 260–269 and 270–279 **D** of both A and C

b The class centre of the first class interval is:
A 224 **B** 224.5 **C** 224.75 **D** 225

c The median of the data is in the interval:
A 230–239 **B** 240–249 **C** 250–259 **D** 260–269

d The estimated mean of the data is:
A 251 **B** 252 **C** 253 **D** 254

UNDERSTANDING

9 A random sample was taken, composed of 30 people shopping at a Coles supermarket on a Tuesday night. The amount of money (to the nearest dollar) spent by each person was recorded as follows:

6, 32, 66, 17, 45, 1, 19, 52, 36, 23, 28, 20, 7, 47, 39
6, 68, 28, 54, 9, 10, 58, 40, 12, 25, 49, 74, 63, 41, 13

a Find the mean and median amount of money spent at the checkout by the people in this sample.
b Group the data into class intervals of 10 and complete the frequency distribution table. Use this table to estimate the mean amount of money spent.
c Add the cumulative frequency column to your table and fill it in. Hence, construct the ogive. Use the ogive to estimate the median.
d Compare the mean and the median of the original data from part **a** with the mean and the median obtained for grouped data in parts **b** and **c**. Were the estimates obtained in parts **b** and **c** good enough? Explain your answer.

10 **a** Add one more number to the set of data 3, 4, 4, 6 so that the mean of a new set is equal to its median.
b Design a set of five numbers so that mean = median = mode = 5.
c In the set of numbers 2, 5, 8, 10, 15, change one number so that the median remains unchanged while the mean increases by 1.

11 Thirty men were asked to reveal the number of hours they spent doing housework each week. The results are detailed below.

1	5	2	12	2	6	2	8	14	18
0	1	1	8	20	25	3	0	1	2
7	10	12	1	5	1	18	0	2	2

a Present the data in a frequency distribution table. (Use class intervals of 0–4, 5–9 etc.)
b Use your table to estimate the mean number of hours that the men spent doing housework.
c Find the median class for hours spent by the men at housework.
d Find the modal class for hours spent by the men at housework.

REASONING

12 The data at right give the age of 25 patients admitted to the emergency ward of a hospital.

18	16	6	75	24
23	82	75	25	21
43	19	84	76	31
78	24	20	63	79
80	20	23	17	19

- **a** Present the data in a frequency distribution table. (Use class intervals of 0−<15, 15−<30 and so on.)
- **b** Draw a histogram of the data.
- **c** What word could you use to describe the pattern of the data in this distribution?
- **d** Use your table to estimate the mean age of patients admitted.
- **e** Find the median class for age of patients admitted.
- **f** Find the modal class for age of patients admitted.
- **g** Draw an ogive of the data.
- **h** Use the ogive to determine the median age.
- **i** Do any of your statistics (mean, median or mode) give a clear representation of the typical age of an emergency ward patient?
- **j** Give some reasons which could explain the pattern of the distribution of data in this question.

13 The batting scores for two cricket players over 6 innings are as follows:

Player A 31, 34, 42, 28, 30, 41

Player B 0, 0, 1, 0, 250, 0

- **a** Find the mean score for each player.
- **b** Which player appears to be better, based upon mean result?
- **c** Find the median score for each player.
- **d** Which player appears to be better when the decision is based on the median result?
- **e** Which player do you think would be the most useful to have in a cricket team and why? How can the mean result sometimes lead to a misleading conclusion?

14 The resting pulse rate of 20 female athletes was measured. The results are detailed below.

50	52	48	52	71	61	30	45	42	48
43	47	51	62	34	61	44	54	38	40

- **a** Construct a frequency distribution table. (Use class sizes of 1−<10, 10−<20 etc.)
- **b** Use your table to estimate the mean of the data.
- **c** Find the median class of the data.
- **d** Find the modal class of the data.
- **e** Draw an ogive of the data. (You may like to use a graphics calculator for this.)
- **f** Use the ogive to determine the median pulse rate.

15 MC In a set of data there is one score that is extremely small when compared to all the others. This outlying value is most likely to:

- **A** have greatest effect upon the mean of the data.
- **B** have greatest effect upon the median of the data.
- **C** have greatest effect upon the mode of the data.
- **D** have very little effect on any of the statistics as we are told that the number is extremely small.

16 The following frequency table gives the number of employees in different salary brackets for a small manufacturing plant.

Position	Salary ($)	Number of employees
Machine operator	18 000	50
Machine mechanic	20 000	15
Floor steward	24 000	10
Manager	62 000	4
Chief executive officer	80 000	1

- **a** Workers are arguing for a pay rise but the management of the factory claims that workers are well paid because the mean salary of the factory is $22 100. Are they being honest?
- **b** Suppose that you were representing the factory workers and had to write a short submission in support of the pay rise. How could you explain the management's claim? Quote some other statistics in favour of your case.

17 Design a set of five numbers with:

- **a** mean = median = mode
- **b** mean > median > mode
- **c** mean < median = mode.

PROBLEM SOLVING

18 The numbers 15, a, 17, b, 22, c, 10 and d have a mean of 14. Find the mean of a, b, c and d.

19 The numbers m, n, p, q, r, and s have a mean of a while x, y and z have a mean of b. Find the mean of all nine numbers.

CHALLENGE 12.1

The mean and median of six two-digit prime numbers is 39 and the mode is 31.
The smallest number is 13.
What are the six numbers?

12.3 Measures of spread

- **Measures of spread** describe how far data values are spread from the centre or from each other.
- A music store proprietor has stores in Newcastle and Wollongong. The number of CDs sold each day over one week is recorded below.

Newcastle:	45, 60, 50, 55, 48, 40, 52
Wollongong:	20, 85, 50, 15, 30, 60, 90

In each of these data sets consider the measures of central tendency.

Newcastle:	Mean = 50	Wollongong:	Mean = 50
	Median = 50		Median = 50
	No mode		No mode

With these measures being the same for both data sets we could come to the conclusion that both data sets are very similar; however, if we look at the data sets, they are very different. We can see that the data for Newcastle are very clustered around the mean, whereas the Wollongong data are spread out more.

- The data from Newcastle are between 45 and 60, whereas the Wollongong data are between 15 and 90.
- **Range** and **interquartile range (IQR)** are both measures of spread.

Range

- The most basic measure of spread is the range. It is defined as the difference between the highest and the lowest values in the set of data.

$$\text{Range} = \text{highest score} - \text{lowest score}$$
$$\Rightarrow \quad \text{Range} = X_{\max} - X_{\min}$$

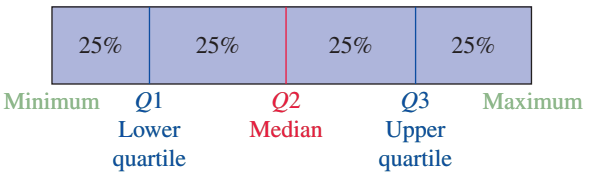
WORKED EXAMPLE 4

Find the range of the given data set: 2.1, 3.5, 3.9, 4.0, 4.7, 4.8, 5.2.

THINK	WRITE
1 Identify the lowest score ($X_{\min}$) of the data set.	Lowest score = 2.1
2 Identify the highest score ($X_{\max}$) of the data set.	Highest score = 5.2
3 Write the rule for the range.	Range = $X_{\max} - X_{\min}$
4 Substitute the known values into the rule.	$= 5.2 - 2.1$
5 Evaluate.	$= 3.1$

Interquartile range

- The interquartile range (IQR) is the range of the middle 50% of all the scores in an ordered set.
  When calculating the interquartile range, the data are first organised into quartiles, each containing 25% of the data. The word 'quartile' comes from the word 'quarter'.



Interquartile range = upper quartile − lower quartile
This can be written as:

$$IQR = Q_{\text{upper}} - Q_{\text{lower}}$$
or
$$IQR = Q_3 - Q_1$$

- The IQR is not affected by extremely large or extremely small data values (**outliers**), so in some circumstances the IQR is a better indicator of the spread of data than the range.

WORKED EXAMPLE 5 TI CASIO

Calculate the interquartile range (IQR) of the following set of data: 3, 2, 8, 6, 1, 5, 3, 7, 6.

THINK	WRITE
1 Arrange the scores in order.	1 2 3 3 5 6 6 7 8
2 Locate the median and use it to divide the data into two halves. *Note:* The median is the 5th score in this data set and should not be included in the lower or upper ends of the data.	1 2 3 3 5 6 6 7 8
3 Find Q_1, the median of the lower half of the data.	$Q_1 = \frac{2+3}{2}$ $= \frac{5}{2}$ $= 2.5$
4 Find Q_3, the median of the upper half of the data.	$Q_3 = \frac{6+7}{2}$ $= \frac{13}{2}$ $= 6.5$
5 Calculate the interquartile range.	$IQR = Q_3 - Q_1$ $= 6.5 - 2.5$ $= 4$

Determining the IQR from a graph

- When data are presented in a frequency distribution table, either ungrouped or grouped, the interquartile range is found by drawing an ogive.

WORKED EXAMPLE 6

The following frequency distribution table gives the number of customers who order different volumes of concrete from a readymix concrete company during the course of a day. Find the interquartile range of the data.

Volume (m^3)	Frequency
0.0–<0.5	15
0.5–<1.0	12
1.0–<1.5	10
1.5–<2.0	8
2.0–<2.5	2
2.5–<3.0	4

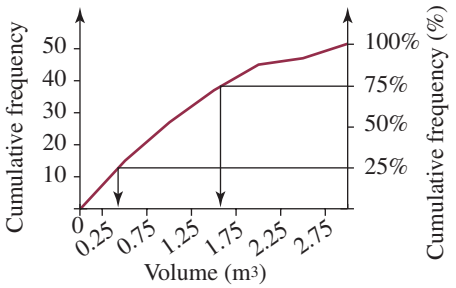
THINK

1 To find the 25th and 75th percentiles from the ogive, first add a class centre column and a cumulative frequency column to the frequency distribution table and fill them in.

WRITE/DRAW

Volume	Class centre	f	cf
0.0–<0.5	0.25	15	15
0.5–<1.0	0.75	12	27
1.0–<1.5	1.25	10	37
1.5–<2.0	1.75	8	45
2.0–<2.5	2.25	2	47
2.5–<3.0	2.75	4	51

2 Draw the ogive. A percentage axis will be useful.



3 Find the upper quartile (75th percentile) and lower quartile (25th percentile) from the ogive.

$Q_3 = 1.6\ m^3$
$Q_1 = 0.4\ m^3$

4 The interquartile range is the difference between the upper and lower quartiles.

$IQR = Q_3 - Q_1$
$= 1.6 - 0.4$
$= 1.2\ m^3$

Exercise 12.3 Measures of spread

assessON

INDIVIDUAL PATHWAYS

PRACTISE
Questions:
1–7, 10, 12

CONSOLIDATE
Questions:
1–8, 10, 11, 12

MASTER
Questions:
1–13

REFLECTION
What do measures of spread tell us about a set of data?

Individual pathway interactivity int-4622 eBookplus

FLUENCY

1 WE4 Find the range for each of the following sets of data.

a 4, 3, 9, 12, 8, 17, 2, 16

b 49.5, 13.7, 12.3, 36.5, 89.4, 27.8, 53.4, 66.8

c $7\frac{1}{2}$, $12\frac{3}{4}$, $5\frac{1}{4}$, $8\frac{2}{3}$, $9\frac{1}{6}$, $3\frac{3}{4}$

2 WE5 Calculate the interquartile range (IQR) for the following sets of data.

a 3, 5, 8, 9, 12, 14

b 7, 10, 11, 14, 17, 23

c 66, 68, 68, 70, 71, 74, 79, 80

d 19, 25, 72, 44, 68, 24, 51, 59, 36

3 The following stem-and-leaf plot shows the mass of newborn babies (rounded to the nearest 100 g). Find the:

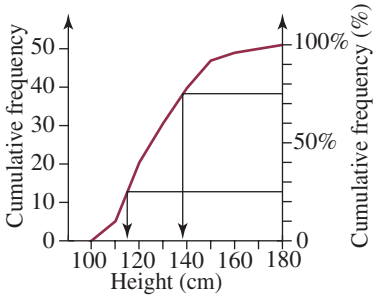
a range of the data

b IQR of the data.

Key: 1* | 9 = 1.9 *kg*

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

4 Use the ogive at right to determine the interquartile range of the data.



5 WE6 The following frequency distribution table gives the amount of time spent by 50 people on shopping for Christmas presents. Estimate the IQR of the data.

Time (h)	0–<0.5	0.5–<1	1–<1.5	1.5–<2	2–<2.5	2.5–<3	3–<3.5	3.5–<4
Frequency	1	2	7	15	13	8	2	2

6 MC Calculate the interquartile range of the following data:

17, 18, 18, 19, 20, 21, 21, 23, 25

A 8 **B** 18 **C** 4 **D** 20

UNDERSTANDING

7 The following frequency distribution table shows the life expectancy in hours of 40 household batteries.

Life (h)	50–<55	55–<60	60–<65	65–<70	70–<75	75–<80
Frequency	4	10	12	8	5	1

a Draw an ogive curve that represents the data in the table at the bottom of the previous page.
b Use the ogive to answer the following questions.
  i What is the median score?
  ii What are the upper and lower quartiles?
  iii What is the interquartile range?
  iv How many batteries lasted less than 60 hours?
  v How many batteries lasted 70 hours or more?

8 Calculate the IQR for the following data.

Class interval	Frequency
120–<130	2
130–<140	3
140–<150	9
150–<160	14
160– < 170	10
170–<180	8
180–<190	6
190–<200	3

9 For each of the following sets of data, state: **i** the range and **ii** the IQR of each set.
a 6, 9, 12, 13, 20, 22, 26, 29
b 7, 15, 2, 26, 47, 19, 9, 33, 38
c 120, 99, 101, 136, 119, 87, 123, 115, 107, 100

REASONING

10 As newly appointed coach of Terrorolo's Meteors netball team, Kate decided to record each player's statistics for the previous season. The number of goals scored by the leading goal shooter was:

1, 3, 8, 18, 19, 23, 25, 25, 25, 26, 27, 28,
28, 28, 28, 29, 29, 30, 30, 33, 35, 36, 37, 40.

a Find the mean of the data.
b Find the median of the data.
c Find the range of the data.
d Find the interquartile range of the data.
e There are three scores that are much lower than most. Explain the effect these scores have on the summary statistics.

11 The following back-to-back stem-and-leaf plot shows the ages of 30 pairs of men and women when entering their first marriage.

Key: 1 | 6 = 16 years old

Leaf: Men	*Stem*	*Leaf: Women*
9 9 8	1	6 7 7 8 9
9 9 8 8 7 6 4 4 3 2 0	2	0 0 1 2 3 4 5 6 7 7 8 9
9 8 8 8 6 5 5 4 3 2	3	0 1 2 2 3 4 7 9
6 3 0 0	4	1 2 4 8
6 0	5	2

a Find the mean, median, range and interquartile range of each set.
b Write a short paragraph comparing the two distributions.

PROBLEM SOLVING

12 Find the mean, median, mode, range and IQR of the following data collected when the temperature of the soil around 25 germinating seedlings was recorded: 28.9, 27.4, 23.6, 25.6, 21.1, 22.9, 29.6, 25.7, 27.4, 23.6, 22.4, 24.6, 21.8, 26.4, 24.9, 25.0, 23.5, 26.1, 23.6, 25.3, 29.5, 23.5, 22.0, 27.9, 23.6.

13 Four positive numbers a, b, c and d have a mean of 12, a median and mode of 9 and a range of 14. Find the values of a, b, c and d.

eBookplus
Digital doc
WorkSHEET 12.1
doc-14595

12.4 Box-and-whisker plots

Five-number summary

- A five-number summary is a list consisting of the lowest score, lower quartile, median, upper quartile and greatest score of a set of data.

X_{min}	Q_1	Median (Q_2)	Q_3	X_{max}

WORKED EXAMPLE 7

From the following five-number summary, find:

a the interquartile range **b the range.**

X_{min}	Q_1	Median (Q_2)	Q_3	X_{max}
29	37	39	44	48

THINK

a The interquartile range is the difference between the upper and lower quartiles.

b The range is the difference between the greatest score and the lowest score.

WRITE

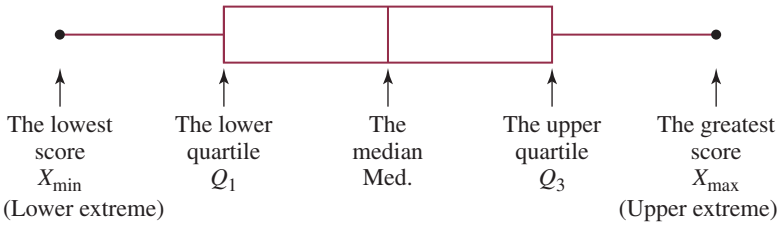
$Q_3 = 44, X_{max} = 48$

a $\text{IQR} = Q_3 - Q_1$
$= 44 - 37$
$= 7$

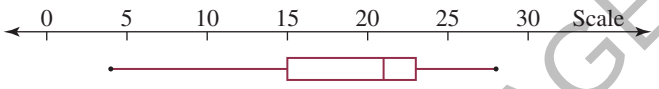
b $\text{Range} = X_{max} - X_{min}$
$= 48 - 29$
$= 19$

Box-and-whisker plots

- A **box-and-whisker plot** (or **boxplot**) is a graph of the five-number summary.
- Box-and-whisker plots consist of a central divided box with attached whiskers.
- The box spans the interquartile range.
- The median is marked by a vertical line drawn inside the box.
- The whiskers indicate the range of scores:



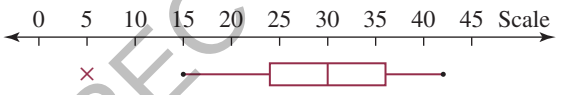
- Box-and-whisker plots are *always drawn to scale*.
- They are presented either with the five-number summary figures attached as labels (diagram at right) or with a scale presented alongside the box-and-whisker plot like the diagram below. They can also be drawn vertically.



Identification of extreme values

- If an extreme value or outlier occurs in a set of data, it can be denoted by a small cross on the box-and-whisker plot. The whisker is then shortened to the next largest (or smallest) figure.

  The box-and-whisker plot below shows that the lowest score was 5. This was an extreme value as the rest of the scores were located within the range 15 to 42.

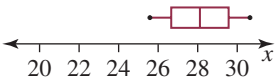


Describing distributions

Symmetry and skewness

- A **symmetrical** plot has data that are evenly spaced around a central point. Examples of a stem-and-leaf plot and a symmetrical boxplot are shown below.

Stem	*Leaf*
26*	6
27	0 1 3
27*	5 6 8 9
28	0 1 1 1 2 4
28*	5 7 8 8
29	2 2 2
29*	5

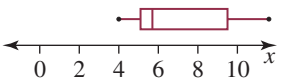


- A **negatively skewed** plot has larger amounts of data at the higher end. This is illustrated by the stem-and-leaf plot below where the leaves increase in length as the data increase in value. It is illustrated on the boxplot when the median is to the right within the box.

Stem	*Leaf*
5	1
6	2 9
7	1 1 2 2
8	1 4 4 5 6 6
9	5 3 4 4 5 6 7 7 7

- A **positively skewed** plot has larger amounts of data at the lower end. This is illustrated on the stem-and-leaf plot below where the leaves increase in length as the data decrease in value. It is illustrated on the boxplot when the median is to the left within the box.

Stem	*Leaf*
5	1 3 4 4 5 6 7 7 7
6	2 4 4 5 6 6
7	1 1 2 2
8	1 6
9	5



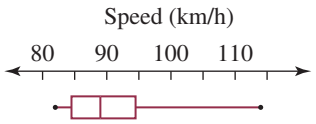
WORKED EXAMPLE 8 TI CASIO

The following stem-and-leaf plot gives the speed of 25 cars caught by a roadside speed camera.

Key: 8 | 2 = 82 km/h, 8* | 6 = 86 km/h

Stem	*Leaf*
8	2 2 4 4 4 4
8*	5 5 6 6 7 9 9 9
9	0 1 1 2 4
9*	5 6 9
10	0 2
10*	
11	4

a Prepare a five-number summary of the data.
b Draw a box-and-whisker plot of the data. (Identify any extreme values.)
c Describe the distribution of the data.

THINK	WRITE
1 First identify the positions of the median and upper and lower quartiles. There are 25 data values. The median is the $\frac{n+1}{2}$th score. The lower quartile is the median of the lower half of the data. The upper quartile is the median of the upper half of the data (each half contains 12 scores).	The median is the $\frac{25+1}{2}$th score — that is, the13th score. Q_1 is the $\frac{12+1}{2}$th score in the lower half — that is, the 6.5th score. That is, halfway between the 6th and 7th scores. Q_3 is halfway between the 6th and 7th scores in the upper half of the data.
2 Mark the positions of the median and upper and lower quartiles on the stem-and-leaf plot.	Key: 8 | 2 = 82 km/h 8* | 6 = 86 km/h *Stem* | *Leaf* 8 | 2 2 4 4 4 4 (Q_1) 8* | 5 5 6 6 7 9 (9) 9 (Median) 9 | 0 1 1 2 4 9* | 5 6 9 (Q_3) 10 | 0 2 10* 11 | 4
a Write the five-number summary: The lowest score is 82. The lower quartile is between 84 and 85; that is, 84.5. The median is 89. The upper quartile is between 94 and 95; that is, 94.5. The greatest score is 114.	**a** Five-number summary:
	X_{min} = 82, Q_1 = 84.5, Median = 89, Q_3 = 94.5, X_{max} = 114
b Draw a labelled axis using an appropriate scale. Plot the points from the five-number summary.	**b** 
c Describe the distribution.	**c** The data are skewed (positively).

Shapes of graphs

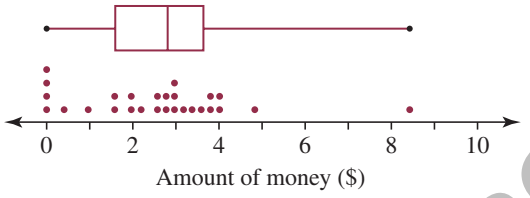
Boxplots and dot plots

- Boxplots are a concise summary of data. A boxplot can be directly related to a dot plot.
- **Dot plots** display each data value represented by a dot placed on a number line.

The following data are the amount of money (in $) that a group of 27 five-year-olds had with them on a day visiting the zoo with their parents.

0	1.65	0	2.60	3	8.45	4	0.55	4.10	3.35	3.25
2	2.85	2.90	1.70	3.65	1	0	0	2.25	2.05	3
3.80	2.65	4.75	3.90	2.95						

- The dot plot below and its comparative boxplot show the distribution of these data.



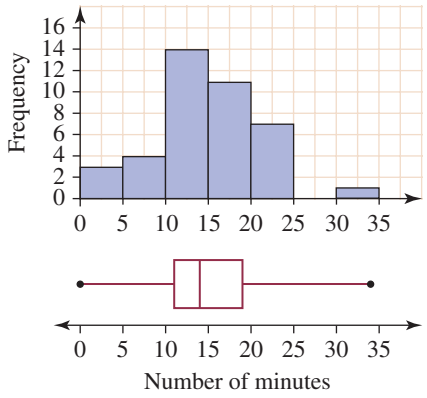
Both graphs indicate that the data are positively skewed. The dot plot clearly shows the presence of the outlier. This is less obvious with the boxplot. However, the boxplot provides an excellent summary of the centre and spread of the distribution.

Boxplots and histograms

- **Histograms** are graphs that display continuous numerical variables and do not retain all original data.
- The following data are the number of minutes, rounded to the nearest minute, that forty Year 10 students take to travel to their school on a particular day.

15	22	14	12	21	34	19	11	13	0	16
4	23	8	12	18	24	17	14	3	10	12
9	15	20	5	19	13	17	11	16	19	24
12	7	14	17	10	14	23				

The data are displayed in the histogram and boxplot shown.



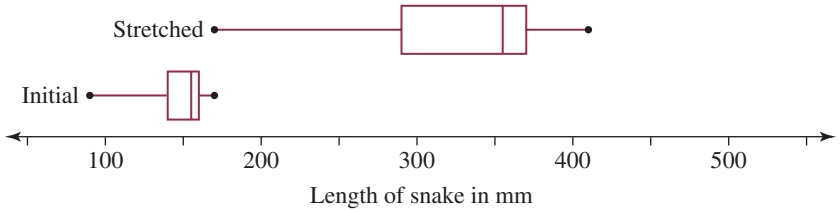
Both graphs indicate that the data are slightly positively skewed. The histogram clearly shows the frequencies of each class interval. Neither graph displays the original values. The histogram does not give precise information about the centre, but the distribution of the data is visible. However, the boxplot provides an excellent summary of the centre and spread of the distribution.

Parallel boxplots

- A major reason for developing statistical skills is to be able to make comparisons between sets of data.
- Consider the following scenario.
  - Each member of a class was given a jelly snake to stretch. They each measured the initial length of their snake to the nearest centimetre and then slowly stretched the snake to make it as long as possible. They then measured the maximum length of the snake by recording how far it had stretched at the time it broke. The results were recorded in the following table.

Initial length (cm)	Stretched length (cm)	Initial length (cm)	Stretched length (cm)
13	29	14	27
14	28	13	27
17	36	15	36
10	24	16	36
14	35	15	36
16	36	16	34
15	37	17	35
16	37	12	27
14	30	9	17
16	33	16	41
17	36	17	38
16	38	16	36
17	38	17	41
14	31	16	33
17	40	11	21

– The data were then displayed on **parallel boxplots**, with the axis displaying in millimetres.
– By drawing the two boxplots on a single axis, it is easy to compare them.



The change in the length of the snake when stretched is evidenced by the increased median and spread shown on the boxplots. The median snake length before being stretched was 150 mm, but the median snake length after being stretched was 350 mm. The range increased after stretching, as did the IQR.

assess on

Exercise 12.4 Box-and-whisker plots

REFLECTION
What advantages and disadvantages do box-and-whisker plots have as a visual form of representing data?

INDIVIDUAL PATHWAYS

PRACTISE	CONSOLIDATE	MASTER
Questions: 1–7, 10, 13, 16, 19	Questions: 1–8, 10–12, 14, 16, 19	Questions: 1–20

Individual pathway interactivity int-4623 eBookplus

FLUENCY

1 WE7 From the following five-number summary find:

X_{min}	Q_1	Median	Q_3	X_{max}
6	11	13	16	32

a the interquartile range
b the range.

2 From the following five-number summary find:

X_{min}	Q_1	Median	Q_3	X_{max}
101	119	122	125	128

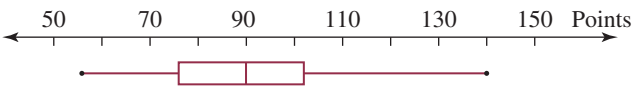
a the interquartile range
b the range.

3 From the following five-number summary find:

X_{min}	Q_1	Median	Q_3	X_{max}
39.2	46.5	49.0	52.3	57.8

a the interquartile range
b the range.

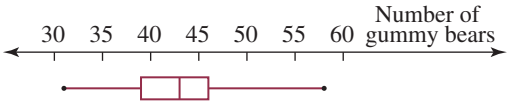
4 The box-and-whisker plot at right shows the distribution of final points scored by a football team over a season's roster.



a What was the team's greatest points score?
b What was the team's least points score?

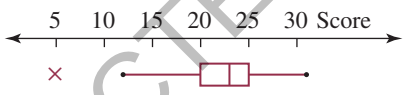
c What was the team's median points score?
d What was the range of points scored?
e What was the interquartile range of points scored?

5 The box-and-whisker plot at right shows the distribution of data formed by counting the number of gummy bears in each of a large sample of packs.



a What was the largest number of gummy bears in any pack?
b What was the smallest number of gummy bears in any pack?
c What was the median number of gummy bears in any pack?
d What was the range of numbers of gummy bears per pack?
e What was the interquartile range of gummy bears per pack?

Questions **6** to **8** refer to the following box-and-whisker plot.



6 MC The median of the data is:
A 20 **B** 23 **C** 25 **D** 31

7 MC The interquartile range of the data is:
A 23 **B** 26 **C** 5 **D** 20 to 25

8 MC Which of the following is not true of the data represented by the box-and-whisker plot?
A One-quarter of the scores are between 5 and 20.
B Half of the scores are between 20 and 25.
C The lowest quarter of the data is spread over a wide range.
D Most of the data are contained between the scores of 5 and 20.

UNDERSTANDING

9 The number of sales made each day by a salesperson is recorded over a 2-week period:

25, 31, 28, 43, 37, 43, 22, 45, 48, 33

a Prepare a five-number summary of the data. (There is no need to draw a stem-and-leaf plot of the data. Just arrange them in order of size.)
b Draw a box-and-whisker plot of the data.

10 The data below show monthly rainfall in millimetres.

J	F	M	A	M	J	J	A	S	O	N	D
10	12	21	23	39	22	15	11	22	37	45	30

a Prepare a five-number summary of the data.
b Draw a box-and-whisker plot of the data.

11 **WE8** The stem-and-leaf plot at right details the age of 25 offenders who were caught during random breath testing.
a Prepare a five-number summary of the data.
b Draw a box-and-whisker plot of the data.
c Describe the distribution of the data.

Key: 1 | 8 = 18 years

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

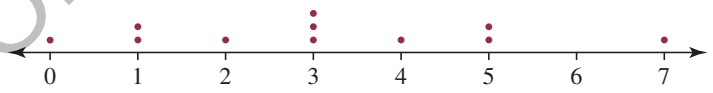
12 The following stem-and-leaf plot details the price at which 30 blocks of land in a particular suburb sold for.

Key: 12 | 4 = $124 000

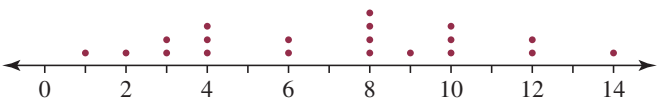
Stem	*Leaf*
12	4 7 9
13	0 0 2 5 5
14	0 0 2 3 5 5 7 9 9
15	0 0 2 3 7 7 8
16	0 2 2 5 8
17	5

a Prepare a five-number summary of the data.
b Draw a box-and-whisker plot of the data.

13 Prepare comparative boxplots for the following dot plots (using the same axis) and describe what each plot reveals about the data.
a Number of sick days taken by workers last year at factory A



b Number of sick days taken by workers last year at factory B



14 An investigation into the transport needs of an outer suburb community recorded the number of passengers boarding a bus during each of its journeys, as follows.
12, 43, 76, 24, 46, 24, 21, 46, 54, 109, 87, 23, 78, 37, 22, 139, 65, 78, 89, 52, 23, 30, 54, 56, 32, 66, 49

Display the data by constructing a histogram using class intervals of 20 and a comparative boxplot on the same axis.

15 At a weight-loss clinic, the following weights (in kilograms) were recorded before and after treatment.

Before	75	80	75	140	77	89	97	123	128	95	152	92
After	69	66	72	118	74	83	89	117	105	81	134	85

Before	85	90	95	132	87	109	87	129	135	85	137	102
After	79	84	90	124	83	102	84	115	125	81	123	94

a Prepare a five-number summary for weight before and after treatment.
b Draw parallel boxplots for weight before and after treatment.
c Comment on the comparison of weights before and after treatment.

REASONING

16 The following data detail the number of hamburgers sold by a fast food outlet every day over a 4-week period.

M	T	W	T	F	S	S
125	144	132	148	187	172	181
134	157	152	126	155	183	188
131	121	165	129	143	182	181
152	163	150	148	152	179	181

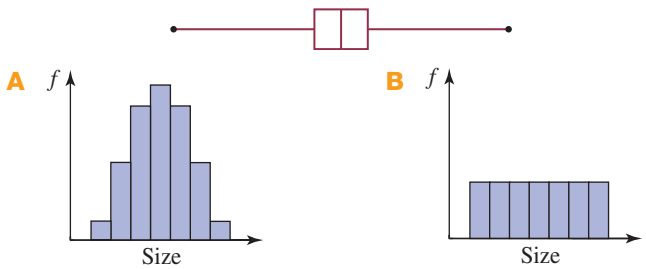
a Prepare a stem-and-leaf plot of the data. (Use a class size of 10.)
b Draw a box-and-whisker plot of the data.
c What do these graphs tell you about hamburger sales?

17 The following data show the ages of 30 mothers upon the birth of their first baby.

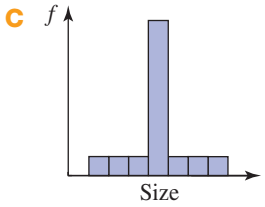
22	21	18	33	17	23	22	24	24	20
25	29	32	18	19	22	23	24	28	20
31	22	19	17	23	48	25	18	23	20

a Prepare a stem-and-leaf plot of the data. (Use a class size of 5.)
b Draw a box-and-whisker plot of the data. Indicate any extreme values appropriately.
c Describe the distribution in words. What does the distribution say about the age that mothers have their first baby?

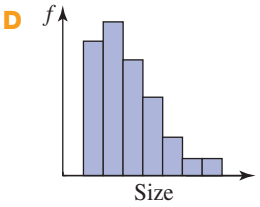
18 MC Match the box-and-whisker plot with its most likely histogram.



C

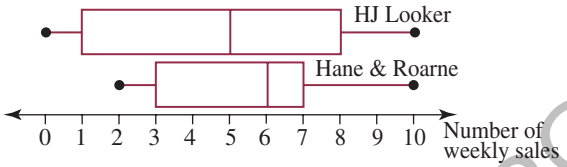


D



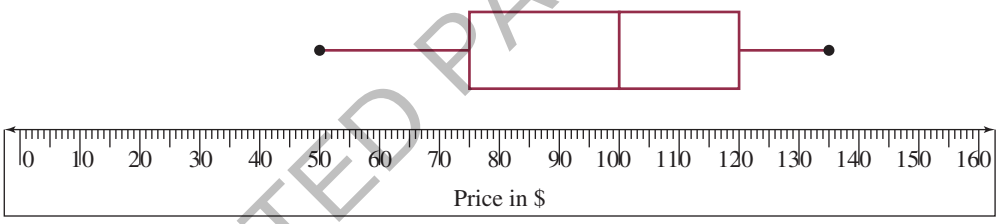
PROBLEM SOLVING

19 Consider the box-and-whisker plot below which shows the number of weekly sales of houses by two real estate agencies.



a What is the median number of weekly sales for each real estate agency?
b Which agency had the greater range of sales?
c Which agency had the greater interquartile range of sales?
d Which agency performed better? Explain your answer.

20 Fifteen French restaurants were visited by three newspaper restaurant reviewers. The average price of a meal for a single person was investigated. The following box-and-whisker plot shows the results.



a What was the price of the cheapest meal?
b What was the price of the most expensive meal?
c What is the median cost of a meal?
d What is the interquartile range for the price of a meal?
e What percentage of the prices were below the median?

12.5 The standard deviation

- The **standard deviation** for a set of data is a measure of how far the data values are spread out (deviate) from the mean.
- **Deviation** is the difference between each data value and the mean $(x - \bar{x})$. The standard deviation is calculated from the square of the deviations.
- Standard deviation is denoted by the Greek letter sigma, σ, and can be calculated by using the formula

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

where $\bar{x}$ is the mean of the data values and n is the number of data values.
- A low standard deviation indicates that the data values tend to be close to the mean.
- A high standard deviation indicates that the data values tend to be spread out over a large range, away from the mean.

- Standard deviation can be calculated using a scientific or graphics calculator, or it can be calculated from a frequency table by following the steps below.

Step 1	Calculate the mean.	Step 2	Calculate the deviations.
Step 3	Square each deviation.	Step 4	Sum the squares.
Step 5	Divide the sum of the squares by the number of data values.	Step 6	Take the square root of the result.

WORKED EXAMPLE 9 TI CASIO

The number of lollies in each of 8 packets is 11, 12, 13, 14, 16, 17, 18, 19. Calculate the mean and standard deviation correct to 2 decimal places.

THINK

1 Calculate the mean.

WRITE

$$\bar{x} = \frac{11 + 12 + 13 + 14 + 16 + 17 + 18 + 19}{8}$$
$$= \frac{120}{8}$$
$$= 15$$

2 To calculate the deviations $(x - \bar{x})$, set up a frequency table as shown and complete.

No. of lollies (x)	$(x - \bar{x})$
11	$11-15 = -4$
12	-3
13	-2
14	-1
16	1
17	2
18	3
19	4
Total	

3 Add another column to the table to calculate the square of the deviations, $(x - \bar{x})^2$. Then sum the results: $\sum (x - \bar{x})^2$.

No. of lollies (x)	$(x - \bar{x})$	$(x - \bar{x})^2$
11	$11-15 = -4$	16
12	-3	9
13	-2	4
14	-1	1
16	1	1
17	2	4
18	3	9
19	4	16
Total		$\sum (x - \bar{x})^2 = 60$

4 To calculate the standard deviation, divide the sum of the squares by the number of data values, then take the square root of the result.

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$
$$= \sqrt{\frac{60}{8}}$$
$$\approx 2.74 \quad \text{(correct to 2 decimal places)}$$

5 Check the result using a calculator.

The calculator returns an answer of $\sigma_n = 2.73861$.
Answer confirmed.

6 Interpret the result.

The average (mean) number of lollies in each pack is 15 with a standard deviation of 2.74, which means that the number of lollies in each pack differs from the mean by an average of 2.74.

- When calculating the standard deviation from a frequency table, the frequencies must be taken into account. Therefore, the following formula is used.

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}}$$

WORKED EXAMPLE 10 TI CASIO

Lucy's scores in her last 12 games of golf were 87, 88, 88, 89, 90, 90, 90, 92, 93, 93, 95 and 97. Calculate the mean score and the standard deviation correct to 2 decimal places.

THINK

1 To calculate the mean, first set up a frequency table.

WRITE

Golf score (x)	Frequency (f)	fx
87	1	87
88	2	176
89	1	89
90	3	270
92	1	92
93	2	186
95	1	95
97	1	97
Total	$\sum f = 12$	$\sum fx = 1092$

2 Calculate the mean.

$$\bar{x} = \frac{\sum fx}{\sum f}$$
$$= \frac{1092}{12}$$
$$= 91$$

3 To calculate the deviations $(x - \bar{x})$, add another column to the frequency table and complete.

Golf score (x)	Frequency (f)	fx	$(x - \bar{x})$
87	1	87	$87 - 91 = -4$
88	2	176	-3
89	1	89	-2
90	3	270	-1
92	1	92	1
93	2	186	2
95	1	95	4
97	1	97	6
Total	$\sum f = 12$	$\sum fx = 1092$	

4 Add another column to the table and multiply the square of the deviations, $(x - \bar{x})^2$, by the frequency $f(x - \bar{x})^2$.
Then sum the results: $\sum f(x - \bar{x})^2$.

Golf score (x)	Frequency (f)	fx	$(x - \bar{x})$	$f(x - \bar{x})^2$
87	1	87	$87 - 91 = -4$	$1 \times (-4)^2 = 16$
88	2	176	-3	18
89	1	89	-2	4
90	3	270	-1	3
92	1	92	1	1
93	2	186	2	8
95	1	95	4	16
97	1	97	6	36
Total	$\sum f = 12$	$\sum fx = 1092$		$\sum f(x - \bar{x})^2 = 102$

5 Calculate the standard deviation using the formula.

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}}$$
$$= \sqrt{\frac{120}{12}}$$
$$\approx 2.92$$
(correct to 2 decimal places)

6 Check the result using a calculator.

The calculator returns an answer of $\sigma_n = 2.91548$.
The answer is confirmed.

7 Interpret the result.

The average (mean) score for Lucy is 91 with a standard deviation of 2.92, which means that her score differs from the mean by an average of 2.92.

Why the deviations are squared

- For large data sets that are symmetrically distributed, the sum of the deviations is usually zero, that is, $\sum(x - \bar{x}) = 0$. When the mean is greater than the data value ($\bar{x} > x$), the deviation is negative. When the mean is smaller than the data value ($\bar{x} < x$), the deviation is positive. The negative and positive deviations cancel each other out; therefore, calculating the sum and average of the deviations is not useful. This explains why the standard deviation is calculated using the squares of the deviations, $(x - \bar{x})^2$, for all data values.

Standard deviations of populations and samples

- So far we have calculated the standard deviation for a population of data, that is, for complete sets of data. There is another formula for calculating standard deviation for samples of data, that is, data that have been randomly selected from a larger population.
- For example, a sample of 100 Year 10 students from New South Wales is taken to determine the amount of time they spend on their mobile phones. In this case, the standard deviation formula, denoted by s, that would apply is

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$$

- The calculator usually displays both values for the standard deviation, so it is important to understand the difference between them. However, in this course we will use the formula for populations, σ.

Effects on standard deviation

- The standard deviation is affected by extreme values.

WORKED EXAMPLE 11

On a particular day Lucy played golf brilliantly and scored 60. The scores in her previous 12 games of golf were 87, 88, 88, 89, 90, 90, 90, 92, 93, 93, 95 and 97 (see Worked example 10). Comment on the effect this latest score has on the standard deviation.

THINK	WRITE
1 Use a calculator to calculate the mean and the standard deviation.	$\bar{x} = 88.6154$ ≈ 88.62 $\quad \sigma = 8.7225$ ≈ 8.72
2 Interpret the result and compare it to the results found in Worked example 10.	In the first 12 games Lucy's mean score was 91 with a standard deviation of 2.92. This implied that Lucy's scores on average were 2.92 either side of her average of 91. Lucy's latest performance resulted in a mean score of 88.62 with a standard deviation of 8.72. This indicates a slightly lower mean score, but the much higher standard deviation indicates that the data are now much more spread out and that the extremely good score of 60 is an anomaly.

Properties of standard deviation

- If a constant c is added to all data values in a set, the deviations $(x - \bar{x})$ will remain unchanged and consequently the standard deviation remains unchanged.
- If all data values in a set are multiplied by a constant k, the deviations $(x - \bar{x})$ will be multiplied by k, that is $k(x - \bar{x})$; consequently the standard deviation is increased by a factor of k.
- Standard deviation can be used to measure consistency.
- When the standard deviation is low we are able to say that the scores in the data set are more consistent with each other.

WORKED EXAMPLE 12

For the data 5, 9, 6, 11, 10, 7:

a calculate the standard deviation

b calculate the standard deviation if 4 is added to each data value. Comment on the effect.

c calculate the standard deviation if all data values are multiplied by 2. Comment on the effect.

THINK

a 1 Calculate the mean.

WRITE

a $\bar{x} = \dfrac{5 + 9 + 6 + 11 + 10 + 7}{6}$

$= 8$

2 Set up a frequency table and enter the squares of the deviations.

(x)	$(x - \bar{x})$	$(x - \bar{x})^2$
5	$5 - 8 = -3$	9
6	-2	4
7	-1	1
9	1	1
10	2	4
11	3	9
Total		$\sum (x - \bar{x})^2 = 28$

3 To calculate the standard deviation, apply the formula for standard deviation.

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{28}{6}}$$

≈ 2.16

(correct to 2 decimal places)

b 1 Add 4 to each data value in the set.

b 9, 13, 10, 15, 14, 11

2 Calculate the mean.

$$\bar{x} = \frac{9 + 13 + 10 + 15 + 14 + 11}{6}$$

$= 12$

3 Set up a frequency table and enter the squares of the deviations.

(x)	$(x - \bar{x})$	$(x - \bar{x})^2$
9	$9 - 12 = -3$	9
10	-2	4
11	-1	1
13	1	1
14	2	4
15	3	9
Total		$\sum (x - \bar{x})^2 = 28$

4 To calculate the standard deviation, apply the formula for standard deviation.

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$
$$= \sqrt{\frac{28}{6}}$$
$$\approx 2.16$$
(correct to 2 decimal places)

5 Comment on the effect of adding of 4 to each data value.

Adding 4 to each data value increased the mean but had no effect on the standard deviation, which remained at 2.16.

c 1 Multiply each data value in the set by 2.

c 10, 18, 12, 22, 20, 14

2 Calculate the mean.

$$\bar{x} = \frac{10 + 18 + 12 + 22 + 20 + 14}{6}$$
$$= 16$$

3 Set up a frequency table and enter the squares of the deviations.

(x)	$(x - \bar{x})$	$(x - \bar{x})^2$
10	$10 - 16 = -6$	36
12	-4	16
14	-2	4
18	2	4
20	4	16
22	6	36
Total		$\sum (x - \bar{x})^2 = 112$

4 To calculate the standard deviation, apply the formula for standard deviation.

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$
$$= \sqrt{\frac{112}{6}}$$
$$\approx 4.32$$
(correct to 2 decimal places)

5 Comment on the effect of multiplying each data value by 2.

Multiplying each data value by 2 doubled the mean and doubled the standard deviation, which changed from 2.16 to 4.32.

Exercise 12.5 The standard deviation

assessON

INDIVIDUAL PATHWAYS

PRACTISE	CONSOLIDATE	MASTER
Questions: 1–7, 9, 10, 13	Questions: 1–11, 13	Questions: 1–14

Individual pathway interactivity int-4624 eBookplus

REFLECTION
What does the standard deviation tell us about a set of data?

FLUENCY

1 WE9 Calculate the standard deviation of each of the following data sets, correct to 2 decimal places.

a 3, 5, 8, 2, 7, 1, 6, 5

b 11, 8, 7, 12, 10, 11, 14

c 25, 15, 78, 35, 56, 41, 17, 24

d 5.2, 4.7, 5.1, 12.6, 4.8

2 WE10 Calculate the standard deviation of each of the following data sets, correct to 2 decimal places.

a

Score (x)	Frequency (f)
1	1
2	5
3	9
4	7
5	3

b

Score (x)	Frequency (f)
16	15
17	24
18	26
19	28
20	27

c

Score (x)	Frequency (f)
8	15
10	19
12	18
14	7
16	6
18	2

d

Score (x)	Frequency (f)
65	15
66	15
67	16
68	17
69	16
70	15
71	15
72	12

3 Complete the following frequency distribution table and use the table to calculate the standard deviation of the data set, correct to 2 decimal places.

Class	Class centre (x)	Frequency (f)
1–10		6
11–20		15
21–30		25
31–40		8
41–50		6

4 First-quarter profit increases for 8 leading companies are given below as percentages.

2.3 0.8 1.6 2.1 1.7 1.3 1.4 1.9

Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

5 The heights in metres of a group of army recruits are given below.

1.8 1.95 1.87 1.77 1.75 1.79 1.81 1.83 1.76 1.80 1.92 1.87 1.85 1.83

Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

6 Times (to the nearest tenth of a second) for the heats in the open 100 m sprint at the school sports are given at right.

Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

Key: 11 | 0 = 11.0 s

Stem	*Leaf*
11	0
11	2 3
11	4 4 5
11	6 6
11	8 8 9
12	0 1
12	2 2 3
12	4 4
12	6
12	9

7 The number of outgoing phone calls from an office each day over a 4-week period is shown on the stem plot at right.

Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

Key: 1 | 3 = 13 calls

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

8 **MC** A new legal aid service has been operational for only 5 weeks. The number of people who have made use of the service each day during this period is set out below.

Key: 1 | 6 = 16 people

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

The standard deviation (to 2 decimal places) of these data is:

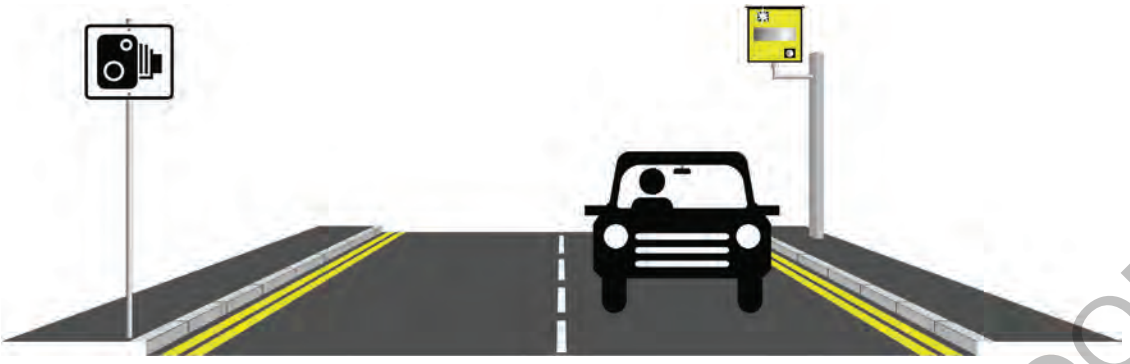
A 6.00 **B** 6.34 **C** 6.47 **D** 15.44

UNDERSTANDING

9 **WE11** The speeds, in km/h, of the first 25 cars caught by a roadside speed camera on a particular day were:

82, 82, 84, 84, 84, 84, 85, 85, 85, 86, 86, 87, 89, 89, 89, 90, 91, 91, 92, 94, 95, 96, 99, 100, 102

The next car that passed the speed camera was travelling at 140 km/h. Comment on the effect of the speed of this last car on the standard deviation for the data.



REASONING

10 WE12 For the data 1, 4, 5, 9, 11:
- **a** calculate the standard deviation
- **b** calculate the standard deviation if 7 is added to each data value. Comment on the effect.
- **c** calculate the standard deviation if all data values are multiplied by 3. Comment on the effect.

11 Show using an example the effect, if any, on the standard deviation of adding a data value to a set of data that is equivalent to the mean.

12 If the mean for a set of data is 45 and the standard deviation is 6, how many standard deviations above the mean is a data value of 57?

PROBLEM SOLVING

13 Five numbers a, b, c, d and e have a mean of 12 and a standard deviation of 4.
- **a** If each number is increased by 3, find the new mean and standard deviation in terms of the original mean and standard deviation.
- **b** If each number is multiplied by 3, find the new mean and standard deviation in terms of the original mean and standard deviation.

14 Twenty-five students sat a test and the results for 24 of the students are given in the following stem-and-leaf plot.

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

- **a** If the average mark for the test was 27.84, determine the mark obtained by the 25th student.
- **b** How many students scored higher than the median score?
- **c** Find the standard deviation of the marks, giving your answer correct to 2 decimal places.

eBookplus
Digital doc
WorkSHEET 12.2
doc-14596

12.6 Comparing data sets

- Besides locating the centre of the data (the mean, median or mode), any analysis of data must measure the extent of the spread of the data (range, interquartile range and standard deviation). Two data sets may have centres that are very similar but be quite differently distributed.
- Decisions need to be made about which measure of centre and which measure of spread to use when analysing and comparing data.

eBookplus
Interactivity
Parallel boxplots
int-2788

- The mean is calculated using every data value in the set. The median is the middle score of an ordered set of data, so it does not include every individual data value in its calculation. The mode is the most frequently occurring data value, so it also does not include every individual data value in its calculation.
- The range is calculated by finding the difference between the maximum and minimum data values, so it includes outliers. It provides only a rough idea about the spread of the data and is inadequate in providing sufficient detail for analysis. It is useful, however, when we are interested in extreme values such as high and low tides or maximum and minimum temperatures.

  The interquartile range is the difference between the upper and lower quartiles, so it does not include every data value in its calculation, but it will overcome the problem of outliers skewing data.

  The standard deviation is calculated using every data value in the set.

WORKED EXAMPLE 13

For the two sets of data 6, 7, 8, 9, 10 and 12, 4, 10, 11, 3:

a calculate the mean

b calculate the standard deviation

c comment on the similarities and differences.

THINK	WRITE
a 1 Calculate the mean of the first set of data.	**a** $\bar{x}_1 = \dfrac{6 + 7 + 8 + 9 + 10}{5}$ $= 8$
2 Calculate the mean of the second set of data.	$\bar{x}_2 = \dfrac{12 + 4 + 10 + 11 + 3}{5}$ $= 8$
b 1 Calculate the standard deviation of the first set of data.	**b** $\sigma_1 = \sqrt{\dfrac{(6-8)^2 + (7-8)^2 + (8-8)^2 + (9-8)^2 + (10-8)^2}{5}}$ ≈ 1.41
2 Calculate the standard deviation of the second set of data.	$\sigma_2 = \sqrt{\dfrac{(12-8)^2 + (4-8)^2 + (10-8)^2 + (11-8)^2 + (3-8)^2}{5}}$ ≈ 3.74
c Comment on the findings.	**c** For both sets of data the mean was the same, 8. However, the standard deviation for the second set (3.74) was much higher than the standard deviation of the first set (1.41), implying that the second set is more widely distributed than the first. This is confirmed by the range, which is $10 - 6 = 4$ for the first set and $12 - 3 = 9$ for the second.

- When multiple data displays are used to display similar sets of data, comparisons and conclusions can then be drawn about the data.
- We can use **back-to-back stem-and-leaf plots** and multiple or parallel box-and-whisker plots to help compare statistics such as the median, range and interquartile range.

WORKED EXAMPLE 14 TI CASIO

Below are the scores achieved by two students in eight Mathematics tests throughout the year.
John: 45, 62, 64, 55, 58, 51, 59, 62
Penny: 84, 37, 45, 80, 74, 44, 46, 50

a Determine the most appropriate measure of centre and measure of spread to compare the performance of the students.
b Which student had the better overall performance on the eight tests?
c Which student was more consistent over the eight tests?

THINK	WRITE
a In order to include all data values in the calculation of measures of centre and spread, calculate the mean and standard deviation.	a John: $\bar{x} = 57, \sigma = 6$ Penny: $\bar{x} = 57.5, \sigma = 17.4$
b Compare the mean for each student. The student with the higher mean performed better overall.	b Penny performed slightly better on average as her mean mark was higher than John's.
c Compare the standard deviation for each student. The student with the lower standard deviation performed more consistently.	c John was the more consistent student because his standard deviation was much lower than Penny's. This means that his test results were closes to his mean score than Penny's were to hers.

Exercise 12.6 Comparing data sets

assessON

INDIVIDUAL PATHWAYS

PRACTISE	CONSOLIDATE	MASTER
Questions: 1–7, 9, 10, 12, 14	Questions: 1–12, 15	Questions: 1–18

Individual pathway interactivity int-4625 eBookplus

REFLECTION
Which data display is best for comparing data sets?

FLUENCY

1 WE13 For the two sets of data 65, 67, 61, 63, 62, 60 and 56, 70, 65, 72, 60, 55:
- **a** calculate the mean
- **b** calculate the standard deviation
- **c** comment on the similarities and differences.

2 A bank surveys the average morning and afternoon waiting times for customers. The figures were taken each Monday to Friday in the morning and afternoon for one month. The stem-and-leaf plot below shows the results.

Key: 1 | 2 = 1.2 minutes

Leaf: Morning	*Stem*	*Leaf: Afternoon*
7	0	7 8 8
8 6 3 1 1	1	1 1 2 4 4 5 6 6 6 7
9 6 6 6 5 5 4 3 3 1	2	2 5 5 8
9 5 2	3	1 6
5	4	
	5	7

a Find the median morning waiting time and the median afternoon waiting time.
b Calculate the range for morning waiting times and the range for afternoon waiting times.
c What conclusions can be made from the display about the average waiting time at the bank in the morning compared with the afternoon?

3 In a class of 30 students there are 15 boys and 15 girls. Their heights are measured (in metres) and are listed below.

Boys: 1.65, 1.71, 1.59, 1.74, 1.66, 1.69, 1.72, 1.66, 1.65, 1.64, 1.68, 1.74, 1.57, 1.59, 1.60

Girls: 1.66, 1.69, 1.58, 1.55, 1.51, 1.56, 1.64, 1.69, 1.70, 1.57, 1.52, 1.58, 1.64, 1.68, 1.67

Display this information in a back-to-back stem-and-leaf plot.

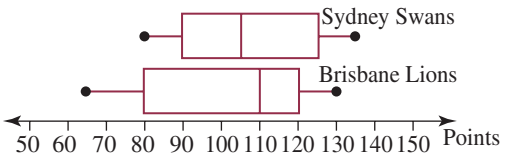
4 The stem-and-leaf plot at right is used to display the number of vehicles sold by the Ford and Holden dealerships in a Sydney suburb each week for a three-month period.

Key: 1 | 5 = 15 vehicles

Leaf: Ford	*Stem*	*Leaf: Holden*
7 4	0	3 9
9 5 2 2 1 0	1	1 1 1 6 6 8
8 5 4 4	2	2 2 7 9
0	3	5

a State the median of both distributions.
b Calculate the range of both distributions.
c Calculate the interquartile range of both distributions.
d Show both distributions on a box-and-whisker plot.

5 The box-and-whisker plot drawn below displays statistical data for two AFL teams over a season.



a Which team had the higher median score?
b What was the range of scores for each team?
c For each team calculate the interquartile range.

UNDERSTANDING

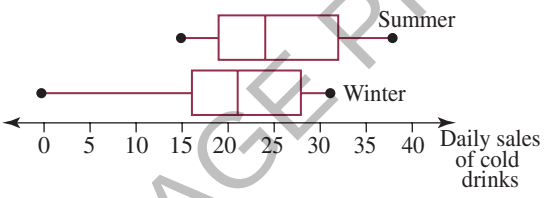
6 Tanya measures the heights (in m) of a group of Year 10 boys and girls and produces the following five-point summaries for each data set.

Boys: 1.45, 1.56, 1.62, 1.70, 1.81

Girls: 1.50, 1.55, 1.62, 1.66, 1.73

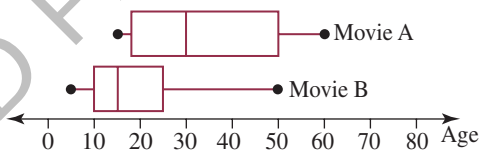
- **a** Draw a box-and-whisker plot for both sets of data and display them on the same scale.
- **b** What is the median of each distribution?
- **c** What is the range of each distribution?
- **d** What is the interquartile range for each distribution?
- **e** Comment on the spread of the heights among the boys and the girls.

7 The box-and-whisker plots at right show the average daily sales of cold drinks at the school canteen in summer and winter.



- **a** Calculate the range of sales in both summer and winter.
- **b** Calculate the interquartile range of the sales in both summer and winter.
- **c** Comment on the relationship between the two data sets, both in terms of measures of centre and measures of spread.

8 MC Andrea surveys the age of people at two movies being shown at a local cinema. The box-and-whisker plot at right shows the results.



Which of the following conclusions could be drawn based on the above information?

- **A** Movie A attracts an older audience than Movie B.
- **B** Movie B attracts an older audience than Movie A.
- **C** Movie A appeals to a wider age group than Movie B.
- **D** Movie B appeals to a wider age group than Movie A.

9 MC *Note:* There may be more than one correct answer.

The figures below show the age of the first 10 men and women to finish a marathon.

Men: 28, 34, 25, 36, 25, 35, 22, 23, 40, 24

Women: 19, 27, 20, 26, 30, 18, 28, 25, 28, 22

Which of the following statements is correct?

- **A** The mean age of the men is greater than the mean age of the women.
- **B** The range is greater among the men than among the women.
- **C** The interquartile range is greater among the men than among the women.
- **D** The standard deviation is greater among the men than among the women.

REASONING

10 WE14 Cory recorded his marks for each test that he did in English and Science throughout the year.

English: 55, 64, 59, 56, 62, 54, 65, 50

Science: 35, 75, 81, 32, 37, 62, 77, 75

- **a** In which subject did Cory achieve the better average mark?
- **b** In which subject was Cory more consistent? Explain your answer.

11 The police set up two radar speed checks on a back street of Sydney and on a main road. In both places the speed limit is 60 km/h. The results of the first 10 cars that have their speed checked are given below.

Back street: 60, 62, 58, 55, 59, 56, 65, 70, 61, 64
Main road: 55, 58, 59, 50, 40, 90, 54, 62, 60, 60

a Calculate the mean and standard deviation of the readings taken at each point.
b On which road are drivers generally driving faster?
c On which road is the spread of the reading taken greater? Justify your answer.

12 Nathan and Timana are wingers in their local rugby league team. The number of tries they have scored in each season are listed below.

Nathan: 25, 23, 13, 36, 1, 8, 0, 9, 16, 20
Timana: 5, 10, 12, 14, 18, 11, 8, 14, 12, 19

a Calculate the mean number of tries scored by each player.
b What is the range of tries scored by each player?
c What is the interquartile range of tries scored by each player?
d Which player would you consider to be the more consistent player? Justify your answer.

13 In boxes of Smarties it is advertised that there are 50 Smarties in each box. Two machines are used to distribute the Smarties into the boxes. The results from a sample taken from each machine are shown in the stem-and-leaf plot below.

Key: 5 | 1 = 51 5*|6 = 56

Leaf: *Machine A*	*Stem*	*Leaf:* *Machine B*
4	4	
9 9 8 7 7 6 6 5	4*	5 7 8 9 9 9 9 9 9 9
4 3 2 2 2 1 1 1 0 0 0 0 0 0	5	0 0 0 0 0 1 1 1 1 1 2 2 3
5 5	5*	9

a Display the data from both machines on parallel box-and-whisker plots.
b Calculate the mean and standard deviation of the number of Smarties distributed from both machines.
c Which machine is the more dependable? Justify your answer.

14 Year 10 students at Merrigong High School sit exams in Science and Maths. The results are shown in the table below.

Mark	Number of students in Science	Number of students in Maths
51–60	7	6
61–70	10	7
71–80	8	12
81–90	8	9
91–100	2	6

a Is either distribution symmetrical?
b If either distribution is not symmetrical, state whether it is positively or negatively skewed.
c Discuss the possible reasons for any skewness.
d State the modal class of each distribution.
e In which subject is the standard deviation greater? Explain your answer.

15 Draw an example of a graph that is:
a symmetrical
b positively skewed with one mode
c negatively skewed with two modes.

16 A new drug for the relief of cold symptoms has been developed. To test the drug, 40 people were exposed to a cold virus. Twenty patients were then given a dose of the drug while another 20 patients were given a placebo. (In medical tests a control group is often given a *placebo* drug. The subjects in this group believe that they have been given the real drug but in fact their dose contains no drug at all.) All participants were then asked to indicate the time when they first felt relief of symptoms. The number of hours from the time the dose was administered to the time when the patients first felt relief of symptoms are detailed below.

Group A (drug)

25	29	32	45	18	21	37	42	62	13
42	38	44	42	35	47	62	17	34	32

Group B (placebo)

25	17	35	42	35	28	20	32	38	35
34	32	25	18	22	28	21	24	32	36

a Detail the data on a back-to-back stem-and-leaf plot.
b Display the data for both groups on a box-and-whisker plot.
c Make comparisons of the data. Use statistics in your answer.

d Does the drug work? Justify your answer.
e What other considerations should be taken into account when trying to draw conclusions from an experiment of this type?

PROBLEM SOLVING

17 The heights of Year 10 and Year 12 students (to the nearest centimetre) are being investigated. The results of some sample data are shown below.

Year 10	160	154	157	170	167	164	172	158	177	180	175	168	159	155	163	163	169	173	172	170
Year 12	160	172	185	163	177	190	183	181	176	188	168	167	166	177	173	172	179	175	174	180

a Draw a back-to-back stem-and-leaf plot.
b Draw a parallel boxplot.
c Comment on what the plots tell you about the heights of Year 10 and Year 12 students.

18 Kloe compares her English and Maths marks. The results of eight tests in each subject are shown below.

English: 76, 64, 90, 67, 83, 60, 85, 37
Maths: 80, 56, 92, 84, 65, 58, 55, 62

a Calculate Kloe's mean mark in each subject.
b Calculate the range of marks in each subject.
c Calculate the standard deviation of marks in each subject.
d Based on the above data, in which subject would you say that Kloe has performed more consistently?

CHALLENGE 12.2

A sample of 50 students was surveyed on whether they owned an iPod or a mobile phone. The results showed that 38 per cent of the students owned both. Sixty per cent of the students owned a mobile phone and there were four students who had an iPod only. What percentage of students did not own a mobile phone or an iPod?

12.7 Review

eBookplus ONLINE ONLY

www.jacplus.com.au

The Maths Quest Review is available in a customisable format for students to demonstrate their knowledge of this topic.

The Review contains:

- **Fluency** questions — allowing students to demonstrate the skills they have developed to efficiently answer questions using the most appropriate methods
- **Problem Solving** questions — allowing students to demonstrate their ability to make smart choices, to model and investigate problems, and to communicate solutions effectively.

A summary of the key points covered and a concept map summary of this topic are available as digital documents.

Review questions

Download the Review questions document from the links found in your eBookPLUS.

eBookplus

Interactivities

Word search
int-2859

Crossword
int-2860

Sudoku
int-3599

Language

It is important to learn and be able to use correct mathematical language in order to communicate effectively. Create a summary of the topic using the key terms below. You can present your summary in writing or using a concept map, a poster or technology.

box-and-whisker plot
cumulative frequency curve
data
data sets
dot plot
extreme values
five-number summary
frequency
frequency distribution
frequency distribution table
grouped data
histogram
interquartile range
interval
mean
measures of central tendency
median
modal class
mode
negatively skewed
ogive
outlier
percentile
positively skewed
quartile
range
score
skewed
skewness
spread
stem-and-leaf plot
symmetrical

Link to assessON for questions to test your readiness **FOR** learning, your progress **AS** you learn and your levels **OF** achievement.

assessON

assessON provides sets of questions for every topic in your course, as well as giving instant feedback and worked solutions to help improve your mathematical skills.

www.assesson.com.au

The story of mathematics is an exclusive Jacaranda video series that explores the history of mathematics and how it helped shape the world we live in today.

Koby's bid to make the Olympic athletics team (eles-1852) explores the use of data analysis to provide valuable information to Koby's coach. When Koby's data are properly analysed, the coach can make conclusions about his performance and suggest ways in which Koby can improve.

RICH TASK

Cricket scores

Data are used to predict, analyse, compare and measure many aspects of the game of cricket. Attendance is tallied at every match. Players' scores are analysed to see if they should be kept on the team. Comparisons of bowling and batting averages are used to select winners for awards. Runs made, wickets taken, no-balls bowled, the number of ducks scored in a game as well as the number of 4s and 6s are all counted and analysed after the game. Data of all sorts are gathered and recorded, and measures of central tendency and spread are then calculated and interpreted.

Sets of data have been made available for you to analyse, and decisions based on the resultant measures can be made.

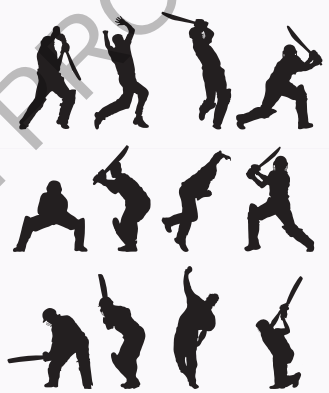
Batting averages

The following table shows the runs scored by four cricketers who are vying for selection to the state team.

Player	Runs in the last 25 matches	Mean	Median	Range	IQR
Allan	13, 18, 23, 21, 9, 12, 31, 21, 20, 18, 14, 16, 28, 17, 10, 14, 9, 23, 12, 24, 0, 18, 14, 14, 20				
Shane	2, 0, 112, 11, 0, 0, 8, 0, 10, 0, 56, 4, 8, 164, 6, 12, 2, 0, 5, 0, 0, 0, 8, 18, 0				
Glenn	12, 0, 45, 23, 0, 8, 21, 32, 6, 0, 8, 14, 1, 27, 23, 43, 7, 45, 2, 32, 0, 6, 11, 21, 32				
Rod	2, 0, 3, 12, 0, 2, 5, 8, 42, 0, 12, 8, 9, 17, 31, 28, 21, 42, 31, 24, 30, 22, 18, 20, 31				

1 Find the mean, median, range and IQR scored for each cricketer.
2 You need to recommend the selection of two of the four cricketers. For each player, write two points as to why you would or would not select them. Use statistics in your comments.
   a Allan _____
   _____
   b Shane _____
   _____
   c Glenn _____
   _____
   d Rod _____
   _____

Bowling averages



The bowling average is the number of runs per wicket taken.

$$\text{Bowling average} = \frac{\text{no. of runs scored}}{\text{no. of wickets taken}}$$

The smaller the average, the better the bowler has performed.
Brad and Dennis were competing for three bowling awards:
- Best in semifinal
- Best in final
- Best overall

The following table gives their scores.

	Semifinal		Final	
	Runs scored	Wickets taken	Runs scored	Wickets taken
Brad	12	5	28	6
Dennis	10	4	15	3

3 Calculate the bowling averages for the following and fill in the table below.
   a Semifinal
   b Final
   c Overall

	Semifinal average	Final average	Overall average
Brad			
Dennis			

4 Explain how Dennis can have the better overall average when Brad has the better average in both the semifinal and final.

CODE PUZZLE

Medical discovery of 1928

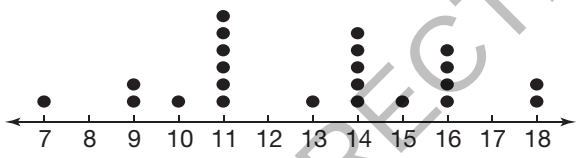
Using technology where appropriate, calculate the required measures of central tendency for each of the following data sets to find the puzzle's code.

x	24	25	26	27	28
f	10	9	8	1	0

A = mean = _____ B = mode = _____

Stem	Leaf
1	5 3 0 1
2	3 0 3 1
3	1 0
4	5

F = mean = _____
G = median = _____
H = mode = _____



M = mean = ___ N = median = ___ O = mode = ___

12	16	19
20	17	15
13	32	16
18	17	40
19	15	16

P = mean = _____
R = median = _____
S = mode = _____

9	9	9	10
10	15	16	18

C = mean = _____
D = median = _____
E = mode = _____

x	*f*
7	4
8	3
9	2
10	1
11	0

I = mean = _____ L = mode = _____

Stem	Leaf
1	5 5 5 6 7 8
2	0 0 1 3
3	
4	0

T = mean = _____
V = median = _____
X = mode = _____

25	7	9	15	25	14	10	9	17	22	7	9	13	8	14	21
10	8	16	12	11	18	9	17	16	19	9	14	8	12	8	7	7	8	14	–
20	23	9	22	8	17	16	20	25	14	20	8	24	8	11	20	8	12		

eBookplus Activities

12.1 Overview

Video

- The story of mathematics (eles-1852)

12.2 Measures of central tendency

Interactivity

- IP interactivity 12.2 (int-4621): Measures of central tendency

Digital docs

- SkillSHEET (doc-5299): Finding the mean of a small data set
- SkillSHEET (doc-5300): Finding the median of a small data set
- SkillSHEET (doc-5301): Finding the mode of a small data set
- SkillSHEET (doc-5302): Finding the mean, median and mode from a stem-and-leaf plot
- SkillSHEET (doc-5303): Presenting data in a frequency distribution table
- SkillSHEET (doc-5304): Drawing statistical graphs

12.3 Measures of spread

Interactivity

- IP interactivity 12.3 (int-4622): Measures of spread

Digital doc

- WorkSHEET 12.1 (doc-14595): Univariate data I

12.4 Box-and-whisker plots

Interactivity

- IP interactivity 12.4 (int-4623): Box-and-whisker plots

12.5 The standard deviation

Interactivity

- IP interactivity 12.5 (int-4624): The standard deviation

Digital doc

- WorkSHEET 12.2 (doc-14596): Univariate data II

12.6 Comparing data sets

Interactivities

- Parallel boxplots (int-2788)
- IP interactivity 12.6 (int-4625): Comparing data sets

12.7 Review

Interactivities

- Word search (int-2859)
- Crossword (int-2860)
- Sudoku (int-3599)

Digital docs

- Topic summary (doc-14597)
- Concept map (doc-14598)

To access eBookPLUS activities, log on to www.jacplus.com.au

Answers

TOPIC 12 Univariate data

Exercise 12.2 — Measures of central tendency

1 a i 7 **ii** 8 **iii** 8
b i 6.875 **ii** 7 **iii** 4, 7
c i 39.125 **ii** 44.5 **iii** No mode
d i 4.857 **ii** 4.8 **iii** 4.8
e i 12 **ii** 12.625 **iii** 13.5

2 Science: mean = 57.6, median = 57, mode = 42, 51
Maths: mean = 69.12, median = 73, mode = 84

3 a i 5.83 **ii** 6 **iii** 6
b i 14.425 **ii** 15 **iii** 15

4 a Mean = 2.5, median = 2.5
b Mean = 4.09, median = 3
c Median

5 a $72\frac{2}{3}$ **b** 73 **c** 70–<80

6 124.83

7 65–<70

8 a B **b** B **c** C **d** D

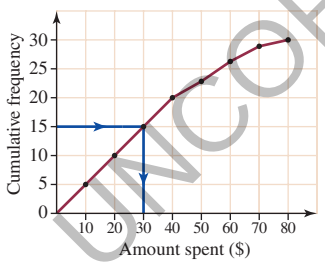
9 a Mean = $32.93, median = $30

b

Class interval	Frequency	Cumulative frequency
0–9	5	5
10–19	5	10
20–29	5	15
30–39	3	18
40–49	5	23
50–59	3	26
60–69	3	29
70–79	1	30
Total	30	

Mean = $32.50, median = $30

c



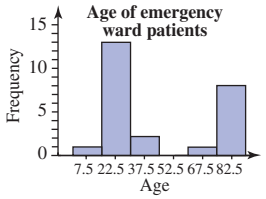
d The mean is slightly underestimated; the median is exact. The estimate is good enough as it provides a guide only to the amount that may be spent by future customers.

10 a 3
b 4, 5, 5, 5, 6 (one possible solution)
c One possible solution is to exchange 15 with 20.

11 a Frequency column: 16, 6, 4, 2, 1, 1
b 6.8
c 0–4 hours
d 0–4 hours

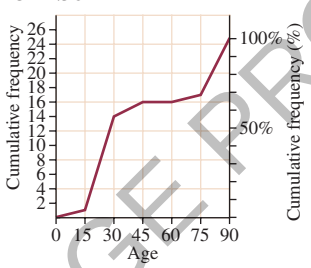
12 a Frequency column: 1, 13, 2, 0, 1, 8

b



c Asymmetrical or bimodal (as if the data come from two separate graphs).
d 44.1
e 15–<30
f 15–<30

g

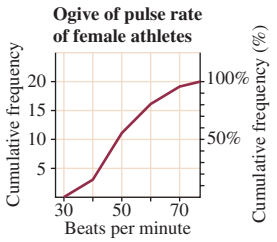


h 28
i No
j Class discussion

13 a Player A mean = 34.33, Player B mean = 41.83
b Player B
c Player A median = 32.5, Player B median = 0
d Player A
e Player A is more consistent. One large score can distort the mean.

14 a Frequency column: 3, 8, 5, 3, 1
b 50.5 **c** 40–<50 **d** 40–<50

e



f Approximately 48 beats/min

15 A

16 Check with your teacher.

17 Answers will vary. Examples given.
a 3, 4, 5, 5, 8 **b** 4, 4, 5, 10, 16 **c** 2, 3, 6, 6, 12

18 12

19 $\frac{2a + b}{3}$

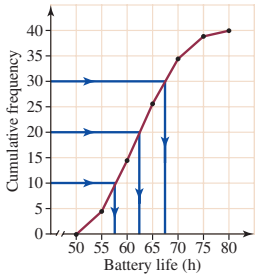
Challenge 12.1

13, 31, 31, 47, 53, 59

Exercise 12.3 — Measures of spread

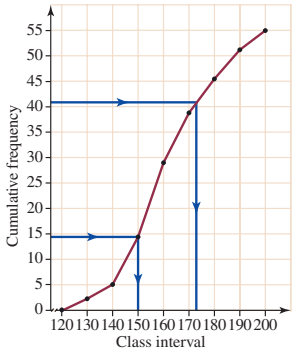
1 a 15 **b** 77.1 **c** 9
2 a 7 **b** 7 **c** 8.5 **d** 39
3 a 3.3 kg **b** 1.5 kg
4 22 cm
5 0.8
6 C

7 a



b i 62.5
ii $Q_1 = 58$, $Q_3 = 67$
iii 9
iv 14
v 6

8



IQR = 27

9 a i Range = 23 **ii** IQR = 13.5
b i Range = 45 **ii** IQR = 27.5
c i Range = 49 **ii** IQR = 20

10 a 25.5 **b** 28 **c** 39 **d** 6
e The three lower scores affect the mean but not the median or mode.

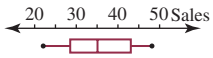
11 a Men: mean = 32.3; median = 32.5; range = 38; IQR = 14
Women: mean = 29.13; median = 27.5; range = 36; IQR = 13
b Typically, women marry younger than men, although the spread of ages is similar.

12 Mean = 25.036, median = 24.9, mode = 23.6, range = 8.5, IQR = 3.4

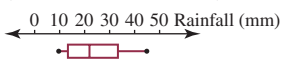
13 $a = 22$, $b = 9$, $c = 9$ and $d = 8$

Exercise 12.4 — Box-and-whisker plots

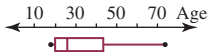
1 a 5 **b** 26
2 a 6 **b** 27
3 a 5.8 **b** 18.6
4 a 140 **b** 56 **c** 90 **d** 84 **e** 26
5 a 58 **b** 31 **c** 43 **d** 27 **e** 7
6 B
7 C
8 D
9 a (22, 28, 35, 43, 48)
b



10 a (10, 13.5, 22, 33.5, 45)
b

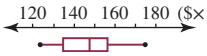


11 a (18, 20, 26, 43.5, 74)
b

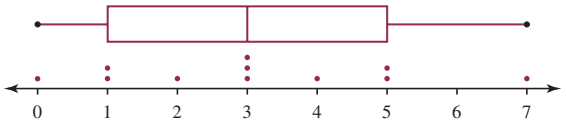


c The distribution is positively skewed, with most of the offenders being young drivers.

12 a (124 000, 135 000, 148 000, 157 000, 175 000)
b

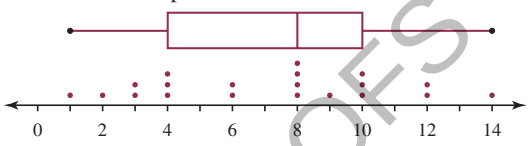


13 a



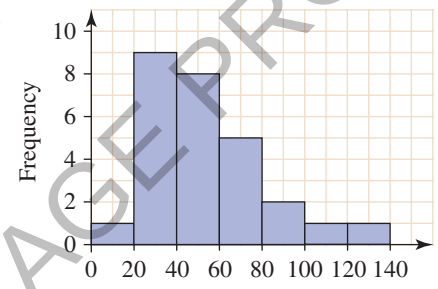
Both graphs indicate that the data is slightly positively skewed. However, the boxplot provides an excellent summary of the centre and spread of the distribution.

b



Both graphs indicate that the data is slightly negatively skewed. However, the boxplot provides an excellent summary of the centre and spread of the distribution.

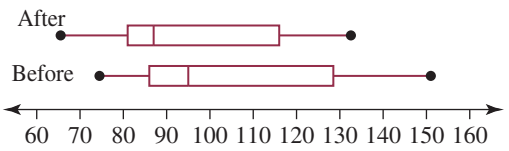
14



15 a

	X_{min}	Q_1	Median	Q_3	X_{max}
Before	75	86	95	128.5	152
After	66	81	87	116	134

b

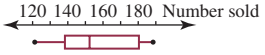


c As a whole, the program was effective. The median weight dropped from 95 kg to 87 kg, a loss of 8 kg. A noticeable shift in the graph shows that after the program 50% of participants weighed between 66 and 87 kg, compared to 25% of participants weighing between 75 and 86 kg before they started. Before the program the range of weights was 77 kg (from 75 kg to 152 kg); after the program the range had decreased to 68 kg. The IQR also diminished from 42.5 kg to 35 kg.

16 a Key: 12 | 1 = 121

Stem	*Leaf*
12	1 5 6 9
13	1 2 4
14	3 4 8 8
15	0 2 2 2 5 7
16	3 5
17	2 9
18	1 1 1 2 3 7 8

b

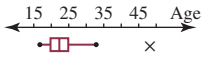


c On most days the hamburger sales are less than 160. Over the weekend the sales figures spike beyond this.

17 a Key: 1* | 7 = 17 years

Stem	*Leaf*
1*	7 7 8 8 8 9 9
2	0 0 0 1 2 2 2 2 3 3 3 3 4 4 4
2*	5 5 8 9
3	1 2 3
3*	
4	
4*	8

b



c The distribution is positively skewed, with first-time mothers being under the age of 30. There is one outlier (48) in this group.

18 C

19 a HJ Looker: median = 5;
Hane and Roarne: median = 6

b HJ Looker

c HJ Looker

d Hane and Roarne had a higher median and a lower spread and so they appear to have performed better.

20 a $50 **b** $135 **c** $100
d $45 **e** 50%

Exercise 12.5 — The standard deviation

1 a 2.29 **b** 2.19 **c** 20.17 **d** 3.07

2 a 1.03 **b** 1.33 **c** 2.67 **d** 2.22

3 10.82

4 0.45%

5 0.06 m

6 0.49 s

7 15.10 calls

8 B

9 The mean of the first 25 cars is 89.24 km/h with a standard deviation of 5.60. The mean of the first 26 cars is 91.19 with a standard deviation of 11.20, indicating that the extreme speed of 140 km/h is an anomaly.

10 a $\sigma \approx 3.58$

b The mean is increased by 7 but the standard deviation remains at $\sigma \approx 3.58$.

c The mean is tripled and the standard deviation is tripled to $\sigma \approx 10.74$.

11 The standard deviation will decrease because the average distance to the mean has decreased.

12 57 is two standard deviations above the mean.

13 a New mean is the old mean increased by 3 but no change to the standard deviation.

b New mean is 3 times the old mean and new standard deviation is 3 times the old standard deviation.

14 a 43 **b** 12 **c** 12.19

Exercise 12.6 — Comparing data sets

1 a The mean of the first set is 63. The mean of the second set is 63.

b The standard deviation of the first set is 2.38. The standard deviation of the second set is 6.53.

c For both sets of data the mean is the same, 63. However, the standard deviation for the second set (6.53) is much higher than the standard deviation of the first set (2.38), implying that the second set is more widely distributed than the first. This is confirmed by the range, which is $67 - 60 = 7$ for the first set and $72 - 55 = 17$ for the second.

2 a Morning: median = 2.45; afternoon: median = 1.6

b Morning: range = 3.8; afternoon: range = 5

c The waiting time is generally shorter in the afternoon. One outlier in the afternoon data causes the range to be larger. Otherwise the afternoon data are far less spread out.

3 Key: 16 | 1 = 1.61 m

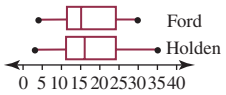
Leaf: Boys	*Stem*	*Leaf: Girls*
9 9 7	15	1 2 5 6 7 8 8
9 8 6 6 5 5 4 0	16	4 4 6 7 8 9 9
4 4 2 1	17	0

4 a Ford: median = 15; Holden: median = 16

b Ford: range = 26; Holden: range = 32

c Ford: QR = 14; Holden: IQR = 13.5

d

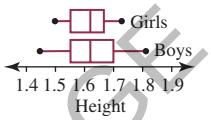


5 a Brisbane Lions

b Brisbane Lions: range = 63;
Sydney Swans: range = 55

c Brisbane Lions: IQR = 40;
Sydney Swans: IQR = 35

6 a



b Boys: median = 1.62; girls: median = 1.62

c Boys: range = 0.36; girls: range = 0.23

d Boys: IQR = 0.14; girls: IQR = 0.11

e Although boys and girls have the same median height, the spread of heights is greater among boys as shown by the greater range and interquartile range.

7 a Summer: range = 23; winter: range = 31

b Summer: IQR = 14; winter: IQR = 11

c There are generally more cold drinks sold in summer as shown by the higher median. The spread of data is similar as shown by the IQR although the range in winter is greater.

8 A

9 A, B, C, D

10 a Cory achieved a better average mark in Science (59.25) than he did in English (58.125).

b Cory was more consistent in English ($\sigma = 4.9$) than he was in Science ($\sigma = 19.7$)

11 a Back street: $\bar{x} = 61$, $\sigma = 4.3$;
main road: $\bar{x} = 58.8$, $\sigma = 12.1$

b The drivers are generally driving faster on the back street.

c The spread of speeds is greater on the main road as indicated by the higher standard deviation.

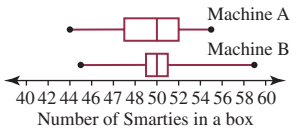
12 a Nathan: mean = 15.1; Timana: mean = 12.3

b Nathan: range = 36; Timana: range = 14

c Nathan: IQR = 15; Timana: IQR = 4

d Timana's lower range and IQR shows that he is the more consistent player.

13 a



b Machine A: mean = 49.88,
standard deviation = 2.87;
Machine B: mean = 50.12,
standard deviation = 2.44

c Machine B is more reliable, as shown by the lower standard deviation and IQR. The range is greater on machine B only because of a single outlier.

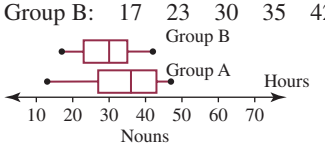
14 a Yes — Maths
b Science: positively skewed
c The Science test may have been more difficult.
d Science: 61–70, Maths: 71–80
e Maths has a greater standard deviation (12.6) compared to Science (11.9).

15 Answers will vary. Check with your teacher.

16 a Key: 2 | 3 = 2.3 hours

Leaf: *Group A*	*Stem*	*Leaf:* *Group B*
8 7 3	1	7 8
9 5 1	2	0 1 2 4 5 5 8 8
8 7 5 4 2 2	3	2 2 2 4 5 5 5 6 8
7 5 4 2 2 2	4	2
	5	
2 2	6	

b Five-point summary
Group A: 13 27 36 43 62
Group B: 17 23 30 35 42



c Student comparison

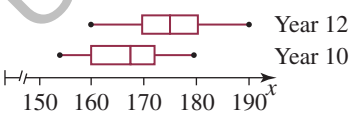
Statistics	Group A	Group B
Five-point summary	13 27 36 43 62	17 23 30 35 42
x	35.85 hours	28.95 hours
Range	49 hours	25 hours
IQR	16 hours	12 hours
σ	13 hours	7 hours

d Student decision, justifying answer
e Class discussion

17 a

Leaf: *Year 10*	*Stem*	*Leaf:* *Year 11*
9 8 7 5 4	15	
9 8 7 4 3 3 0	16	0 3 6 7 8
7 5 3 2 2 0 0	17	2 2 3 4 5 6 7 7 9
0	18	0 1 3 5 8
	19	0

b



c On average, the Year 12 students are about 6–10 cm taller than the Year 10 students. The heights of the majority of Year 12 students are between 170 cm and 180 cm, whereas the majority of the Year 10 students are between 160 and 172 cm in height.

18 a English: mean = 70.25; Maths: mean = 69
b English: range = 53; Maths: range = 37
c English: $\sigma = 16.1$; Maths: $\sigma = 13.4$
d Kloe has performed more consistently in Maths as the range and standard deviation are both lower.

Challenge 12.2

32%

Investigation — Rich task

1

Player	Runs in the last 25 matches	Mean	Median	Range	IQR
Allan	13, 18, 23, 21, 9, 12, 31, 21, 20, 18, 14, 16, 28, 17, 10, 14, 9, 23, 12, 24, 0, 18, 14, 14, 20	16.76	17	31	8.5
Shane	2, 0, 112, 11, 0, 0, 8, 0, 10, 0, 56, 4, 8, 164, 6, 12, 2, 0, 5, 0, 0, 0, 8, 18, 0	17.04	4	164	10.5
Glenn	12, 0, 45, 23, 0, 8, 21, 32, 6, 0, 8, 14, 1, 27, 23, 43, 7, 45, 2, 32, 0, 6, 11, 21, 32	16.76	12	45	25.5
Rod	2, 0, 3, 12, 0, 2, 5, 8, 42, 0, 12, 8, 9, 17, 31, 28, 21, 42, 31, 24, 30, 22, 18, 20, 31	16.72	17	42	25

2 a Allan: has a similar mean and median, which shows he was fairly consistent. The range and IQR values are low indicating that his scores remain at the lower end with not much deviation for the middle 50%.
b Shane: has the best average but a very low median indicating his scores are not consistent. The range is extremely high and the IQR very low in comparison showing he can score very well at times but is not a consistent scorer.
c Glenn: has a similar mean to Allan and Rod but a lower median, indicating his scores are sometimes high but generally are lower than the average. The range and IQR show a consistent batting average and spread with only a few higher scores and some lower ones.
d Rod: has a similar mean and median which shows he was a consistent player. The range and IQR show a consistent batting average and spread.

Players to be selected:
Would recommend **Allan** if the team needs someone with very consistent batting scores every game but no outstanding runs.

Would recommend **Shane** if the team needs someone who might score very high occasionally but in general fails to score many runs.

Would recommend **Glenn** if the team needs someone who is fairly consistent but can score quite well at times and the rest of the time does OK.

Would recommend **Rod** if the team needs someone who is fairly consistent but can score quite well at times and the rest of the time has a better median than Glenn.

3

	Semifinal average	Final average	Overall average
Brad	2.4	4.67	3.64
Dennis	2.5	5	3.57

a Brad was best in the semifinal.
b Brad was best in the final.
c Dennis was best overall.

4 In the final, wickets were more costly than in the semifinal. Brad therefore conceded many runs in getting his six wickets. This affected the overall mean. In reality Brad was the most valuable player overall, but this method of combining the data of the two matches led to this unexpected result.

Code puzzle

Alexander Fleming discovers penicillin — the first antibiotic.