8

Box plots

This chapter deals with the construction and interpretation of box plots.

At the end of this chapter you should be able to:

- find the upper and lower extremes, the median, and the upper and lower quartiles for sets of numerical data
- calculate the range and interquartile range
- compare the relative merits of range and interquartile range as measures of spread
- construct a box plot using the median, upper and lower quartiles, and upper and lower extremes of a set of data
- compare two or more sets of data using parallel box plots
- determine quartiles from data displayed in histograms and dot plots and use these to draw box plots
- identify skewed and symmetrical sets of data displayed in histograms and dot plots
- evaluate survey data reported in the digital media and elsewhere.

NSW Syllabus references: 5.2 S&P Single variable data analysis
Outcomes: MA5.2-1WM, MA5.2-3WM, MA5.2-15SP
Statistics & probability – ACMSP227, ACMSP248, ACMSP249, ACMSP250

Diagnostic test

Use the table below to answer questions **1** to **4**.

Score	5	6	7	8	9	10
Frequency	4	7	12	17	11	5

1 The mean of the data in the table is closest to:
A 7 **B** 7.7 **C** 8 **D** 7.5

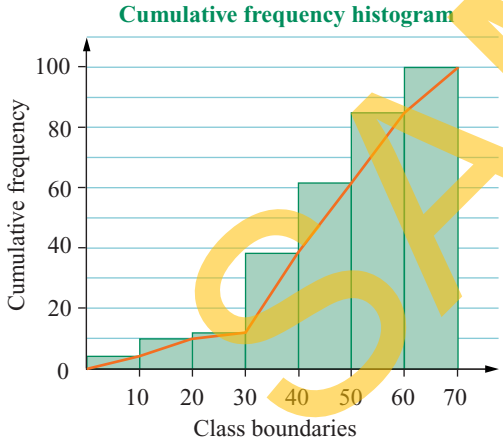
2 The median of the data in the table is closest to:
A 17 **B** 7.7 **C** 8 **D** 7.5

3 The mode of the data in the table is closest to:
A 17 **B** 7.7 **C** 8 **D** 7.5

4 The cumulative frequency for the score of 8 is:
A 40 **B** 17 **C** 8 **D** 23

5 The ogive is the:
A frequency polygon
B frequency histogram
C cumulative frequency polygon
D cumulative frequency histogram

Use this cumulative frequency histogram and polygon to answer questions **6** and **7**.



6 The cumulative frequency for the 40–49 class is:
A 61 **B** 44.5 **C** 26
D It is impossible to determine without the exact scores.

7 An estimate for the median is closest to:
A 40–49 class **B** 50 **C** 45
D It is impossible to determine without the exact scores.

8 Which frequency distribution table represents the following scores?
12, 30, 38, 49, 13, 28, 33, 17, 21, 31, 23, 32, 25, 26, 39, 36, 42, 46, 36, 50, 48, 32, 45, 57, 43, 51, 49, 53, 42, 33

A

x	*f*
10–19	3
20–29	6
30–39	9
40–49	8
50–59	4

B

x	*f*
10–19	3
20–29	5
30–39	10
40–49	8
50–59	4

C

x	*f*
10–19	4
20–29	5
30–39	9
40–49	9
50–59	3

D

x	*f*
10–19	4
20–29	4
30–39	10
40–49	7
50–59	5

9 The class centre for the 32–38 class is:
A 32 **B** 38 **C** 70 **D** 35

Use this table to answer questions **10** to **12**.

Class	Class centre	Frequency	*fx*
10–16		3	
17–23		15	
24–30		8	
31–37		12	
38–44		5	

10 The class centre for the 38–44 class is:
A 84 **B** 41 **C** 5 **D** 10–44

11 The mean for the data is closest to:
A 27 **B** 10 **C** 8.6
D It is impossible to determine without the exact scores.

12 The modal class is:
A 24–30 **B** 17–23
C 24–30 **D** 31–37

The diagnostic test questions refer to outcomes ACMSP170 and ACMSP171.

A Mean, mode, median and range

This section reviews the three measures of central tendency, mean, mode and median, and the measure of spread or range.

Mean

The **mean** is the statistical term most thought of when the word 'average' is used. The mean of a set of scores is calculated by adding all the scores and dividing this sum by the number of scores. $\bar{x}$ is the symbol used to represent the mean.

For example, for the scores 3, 7, 8, 9 and 9:

$$\bar{x} = \frac{3 + 7 + 8 + 9 + 9}{5} = \frac{36}{5} = 7.2$$

Mode

The **mode** is the score that occurs most often; that is, it is the score with the highest frequency. It is the most commonly occurring score. For example, for the scores 3, 7, 8, 9 and 9, the mode is 9 (as it occurs more frequently that any other score).

A set of scores may be **bimodal**; that is, have two modes. For example, 2, 3, 3, 3, 4, 4, 4, 5, 6, 8 is bimodal as it has two modes; namely, 3 and 4.

Median

The **median** of a set of scores is the middle score (or the average of the two middle scores) after the scores have been arranged in ascending order (that is from smallest to largest).

- For an *odd number* of scores, there is one middle score.
  If there are n scores in ascending order, the median is the value of the score in the $\left(\frac{n+1}{2}\right)$th position.
  For example, for the scores 7, 9, 3, 9, 8, the ascending order is 3, 7, 8, 9, 9.
  $$n = 5, \text{ so } \frac{n+1}{2} = 3$$
  The median is the 3rd score; that is, the median is 8.

  There are two numbers below the median and two numbers above the median.

- For an *even number* of scores, there are two middle scores so the median is not always one of the scores.
  For example, consider the scores 4, 7, 9, 6, 5, 9, 3, 7.
  The ascending order is

3	4	5	6	7	7	9	9
1st	2nd	3rd	4th	5th	6th	7th	8th

  As $n = 8$, the median is the number midway between the 4th and 5th scores.
  The median $= \frac{6+7}{2}$ (the average of the 4th and 5th scores) $= 6.5$.

  There are four scores below the median and four scores above the median.

Range

The range of a set of data is a measure of its spread. It is found by subtracting the lowest score from the highest score.

$$\text{Range} = \text{highest score} - \text{lowest score}$$

Consider these two data sets:

Set A: 3, 5, 7, 9, 11

Set B: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

Both sets have a mean of 7. However, set B is obviously more spread out and the range highlights this fact.

For set A: range $= 11 - 3 = 8$

For set B: range $= 13 - 1 = 12$

EXAMPLE 1

For the scores 12, 13, 14, 14, 14, 15, 16, find the:

a mean **b** mode **c** median **d** range.

	Solve	Think	Apply
a	Mean = 14	Mean $= \frac{12 + 13 + 14 + 14 + 14 + 15 + 16}{7} = \frac{98}{7}$ $\bar{x} = 14$	Use the definitions for mean, mode, median and range to find these measures.
b	Mode = 14	As 14 occurs more frequently than any other score, the mode = 14.	
c	Median = 14	To find the median, arrange the scores in ascending order. Find the middle score by crossing out equal numbers of scores from each end. ~~12~~ ~~13~~ ~~14~~ (14) ~~14~~ ~~15~~ ~~16~~ Median = 14	
d	Range = 4	Range = highest score − lowest score = 16 − 12 = 4	

EXAMPLE 2

For the scores 110, 106, 114, 109, 114, 107, find the:

a mean **b** mode **c** median **d** range.

	Solve	Think	Apply
a	Mean = 110	Mean $= \frac{110 + 106 + 114 + 109 + 114 + 107}{6} = \frac{660}{6}$ $\bar{x} = 110$	Use the definitions for mean, mode, median and range to find these measures.
b	Mode = 114	As 114 occurs more frequently than any other score, the mode = 114.	
c	Median = 109.5	To find the median, first arrange the scores in ascending order. Cross out scores from each end until there are two numbers left. The median is the average of them. ~~106~~ ~~107~~ (109) (110) ~~114~~ ~~114~~ Median $= \frac{109 + 110}{2} = 109.5$	
d	Range = 8	Range = 114 − 106 = 8	

Exercise 8A

1 Find the mean of the following sets of data. Answer to 1 decimal place if necessary.

a 1, 3, 4, 5, 8, 8, 9 **b** 1, 2, 2, 3, 4, 5, 6, 8, 9 **c** 10, 12, 13, 15, 16

d 20, 20, 20, 23, 25, 27 **e** 51, 52, 54, 55, 57, 57, 58, 59 **f** 0, 0, 1, 3, 3, 3, 5, 6, 6, 7, 8, 9

g 105, 101, 104, 101, 101, 102 **h** 3, 4, 9, 5, 1, 8, 3, 2, 0 **i** 6, 4, 5, 8, 9, 3, 5, 5, 4, 9

2 Find the mode, if there is one, of the scores in question **1**.

A set of scores with two modes is called bimodal.

3 Find the median of the scores in question **1**.

4 Find the range of the scores in question **1**.

5 The number of lollies in 10 packets were 15, 18, 17, 15, 16, 14, 15, 18, 16, 19. Find the:

a range
b mean number of lollies per packet
c modal number of lollies per packet
d median number of lollies per packet.

6 **a** **i** Find the mean of the first eight counting numbers: 1, 2, 3, 4, 5, 6, 7, 8.
ii Subtract 3 from each of the first eight counting numbers and then find the mean of these numbers.
iii How has the mean changed?
b **i** Add 20 to each of the first eight counting numbers and then find the mean of these numbers.
ii How has the mean changed?
c **i** Find the median of the first eight counting numbers.
ii Add 3 to each of the first eight counting numbers and then find the median of these numbers.
iii How has the median changed?
d **i** Find the range of the first eight counting numbers.
ii Add 3 to each of the first eight counting numbers and find the range.
iii What has happened to the range?

7 A shoe store had a special on women's running shoes and sold the following sizes: 6, 10, 4, 7, 8, 7, 6, 5, 7, 8, 7, 5, 6, 4, 3, 7.

The average female shoe size is 7. This refers to the mode not the mean.

a Find the:
i range **ii** mean **iii** mode **iv** median.
b Which of the mean, mode and median would be of most use to the shop owner?

8 The Lighthouse Lamp Company has a total of 30 employees whose annual salaries are listed below.

1 general manager	$260 000
1 marketing manager	$100 000
1 accountant	$120 000
1 engineer	$100 000
1 warehouse manager	$100 000
15 production workers	$50 000 each
10 tradespeople	$60 000 each

a What is the total annual wages bill for this company?
b Calculate the mean wage for the employees.
c How many employees earn:
i less than this mean wage?
ii more than this mean wage?
d What is the median wage?
e What is the modal wage?
f In a wage determination case for the employees of this company, which measure of central tendency would you use to support your argument if you were the representative for:
i the general manager? **ii** the production workers?
g Which measure of central tendency is the most appropriate to represent the wages of the employees of this company? Give reasons for your answer.

9 **a** Find the mean and median of these scores:
15, 16, 16, 17, 19, 20, 430.

b Find the mean and median leaving out the score 430.

c The score 430 is called an **outlier** score because it is an extremely large score compared with all the other scores. An outlier can also be a very small score when compared with the other scores. Which measure, mean or median, is most affected by the outlier?

10 A batsman's scores for six innings are 55, 73, 96, 88, 34, 64.

a Find the mean and median.

b In his next innings he scores 0 runs. Find the new mean and median.

c In cricket a score of 0 is called a 'duck'. For this batsman's scores, what statistical name would you give the score of 0?

d Is the mean or median most affected by the score of 0?

EXAMPLE 3

The mean of five scores is 12.2. What is the sum of the scores?

Solve/Think	Apply
Let S = sum of scores $\frac{S}{5} = 12.2$ $S = 12.2 \times 5 = 61$ The sum of the scores is 61.	Substitute the given information into the formula: $\text{Mean} = \frac{\text{sum of scores}}{\text{number of scores}}$ Solve the resulting equation.

11 **a** The mean of eight scores is 7.5. What is the sum of the scores?

b The mean of nine scores is 11.6. What is the sum of the scores?

c While on an outback trip, Bill drove, on average, 262 km per day for a period of 12 days. How far did Bill drive in total while on the trip?

d The mean monthly sales of a clothing store is $15 467. Calculate the total sales of the store for the year.

EXAMPLE 4

Find x if 10, 7, 3, 6 and x have a mean of 8.

Solve/Think	Apply
$\frac{10 + 7 + 3 + 6 + x}{5} = 8$ $\frac{26 + x}{5} = 8$ $26 + x = 40$ $x = 14$	Write an algebraic expression for the sum of the scores and then substitute the given information into the formula. $\text{Mean} = \frac{\text{sum of scores}}{\text{number of scores}}$ Solve the resulting equation.

12 **a** Find x if 8, 11, 5, 7 and x have a mean of 8.

b Find x if 3, 15, 7, 9, 11 and x have a mean of 10.

c Find x if 5, 9, 11, 12, 13, 14, 17 and x have a mean of 12.

d Find a, given that 3, 0, a, a, 4, a, 6, a and 3 have a mean of 4.

e Over the complete assessment period, Jenny averaged 35 out of a possible 40 marks for her eight Mathematics tests. However, when checking her files, she could only find seven of the tests. For these she scored 29, 36, 32, 38, 35, 34 and 39. Can you determine how many marks she scored for the eighth test?

EXAMPLE 5

A cricketer played 12 innings and had a mean of 38.5 runs. He then scored 12 and 71 runs in the next two innings. Find the cricketer's new mean number of runs.

Solve/Think	Apply
Let S = sum of scores $\frac{S}{12} = 38.5$ $S = 462$ New mean $= \frac{462 + 12 + 71}{14}$ $= \frac{545}{14} \approx 38.9$	Find the sum of the scores for the first 12 innings. Use this to determine the sum of the cricketer's 14 scores. Calculate the mean of these 14 scores. *There are 12 + 2 = 14 scores in total.* !

13 **a** A netballer played 14 matches and had a mean of 16.5 goals per game. In the next two matches she threw 21 goals and 24 goals. Calculate her new mean.

b A cricketer played 11 matches and had a mean of 23 runs per game. In the next two games she scored 41 and 35 runs. Calculate her new mean.

c A tennis player averaged 8 aces per match in her first six matches. In the next three matches she served 6, 11 and 13 aces. Calculate her new average.

14 A sample of 12 measurements has a mean of 16.5, and a sample of 15 measurements has a mean of 18.6. What is the mean of all 27 measurements?

15 Fifteen of 31 measurements are below 10 cm and 12 measurements are above 11 cm. Find the median if the other four measurements are 10.1 cm, 10.4 cm, 10.7 cm and 10.9 cm.

16 The mean and the median of a set of nine measurements are both 12. If seven of the measurements are 7, 9, 11, 13, 14, 17 and 19, find the other two measurements.

17 Write your own group of seven scores for which the following measure of central tendency is not appropriate.

a the mean **b** the mode **c** the median

18 Write a group of seven scores for which:

a the mean is less than the median

b the mean is greater than the median

c the mean and the median are equal

d the mean, the mode and the median are equal.

Summary: choosing the appropriate measure

The mean, mode and median can be used to indicate the middle of a set of numbers. Which of these values is the most appropriate measure to use will depend on the type of data under consideration. For example, when reporting on shoe sizes stocked by a shoe store, the average or mean size would be a useless measure of the stock. In this case, the mode would be the most useful measure.

When selecting which of the three measures of central tendency to use as a representative figure for a set of data, you should keep the following advantages and disadvantages of each measure in mind.

Mean

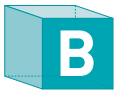
- The mean's main advantage is that it is commonly used, easy to understand and easy to find.
- The mean's main disadvantage is that it is affected by extreme values within a set of data and may give a distorted impression of the data. For example, consider the data set 4, 6, 7, 8, 19, 111. The total of these six numbers is 155, so the mean is approximately 25.8. Is 25.8 a representative figure for the data? The outlier of 111 has distorted the mean in this case.

Mode

- The mode's main advantage is that it is the most usual or typical value within a set of data.
- The mode has an advantage over the mean in that it is not affected by extreme values contained in the data.
- The mode's main disadvantage is that it does not take into account all the values within the data. For example, the mode for the scores 2, 2, 2, 5, 5, 7, 8, 8, 9, 9, 10, 11, 12, 12 is 2. This is not representative of the rest of the data.

Median

- The median's main advantage is that it is easily calculated and is the middle value of the data.
- Unlike the mean, the median is not affected by extreme values.
- The median's main disadvantage is that it ignores all values outside the middle range. For example, the median of the scores 1, 1, 2, 2, 3, 104, 108, 110, 135 is 3, but this is not necessarily representative of the sample.



B Quartiles

The median divides a set of data into two parts with an equal number of scores in each.

In the same way, a set of data can be divided into four parts with an equal number of scores in each. These scores are called quartiles. Each set of data then has three quartiles, the lower, the middle (called the median) and the upper, usually denoted as Q_1, Q_2 and Q_3 respectively.

- The median (Q_2) is the middle score. It divides the data into two equal groups.
- The upper quartile (Q_3) is the middle score of the upper group.
- The lower quartile (Q_1) is the middle score of the lower group.
- The interquartile range (IQR) = upper quartile − lower quartile, or IQR = $Q_3 - Q_1$.

For a large number of scores:

- 25% of the scores $< Q_1$ and 75% of the scores $> Q_1$
- 50% of the scores $< Q_2$ and 50% of the scores $> Q_2$
- 75% of the scores $< Q_3$ and 25% of the scores $> Q_3$.

Hence the interquartile range is a measure of the spread of the middle 50% of the data. It is often a better measure of dispersion (the spread of the scores) than the range because it is not affected by outliers in the data.

EXAMPLE 1

Find the lower, middle and upper quartiles of these scores and then find the interquartile range.

21 22 22 23 25 26 27

Solve/Think	Apply
There are seven scores so the median is the fourth score $\left(\frac{7+1}{2} = 4\right)$. ~~21~~ ~~22~~ ~~22~~ 23 ~~25~~ ~~26~~ ~~27~~ ↑ Q_2 The median is 23. The data is divided into two parts each with three scores. ~~21~~ 22 ~~22~~ 23 ~~25~~ 26 ~~27~~ ↑ ↑ ↑ Q_1 Q_2 Q_3 Cross out in each half to find the quartiles: $Q_1 = 22$ and $Q_3 = 26$. Interquartile range $= Q_3 - Q_1$ $= 26 - 22 = 4$	Divide the scores into two parts with equal numbers of scores in each by finding the median. The median is the middle quartile. The lower quartile is the middle score of the lower group. The upper quartile is the middle score of the upper group.

Exercise 8B

1 For the scores 19, 20, 22, 27, 28, 30, 31, find:

a the median **b** the lower quartile

c the upper quartile **d** the interquartile range.

2 For the scores below, find:

i the median **ii** the lower quartile

iii the upper quartile **iv** the interquartile range.

a 1, 3, 4, 7, 9, 10, 11 **b** 14, 14, 15, 16, 16, 18, 19, 19, 20, 20, 21

EXAMPLE 2

Find the lower, middle and upper quartiles for these scores, and then find the interquartile range.

3 3 4 4 5 5 5 6 8 10

Solve/Think	Apply
There are ten scores, so the median is the average of the 5th and 6th scores. ~~3~~ ~~3~~ ~~4~~ ~~4~~ 5 | 5 ~~5~~ ~~6~~ ~~8~~ ~~10~~ Median is $Q_2 = \frac{5+5}{2} = 5$ The median divides the scores into two equal groups of five scores. ~~3~~ ~~3~~ 4 ~~4~~ 5 | 5 ~~5~~ 6 ~~8~~ ~~10~~ ↑ ↑ ↑ Q_1 Q_2 Q_3 Cross out in each half to find the quartiles: $Q_1 = 4$ and $Q_3 = 6$. Interquartile range $= Q_3 - Q_1$ $= 6 - 4 = 2$	Find the upper and lower quartiles and hence the interquartile range.

3 For the scores below, find:

i the median **ii** the lower quartile

iii the upper quartile **iv** the interquartile range.

a 8, 8, 9, 10, 11, 11, 11, 11, 14, 15

b 15, 17, 20, 22, 22, 24

c 12, 12, 15, 16, 17, 20, 21, 23, 25, 25, 27, 27

d 13, 15, 19, 25, 28, 31, 42, 45

4 Find the interquartile range for the following scores.

a 15, 16, 16, 20, 22, 23, 25

b 11, 13, 13, 14, 14, 15, 18

c 30, 32, 35, 35, 35, 37, 38

d 2, 3, 3, 4, 5, 5, 6, 7

e 50, 50, 52, 55, 55, 57, 57, 58, 60, 60

f 15, 15, 16, 17, 17, 18, 20, 21, 21, 22

g 23, 23, 23, 24, 25, 26, 28, 28, 29, 32

h 33, 35, 38, 42, 43, 44, 52, 53, 55, 58, 61, 64, 66, 67

i 11, 12, 14, 18, 18, 20, 22, 25, 25, 26, 30

j 46, 50, 50, 53, 54, 54, 58, 58, 58, 60, 62, 62, 66, 66

EXAMPLE 3

Find the interquartile range for the following scores.

a 30, 32, 32, 33, 35, 40, 41, 42, 45 **b** 9, 5, 7, 11, 10, 4, 14, 7

	Solve/Think	Apply
a	Find the median by crossing off from each end. ~~30~~ ~~32~~ ~~32~~ ~~33~~ 35 ~~40~~ ~~41~~ ~~42~~ ~~45~~ Median = 35 The middle score in the set 30, 32, 32, 33 is $Q_1 = \frac{32 + 32}{2} = 32$ The middle score in the set 40, 41, 42, 45 is $Q_3 = \frac{41 + 42}{2} = 41.5$ Interquartile range = 41.5 − 32 = 9.5	Find the upper and lower quartiles and hence the interquartile range.
b	Arrange in ascending order to find the median. ~~4~~ ~~5~~ ~~7~~ 7 9 10 11 ~~14~~ Median $= \frac{7 + 9}{2} = 8$ The middle score in the set 4, 5, 7, 7 is $Q_1 = \frac{5 + 7}{2} = 6$ The middle score in the set 9, 10, 11, 14 is $Q_3 = \frac{10 + 11}{2} = 10.5$ Interquartile range = 10.5 − 6 = 4.5	

5 Find the interquartile range for the following scores.

a 42, 45, 45, 48, 53, 61, 64, 68, 71

b 170, 170, 170, 185, 188, 189, 194, 196, 203

c 2, 3, 5, 8, 8, 8, 9, 9, 9, 11, 15, 17, 18

d 15, 17, 20, 23, 28, 35, 42, 44

e 95, 102, 95, 89, 92, 103, 90, 98

f 5, 2, 3, 9, 7, 11, 1, 5, 7, 13

EXAMPLE 4

Find the interquartile range from this stem-and-leaf plot.

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

EXAMPLE 5 CONTINUED

Solve/Think	Apply
There are 23 scores so the median is the 12th score. Alternatively cross off the numbers to find the median. The median is 64. There are 11 scores above the median and 11 scores below the median. The lower quartile is the 6th score; that is, 57. The upper quartile is the 12th + 6 = 18th score; that is, 77. Interquartile range = 77 − 57 = 20	Find the upper and lower quartiles and hence the interquartile range.

6 Find the interquartile range from these stem-and-leaf plots.

a

Stem	Leaf
7	1 1 2 2 3 3 7 8 9
8	2 3 3 6 8 8

b

Stem	Leaf
10	4 5 6 8 9
11	2 2 2 3 5 6 8 8 9
12	0 0 2 2 5 6 7 7
13	1 1 1 6

c

Stem	Leaf
85	5 8 4 5 3 4 2
86	1 2 1 1 1 2 3 4 7 6
87	2 3 4 8 4 5 1 3 6 9 9
88	5 3 6 8 7 4 3 0 3

For part c put in order first.

7 Consider the scores 8, 8, 9, 10, 12, 14, 148.

a Determine:

i the range **ii** the interquartile range.

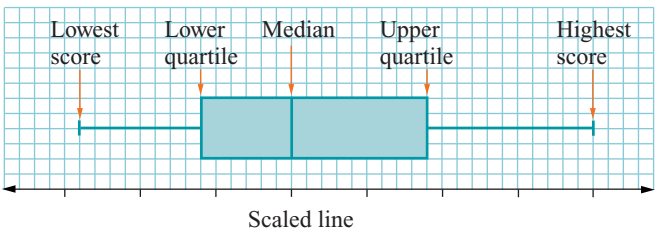
b Explain why the interquartile range is a better measure of spread than the range for this set of data.

C Box plots

A box plot uses five especially selected numbers to display information about numerical scores in a graphical form. The numbers used are the extremes (the highest and lowest scores), the median (the middle score) and the upper and lower quartiles. These five numbers make up the five-number summary.

A box plot is used to show the range and middle half of ranked data. Ranked data is numerical data such as numbers. The middle half of the data is represented by the box. The highest and lowest scores are joined to the box by straight lines. The regions above the upper quartile and below the lower quartile each contain 25% of the data.

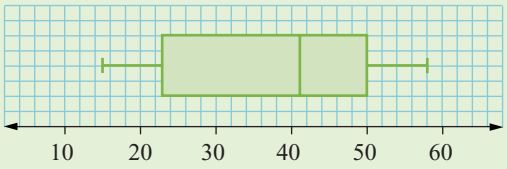
The five-number summary is shown in the diagram.



EXAMPLE 1

From this box plot, find the following.

a i highest score
ii lowest score
iii range of the scores
b median
c i upper quartile ii lower quartile iii interquartile range



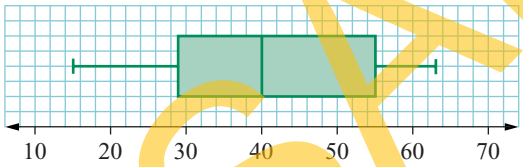
	Solve/Think	Apply
a i	Highest score = 58	Read the values of the quartiles and extremes from the box plot. Calculate the range and interquartile range from these values. *Note:* From these results we can say that: • The bottom 25% of the scores take values from 15 up to, but less than, 23. • The top 25% of the scores take values from, but not including, 50 up to 58. • The middle 50% of the scores lie between 23 and 50. The median is closer to the upper quartile than to the lower quartile, so the top half of the scores are clustered closer to the median than the bottom half.
ii	Lowest score = 15	
iii	Range = 58 − 15 = 43	
b	Median = 41	
c i	$Q_3 = 50$	
ii	$Q_1 = 23$	
iii	Interquartile range = 50 − 23 = 27	

Exercise 8C

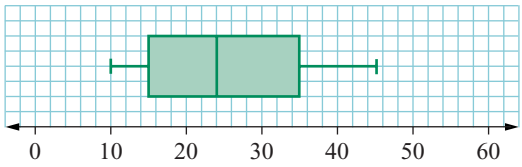
1 From the box plots shown below, find the following.

i highest score ii lowest score iii range of the scores iv median
v upper quartile vi lower quartile vii interquartile range

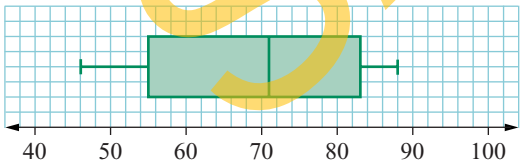
a



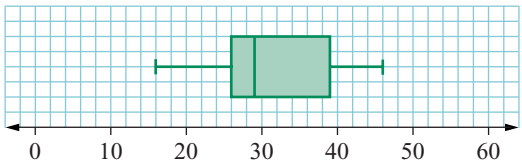
b



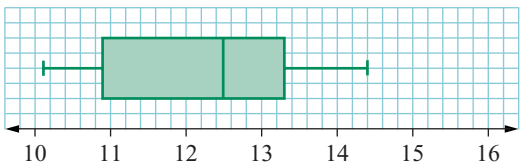
c



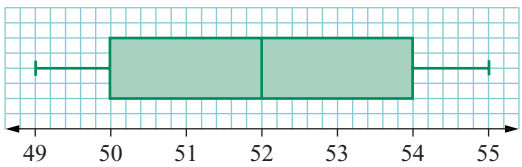
d



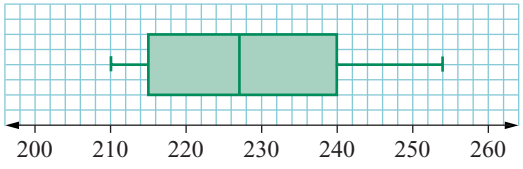
e



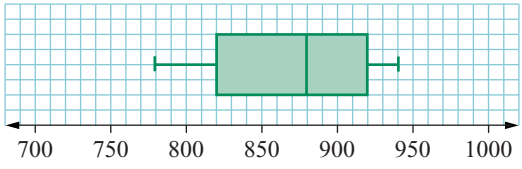
f



g

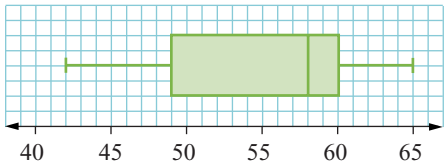


h



EXAMPLE 2

Draw a box plot for ranked data with highest score 65, lowest score 42, median 58, upper quartile 60 and lower quartile 49.

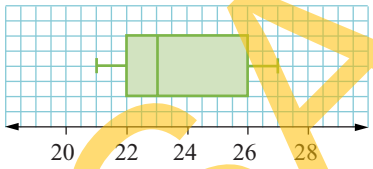
Solve/Think	Apply
	Use the three quartiles to form the box. Plot the extreme values and join them to the box.

2 Draw box plots for ranked data with the following values.

	Highest score	Lowest score	Median	Upper quartile	Lower quartile
a	40	15	28	32	23
b	153	130	141	148	139
c	28	6	10	18	7
d	83	71	78	80	73
e	9	1	5	7	3

EXAMPLE 3

Draw a box plot for the scores 21, 22, 22, 23, 25, 26, 27.

Solve	Think	Apply
	Median is 23. Upper quartile is 26. Lower quartile is 22. Highest score is 27. Lowest score is 21.	Determine the values of the quartiles and extremes. Draw the box plot.

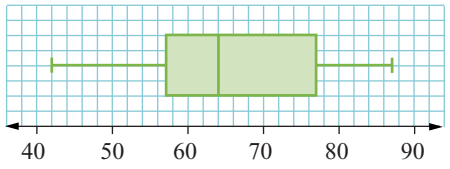
3 Draw box plots for the following scores.

a 34, 35, 36, 36, 37, 38, 39, 39, 39, 40
b 4, 5, 8, 8, 10, 12, 12, 14, 15, 19
c 21, 21, 23, 24, 24, 24, 26, 28, 30
d 89, 90, 92, 95, 95, 98, 102, 103
e 18, 20, 22, 23, 25, 29, 30, 30, 30, 31
f 1, 3, 4, 4, 5, 5, 5, 7, 11, 15

EXAMPLE 4

Construct a box plot for the data in this stem-and-leaf plot.

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

Solve	Think	Apply
	There are 23 leaves and hence 23 scores. So median = 12th score = 64 Q_1 (lower quartile) = 6th score = 57 Q_3 (upper quartile) = 18th score = 77 Highest score = 87, lowest score = 42	Determine the values of the quartiles and extremes. Draw the box plot.

4 Draw box plots for the data in the following stem-and-leaf plots.

a

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

b

Stem	Leaf
10	8 8 9
11	1 1 2 3 3 3 4 4 5 8 8 9
12	0 2 2 3 4 4 4 5 5 8 8 8 9
13	0 0 0 1 1

c

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

d

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

e

Stem	Leaf
33	7 7 7 7
34	3 4 4 5 5 5 5 6 9
35	0 0 3 7 7 8 8 9
36	0 1 2 3 3 4 7 7 9
37	1 1 1 2

f

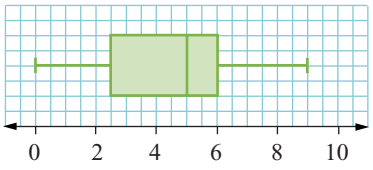
Stem	Leaf
21	0 3 3 3 4 5 6 7 8
22	1 1 1 1 2 2 3 4 6 7
23	2 3 4 4 5 5 8
24	1 2 3 3 4 4 5 6 8 9 9

EXAMPLE 5

Draw a box plot for the data in this frequency distribution table.

Score	Frequency	Cumulative frequency
0	1	1
1	3	4
2	4	8
3	3	11
4	3	14
5	7	21
6	4	25
7	5	30
8	1	31
9	1	32

EXAMPLE 5 CONTINUED

Solve	Think	Apply
	As there are 32 scores, the median is the average of the 16th and 17th scores. Median $= \frac{16\text{th score} + 17\text{th score}}{2}$ $= \frac{5+5}{2} = 5$ There are 16 scores above and 16 scores below the median. The 16th score is found by looking for the first score in the cumulative frequency column that is equal to or greater than 16. Go across from this value to find the score required, in this case 5. All other scores are found in this way. The lower quartile is the average of the 8th and 9th scores. $Q_1 = \frac{8\text{th score} + 9\text{th score}}{2}$ $= \frac{2+3}{2} = 2.5$ The upper quartile is the average of the 24th and 25th scores. $Q_3 = \frac{24\text{th score} + 25\text{th score}}{2}$ $= \frac{6+6}{2} = 6$ Highest score = 9 Lowest score = 0	If it is not given, complete a cumulative frequency column to determine the values of the quartiles and extremes. Draw the box plot.

5 Copy these tables and add a cumulative frequency column. Calculate the necessary information and draw a box plot for each data set.

a

Score	Frequency
12	20
13	18
14	15
15	15
16	17
17	15

b

Score	Frequency
53	15
54	30
55	13
56	3
57	9
58	30

c

Score	Frequency
110	5
111	22
112	26
113	25
114	17
115	5

d

Score	Frequency
32	6
33	8
34	9
35	13
36	9
37	3
38	2

e

Score	Frequency
47	4
48	7
49	12
50	21
51	10
52	6

f

Score	Frequency
0	1
1	2
2	3
3	4
4	2
5	5
6	4

D Comparing data sets

EXAMPLE 1

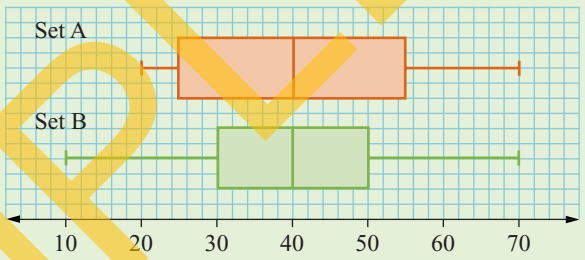
Two data sets are shown in these parallel box plots.

a Describe any similarities in the data sets.

b Compare the range of set A with that of set B.

c For which data set is the middle 50% clustered more closely to the median?

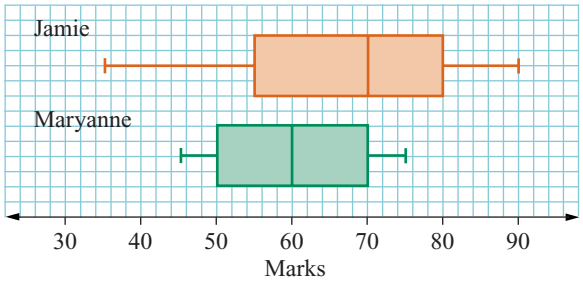
d In which data set is the top 50% of scores more closely clustered to the median?



	Solve	Think	Apply
a	The greatest score and the median are the same for both data sets.	Greatest score of both is 70. Median of both is 40.	Use the known proportion of values between the quartiles as well as the extreme values to analyse the data.
b	The range of set A is less than the range of set B.	Range of set A = 70 − 20 = 50 Range of set B = 70 − 10 = 60 Set B has the greater spread of scores.	
c	Set B	IQR of set A = 55 − 25 = 30 IQR of set B = 50 − 30 = 20 The spread of scores about the median is less for set B than set A.	
d	Set B	As the top 50% of scores are spread over the same interval, the scores between Q_2 and Q_3 will show any clustering. Q_3 is closer to the median for set B than for set A. The 25% of scores between Q_2 and Q_3 (and hence the top 50%) for set B are closer to their median.	

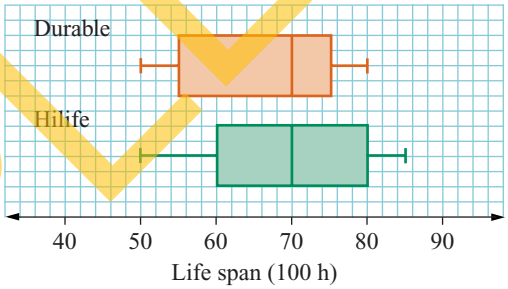
Exercise 8D

1 A class has eight assessment tasks over a year. The parallel box plots show a summary of the marks for the assessments for two students, Jamie and Maryanne.



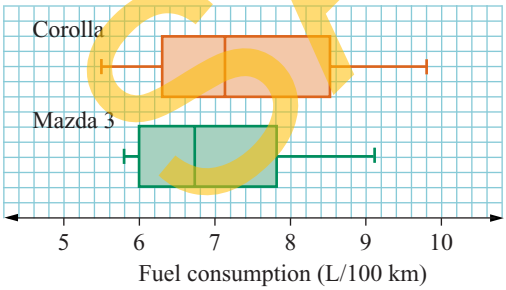
- **a** **i** Who scored the highest mark?
  ii Who scored the lowest mark?
- **b** **i** What was the range of marks for each student?
  ii Who had the greater spread of marks?
- **c** **i** What was the interquartile range of marks for each student?
  ii Whose marks were the more consistent?
- **d** Who had more marks over 70?
- **e** Assuming each assessment task had the same weighting, who do you think finished the year with the higher overall assessment? Give reasons for your answer.

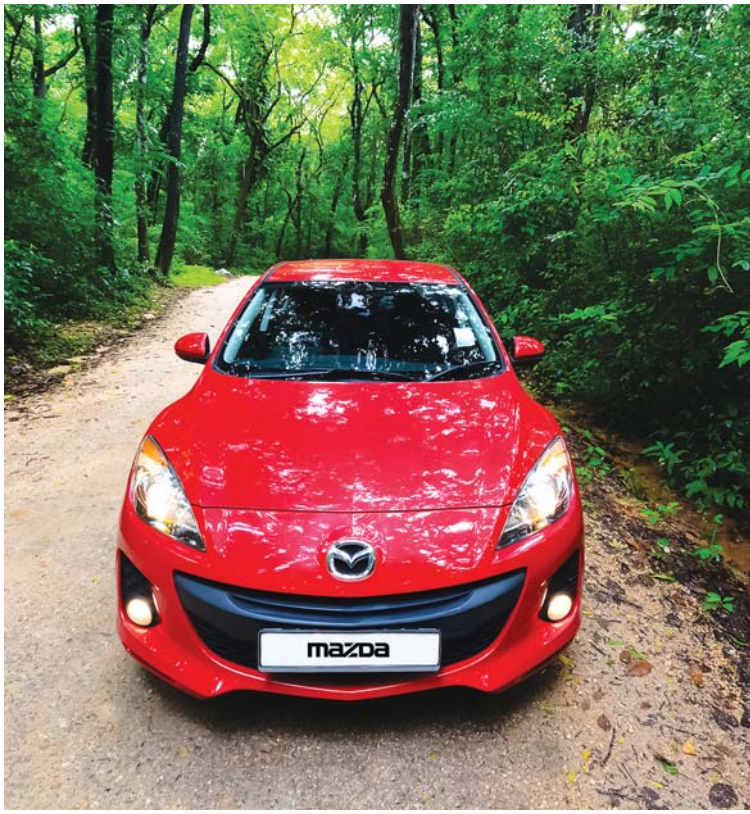
2 These parallel box plots show the life span of two brands of light globes (× 100 hours).



- **a** Describe any similarities in the data.
- **b** Which brand had the globe with the:
  i greatest life span? **ii** shortest life span?
- **c** **i** What was the range of life spans for each brand?
  ii Which brand had the greater spread of life spans?
- **d** **i** What was the interquartile range for each brand?
  ii What does this indicate about the middle 50% of life spans for each brand?
- **e** Which brand lasts longer? Give reasons for your answer.

3 A new Toyota Corolla and Mazda 3 were each taken for ten test drives over the same routes. These parallel box plots show a summary of the fuel consumption, in L/100 km, of each vehicle over these routes.

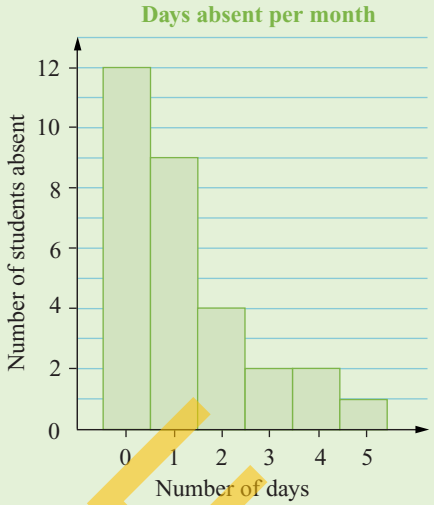




- **a** Which car recorded the:
  i highest fuel consumption?
  ii lowest fuel consumption?
- **b** Which car had the greater range of results?
- **c** Which car demonstrated the more consistent fuel consumption over all routes? Give a reason.
- **d** Which car used less than 7 L/100 km more often?
- **e** Which car had the better overall fuel consumption. Give reasons.

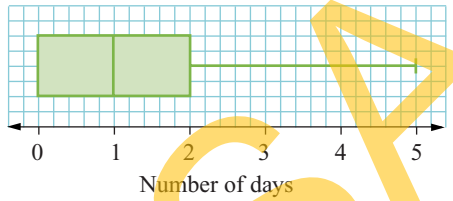
EXAMPLE 2

The histogram shows the number of days in a month on which students in a Year 10 class were absent. Draw a box plot for this data



Solve

Number of days	Number of students absent	Cumulative frequency
0	12	12
1	9	21
2	4	25
3	2	27
4	2	29
5	1	30



Think

The frequency of each score can be found from the histogram and put in a frequency distribution table. Add a cumulative frequency column.

From the cumulative frequency column:

$Q_2 = \frac{15\text{th} + 16\text{th scores}}{2} = 1$

$Q_1 = 8\text{th score} = 0$

$Q_3 = 23\text{rd score} = 2$

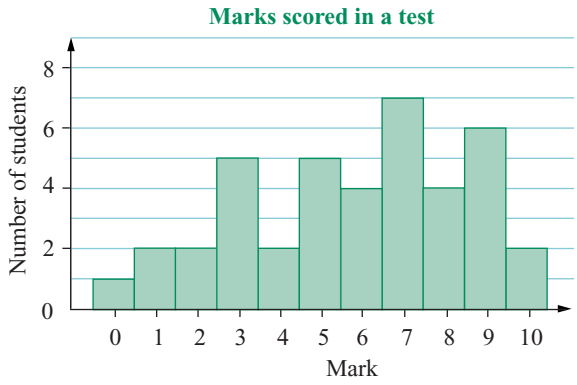
Lowest score = 0

Highest score = 5

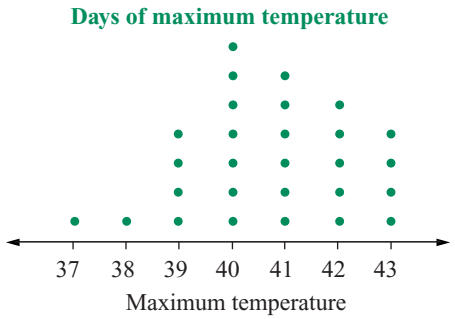
Apply

Put the information shown in the histogram into a frequency distribution table and add a cumulative frequency column. Use the cumulative frequency column to find the quartiles and add the extreme scores to make a five-number summary for the data. Draw the box plot.

4 The histogram shows the marks scored by a class in a test. Draw a box plot for this data.



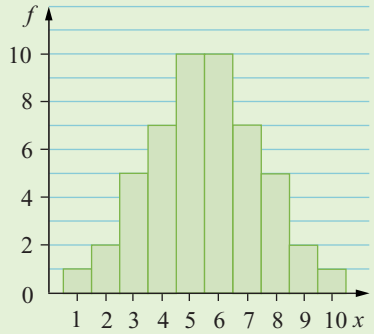
5 The dot plot shows the maximum daily temperatures for February. Draw a box plot for this information.



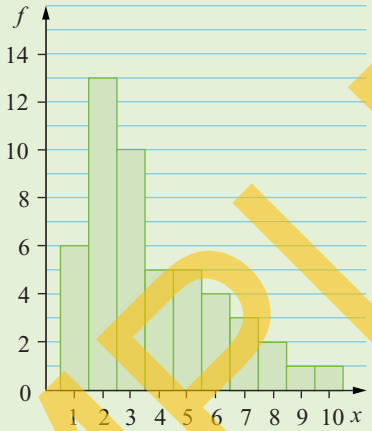
EXAMPLE 3

i Draw a box plot for the data shown in each of the histograms below.
ii Describe and compare the features of each histogram and its corresponding box plot.

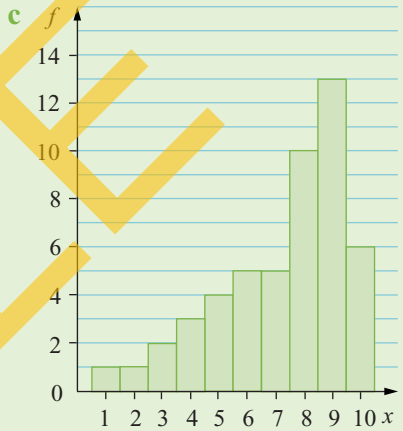
a



b



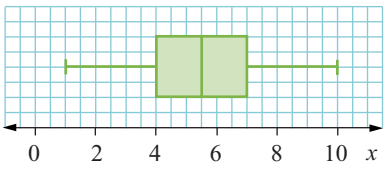
c



Solve	Think	Apply
a i From the histogram:	Using the information from the histogram, first compile a cumulative frequency table. Determine the position of the box and the relative spread between the quartiles and extremes in the box plot.	Put the information shown in the histogram into a frequency distribution table and add a cumulative frequency column. Use the cumulative frequency column to find the quartiles and add the extreme scores to make a five-number summary for the data. Draw the box plot.

x	f	cf
1	1	1
2	2	3
3	5	8
4	7	15
5	10	25
6	10	35
7	7	42
8	5	47
9	2	49
10	1	50

So the box plot is:

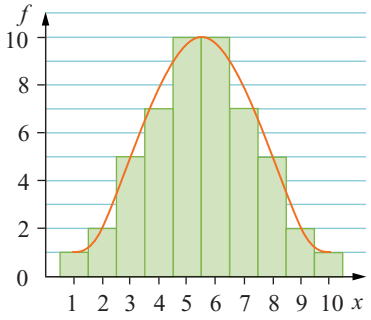
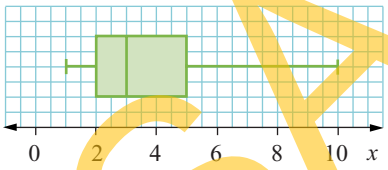
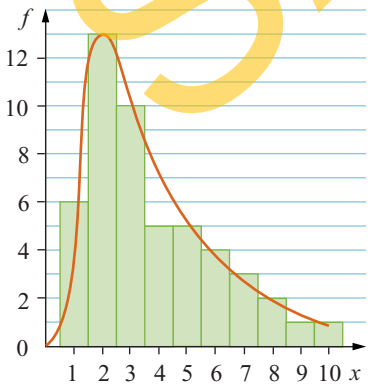


For this data:

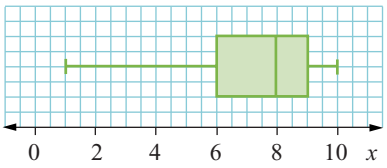
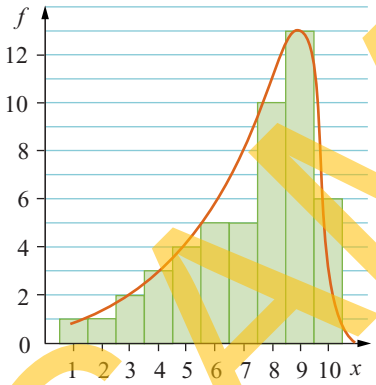
$$\text{Median} = \frac{25\text{th} + 26\text{th scores}}{2}$$
$$= \frac{5 + 6}{2}$$
$$= 5.5$$

Upper quartile is 38th score = 7
Lower quartile is 13th score = 4
Highest score = 10
Lowest score = 1
Q_1 and Q_3 are equally spaced from the median. The lowest and highest scores are equally spaced from Q_1 and Q_3 respectively.

EXAMPLE 3 CONTINUED

	Solve	Think	Apply
ii	 The histogram shows that the data is symmetrical. The corresponding box plot is also symmetrical.		Determine the position of the box and the relative spread between the quartiles and extremes in the box plot.
b i	From the histogram: x | f | cf 1 | 6 | 6 2 | 13 | 19 3 | 10 | 29 4 | 5 | 34 5 | 5 | 39 6 | 4 | 43 7 | 3 | 46 8 | 2 | 48 9 | 1 | 49 10 | 1 | 50 So the box plot is: 	Using the information from the histogram, first compile a cumulative frequency table. For this data: Median = 3 Upper quartile = 5 Lower quartile = 2 Highest score = 10 Lowest score = 1	Use the cumulative frequency column to find the quartiles and add the extreme scores to make a five-number summary for the data. Draw the box plot.
ii	 The histogram shows that the data is positively skewed. This is equivalent to saying that most of the data is on the left of the data set. The box plot shows the 'box' is to the far left of the data set.	The tail of the frequency curve is on the right. The lower 50% of the data is compressed between the scores 1 and 3. the upper 50% is widely spread between 3 and 10. Q_1 is closer than Q_3 to the median, showing the relative compression of this 25% of scores.	

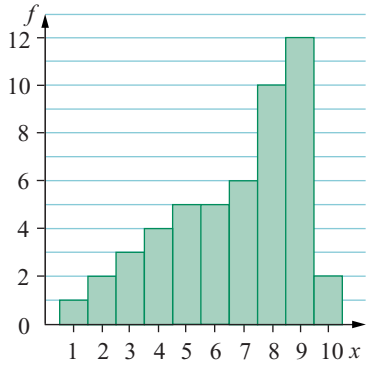
EXAMPLE 3 CONTINUED

	Solve	Think	Apply
c i	From the histogram: x | f | cf 1 | 1 | 1 2 | 1 | 2 3 | 2 | 4 4 | 3 | 7 5 | 4 | 11 6 | 5 | 16 7 | 5 | 21 8 | 10 | 31 9 | 13 | 44 10 | 6 | 50 So the box plot is:  	Using the information from the histogram, first compile a cumulative frequency table. For this data: Median = 8 Upper quartile = 9 Lower quartile = 6 Highest score = 10 Lowest score = 1	Use the cumulative frequency column to find the quartiles and add the extreme scores to make a five-number summary for the data. Draw the box plot.
ii	 The histogram shows that the data is negatively skewed. The corresponding box plot shows the 'box' is to the far right of the data set.	The tail of the frequency curve is on the left as most of the data is on the right of the data set. The upper 50% of the data is compressed between the scores 8 and 10. The lower 50% is widely spread between 1 and 8. Also Q_3 is closer than Q_1 to the median, showing the relative compression of this 25% of scores.	Determine the position of the box and the relative spread between the quartiles and extremes in the box plot.

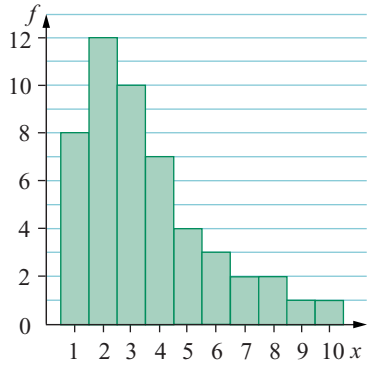
6 **i** Draw a box plot for the data shown in each of the histograms below.

ii Describe and compare the features of each histogram and its corresponding box plot.

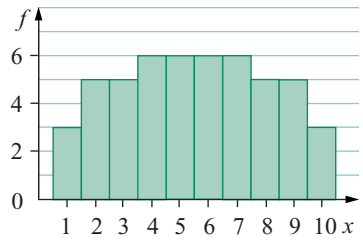
a



b

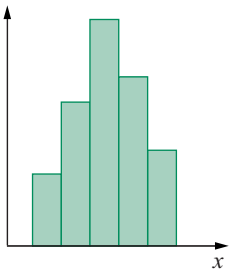


c

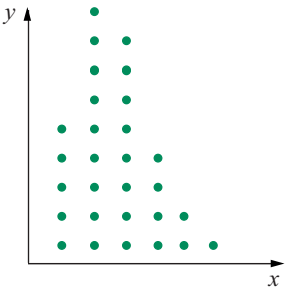


7 Match each histogram or dot plot with its corresponding box plot.

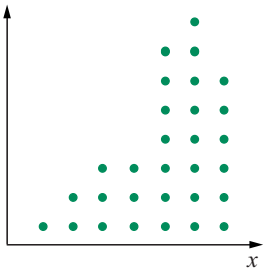
a



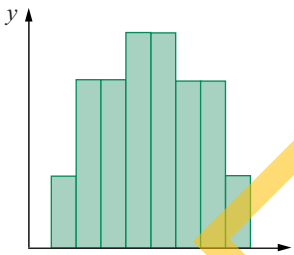
b



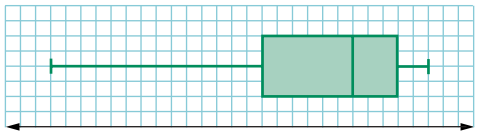
c



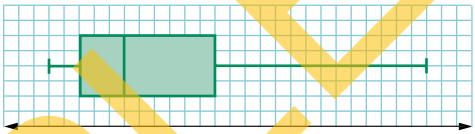
d



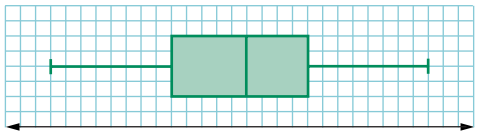
A



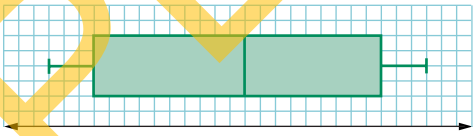
B



C



D



Investigation 1 Statistical reports in the media

1 Investigate survey data reported in the digital media and elsewhere to critically evaluate the reliability and validity of the source of the data and its usefulness. Describe bias that may exist due to the way in which the data was obtained. These are questions to consider:

- **a** Who instigated and/or funded the research?
- **b** Is the sample being used representative of the population?
- **c** Is the sample big enough?
- **d** Do the questions contain bias?
- **e** Is the research recent?

Language in mathematics

1 Insert vowels to complete these terms

a m__ __n **b** q__ __rt_l_ **c** sk_w_d d_str_b_t_ _n

d _pp_r q_ _rt_l_ **e** b_x pl_ts

2 **a** Describe the difference between the range and the interquartile range of a set of scores.

b When would it be better to use the interquartile range rather than the range?

3 Rearrange these words to form a sentence. The first word has a capital letter.

a by is range The unaffected outliers interquartile

b spread a is The of measure range

c score middle median order in the The is when arranged scores are the

4 Use every third letter to reveal a sentence about statistics.

E F T G T H H Y E U J M I K E A S A D F N E F A K G N Q E D D C M V G E B
H D P O I I U A Y T N T G A H J R K K E L L M O O E I I A W S S C F U B G
R Y A E X G S E T O U T F Y E C Q T E A V N E F T H R R J K A E D L W S T
Q A E Z X N D F D B H E J U N T G C B N Y M K W L O H O I I I U L U Y E T
R T Q W H S D E R V R G B A H N N J U G Y T E T S A F E N E C D E T I E D
N E T T I Y E A S R D V Q G T U Q B A J O R R F T D G I B M L A E E T G R
R F A E D N S E G E T E V H A U I R Y T E E W M S Q E A Z A X V S D F U G
W R Q X E C T S Y N O M I F R E S A G P C T R Q Z E X C A V G D H J T I K
H I O E O L I U Y N W E T F G E B H R U J Q M W U S Y A T O R D V T Q S I
C G L J I E L I R P O A T T N R E G C E E S E I A E S Q S T C V H H Y E I
U O T R N F H L E I Y O P M E D E C E A V E S F E U R G R T H E J K O A S
F S A S E S P E A R S E E D E A F E D R G T H J H Q S A A Z T C G I U O S
J K N R F O B T T H J A E D F N J F O I E O K C O L T P O E U Y D E R B F
G Y V F O E R U S D T V R L T H I U K E O I R L I S

Terms

bimodal	box plots	data sets	five-number summary	highest score
interquartile range	lower quartile	lowest score	mean	median
mode	normal distribution	outlier	quartile	range
skewed distribution	upper quartile			

Check your skills

1 The mean of the scores 8, 11, 11, 12, 14, 15, 15, 15, 16, 17, 20 is:

A 15 **B** 14 **C** 12 **D** 11

2 The range of the scores 8, 11, 11, 12, 14, 15, 15, 15, 16, 17, 20 is:

A 15 **B** 14 **C** 12 **D** 11

3 The median of the scores 8, 11, 11, 12, 14, 15, 15, 15, 16, 17, 20 is:

A 15 **B** 14 **C** 12 **D** 11

4 The mode of the scores 8, 11, 11, 12, 14, 15, 15, 15, 16, 17, 20 is:

A 15 **B** 14 **C** 12 **D** 11

5 The mean of 11, 15, 16, 19, 21 and x is 17. The value of x is:

A 20 **B** 21 **C** 16.4 **D** 10

Use the scores 15, 16, 17, 18, 18, 18, 20, 21, 21, 25 to answer questions **6** to 8.

6 The lower quartile is:

A 25 **B** 21 **C** 17 **D** 15

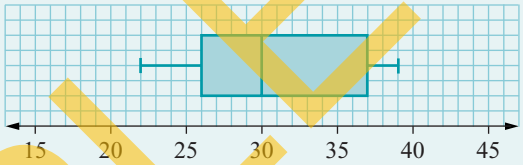
7 The upper quartile is:

A 21 **B** 18 **C** 17 **D** 10

8 The interquartile range is:

A 10 **B** 18 **C** 5 **D** 4

Use the information in this box plot to answer questions **9** to **11**.



9 The range is:

A 39 **B** 17

C 11 **D** 30

10 The median is:

A 39 **B** 17 **C** 11 **D** 30

11 The interquartile range is:

A 39 **B** 17 **C** 11 **D** 30

Use the data in this stem-and-leaf plot to answer questions **12** to **14**.

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

12 The median is:

A 86 **B** 75.5

C 65 **D** 50

13 The interquartile range is:

A 50 **B** 86

C 21 **D** 11

14 The lowest and highest scores are:

A 48 and 98 **B** 65 and 86 **C** 0 and 9 **D** 48 and 75

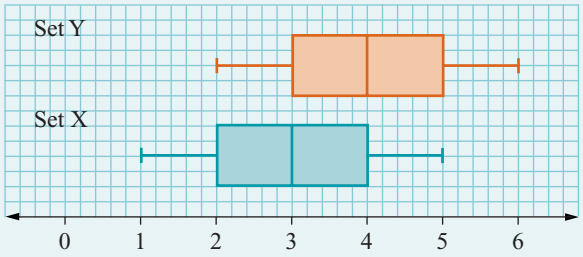
15 For the parallel box plots shown on the right, which statement is not true?

A The range is the same for both data sets.

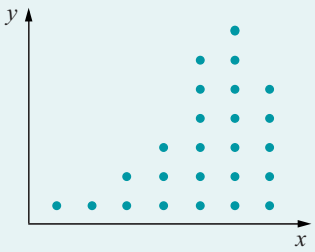
B The interquartile range is the same for both data sets.

C The median of set X is greater than the median of set Y.

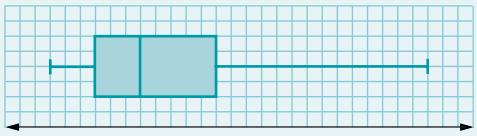
D Both data sets are symmetrical.



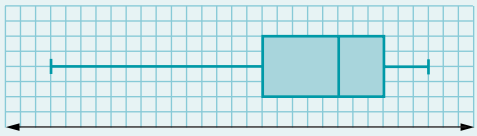
16 The box plot that best matches the given dot plot is:



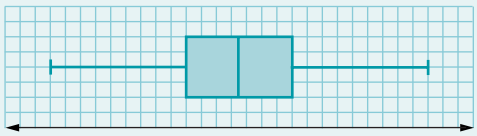
A



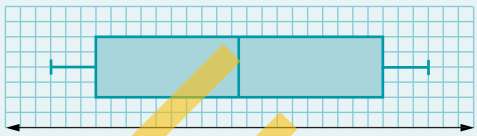
B



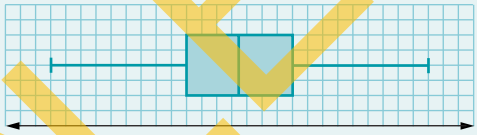
C



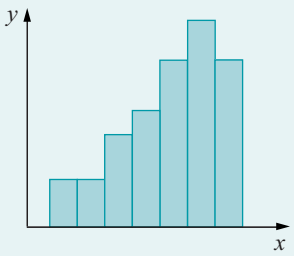
D



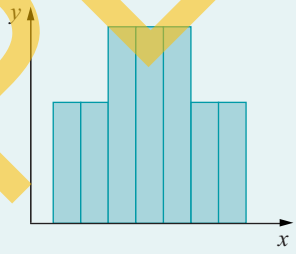
17 The histogram that best matches this box plot is:



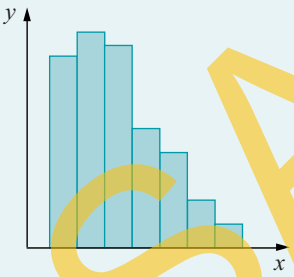
A



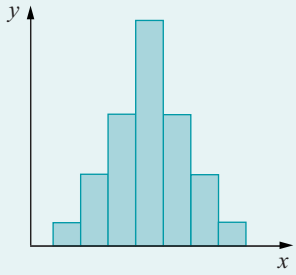
B



C



D



If you have any difficulty with these questions, refer to the examples and questions in the sections listed in the table.

Question	1–5	6–8	9–14	15–17
Section	A	B	C	D

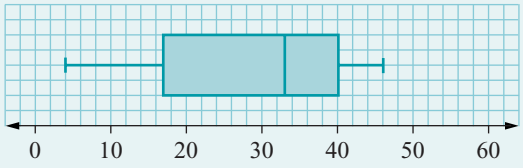
8A Review set

1 For the scores 6, 7, 7, 9, 10, 11, 14, find:

a the mean
b the median
c the range
d the interquartile range.

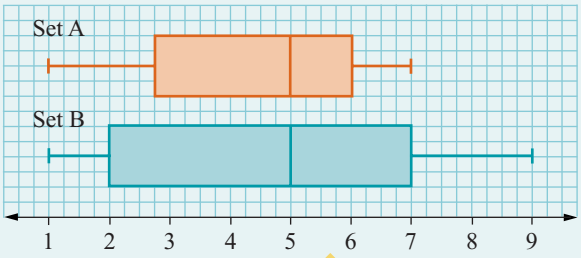
2 Find x when the mean of 7, 12, 18, 16 and x is 15.

3 From the box plot shown, find:



- **a** the highest score
- **b** the lowest score
- **c** the range
- **d** the median
- **e** the upper quartile
- **f** the lower quartile
- **g** the interquartile range.

4 The diagram shows parallel box plots for the data in sets A and B.



- **a** What are the similarities between these sets of data?
- **b** Which data set has the greater range?
- **c** Which data set has the greater spread of the middle 50% of its scores?
- **d** Compare the spread of the lower 50% of scores in each data set.
- **e** If the box plots represent the marks of two classes on a test, which class do you think was more consistent?

8B Review set

1 For the scores 2, 4, 6, 9, 9, 10, find:

- **a** the mean
- **b** the median
- **c** the range
- **d** the interquartile range.

2 **a** The mean of six scores is 14. What is the sum of the scores?

b If 11, 15, 12, 11, 8 and x have a mean of 13, find x.

3 **a** Find the range of the scores in these frequency distribution tables.

b What is the interquartile range?

i

Score	Frequency
9	6
10	5
11	9
12	11
13	3
14	6

ii

Score	Frequency
25	6
26	10
27	10
28	13
29	6
30	2

4 Draw box plots for the following data sets.

a 3, 4, 7, 7, 9, 11, 11, 13, 14, 18

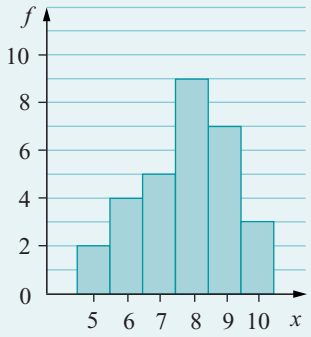
b

Score	Frequency
15	18
16	16
17	13
18	13
19	15
20	13

c

Stem	Leaf
6	3 3 3 3
7	3 4 4 5 5 5 6 6 8
8	0 0 2 5 5 9 9 9
9	0 1 3 3 3 4 7 7 9
10	3 3 3 4

5 **a** Draw a box plot for the data shown in the histogram.
b Describe how the features of the histogram are shown in the corresponding box plot.

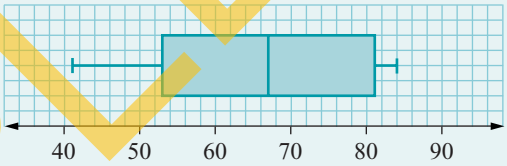


8C Review set

1 For the scores 11, 11, 12, 13, 15, 15, 15, 16, 19, 20, 21, 21, find:
a the mean
b the median
c the range
d the interquartile range.

2 Find x when the mean of 17, 22, 38, 36 and x is 30.

3 From the box plot, find:
a the highest score
b the lowest score
c the range
d the median
e the upper quartile
f the lower quartile
g the interquartile range.

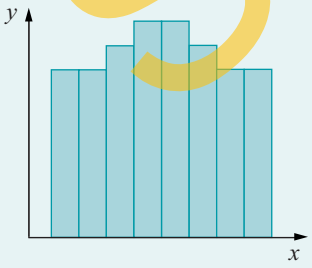


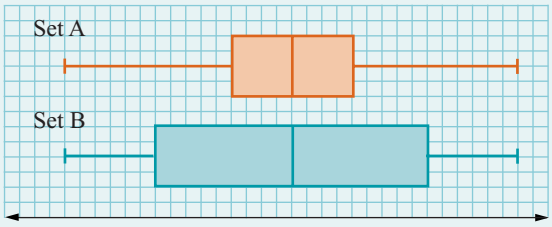
4 **a** Find the range and interquartile range for the scores in this frequency distribution table.
b Draw a histogram for the scores in the table.
c Comment on the shape of the distribution.

Score	Frequency
16	4
17	6
18	8
19	15
20	23
21	14

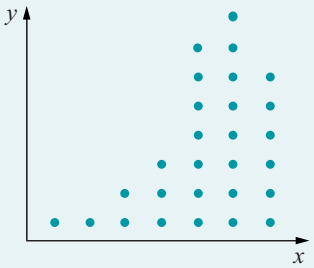
5 Select the data set in the box plot that best matches the given histogram or dot plot.

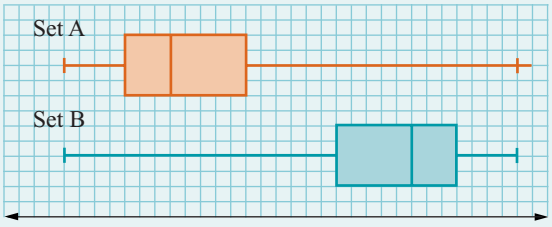
a





b





8D Review set

1 For the scores 65, 61, 64, 61, 61, 62, find:

a the mean

b the median

c the range

d the interquartile range.

2 **a** If 9, 6, 2, 5 and x have a mean of 7, find x.

b The mean of eight scores is 5.25. What is the sum of the scores?

3 **a** For this stem-and-leaf plot, find:

i the mean **ii** the range

iii the median **iv** the interquartile range.

b Is this distribution symmetrical or skewed? Explain.

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

4 Draw box plots for the following data sets.

a 1, 1, 3, 4, 5, 5, 5, 6, 7, 10, 11

b

Score	Frequency
21	23
22	28
23	15
24	31
25	12

c

Stem	Leaf
18	1 2 3 3 5 6 7 8
19	1 1 1 1 2 2 3 6 6 7 9
20	1 3 4 5 6 6 7 8
21	0 0 0 1 3 4 5 9

5 The box plot shows the mean daily maximum temperatures in Sydney and Melbourne for the month of January. Compare and describe the features of the weather illustrated by these displays.