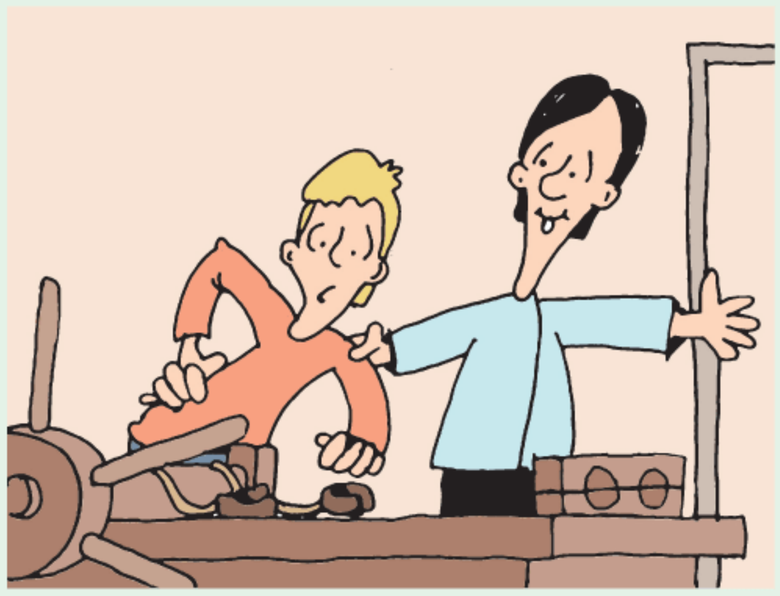
Chapter 12

Statistics

Contents:

OPENING PROBLEM

Nick believes he has devised a series of stretches which can help relieve back pain. He invites people with back pain to perform the stretches for several weeks.



The participants rate their level of back pain on a scale of 1 to 10 (10 being the greatest) before and after the experiment:

Before:	7	9	5	6	9	7	10	6	8	9
	8	7	9	8	10	4	8	6	7	8
After:	4	7	6	3	8	5	9	4	5	8
	7	6	5	4	7	2	5	3	4	6

Things to think about:

a What statistics can we calculate to measure the *centre* of each data set?

b How can we use a graph to make a visual comparison between the data sets?

c Do you believe that Nick's stretching exercises reduce back pain? Explain your answer.

In the previous Chapter, we looked at how data can be collected, organised, and displayed. By looking at appropriate graphs, we can get an idea of a data set's **distribution**.

We can get a better understanding of a data set if we can locate its **middle** or **centre**, and measure its **spread** or dispersion. Knowing one of these without the other is often of little use.

However, whatever statistics we calculate, it is essential to view and interpret them in the context of what we are studying.

A MEASURING THE CENTRE OF DATA

There are three statistics that are used to measure the **centre** of a data set. These are the **mode**, the **mean**, and the **median**.

THE MODE

In the previous Chapter we saw that:

- For discrete data, the **mode** is the most frequently occurring value in the data set.
- For continuous data, we cannot talk about a mode in this way because no two data values will be *exactly* equal. Instead we talk about a **modal class**, which is the class or group that has the highest frequency.

If a data set has two values which both occur most frequently, we say it is **bimodal**.

If a data set has three or more values which all occur most frequently, the mode is not an appropriate measure of centre to use.

THE MEAN

The **mean** of a data set is the statistical name for its arithmetic average.

For the data set $\{x_1, x_2, x_3, ..., x_n\}$,

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

$$= \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$

We use $\overline{x}$ to represent the mean of a **sample**, and μ to represent the mean of a **population**.

In many cases we do not have data from all of the members of a population, so the exact value of μ is unknown. Instead we collect data from a sample of the population, and use the mean of the sample $\overline{x}$ as an approximation for μ.

THE MEDIAN

The **median** is the *middle value* of an ordered data set.

An ordered data set is obtained by listing the data from the smallest to the largest value.

The median splits the data in halves. Half of the data values are less than or equal to the median, and half are greater than or equal to it.

For example, if the median mark for a test is 73% then you know that half the class scored less than or equal to 73% and half scored greater than or equal to 73%.

For an **odd number** of data values, the median is one of the original data values.

For an **even number** of data values, the median is the average of the two middle values, and hence may not be in the original data set.

If there are n data values listed in order from smallest to largest, the median is the $\left(\frac{n+1}{2}\right)$th data value.

For example:

If $n = 13$, $\frac{n+1}{2} = 7$, so the median is the 7th ordered data value.

If $n = 14$, $\frac{n+1}{2} = 7.5$, so the median is the average of the 7th and 8th ordered data values.

Example 1 **Self Tutor**

The numbers of faulty products returned to an electrical goods store each day over a 21 day period are:

3 4 4 9 8 8 6 4 7 9 1 3 5 3 5 9 8 6 3 7 1

a For this data set, find:

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

b On the 22nd day there were 9 faulty products returned. How does this affect the measures of the centre?

a **i** $\text{mean} = \dfrac{3+4+4+....+3+7+1}{21}$ ← sum of all the data values
← 21 data values

$= \frac{113}{21}$

≈ 5.38 faulty products

ii As $n = 21$, $\dfrac{n+1}{2} = 11$

The ordered data set is: ~~1 1 3 3 3 3 4 4 4 5~~ 5 ~~6 6 7 7 8 8 8 9 9 9~$

11th value

$\therefore$ median $= 5$ faulty products

iii 3 is the data value which occurs most often, so the mode is 3 faulty products.

b We expect the mean to increase since the new data value is greater than the old mean.

In fact, the new mean $= \dfrac{113+9}{22} = \dfrac{122}{22} \approx 5.55$ faulty products.

Since $n = 22$, $\dfrac{n+1}{2} = 11.5$

The new ordered data set is:

~~1 1 3 3 3 3 4 4 4 5~~ 5 6 ~~6 7 7 8 8 8 9 9 9 9~~

two middle data values

$\therefore$ the new median $= \dfrac{5+6}{2} = 5.5$ faulty products.

The new data set has two modes which are 3 and 9 faulty products.

You can use your **graphics calculator** or the **statistics package** to find measures of centre.

GRAPHICS CALCULATOR INSTRUCTIONS

STATISTICS PACKAGE

EXERCISE 12A

1 Phil kept a record of the number of cups of coffee he drank each day for 15 days:

2, 3, 1, 1, 0, 0, 4, 3, 0, 1, 2, 3, 2, 1, 4

Without using technology, find the **a** mode **b** median **c** mean of the data.

2 For each data set, find the: **i** mean **ii** median **iii** mode.

a 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9

b 10, 12, 12, 15, 15, 16, 16, 17, 18, 18, 18, 18, 19, 20, 21

c 22.4, 24.6, 21.8, 26.4, 24.9, 25.0, 23.5, 26.1, 25.3, 29.5, 23.5

Check your answers using technology.

3 The sum of 7 scores is 63. What is their mean?

4 The scores obtained by two ten-pin bowlers over a 10 game series are:

Gordon:	160,	175,	142,	137,	151,	144,	169,	182,	175,	155
Ruth:	157,	181,	164,	142,	195,	188,	150,	147,	168,	148

Who had the higher mean score?

5 Consider the two data sets:

Data set A: 3, 4, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 10

Data set B: 3, 4, 4, 5, 6, 6, 7, 7, 7, 8, 8, 9, 15

a Find the mean of both data set A and data set B.

b Find the median of both data set A and data set B.

c Comment on your answers to **a** and **b**.

6 An Indian dessert shop keeps a record of how many motichoor ladoo and malai jamun they sell each day for a month:

Motichoor ladoo

62	76	55	65	49	78	71	82
79	47	60	72	58	82	76	67
50	61	70	85	77	69	48	74
63	56	81	75	63	74	54	

Malai jamun

37	52	71	59	63	47	56	68
43	67	38	73	54	55	61	49
50	48	53	39	45	60	46	51
38	57	41	72	50	44	76	

a Find the:

i mean number of motichoor ladoo and malai jamun sold

ii median number of motichoor ladoo and malai jamun sold.

b Which item was more popular? Explain your answer.

7 A bus and tram travel the same route many times during the day. The drivers counted the number of passengers on each trip one day, as listed below.

Bus

30	43	40	53	70	50	63
41	38	21	28	23	43	48
20	26	35	48	41	33	

Tram

58	68	43	45	70	79
38	23	30	22	63	73
25	35	60	53		

a Use technology to calculate the mean and median number of passengers for both the *Bus* and *Tram* data.

b Which method of transport do you think is more popular? Explain your answer.

8 A basketball team scored 43, 55, 41, and 37 points in their first four matches.

a Find the mean number of points scored for these four matches.

b What score does the team need to shoot in their next match to maintain the same mean score?

c The team scores only 25 points in the fifth match.

i Will this increase or decrease their overall mean score? Explain your answer.

ii Find the mean number of points scored for the five matches.

Example 2 **Self Tutor**

If 6 people have a mean mass of 53.7 kg, find their total mass.

$$\frac{\text{sum of masses}}{6} = 53.7 \text{ kg}$$

$\therefore$ sum of masses $= 53.7 \times 6$

$\therefore$ the total mass $= 322.2$ kg

9 This year, the mean monthly sales for a clothing store have been €15 467. Calculate the total sales for the store for the year.

10 While on a 12 day outback safari, Bill drove an average of 262 km per day. How far did Bill drive in total while on the safari?

11 Given $\overline{x} = 11.6$ and $n = 10$, calculate $\sum_{i=1}^{10} x_i$.

12 Towards the end of a season, a netballer had played 14 matches and scored an average of 16.5 goals per game. In the final two matches of the season she scored 21 goals and 24 goals. Find the netballer's average for the whole season.

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

14 Find a if 3, 0, a, a, 4, a, 6, a, and 3 have a mean of 4.

15 Over the entire assessment period, Aruna averaged 35 out of a possible 40 marks for her Mathematics tests. However, when checking her files, she could only find 7 of the 8 tests. For these she scored 29, 36, 32, 38, 35, 34, and 39. How many marks out of 40 did she score for the eighth test?

16 A sample of 10 measurements has a mean of 15.7, and a sample of 20 measurements has a mean of 14.3. Find the mean of all 30 measurements.

17 The mean and median of a set of 9 measurements are both 12. Seven of the measurements are 7, 9, 11, 13, 14, 17, and 19. Find the other two measurements.

INVESTIGATION 1 EFFECTS OF OUTLIERS

We have seen that an **outlier** or **extreme value** is a value which is much greater than, or much less than, the other values.

Your task is to examine the effect of an outlier on the three measures of centre.

What to do:

1 Consider the set of data: 4, 5, 6, 6, 6, 7, 7, 8, 9, 10. Calculate:

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

2 Suppose we introduce the extreme value 100 to the data, so the data set is now: 4, 5, 6, 6, 6, 7, 7, 8, 9, 10, 100. Calculate:

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

3 Comment on the effect that the extreme value has on:

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

4 Which of the three measures of centre is most affected by the inclusion of an outlier?

5 Discuss situations with your class when it would *not* be appropriate to use a particular measure of centre of a data set.

B CHOOSING THE APPROPRIATE MEASURE

The mean, mode, and median can all be used to indicate the centre of a set of numbers. The most appropriate measure will depend upon the type of data under consideration. When selecting which one to use for a given set of data, you should keep the following properties in mind.

Statistic	*Properties*
Mode	• gives the most usual value • only takes common values into account • not affected by extreme values
Mean	• commonly used and easy to understand • takes all values into account • affected by extreme values
Median	• gives the halfway point of the data • only takes middle values into account • not affected by extreme values

For example:

- A shoe store is investigating the sizes of shoes sold over one month. The mean shoe size is not useful to know, since it probably will not be an actual shoe size. However, the mode shows at a glance which size the store most commonly has to restock.
- On a particular day a computer shop makes sales of \$900, \$1250, \$1000, \$1700, \$1140, \$1100, \$1495, \$1250, \$1090, and \$1075. In this case the mode is meaningless, the median is \$1120, and the mean is \$1200. The mean is the best measure of centre as the salesman can use it to predict average profit.
- When looking at real estate prices, the mean is distorted by the few sales of very expensive houses. For a typical house buyer, the median will best indicate the price they should expect to pay in a particular area.

EXERCISE 12B

1 The selling prices of the last 10 houses sold in a certain district were as follows:

\$346 400, \$327 600, \$411 000, \$392 500, \$456 400,
\$332 400, \$348 000, \$329 500, \$331 400, \$362 500

a Calculate the mean and median selling prices. Comment on your results.

b Which measure would you use if you were:

i a vendor wanting to sell your house

ii looking to buy a house in the district?

2 The annual salaries of ten office workers are:

$$\$33\,000,\ \$56\,000,\ \$33\,000,\ \$48\,000,\ \$34\,000,$$
$$\$33\,000,\ \$33\,000,\ \$48\,000,\ \$33\,000,\ \$42\,000$$

a Find the mode, mean, and median salaries of this group.

b Explain why the mode is an unsatisfactory measure of the centre in this case.

c Is the median a satisfactory measure of the centre of this data set?

3 The following raw data is the daily rainfall, to the nearest millimetre, for a month:

3, 1, 0, 0, 0, 0, 0, 2, 0, 0, 3, 0, 0, 0, 7, 1, 1, 0, 3, 8, 0, 0, 0, 42, 21, 3, 0, 3, 1, 0, 0

a Use technology to find the mean, median, and mode of the data.

b Explain why the median is not the most suitable measure of centre for this set of data.

c Explain why the mode is not the most suitable measure of centre for this set of data.

d Identify the outliers in this data set.

e The outliers are genuine pieces of data and not the result of recording errors. Should they be removed before calculating statistics?

4 Esmé runs a day-tour business in Amsterdam. She wants to offer a "family package" that includes the charges for two adults and their children. To investigate the number of children she should include in the package, she asks 30 randomly selected customers with children how many children they have. Their responses are:

2	2	2	3	4	1	1	2	1	1	1	2	2	3	4
1	4	4	2	3	1	1	1	2	1	1	2	2	3	2

a Calculate the mean, median, and modal number of children per family.

b Is the mode a useful statistic in this case?

c Suggest how many children Esmé should include in the package, giving reasons for your answer.

THEORY OF KNOWLEDGE

We have seen that the mean, the median, and mode are all statistics that give an *indication* of a data set's centre. The actual things that they measure are quite different!

- The mode is the value with the highest frequency. It is a measure of centre in terms of *frequency*.
- The median divides the data into halves. It is a measure of centre in terms of *proportion*.
- The mean is the arithmetic average. It can be thought of as the "balancing point" of the data set's distribution.

Other less commonly used measures for a data set $\{x_1, x_2, ..., x_n\}$ include:

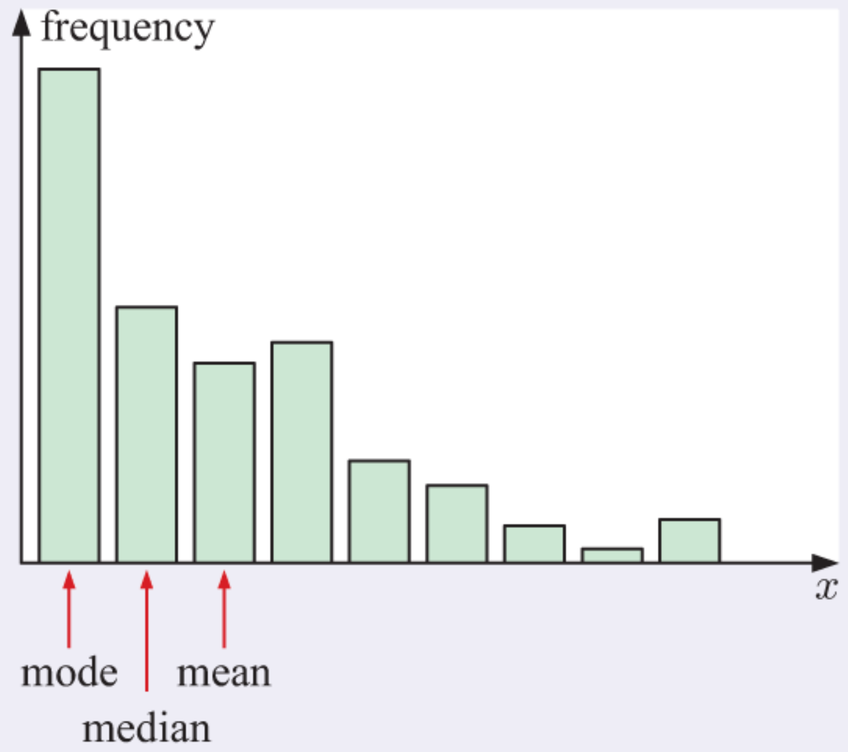
- the **geometric mean** $= \sqrt[n]{x_1 \times x_2 \times .... \times x_n}$
- the **mid-range value** $= \dfrac{\text{maximum} + \text{minimum}}{2}$.

We have seen that the most appropriate measure of centre will depend on what we are investigating. In a way, we change how the "centre" of a data set is defined to suit the purpose of our investigation.

1 When we have data that is heavily skewed, the mode will be on the far left or far right on its column graph.

a Does the mode give an accurate indication of a data set's centre in these cases?

b Is the relationship between mode and the centre of a data set purely coincidental?



2 For what kinds of data sets would the geometric mean and the mid-range value be useful?

3 How would *you* define the "centre" of a data set?

4 What makes a measure of centre objectively "better" than another measure?

5 Is there a *canonical* measure of centre, which means a measure of centre that is "better" than any other in all cases?

C USING FREQUENCY TABLES

We have already seen how to organise data into a **frequency table** like the one alongside.

The mode of the data is found directly from the *Frequency* column.

Value	*Frequency*
3	1
4	1
5	3
6	7
7	15
8	8
9	5

mode →

THE MEAN

Adding a "Product" column to the table helps to add the data values.

For example, the value 7 occurs 15 times, and these add to $15 \times 7 = 105$.

Value (x)	*Frequency* (f)	*Product* (xf)
3	1	$3 \times 1 = 3$
4	1	$4 \times 1 = 4$
5	3	$5 \times 3 = 15$
6	7	$6 \times 7 = 42$
7	15	$7 \times 15 = 105$
8	8	$8 \times 8 = 64$
9	5	$9 \times 5 = 45$
Total	$\sum f = 40$	$\sum xf = 278$

Since the mean $= \dfrac{\text{sum of all data values}}{\text{the number of data values}}$, we find

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_k f_k}{f_1 + f_2 + f_3 + \ldots + f_k}$$ where k is the number of *different* values in the data.

$$\therefore \quad \overline{x} = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j} \quad \text{which we often abbreviate as} \quad \frac{\sum xf}{\sum f}.$$

In this case the mean $= \dfrac{278}{40} = 6.95$.

THE MEDIAN

Since $\dfrac{n+1}{2} = \dfrac{41}{2} = 20.5$, the median is the average of the 20th and 21st ordered data values.

In the table, the blue numbers show the accumulated frequency values, or **cumulative frequency**.

We can see that the 20th and 21st ordered data values are both 7s.

$\therefore$ the median $= \dfrac{7+7}{2} = 7$

Value	*Frequency*	*Cumulative frequency*
3	1	1 ← one number is 3
4	1	2 ← two numbers are 4 or less
5	3	5 ← five numbers are 5 or less
6	7	12 ← 12 numbers are 6 or less
7	15	27 ← 27 numbers are 7 or less
8	8	35 ← 35 numbers are 8 or less
9	5	40 ← all numbers are 9 or less
Total	40	

Example 3 **Self Tutor**

The table below shows the number of aces served by a sample of tennis players in their first sets of a tournament.

Number of aces	1	2	3	4	5	6
Frequency	4	11	18	13	7	2

Determine the: **a** mean **b** median **c** mode for this data.

Number of aces (x)	*Frequency* (f)	*Product* (xf)	*Cumulative frequency*
1	4	4	4
2	11	22	15
3	18	54	33
4	13	52	46
5	7	35	53
6	2	12	55
Total	$\sum f = 55$	$\sum xf = 179$	

a $\overline{x} = \dfrac{\sum xf}{\sum f}$

$= \dfrac{179}{55}$

≈ 3.25 aces

In this case $\dfrac{\sum xf}{\sum f}$ is short for $\dfrac{\sum_{j=1}^{6} x_j f_j}{\sum_{j=1}^{6} f_j}$.

b There are 55 data values, so $n = 55$. $\dfrac{n+1}{2} = 28$, so the median is the 28th ordered data value. From the cumulative frequency column, the 16th to 33rd ordered data values are 3 aces.

$\therefore$ the 28th ordered data value is 3 aces.

$\therefore$ the median is 3 aces.

c Looking down the frequency column, the highest frequency is 18. This corresponds to 3 aces, so the mode is 3 aces.

EXERCISE 12C

1 The table alongside shows the number of people in cars on a road.

Calculate the:

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

Number of people	*Frequency*
1	13
2	8
3	4
4	5
Total	30

Check your answers using your graphics calculator.

GRAPHICS CALCULATOR INSTRUCTIONS

2 The frequency table alongside shows the number of phone calls made in a day by 50 fifteen-year-olds.

a For this data set, find the:

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

b Construct a column graph for the data and show the position of the mean, median, and mode on the horizontal axis.

c Describe the distribution of the data.

d Why is the mean larger than the median?

e Which measure of centre would be the most suitable for this data set?

Number of phone calls	*Frequency*
0	5
1	8
2	13
3	8
4	6
5	3
6	3
7	2
8	1
11	1

3

Number of matches	*Frequency*
47	5
48	4
49	11
50	6
51	3
52	1
Total	30

A company claims that their match boxes contain 50 matches on average. The Consumer Protection Society conducts a survey to assess the company's claim. The results of the survey are shown alongside.

a Calculate the:

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

b Do the results support the company's claim?

c In a court for "false advertising", the company won their case against the Consumer Protection Society. Suggest how they did this.

4 Families at a school in Manchester were surveyed, and the number of children in each family was recorded. The results of the survey are shown alongside.

Number of children	*Frequency*
1	5
2	28
3	15
4	8
5	2
6	1
Total	59

a Calculate the:

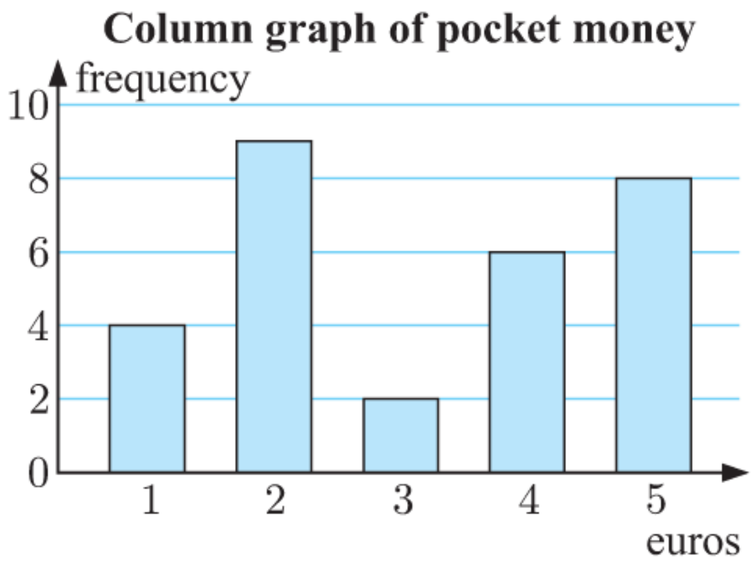
i mean **ii** mode **iii** median.

b The average British family has 2.075 children. How does this school compare to the national average?

c Describe the skewness of the data.

d How has the skewness of the data affected the measures of the centre of the data set?

5 The column graph shows the weekly pocket money for a class of children.



a Construct a frequency table from the graph.

b Determine the total number of children in the class.

c Find the:

i mean **ii** median **iii** mode of the data.

d Which of the measures of centre can be found easily using the graph only?

6 Out of 31 measurements, 15 are below 10 cm and 12 are above 11 cm. Find the median if the other 4 measurements are 10.1 cm, 10.4 cm, 10.7 cm, and 10.9 cm.

7 In an office of 20 people there are only 4 salary levels paid:

\$100 000 (1 person), \$84 000 (3 people),
\$70 000 (6 people), \$56 000 (10 people)

a Calculate:

i the median salary **ii** the modal salary

iii the mean salary.

b Which measure of central tendency might be used by the boss who is against a pay rise for the other employees?

8 The table shows the test scores for a class of students. A pass is a score of 5 or more.

Score	2	3	4	5	6	7	8
Frequency	0	2	3	5	x	4	1

a Given that the mean score was 5.45, find x.

b Find the percentage of students who passed.

D GROUPED DATA

When information has been gathered in groups or classes, we use the **midpoint** or **mid-interval value** to represent all data values within each interval.

We are assuming that the data values within each class are evenly distributed throughout that interval. The mean calculated is an **approximation** of the actual value, and we cannot do better than this without knowing each individual data value.

INVESTIGATION 2 MID-INTERVAL VALUES

When mid-interval values are used to represent all data values within each interval, what effect will this have on estimating the mean of the grouped data?

The table alongside summarises the marks out of 50 received by students who sat a Physics examination. The exact results for each student have been lost.

Marks	*Frequency*
0 - 9	2
10 - 19	31
20 - 29	73
30 - 39	85
40 - 49	28

What to do:

1 Suppose that all of the students scored the lowest possible result in their class interval, so 2 students scored 0, 31 students scored 10, and so on.
Calculate the mean of these results, and hence complete:
"The mean Physics examination mark must be *at least* ....."

2 Now suppose that all of the students scored the highest possible result in their class interval.
Calculate the mean of these results, and hence complete:
"The mean Physics examination mark must be *at most* ....."

3 We now have two extreme values between which the actual mean must lie.
Now suppose that all of the students scored the mid-interval value in their class interval. We assume that 2 students scored 4.5, 31 students scored 14.5, and so on.

a Calculate the mean of these results.

b How does this result compare with lower and upper limits found in **1** and **2**?

c Copy and complete:
"The mean Physics examination mark was approximately ....."

4 Discuss with your class how accurate you think an estimate of the mean using mid-interval values will be. How is this accuracy affected by the number and width of the class intervals?

Example 4 Self Tutor

The table below shows the ages of bus drivers. Estimate the mean age, to the nearest year.

Age (years)	21 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 - 50	51 - 55
Frequency	11	14	32	27	29	17	7

Age (years)	*Frequency* (f)	*Midpoint* (x)	xf
21 - 25	11	23	253
26 - 30	14	28	392
31 - 35	32	33	1056
36 - 40	27	38	1026
41 - 45	29	43	1247
46 - 50	17	48	816
51 - 55	7	53	371
Total	$\sum f = 137$		$\sum xf = 5161$

$$\overline{x} = \frac{\sum xf}{\sum f}$$
$$= \frac{5161}{137}$$
$$\approx 37.7$$

$\therefore$ the mean age of the drivers is about 38 years.

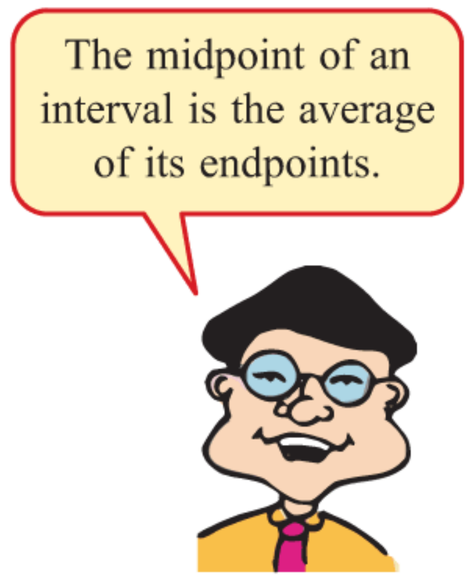
EXERCISE 12D

1 Simone recorded the lengths of her phone calls for one week. The results are shown in the table alongside.

a How many phone calls did she make during the week?

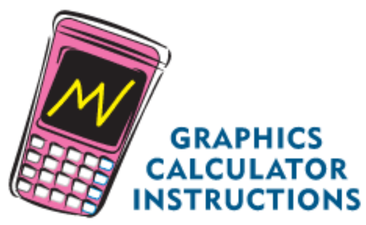
b Estimate the mean length of the calls.

Time (t min)	*Frequency*
$0 \leqslant t < 10$	17
$10 \leqslant t < 20$	10
$20 \leqslant t < 30$	9
$30 \leqslant t < 40$	4



2 50 students sat a Mathematics test. Estimate the mean score given these results:

Score	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
Frequency	2	5	7	27	9



Check your answers using your calculator.

3 A teacher recorded the number of children who used the school's playground each day for 50 days.

a On how many days was the playground used by more than 40 children?

b Find the modal class.

c Estimate the mean of the data.

Number of children	*Frequency*
21 - 30	8
31 - 40	16
41 - 50	14
51 - 60	12
Total	50

4 The table shows the petrol sales in one day by a number of city service stations.

a How many service stations were involved in the survey?

b Estimate the total amount of petrol sold for the day by the service stations.

c Estimate the mean amount of petrol sold for the day.

d Find the modal class for this distribution. Explain your answer.

Amount of petrol (P L)	*Frequency*
$2000 < P \leqslant 3000$	4
$3000 < P \leqslant 4000$	4
$4000 < P \leqslant 5000$	9
$5000 < P \leqslant 6000$	14
$6000 < P \leqslant 7000$	23
$7000 < P \leqslant 8000$	16

5 The data below shows the runs scored by Jeff over an entire cricket season.

17	5	22	13	6	0	15	20
14	7	28	36	13	28	9	18
2	23	12	27	5	22	3	0
32	8	13	25	9			

a Organise the data into the groups 0 - 9, 10 - 19, 20 - 29, 30 - 39.

b Use your grouped data to estimate the mean number of runs scored.

c Use the raw data to find the exact mean number of runs scored. How accurate was your estimate in **b**?

6 The manager of a bank decides to investigate the time customers wait to be served. The results for 300 customers are shown in the table alongside.

a Determine the value of p.

b Estimate the mean waiting time.

c What percentage of customers waited for at least 5 minutes?

Waiting time (t min)	*Frequency*
$0 \leqslant t < 1$	p
$1 \leqslant t < 2$	42
$2 \leqslant t < 3$	50
$3 \leqslant t < 4$	78
$4 \leqslant t < 5$	60
$5 \leqslant t < 6$	30
$6 \leqslant t < 7$	16

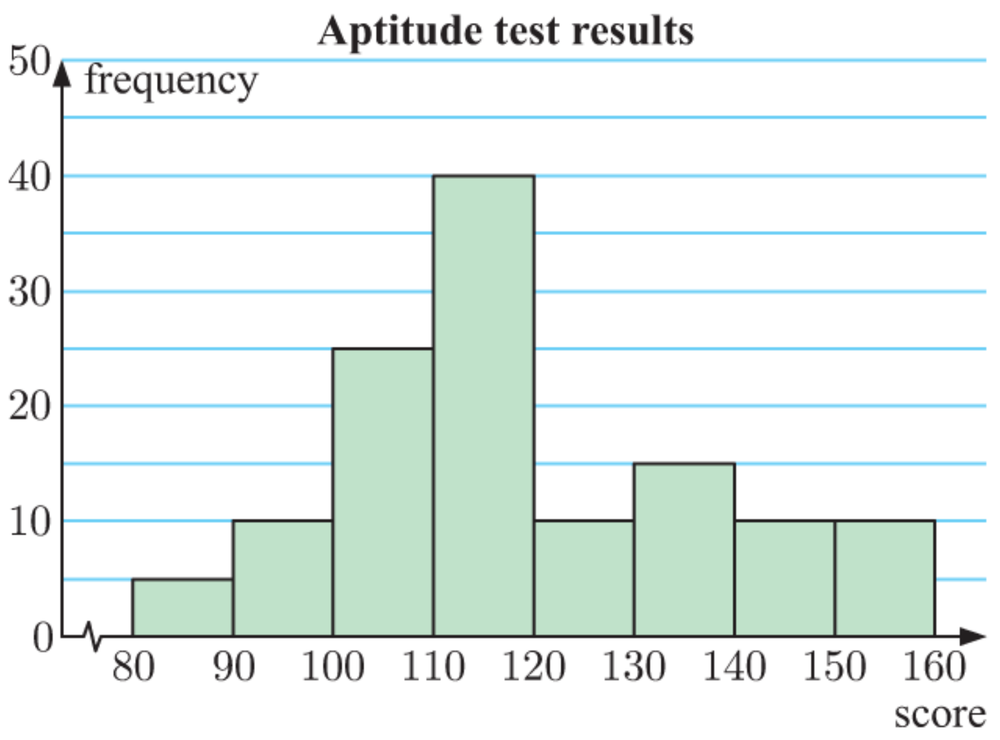
7 This frequency histogram illustrates the results of an aptitude test given to a group of people seeking positions in a company.

a How many people took the test?

b Estimate the mean score for the test.

c What fraction of the people scored less than 100 for the test?

d What percentage of the people scored more than 130 for the test?



E MEASURING THE SPREAD OF DATA

Consider the following statements:

- The mean height of 20 boys in a Year 12 class was found to be 175 cm.
- A carpenter used a machine to cut 20 planks of length 175 cm.

Even though the means of the two data sets are the same, there will clearly be a greater *variation* in the heights of boys than in the lengths of the planks.

Commonly used statistics that measure the spread of a data set are:

- the **range**
- the **variance**
- the **interquartile range**
- the **standard deviation**.

We will look at variance and standard deviation later in the Chapter.

THE RANGE

The **range** is the difference between the **maximum** data value and the **minimum** data value.

$$\textbf{range} = \textbf{maximum} - \textbf{minimum}$$

As a statistic for discussing the spread of a data set, the range is not considered to be particularly reliable. This is because it only uses two data values. It may be influenced by extreme values or outliers.

However, the range is useful for purposes such as choosing class intervals.

Example 5 **Self Tutor**

The weight, in kilograms, of the pumpkins in Herb's crop are:
2.3, 3.1, 2.7, 4.1, 2.9, 4.0, 3.3, 3.7, 3.4, 5.1, 4.3, 2.9, 4.2

Find the range of the data.

Range = maximum − minimum
= 5.1 − 2.3
= 2.8 kg

THE INTERQUARTILE RANGE

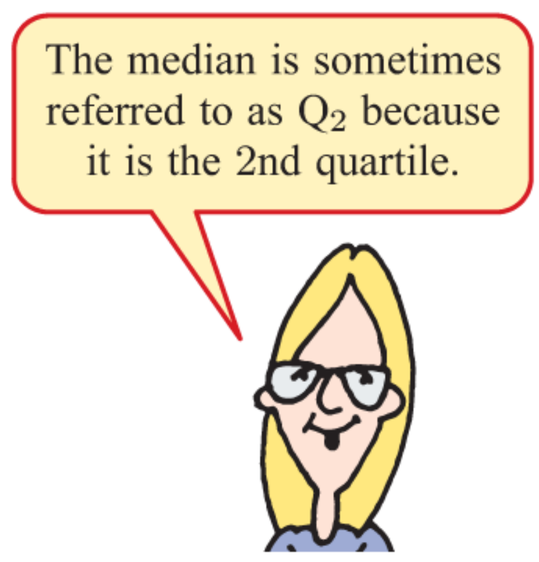
The median divides the ordered data set into two halves, and these halves are divided in half again by the **quartiles**.

The middle value of the *lower* half is called the **lower quartile** (Q_1).

The middle value of the *upper* half is called the **upper quartile** (Q_3).

The **interquartile range** (**IQR**) is the range of the middle half of the data.

interquartile range = upper quartile − lower quartile
$$\mathbf{IQR = Q_3 - Q_1}$$



Example 6 **Self Tutor**

For the data set 5 5 7 3 8 2 3 4 6 5 7 6 4, find:

a the median **b** Q_1 and Q_3 **c** the interquartile range.

The ordered data set is: 2 3 3 4 4 5 5 5 6 6 7 7 8 (13 data values)

a Since $n = 13$, $\frac{n+1}{2} = 7$ ∴ the median is the 7th data value.

~~2 3 3 4 4 5~~ 5 ~~5 6 6 7 7 8~~

∴ median = 5

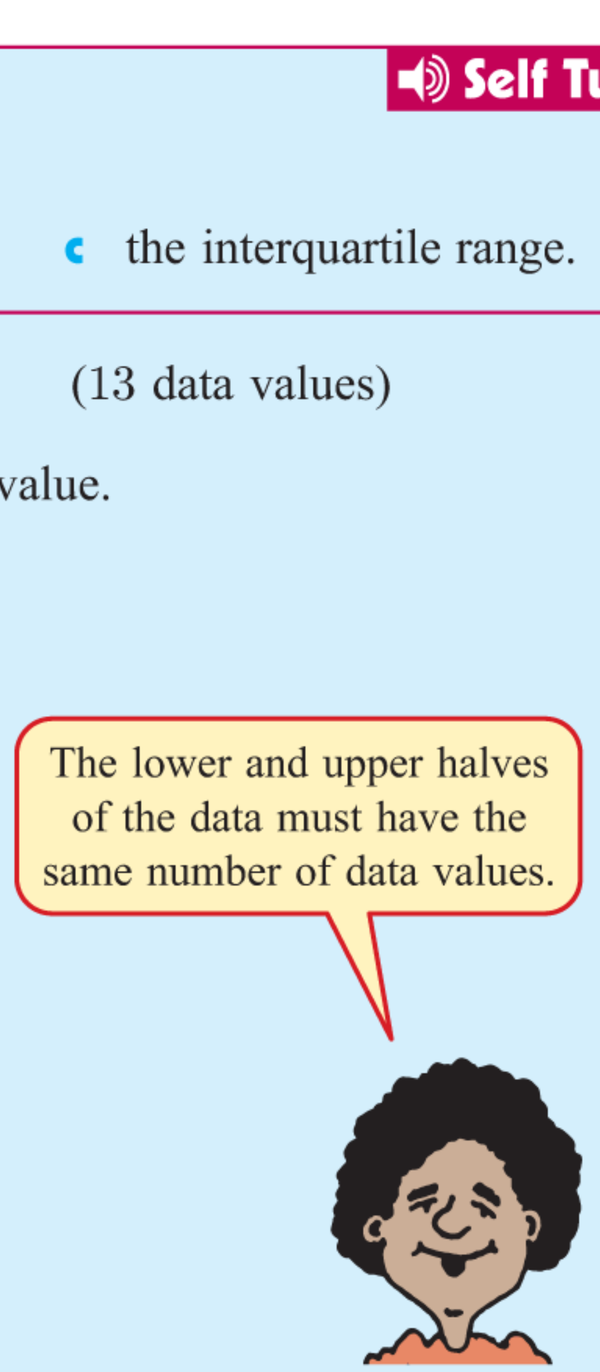
b Since the median is a data value we now ignore it and split the remaining data into two:

lower half: 2 3 3 4 4 5 upper half: 5 6 6 7 7 8

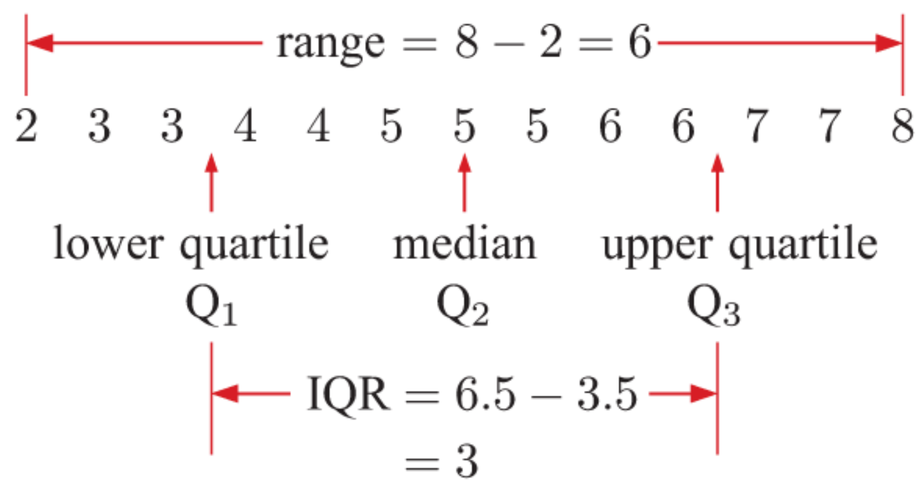
Q_1 = median of lower half = $\frac{3+4}{2} = 3.5$

Q_3 = median of upper half = $\frac{6+7}{2} = 6.5$

c IQR = $Q_3 - Q_1$
= 6.5 − 3.5
= 3



Notice how the data set in **Example 6** can be summarised:



Example 7 **Self Tutor**

For the data set 12 24 17 10 16 29 22 18 32 20, find:

a the median **b** Q_1 and Q_3 **c** the interquartile range.

The ordered data set is: 10 12 16 17 18 20 22 24 29 32 (10 data values)

a Since $n = 10$, $\frac{n+1}{2} = 5.5$ $\therefore$ the median is the average of the 5th and 6th data values.

~~10 12 16 17~~ 18 20 ~~22 24 29 32~~

$\therefore$ median $= \frac{\text{5th value} + \text{6th value}}{2} = \frac{18+20}{2} = 19$

b We have an even number of data values, so we include all data values when we split the data set into two:

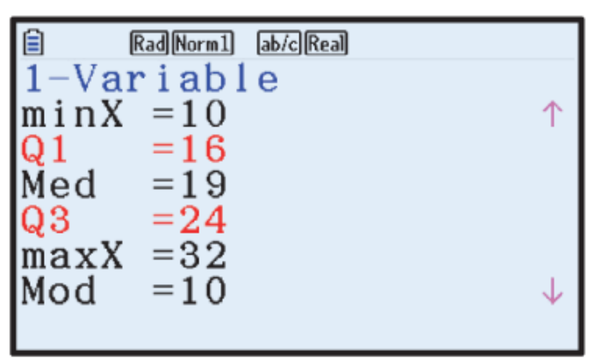
lower half: 10 12 16 17 18 upper half: 20 22 24 29 32

Q_1 = median of lower half = 16

Q_3 = median of upper half = 24

c IQR $= Q_3 - Q_1 = 24 - 16 = 8$

Technology can be used to help calculate the interquartile range. Your graphics calculator gives the values of Q_1 and Q_3, from which the interquartile range is found using $\text{IQR} = Q_3 - Q_1$.



GRAPHICS CALCULATOR INSTRUCTIONS

STATISTICS PACKAGE

EXERCISE 12E

1 For each of the following data sets, make sure the data is ordered and then find:

i the median **ii** the lower and upper quartiles

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

a 5, 6, 9, 10, 11, 13, 15, 16, 18, 20, 21

b 7, 7, 10, 13, 14, 15, 18, 19, 21, 21, 23, 24, 24, 26

c 21, 24, 19, 32, 15, 43, 38, 29

d 32, 45, 26, 28, 52, 57, 41, 69, 33, 20

Check your answers using your graphics calculator.

2 Natalie and Karen scored the following numbers of goals in their last 12 netball games:

Natalie:	29	48	34	39	26	49	28	36	41	46	29	46
Karen:	28	50	38	24	12	47	25	20	44	21	48	22

a Find the range and interquartile range for:

i Natalie **ii** Karen.

b Who was the more consistent netball player?

3 Jane and Ashley's monthly telephone bills are shown below:

Jane:	\$35,	\$47,	\$29,	\$38,	\$29,	\$34,	\$42,	\$29,	\$36,	\$40,	\$36,	\$31
Ashley:	\$19,	\$24,	\$26,	\$19,	\$23,	\$40,	\$35,	\$59,	\$32,	\$42,	\$26,	\$24

a Find the mean and median for each data set.

b Find the range and interquartile range for each data set.

c Which person generally pays more for their telephone bills?

d Which person has the greater variability in their telephone bills?

4 **a** Find the range and interquartile range for the data set:

9	12	7	15	14	22	18	11	20	15
20	10	13	67	25	18	11	7	14	19

b Identify the outlier in the data set.

c Recalculate the range and interquartile range with the outlier removed.

d Which measure of spread is more affected by the outlier?

5 Derrick and Gareth recorded the number of minutes they slept each night for 15 nights:

Derrick:	420, 435, 440, 415, 380, 400, 430, 450, 210, 445, 425, 445, 450, 420, 425
Gareth:	360, 420, 460, 430, 480, 340, 450, 490, 500, 460, 330, 470, 340, 480, 370

a Calculate the range and interquartile range for each data set.

b Which person's data has the lower:

i range **ii** interquartile range?

c Which measure of spread is more appropriate for determining who is generally the more consistent sleeper? Explain your answer.

6 a, b, c, d, e, f, g, h, i, j, k, l, and m are 13 data values which have been arranged in *ascending* order.

a Which variable represents the median?

b Write down an expression for:

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

7 A data set has the following known measures of centre and spread:

Measure	median	mode	range	interquartile range
Value	9	7	13	6

Find the new value of each of these measures if every member of the data set is:

a increased by 2 **b** doubled.

DISCUSSION

Consider the data set:

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

1 Calculate Q_1, Q_3, and the IQR:

- by hand
- using a spreadsheet.
- using your graphics calculator

Do you get the same answers in all 3 cases?

2 If there is an odd number of data values, some statistical packages calculate quartiles by *including* the median in each half of the data.

a Check to see whether your spreadsheet calculates quartiles this way.

b Does this method necessarily change the *interpretation* of the calculated values?

c Are statistical packages that do this necessarily "wrong"?

F BOX AND WHISKER DIAGRAMS

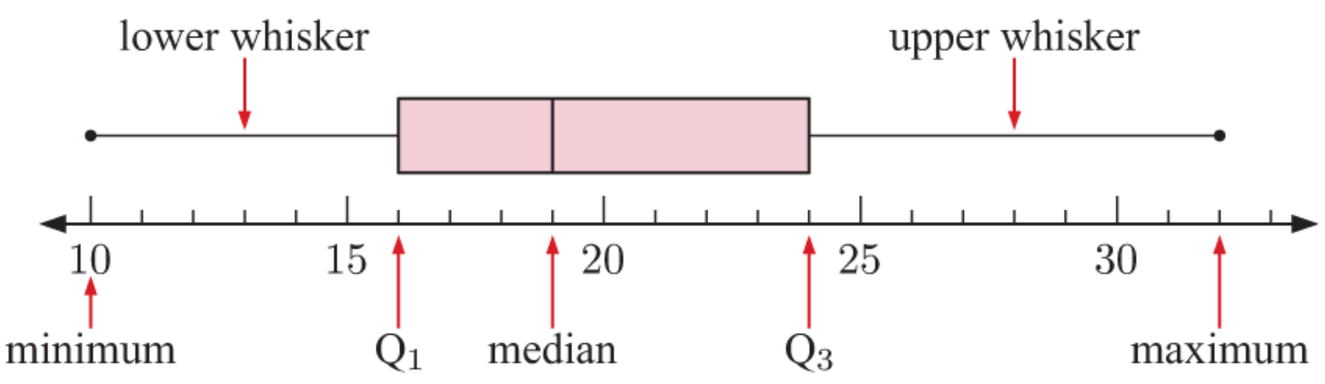
A **box and whisker diagram** or simply **box plot** is a visual display of some of the descriptive statistics of a data set. It shows:

- the minimum value
- the lower quartile (Q_1)
- the median (Q_2)
- the upper quartile (Q_3)
- the maximum value

} These five numbers form the **five-number summary** of the data set.

For the data set in **Example 7** on page **321**, the five-number summary and box plot are:

minimum = 10
Q_1 = 16
median = 19
Q_3 = 24
maximum = 32

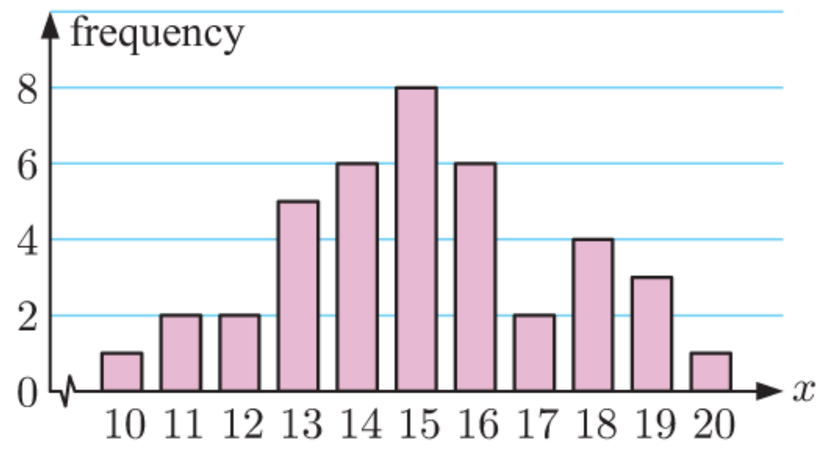


You should notice that:

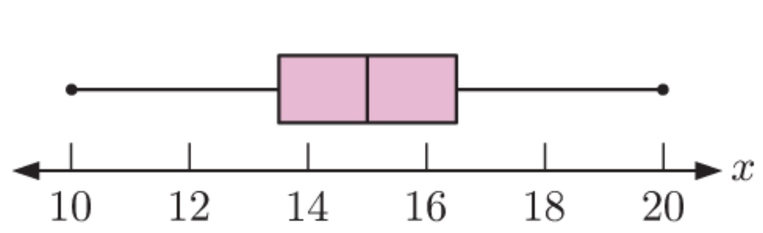
- The rectangular box represents the "middle" half of the data set.
- The lower whisker represents the 25% of the data with smallest values.
- The upper whisker represents the 25% of the data with greatest values.

INTERPRETING A BOX PLOT

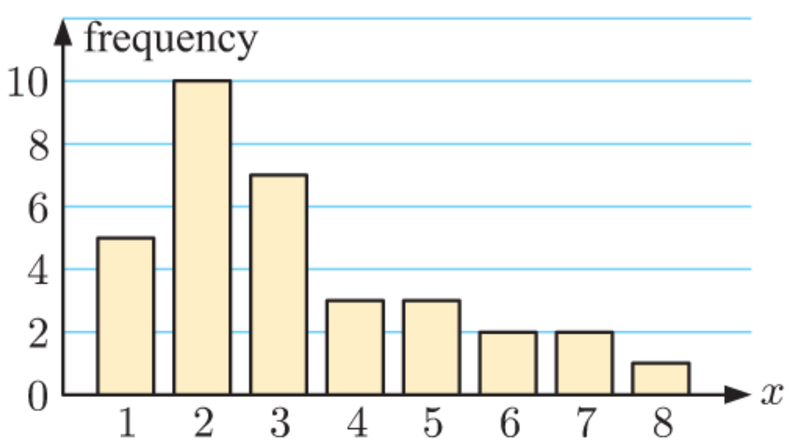
A set of data with a **symmetric distribution** will have a symmetric box plot.



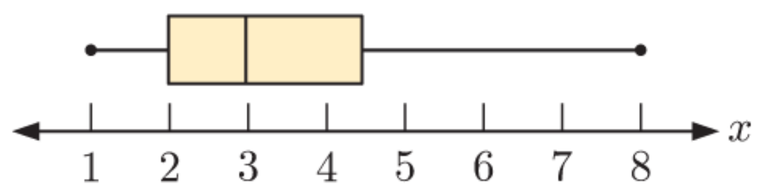
The whiskers of the box plot are the same length and the median line is in the centre of the box.



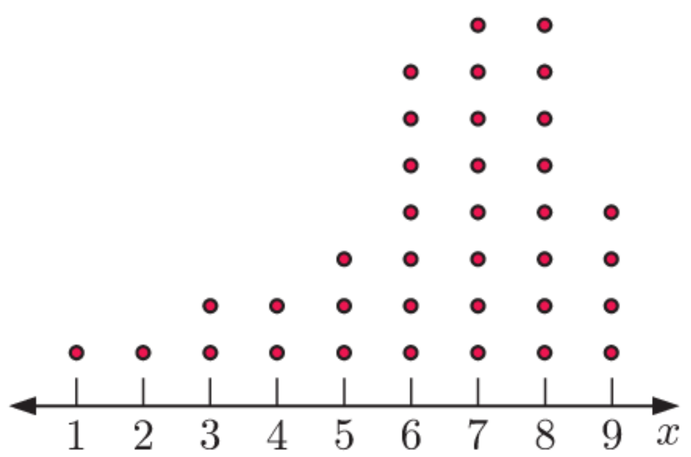
A set of data which is **positively skewed** will have a positively skewed box plot.



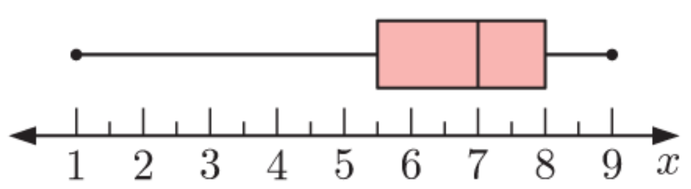
The upper whisker is longer than the lower whisker and the median line is closer to the left hand side of the box.



A set of data which is **negatively skewed** will have a negatively skewed box plot.



The lower whisker is longer than the upper whisker and the median line is closer to the right hand side of the box.



Example 8 — **Self Tutor**

Consider the data set: 8 2 3 9 6 5 3 2 2 6 2 5 4 5 5 6

STATISTICS PACKAGE

a Find the five-number summary for this data.

b Draw a box plot for the data.

c Find the:

i range **ii** interquartile range.

d Find the percentage of data values which are less than 3.

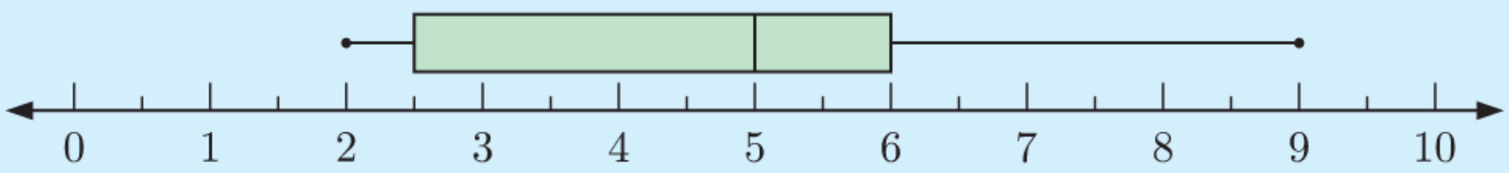
a The ordered data set is:

2 2 2 2 3 3 4 5 5 5 5 6 6 6 8 9 {16 data values}

$Q_1 = 2.5$ median $= 5$ $Q_3 = 6$

The five-number summary is: minimum $= 2$, $Q_1 = 2.5$; median $= 5$, $Q_3 = 6$; maximum $= 9$

b



c **i** range $=$ maximum $-$ minimum
$= 9 - 2$
$= 7$

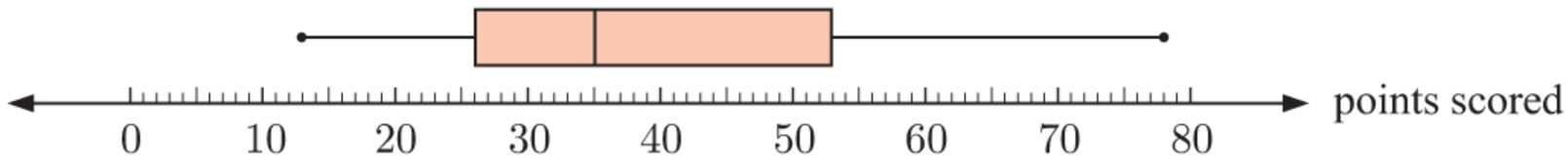
ii IQR $= Q_3 - Q_1$
$= 6 - 2.5$
$= 3.5$

d Using the ordered data set in **a**, 4 out of 16 data values are less than 3.
$\therefore$ 25% of the data values are less than 3.

Part **d** can be seen from the original data set. We cannot read it straight from the box plot because the box plot does not tell us that all of the data values are integers.

EXERCISE 12F

1 The box plot below summarises the points scored by a basketball team.



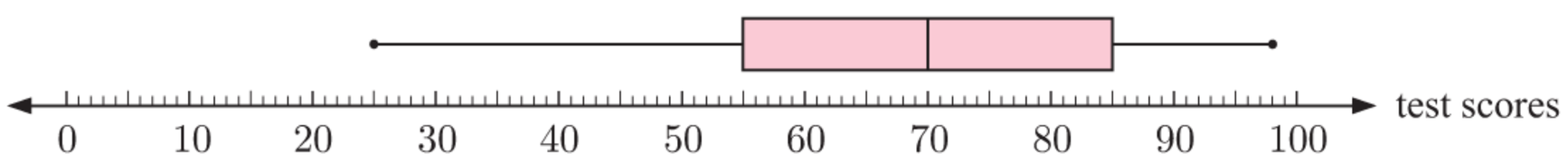
a Locate:

i the median **ii** the maximum value **iii** the minimum value

iv the upper quartile **v** the lower quartile.

b Calculate: **i** the range **ii** the interquartile range.

2 The box plot below summarises the class results for a test out of 100 marks.



a Copy and complete the following statements about the test results:

i The highest mark scored for the test was ..... and the lowest mark was .....

ii Half of the class scored a mark greater than or equal to .....

iii The top 25% of the class scored at least ..... marks.

iv The middle half of the class had scores between ..... and .....

b Find the range of the data set.

c Find the interquartile range of the data set.

3 For the following data sets:

STATISTICS PACKAGE

GRAPHICS CALCULATOR INSTRUCTIONS

i Construct a five-number summary.
ii Draw a box plot.
iii Find the range.
iv Find the interquartile range.

a 3, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 10

b 3, 7, 0, 1, 4, 6, 8, 8, 8, 9, 7, 5, 6, 8, 7, 8, 8, 2, 9

c 23, 44, 31, 33, 26, 17, 30, 35, 47, 31, 51, 47, 20, 31, 28, 49, 26, 49

4 Enid counts the number of beans in 33 pods. Her results are:

5, 8, 10, 4, 2, 12, 6, 5, 7, 7, 5, 5, 5, 13, 9, 3, 4, 4, 7, 8, 9, 5, 5, 4, 3, 6, 6, 6, 6, 9, 8, 7, 6

a Find the median, lower quartile, and upper quartile of the data set.

b Find the interquartile range of the data set.

c Draw a box plot of the data set.

5 Ranji counts the number of bolts in several boxes and tabulates the data as follows:

Number of bolts	33	34	35	36	37	38	39	40
Frequency	1	5	7	13	12	8	0	1

a Find the five-number summary for this data set.

b Find the: **i** range **ii** IQR.

c Draw a box plot of the data set.

GAME

Click on the icon to play a card game about box plots.

CARD GAME

G OUTLIERS

We have seen that **outliers** are extraordinary data that are separated from the main body of the data. However, we have so far identified outliers rather informally by looking at the data directly, or at a column graph of the data.

A commonly used test to identify outliers involves the calculation of upper and lower boundaries:

- **upper boundary = upper quartile + 1.5 × IQR**
  Any data larger than the upper boundary is an outlier.
- **lower boundary = lower quartile − 1.5 × IQR**
  Any data smaller than the lower boundary is an outlier.

Outliers are marked with an asterisk on a box plot. It is possible to have more than one outlier at either end.

Each whisker extends to the last value that is not an outlier.

Example 9 **Self Tutor**

Test the following data for outliers. Hence construct a box plot for the data.

3, 7, 8, 8, 5, 9, 10, 12, 14, 7, 1, 3, 8, 16, 8, 6, 9, 10, 13, 7

The ordered data set is:

1 3 3 5 6 7 7 7 8 8 8 8 9 9 10 10 12 13 14 16 $\{n = 20\}$

$\text{min} = 1$ $\quad Q_1 = 6.5$ $\quad \text{median} = 8$ $\quad Q_3 = 10$ $\quad \text{max} = 16$

$\text{IQR} = Q_3 - Q_1 = 3.5$

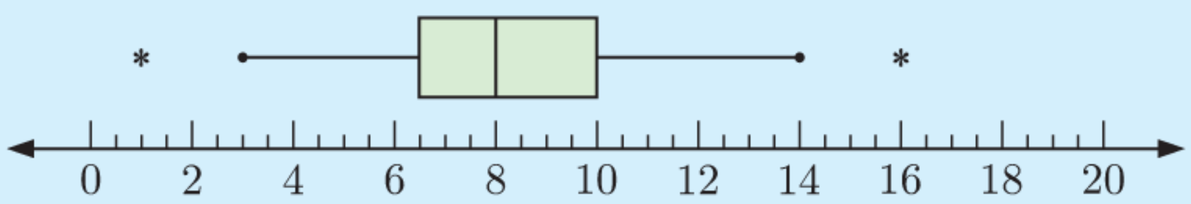
Test for outliers:

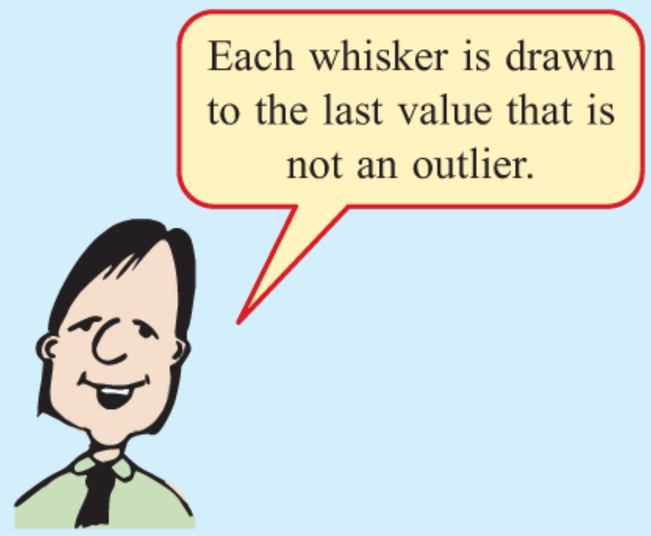
upper boundary $= \text{upper quartile} + 1.5 \times \text{IQR}$
$= 10 + 1.5 \times 3.5$
$= 15.25$

and

lower boundary $= \text{lower quartile} - 1.5 \times \text{IQR}$
$= 6.5 - 1.5 \times 3.5$
$= 1.25$

16 is above the upper boundary, so it is an outlier.
1 is below the lower boundary, so it is an outlier.





EXERCISE 12G

1 A data set has lower quartile $= 31.5$, median $= 37$, and upper quartile $= 43.5$.

a Calculate the interquartile range for this data set.

b Calculate the boundaries that identify outliers.

c The smallest values of the data set are 13 and 20. The largest values are 52 and 55. Which of these are outliers?

d Draw a box plot of the data set.

2 James goes bird watching for 25 days. The number of birds he sees each day are:

12, 5, 13, 16, 8, 10, 12, 18, 9, 11, 14, 14,
22, 9, 10, 7, 9, 11, 13, 7, 10, 6, 13, 3, 8

a Find the median, lower quartile, and upper quartile of the data set.

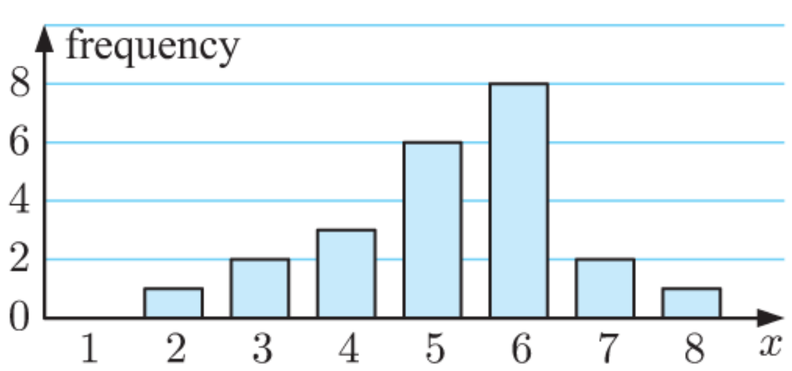
b Find the interquartile range of the data set.

c Find the lower and upper boundaries, and hence identify any outliers.

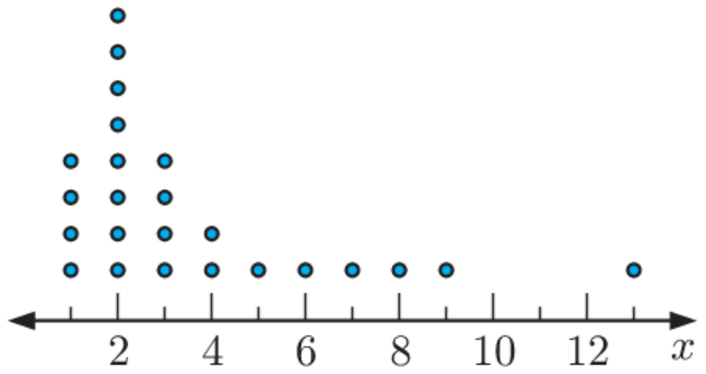
d Draw a box plot of the data set.

3 Match each graph with its box plot:

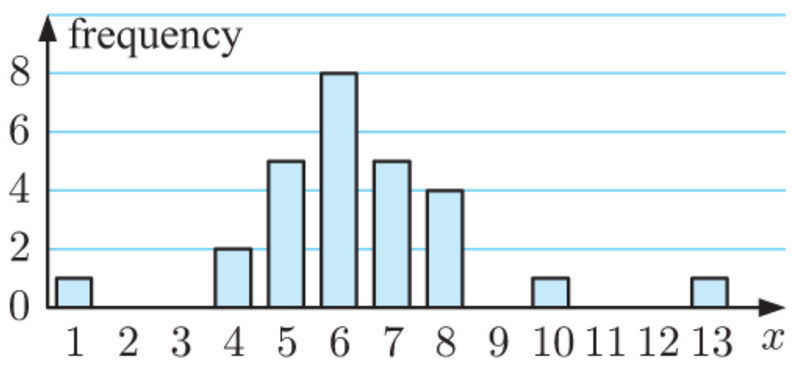
a



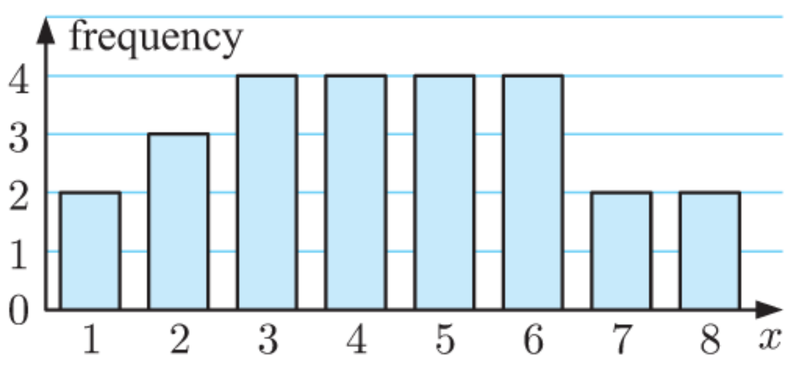
b



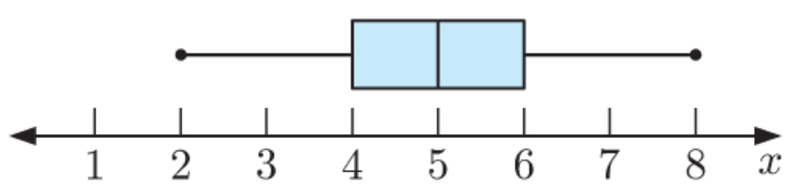
c



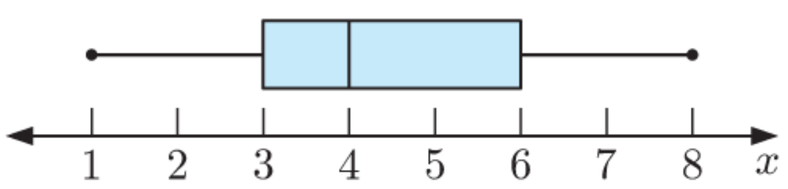
d



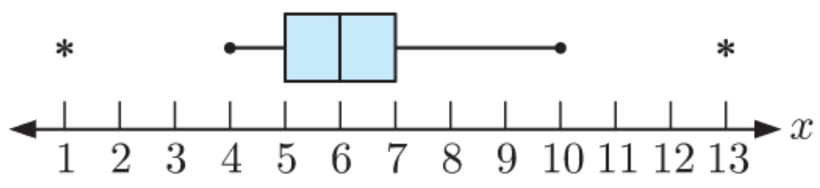
A



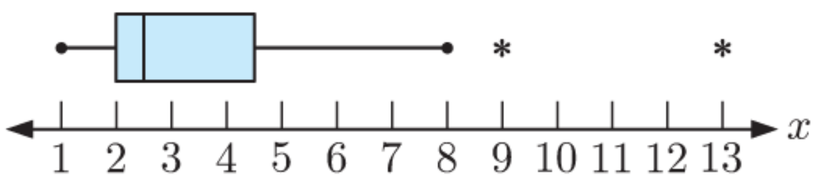
B



C



D



4 The data below shows the number of properties sold by a real estate agent each week in 2018:

2	2	1	3	2	2	2	2	1	4	1	5	1
1	1	2	2	2	3	1	7	2	2	2	0	2
2	4	4	3	3	1	0	2	4	1	2	1	3
0	2	3	1	2	1	3	4	2	2	2	1	3

a Draw a column graph to display the data.

b From the column graph, does the data appear to have any outliers?

c Calculate the upper and lower boundaries to test for outliers and hence check your answer to **b**.

d Construct a box plot for the data.

DISCUSSION

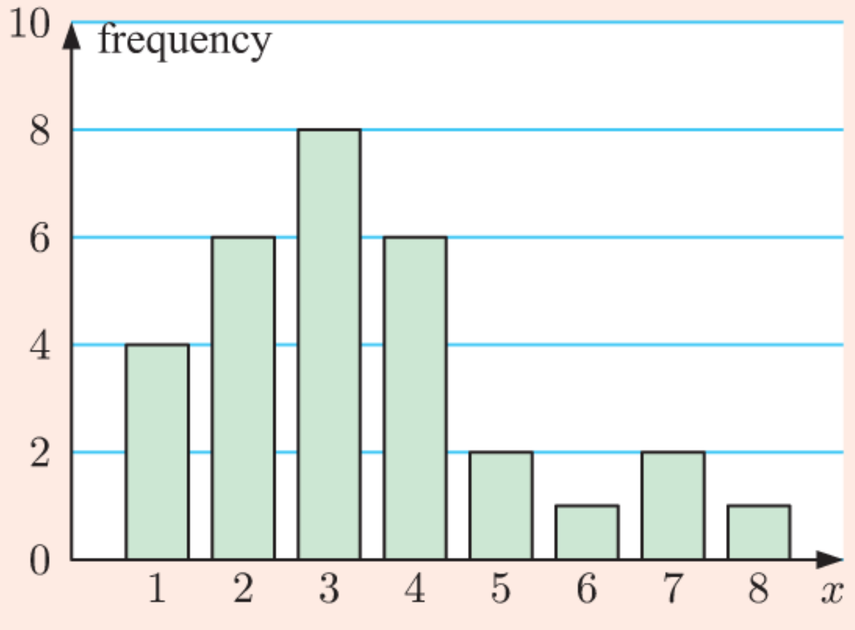
Consider the data in the column graph alongside.

The data has $Q_1 = 2$ and $Q_3 = 4$.

1 Calculate the IQR and the upper and lower boundaries for outliers.

2 According to the upper and lower boundaries, the data value "8" is an outlier. Do you agree that "8" should be considered an outlier of this data set?

3 Do you think the given rule for detecting outliers will be effective for data that is heavily skewed?



H PARALLEL BOX AND WHISKER DIAGRAMS

A **parallel box and whisker diagram** or **parallel box plot** enables us to make a *visual comparison* of the distributions of two data sets. We can easily compare descriptive statistics such as their median, range, and interquartile range.

Example 10 **Self Tutor**

A hospital trialling a new anaesthetic has collected data on how long the new and old drugs take before the patient becomes unconscious. They wish to know which drug acts faster and which is more predictable.

Old drug times (s): 8, 12, 9, 8, 16, 10, 14, 7, 5, 21, 13, 10, 8, 10, 11, 8, 11, 9, 11, 14

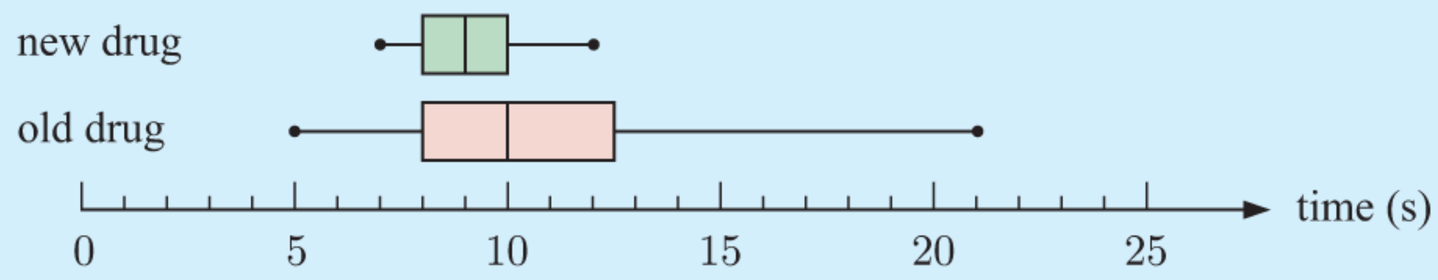
New drug times (s): 8, 12, 7, 8, 12, 11, 9, 8, 10, 8, 10, 9, 12, 8, 8, 7, 10, 7, 9, 9

Draw a parallel box plot for the data sets and use it to compare the two drugs.

The five-number summaries are:

For the old drug: $\min = 5$, $Q_1 = 8$, median $= 10$, $Q_3 = 12.5$, $\max = 21$

For the new drug: $\min = 7$, $Q_1 = 8$, median $= 9$, $Q_3 = 10$, $\max = 12$



Using the median, 50% of the time the new drug takes 9 seconds or less, compared with 10 seconds for the old drug. So, the new drug is generally a little quicker.

Comparing the spreads:

range for old drug $= 21 - 5 = 16$

range for new drug $= 12 - 7 = 5$

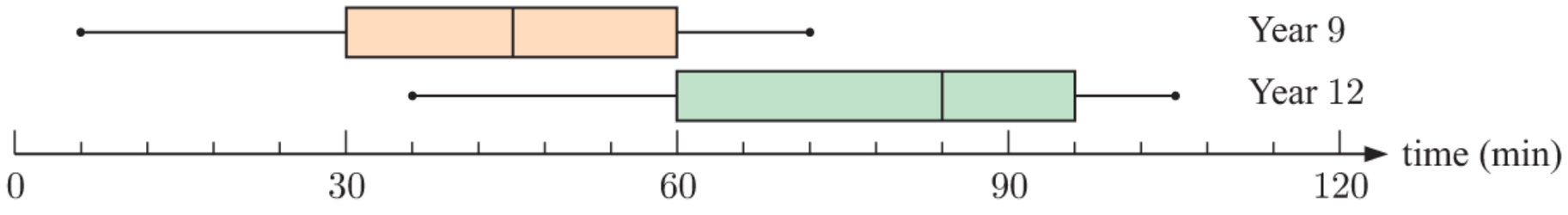
IQR for old drug $= Q_3 - Q_1 = 12.5 - 8 = 4.5$

IQR for new drug $= Q_3 - Q_1 = 10 - 8 = 2$

The new drug times are less "spread out" than the old drug times, so the new drug is more predictable.

EXERCISE 12H

1 The following parallel box plots compare the times students in Years 9 and 12 spend on homework.



a Copy and complete:

Statistic	*Year 9*	*Year 12*
minimum		
Q_1		
median		
Q_3		
maximum		

b For each group, determine the:

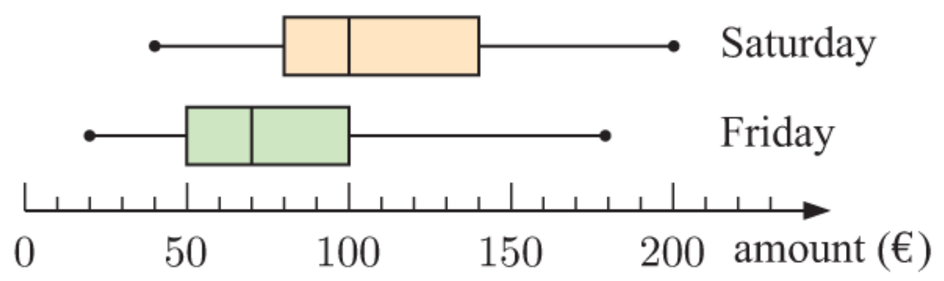
i range **ii** interquartile range.

c Determine whether the following statements are true or false, or if there is not enough information to tell:

i On average, Year 12 students spend about twice as much time on homework as Year 9 students.

ii Over 25% of Year 9 students spend less time on homework than all Year 12 students.

2 The amounts of money withdrawn from an ATM were recorded on a Friday and on a Saturday. The results are displayed on the parallel box plot shown.

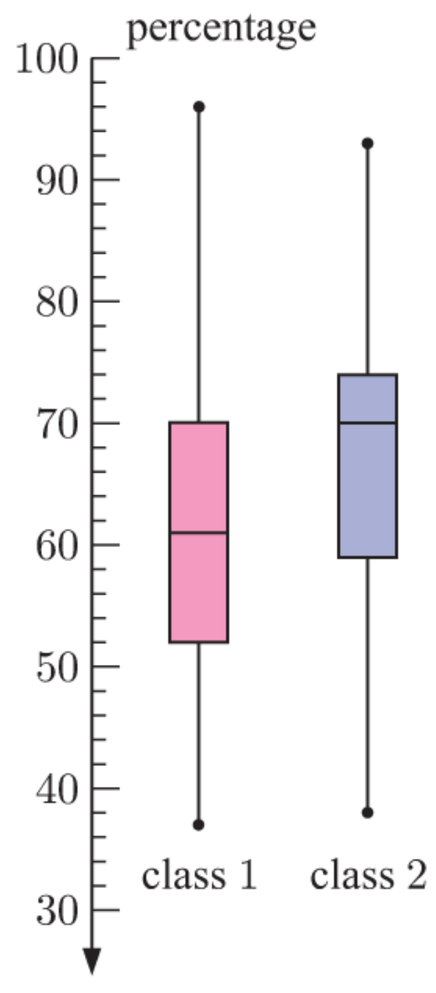


a Find the five-number summary for each data set.

b For each data set, determine the

i range **ii** interquartile range.

3 After the final examination, the results of two classes studying the same subject were compiled in this parallel box plot.



a In which class was:

i the highest mark **ii** the lowest mark

iii there a larger spread of marks?

b Find the interquartile range of class 1.

c Find the range of class 2.

d Students who scored at least 70% received an achievement award. Find the percentage of students who received an award in:

i class 1 **ii** class 2.

e Describe the distribution of marks in:

i class 1 **ii** class 2.

f Copy and complete:
The students in class ..... generally scored higher marks.
The marks in class ..... were more varied.

4 The data below are the durations, in minutes, of Kirsten and Erika's last 25 phone calls.

Kirsten:	1.7	2.0	3.9	3.4	0.9	1.4	2.5	1.1	5.1	4.2	1.5	2.6	0.8
	4.0	1.5	1.0	2.9	3.2	2.5	0.8	1.8	3.1	6.9	2.3	1.2	
Erika:	2.0	4.8	1.2	7.5	3.2	5.7	3.9	0.2	2.7	6.8	3.4	5.2	3.2
	7.2	1.7	11.5	4.0	2.4	3.7	4.2	10.7	3.0	2.0	0.9	5.7	

a Find the five-number summary for each data set.

b Display the data in a parallel box plot.

c Compare and comment on the distributions of the data.

5 Emil and Aaron play in the same handball team and are fierce but friendly rivals when it comes to scoring. During a season, the numbers of goals they scored in each match were:

Emil:	1	6	2	0	3	4	1	4	2	3	0	3	2	4	3	4	3	3
	3	4	2	4	3	2	3	3	0	5	3	5	3	2	4	3	4	3
Aaron:	7	2	4	8	1	3	4	2	3	0	5	3	5	2	3	1	2	0
	4	3	4	0	3	3	0	2	5	1	1	2	2	5	1	4	0	1

a Is the variable discrete or continuous?

b Enter the data into a graphics calculator or statistics package.

c Produce a column graph for each data set.

d Describe the shape of each distribution.

e Compare the measures of the centre of each distribution.

f Compare the spreads of each distribution.

g Draw a parallel box plot for the data.

h What conclusions can be drawn from the data?

6 A manufacturer of light globes claims that their new design has a 20% longer life than those they are presently selling. Forty of each globe are randomly selected and tested. Here are the results to the nearest hour:

Old type:	103	96	113	111	126	100	122	110	84	117	103	113	104	104
	111	87	90	121	99	114	105	121	93	109	87	118	75	111
	87	127	117	131	115	116	82	130	113	95	108	112		
New type:	146	131	132	160	128	119	133	117	139	123	109	129	109	131
	191	117	132	107	141	136	146	142	123	144	145	125	164	125
	133	124	153	129	118	130	134	151	145	131	133	135		

a Is the variable discrete or continuous?

b Enter the data into a graphics calculator or statistics package. Compare the measures of centre and spread.

c Draw a parallel box plot.

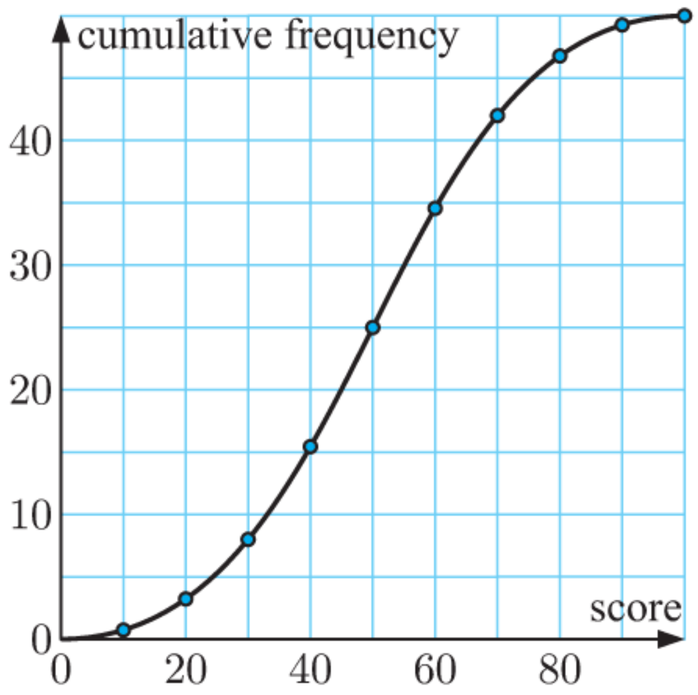
d Describe the shape of each distribution.

e What conclusions, if any, can be drawn from the data?

I CUMULATIVE FREQUENCY GRAPHS

If we want to know the number or proportion of scores that lie above or below a particular value, we add a **cumulative frequency** column to a **frequency table**, and use a graph called a **cumulative frequency graph** to represent the data.

The cumulative frequencies are plotted and the points joined by a smooth curve. This differs from an ogive or cumulative frequency polygon where neighbouring points are joined by straight lines.



PERCENTILES

A **percentile** is the score below which a certain percentage of the data lies.

For example:

- the 85th percentile is the score below which 85% of the data lies.
- If your score in a test is the 95th percentile, then 95% of the class have scored less than you.

Notice that:

- the **lower quartile** ($\mathbf{Q_1}$) is the 25th percentile
- the **median** ($\mathbf{Q_2}$) is the 50th percentile
- the **upper quartile** ($\mathbf{Q_3}$) is the 75th percentile.

A cumulative frequency graph provides a convenient way to find percentiles.

Example 11 **Self Tutor**

The data shows the results of the women's marathon at the 2008 Olympics, for all competitors who finished the race.

a Add a cumulative frequency column to the table.

b Represent the data on a cumulative frequency graph.

c Use your graph to estimate the:

i median finishing time

ii number of competitors who finished in less than 155 minutes

iii percentage of competitors who took more than 159 minutes to finish

iv time taken by a competitor who finished in the top 20% of runners completing the marathon.

Time (t min)	*Frequency*
$146 \leqslant t < 148$	8
$148 \leqslant t < 150$	3
$150 \leqslant t < 152$	9
$152 \leqslant t < 154$	11
$154 \leqslant t < 156$	12
$156 \leqslant t < 158$	7
$158 \leqslant t < 160$	5
$160 \leqslant t < 168$	8
$168 \leqslant t < 176$	6

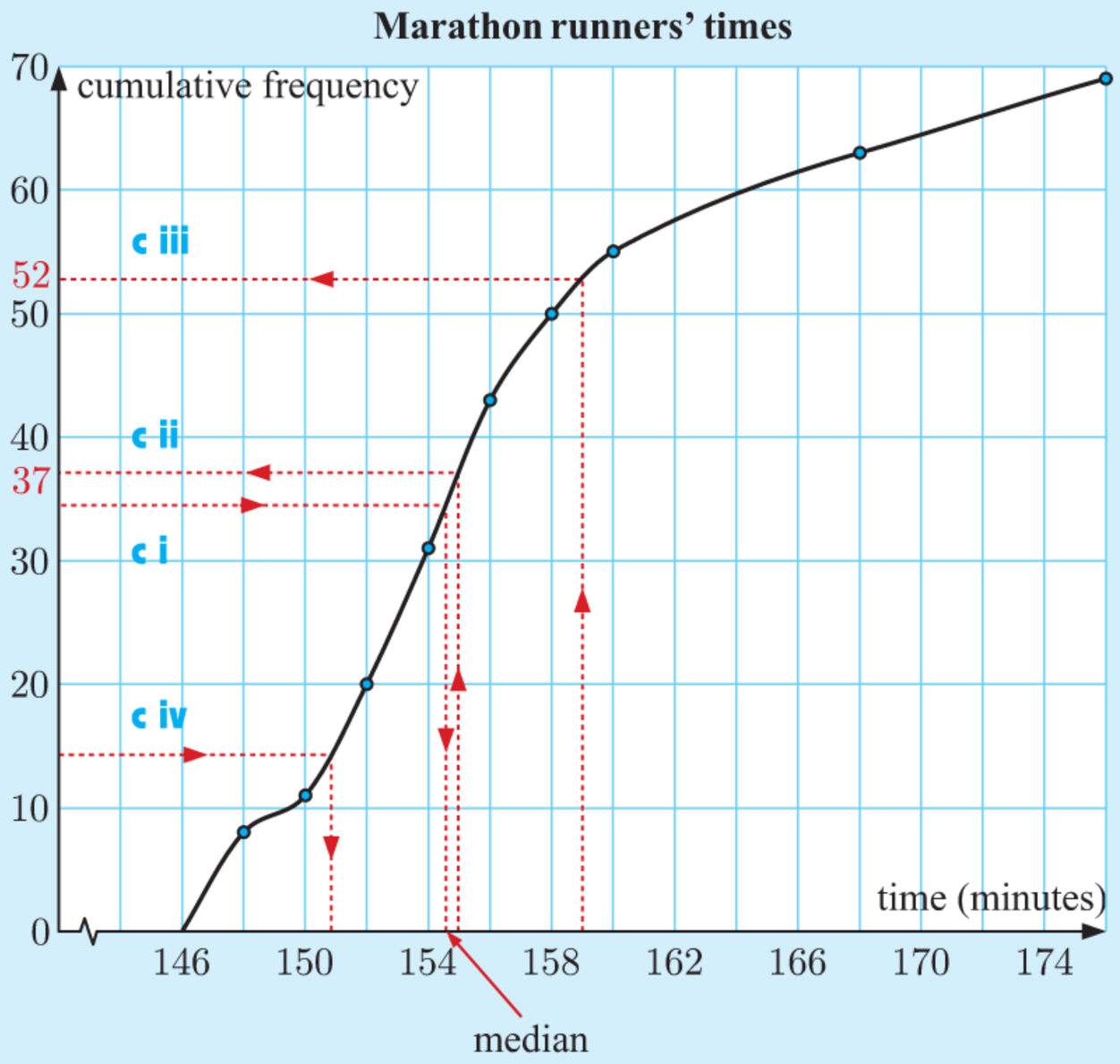
a

Time (t min)	*Frequency*	*Cumulative frequency*
$146 \leqslant t < 148$	8	8
$148 \leqslant t < 150$	3	11
$150 \leqslant t < 152$	9	20
$152 \leqslant t < 154$	11	31
$154 \leqslant t < 156$	12	43
$156 \leqslant t < 158$	7	50
$158 \leqslant t < 160$	5	55
$160 \leqslant t < 168$	8	63
$168 \leqslant t < 176$	6	69

$8 + 3 = 11$ competitors completed the marathon in less than 150 minutes.

50 competitors completed the marathon in less than 158 minutes.

b



The cumulative frequency gives a *running total* of the number of runners finishing by a given time.

c **i** The median is the 50th percentile. As 50% of 69 is 34.5, we start with the cumulative frequency 34.5 and find the corresponding time.

The median ≈ 154.5 min.

ii Approximately 37 competitors took less than 155 min to complete the race.

iii $69 - 52 = 17$ competitors took more than 159 min.

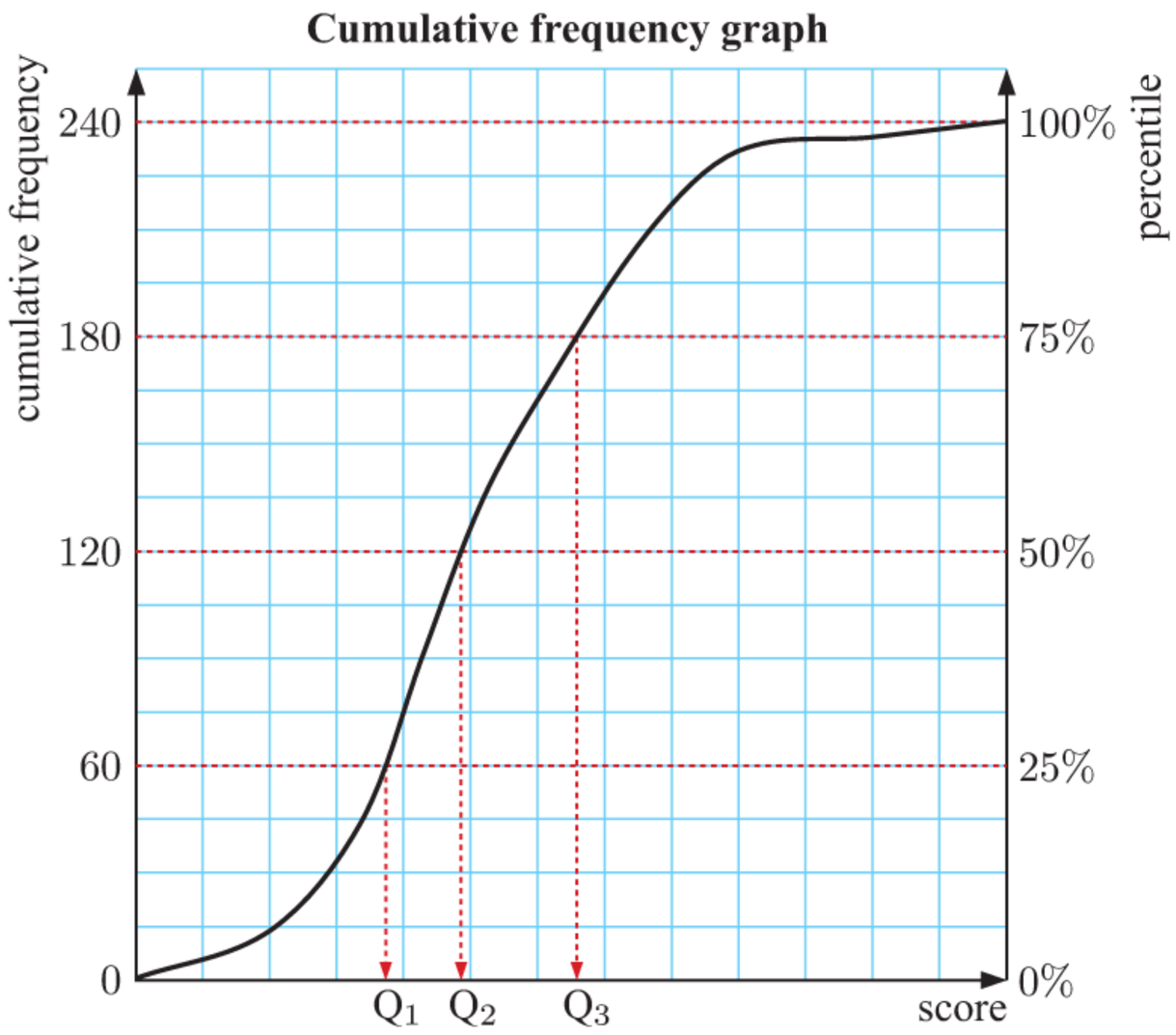
$\therefore \quad \frac{17}{69} \approx 24.6\%$ took more than 159 min.

iv As 20% of 69 is 13.8, we start with the cumulative frequency 14 and find the corresponding time.

The top 20% of competitors took less than 151 min.

Another way to calculate percentiles is to add a separate scale to the cumulative frequency graph.

For example, on the graph alongside, the cumulative frequency is read from the axis on the left side, and each value corresponds to a percentile on the right side.



EXERCISE 12I

1 The examination scores of a group of students are shown in the table.

Score (x)	*Frequency*
$10 \leqslant x < 20$	2
$20 \leqslant x < 30$	5
$30 \leqslant x < 40$	7
$40 \leqslant x < 50$	21
$50 \leqslant x < 60$	36
$60 \leqslant x < 70$	40
$70 \leqslant x < 80$	27
$80 \leqslant x < 90$	9
$90 \leqslant x < 100$	3

a Draw a cumulative frequency graph for the data.

b Find the median examination mark.

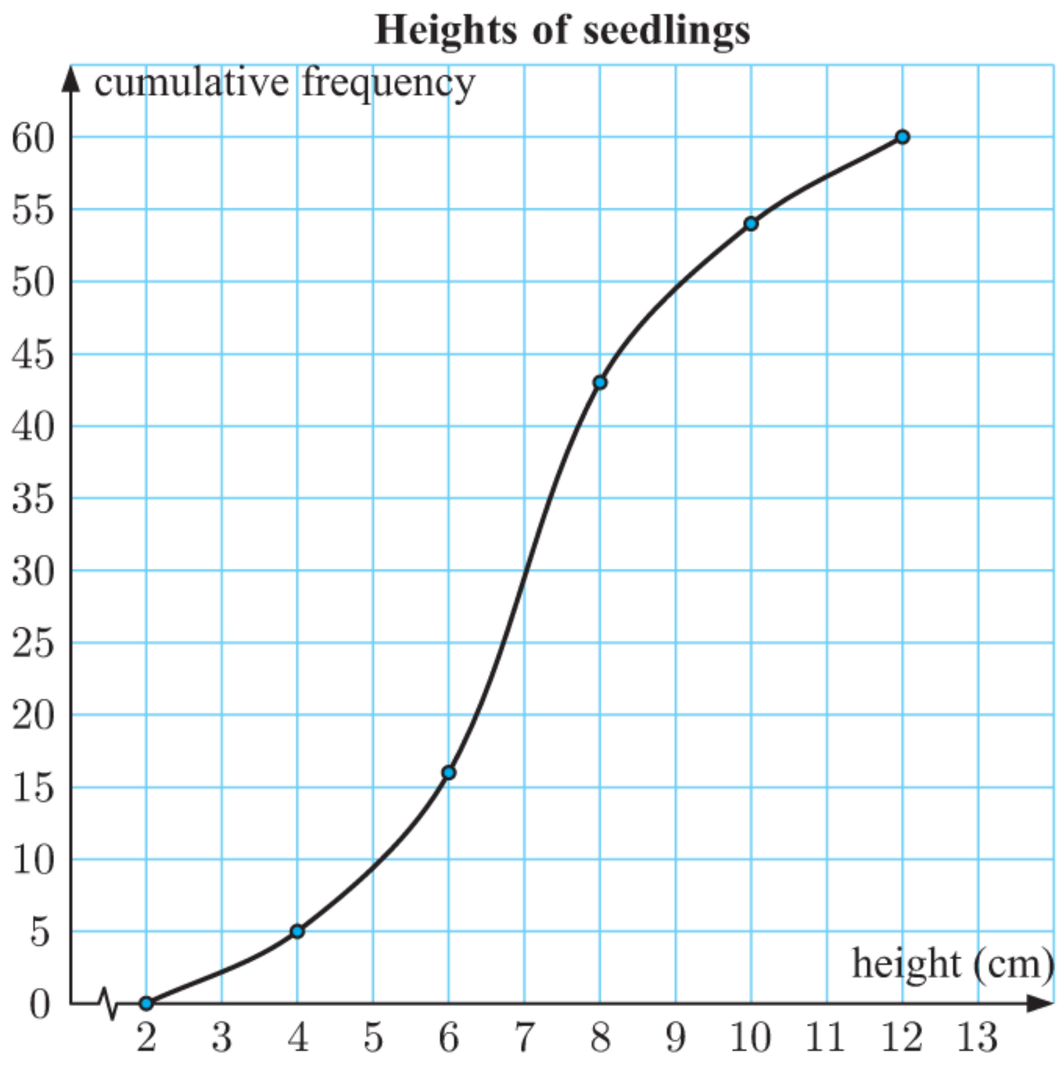
c How many students scored 65 marks or less?

d How many students scored at least 50 but less than 70 marks?

e If the pass mark was 45, how many students failed?

f If the top 16% of students were awarded credits, what was the credit mark?

2 A botanist has measured the heights of 60 seedlings and has presented her findings on this cumulative frequency graph.



a How many seedlings have heights of 5 cm or less?

b What percentage of seedlings are taller than 8 cm?

c Find the median height.

d Find the interquartile range for the heights.

e Find the 90th percentile for the data and explain what this value represents.

3 The following table summarises the age groups of car drivers involved in accidents in a city for a given year.

Age (x years)	*Number of accidents*
$16 \leqslant x < 20$	59
$20 \leqslant x < 25$	82
$25 \leqslant x < 30$	43
$30 \leqslant x < 35$	21
$35 \leqslant x < 40$	19
$40 \leqslant x < 50$	11
$50 \leqslant x < 60$	24
$60 \leqslant x < 80$	41

a Draw a cumulative frequency graph for the data.

b Estimate the median age of the drivers involved in accidents.

c Estimate the percentage of drivers involved in accidents who had an age of 23 or less.

d Estimate the probability that a driver involved in an accident is aged:

i 27 years or less **ii** 27 years.

4 The following data are the lengths of 30 trout caught in a lake during a fishing competition. The measurements were rounded *down* to the next centimetre.

31	38	34	40	24	33	30	36	38	32	35	32	36	27	35
40	34	37	44	38	36	34	33	31	38	35	36	33	33	28

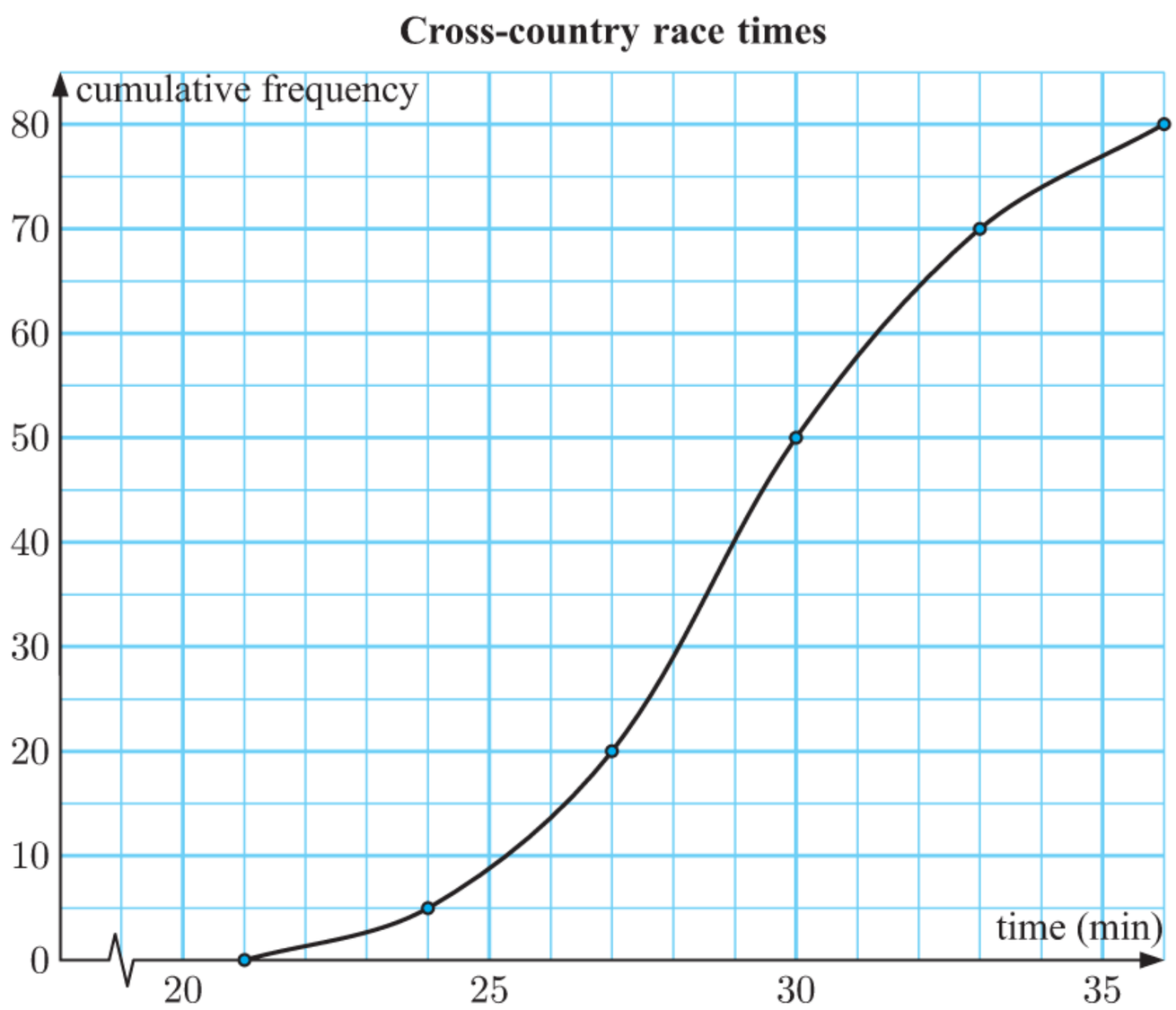
a Construct a cumulative frequency table for trout lengths, x cm, using the intervals $24 \leqslant x < 27$, $27 \leqslant x < 30$, and so on.

b Draw a cumulative frequency graph for the data.

c Hence estimate the median length.

d Use the original data to find its median and compare your answer with **c**.

5 The following cumulative frequency graph displays the performances of 80 competitors in a cross-country race.



a Find the lower quartile.

b Find the median.

c Find the upper quartile.

d Find the IQR.

e Estimate the 40th percentile.

f Use the cumulative frequency curve to complete the following table:

Time (t min)	$21 \leqslant t < 24$	$24 \leqslant t < 27$	$27 \leqslant t < 30$	$30 \leqslant t < 33$	$33 \leqslant t < 36$
Number of competitors					

6 The table shows the lifetimes of a sample of electric light globes.

Life (l hours)	*Number of globes*
$0 \leqslant l < 500$	5
$500 \leqslant l < 1000$	17
$1000 \leqslant l < 2000$	46
$2000 \leqslant l < 3000$	79
$3000 \leqslant l < 4000$	27
$4000 \leqslant l < 5000$	4

a Draw a cumulative frequency graph for the data.

b Estimate the median life of a globe.

c Estimate the percentage of globes which had a life of 2700 hours or less.

d Estimate the number of globes which had a life between 1500 and 2500 hours.

7 The following frequency distribution was obtained by asking 50 randomly selected people to measure the length of their feet. Their answers are given to the nearest centimetre.

Foot length (cm)	20	21	22	23	24	25	26	27	28	29	30
Frequency	1	1	0	3	5	13	17	7	2	0	1

a Between what limits are lengths rounded to 20 cm?

b Rewrite the frequency table to show the data in the class intervals you have just described.

c Hence draw a cumulative frequency graph for the data.

d Estimate:

i the median foot length

ii the number of people with foot length 26 cm or more.

J VARIANCE AND STANDARD DEVIATION

The problem with using the range and the IQR as measures of spread or dispersion is that both of them only use two values in their calculation. As a result, some data sets can have their spread characteristics hidden when only the range or IQR are quoted.

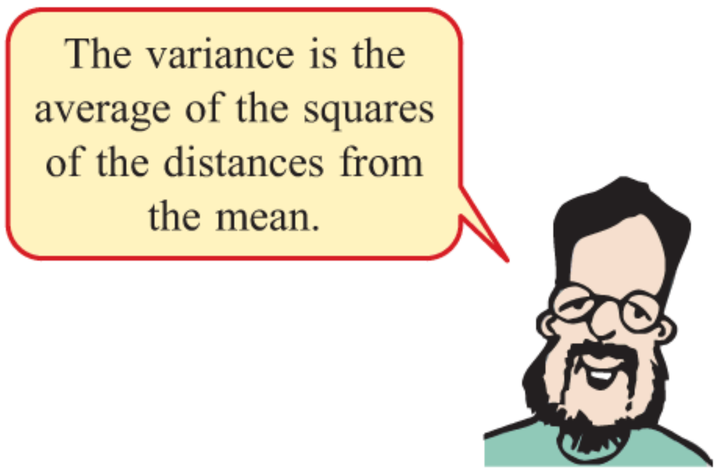
So we need to consider alternative measures of spread which take into account all data values of a data set. We therefore turn to the **variance** and **standard deviation**.

POPULATION VARIANCE AND STANDARD DEVIATION

The **population variance** of a data set $\{x_1, x_2, x_3, ..., x_n\}$ is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

where μ is the population mean
and n is the number of data values.

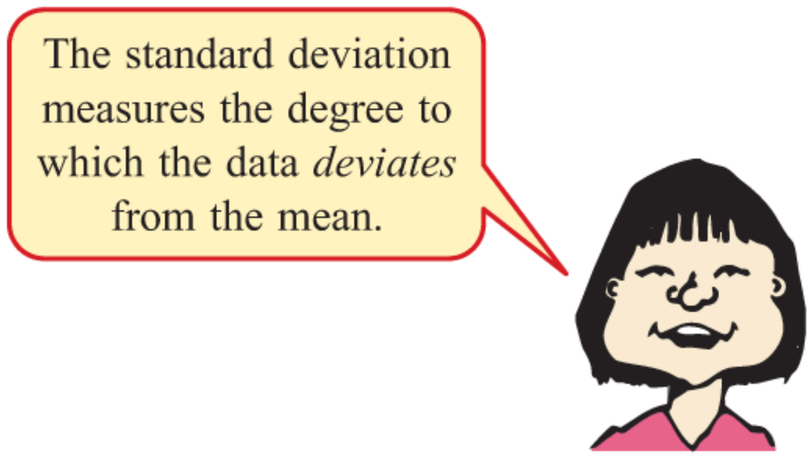


We observe that if the data values x_i are situated close together around the mean μ, then the values $(x_i - \mu)^2$ will be small, and so the variance will be small.

The **standard deviation** is the square root of the variance.

The **population standard deviation** of a data set $\{x_1, x_2, x_3, ..., x_n\}$ is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}.$$

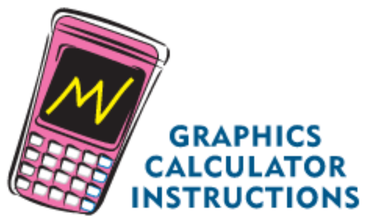


The square root in the standard deviation is used to correct the units. For example, if x_i is the weight of a student in kg, the variance σ^2 would be in kg^2, and σ would be in kg.

The standard deviation is a **non-resistant** measure of spread. This is due to its dependence on the mean and because extreme data values will give large values for $(x_i - \mu)^2$. It is only a useful measure if the distribution is approximately symmetrical.

The IQR and percentiles are more appropriate tools for measuring spread if the distribution is considerably skewed.

In this course you are only expected to use technology to calculate variance and standard deviation. However, we present both methods in the following Example to help you understand standard deviations better.



Example 12 **Self Tutor**

Find the population variance and standard deviation for the data set:

3 12 8 15 7

The mean $\mu = \dfrac{3+12+8+15+7}{5} = 9$

The population variance $\sigma^2 = \dfrac{\sum(x-\mu)^2}{n}$

$= \dfrac{86}{5}$

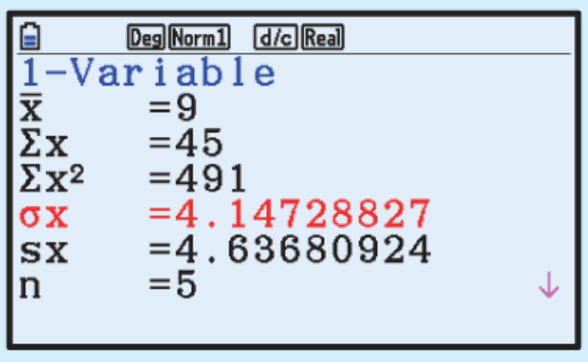
$= 17.2$

The population standard deviation $\sigma = \sqrt{17.2}$

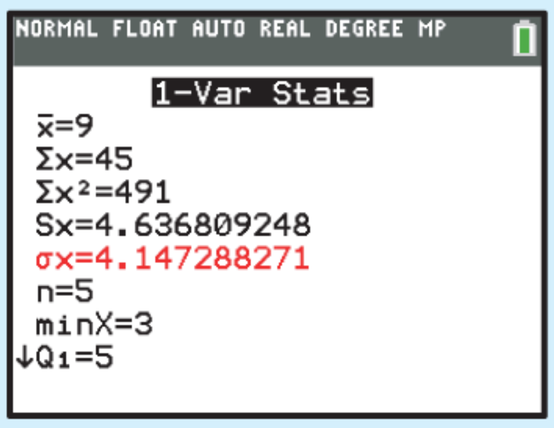
≈ 4.15

x	$x - \mu$	$(x-\mu)^2$
3	-6	36
12	3	9
8	-1	1
15	6	36
7	-2	4
	Total	86

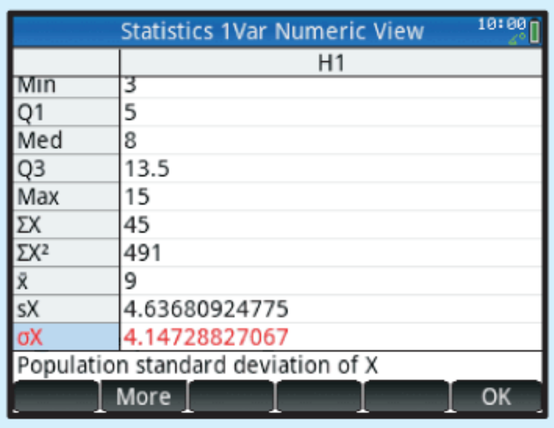
Casio fx-CG50



TI-84 Plus CE



HP Prime



SAMPLE VARIANCE AND STANDARD DEVIATION

In the screenshots on the previous page, you can see that near the population standard deviation σ, there is a statistic s with a similar value.

Technically, if we have data which is a *sample* from a large population, the **sample standard deviation** s provides a better estimate for the actual population standard deviation than if we use the formula for σ on the sample. However, this is beyond the scope of this course.

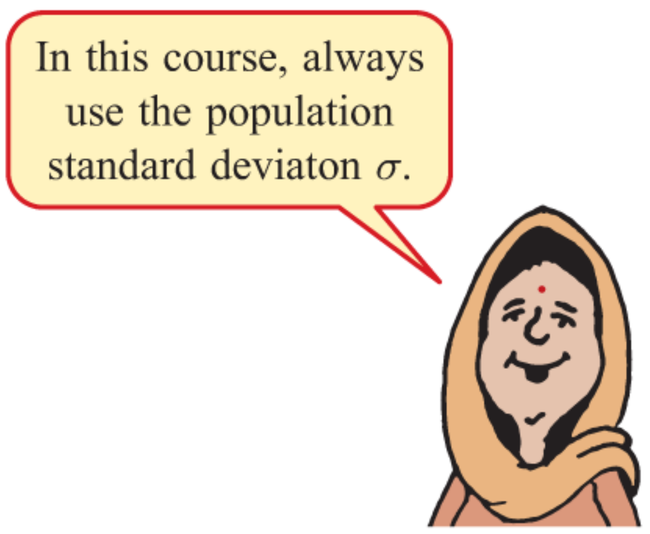
In this course you are expected to calculate all standard deviations as though they were populations. The important thing is that you recognise that the two statistics exist, and that you are using the correct one.

EXERCISE 12J

1 Consider the following data sets:

Data set A: 10 7 5 8 10

Data set B: 4 12 11 14 1 6



a Show that each data set has mean 8.

b Which data set appears to have the greater spread? Explain your answer.

c Find the population variance and standard deviation of each data set. Use technology to check your answers.

2 Skye recorded the number of pets owned by each student in her class.

0 2 3 1 2 4 0 0 1 5 2 3 6
2 3 1 1 0 4 1 1 0 2 1 2 0

a Use technology to find the population standard deviation of the data.

b Find the population variance of the data.

3 The ages of members of an Olympic water polo team are: 22, 25, 23, 28, 29, 21, 20, 26.

a Calculate the mean and population standard deviation for this group.

b The same team members are chosen to play in the next Olympic Games 4 years later. Calculate the mean and population standard deviation of their ages at the next Olympic Games.

c Comment on your results in general terms.

4 A hospital selected a sample of 20 patients and asked them how many glasses of water they had consumed that day. The results were:

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

Find the population standard deviation of the data.

5 Kylie is interested in the ages of spectators at a rugby match. She selects a sample of 30 spectators and records their ages.

17 24 30 10 42 48 37 19 28 53 29 40 11 21 9
43 22 59 46 52 31 13 7 26 32 47 22 15 26 42

Calculate the mean and population standard deviation of the data.

6 Danny and Jennifer recorded how many hours they spent on homework each day for 14 days.

Danny: $3\frac{1}{2}$, $3\frac{1}{2}$, 4, $2\frac{1}{2}$, 3, $3\frac{1}{2}$, 3, $1\frac{1}{2}$, 3, 4, $2\frac{1}{2}$, 4, 4, 3

Jennifer: $2\frac{1}{2}$, 1, $2\frac{1}{2}$, 2, 2, $2\frac{1}{2}$, $1\frac{1}{2}$, 2, 2, $2\frac{1}{2}$, 2, 2, 2, $1\frac{1}{2}$

a Calculate the mean number of hours each person spent on homework.

b Which person generally studies for longer?

c Calculate the population standard deviation σ for each data set.

d Which person studies more consistently?

7 Tyson wants to compare the swimming speeds of boys and girls at his school. He randomly selects 10 boys and 10 girls, and records the time, in seconds, each person takes to swim two laps of the 25 m school pool.

Boys: 32.2, 26.4, 35.6, 30.8, 28.5, 40.2, 27.3, 38.9, 29.0, 31.3

Girls: 36.2, 33.5, 28.1, 39.8, 31.6, 35.7, 37.3, 36.0, 39.7, 29.8

a Copy and complete the table:

	Boys	*Girls*
Mean $\overline{x}$		
Median		
Standard deviation σ		
Range		

b Which group:

i generally swims faster **ii** has the greater spread of swimming speeds?

c How could Tyson improve the reliability of his findings?

8 Two baseball coaches compare the number of runs scored by their teams in their last ten games:

Rockets	0	10	1	9	11	0	8	5	6	7
Bullets	4	3	4	1	4	11	7	6	12	5

a Show that the two teams have the same mean and range of runs scored.

b Which team's performance do you suspect is more variable over the period? Check your answer by finding the population standard deviation for each data set.

c Does the range or the standard deviation give a better indication of variability?

9 The number of visitors to a museum and an art gallery each day during December are shown.

Museum: 1108, 1019, 850, 1243, 1100, 923, 964, 847, 918, 820, 781
963, 814, 881, 742, 911, 1101, 952, 864, 943, 1087, 1132
906, 1050, 0, 826, 986, 1040, 1127, 1084, 981

Art gallery: 1258, 1107, 1179, 1302, 1236, 1386, 1287, 1313, 1269, 1332, 1094
1153, 1275, 1168, 1086, 1276, 1342, 1153, 1227, 1305, 1187, 1249
1300, 1156, 1074, 1168, 1299, 1257, 1134, 1259, 1366

a For each data set, calculate the:

i mean **ii** population standard deviation.

b Which place had the greater spread of visitor numbers?

c **i** Identify the outlier in the *Museum* data.

ii Give a reason why this outlier may have occurred.

iii Do you think it is reasonable to remove the outlier when comparing the numbers of visitors to these places? Explain your answer.

iv Recalculate the mean and population standard deviation with the outlier removed.

v Discuss the effect of the outlier on the population standard deviation.

10 Find the population standard deviation of this data set.

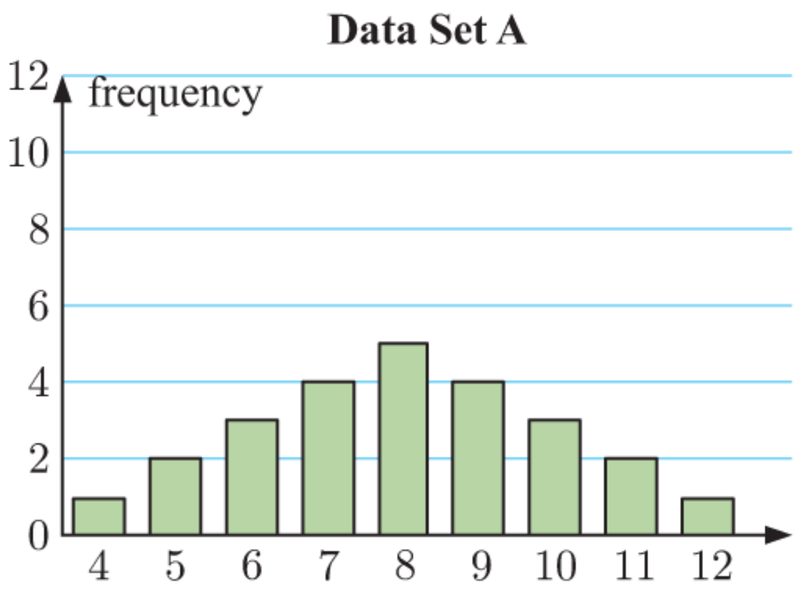
Value	*Frequency*
3	1
4	3
5	11
6	5

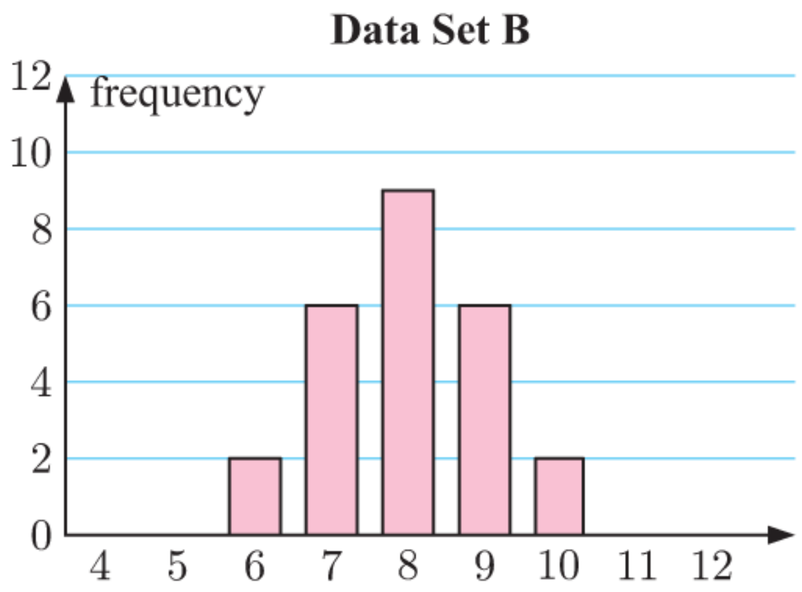
11 The table shows the ages of squash players at the Junior National Squash Championship.

Age	11	12	13	14	15	16	17	18
Frequency	2	1	4	5	6	4	2	1

Find the mean and population standard deviation of the ages.

12 The column graphs show two distributions:





a By looking at the graphs, which distribution appears to have wider spread?

b Find the mean of each data set.

c Find the population standard deviation for each data set. Comment on your answers.

d The other measures of spread for the two data sets are given in the table. In what way does the standard deviation give a better description of how the data is distributed?

Data set	*Range*	*IQR*
A	8	3
B	4	2

13 The table alongside shows the results obtained by female and male students in a test out of 20 marks.

Score	*Females*	*Males*
12	0	1
13	0	0
14	0	2
15	0	3
16	2	4
17	6	2
18	5	0
19	1	1
20	1	0

a Looking at the table:

i Which group appears to have scored better in the test?

ii Which group appears to have a greater spread of scores?

Justify your answers.

b Calculate the mean and population standard deviation for each group.

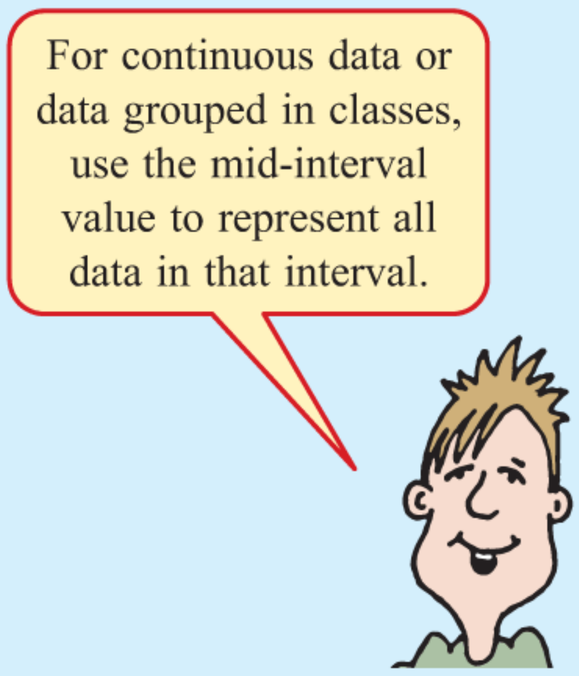
14 Brianna and Jess are conducting a survey of their class.
Brianna asked every student (including Jess and herself) how many children there are in their family.
Jess asked every student (including Brianna and herself) how many siblings or step-siblings they have.
How will their results compare in terms of mean and standard deviation? Explain your answer.

Example 13 **Self Tutor**

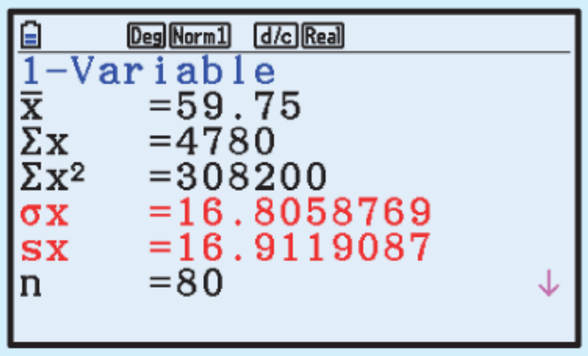
Estimate the standard deviation for this distribution of examination scores:

Mark	*Frequency*	*Mark*	*Frequency*
0 - 9	1	50 - 59	16
10 - 19	1	60 - 69	24
20 - 29	2	70 - 79	13
30 - 39	4	80 - 89	6
40 - 49	11	90 - 99	2

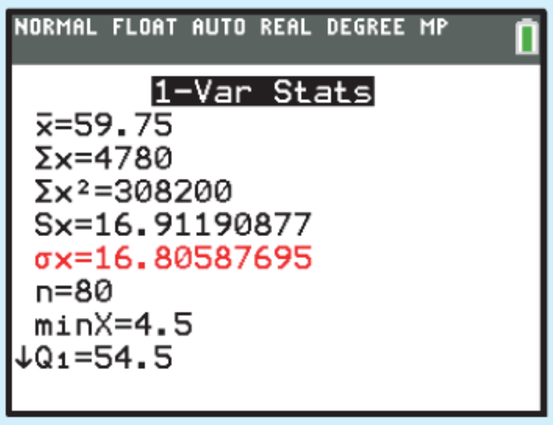
Class interval	*Mid-interval value*	*Frequency*
0 - 9	4.5	1
10 - 19	14.5	1
20 - 29	24.5	2
30 - 39	34.5	4
40 - 49	44.5	11
50 - 59	54.5	16
60 - 69	64.5	24
70 - 79	74.5	13
80 - 89	84.5	6
90 - 99	94.5	2



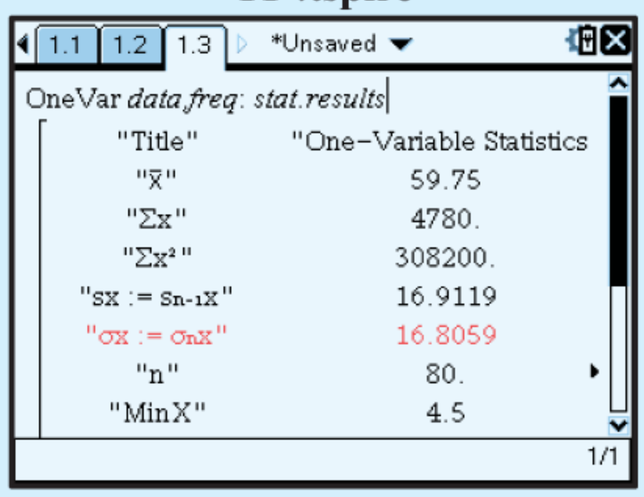
Casio fx-CG50



TI-84 Plus CE



TI-*n*spire



The standard deviation ≈ 16.8.

15 The lengths of 30 randomly selected 12-day old babies were measured and the following data obtained.

For the given data, estimate the:

a mean **b** standard deviation.

Length (L cm)	*Frequency*
$40 \leqslant L < 42$	1
$42 \leqslant L < 44$	1
$44 \leqslant L < 46$	3
$46 \leqslant L < 48$	7
$48 \leqslant L < 50$	11
$50 \leqslant L < 52$	5
$52 \leqslant L < 54$	2

16 A traffic survey revealed that the following numbers of vehicles passed through a suburban intersection in 5 minute intervals during the day.

Vehicles	1 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 - 30	31 - 35	36 - 40
Frequency	4	16	22	28	14	9	5	2

For the given data, estimate the:

a mean **b** standard deviation.

17 The weekly wages of 200 randomly selected steel workers are given alongside:

For the given data, estimate the:

a mean **b** standard deviation.

Wage (W)	*Number of workers*
$720 \leqslant W < 740$	17
$740 \leqslant W < 760$	38
$760 \leqslant W < 780$	47
$780 \leqslant W < 800$	57
$800 \leqslant W < 820$	18
$820 \leqslant W < 840$	10
$840 \leqslant W < 860$	10
$860 \leqslant W < 880$	3

18 The hours worked last week by 40 employees of a local clothing factory were as follows:

38 40 46 32 41 39 44 38 40 42 38 40 43 41
47 36 38 39 34 40 48 30 49 40 40 43 45 36
35 39 42 44 48 36 38 42 46 38 39 40

a Calculate the mean and standard deviation for this data.

b Now group the data into classes 30 - 33, 34 - 37, and so on. Calculate the mean and standard deviation using these groups. Examine any differences between the two sets of answers.

INVESTIGATION 3 TRANSFORMING DATA

In this Investigation we will explore the effects of transforming a data set on its mean and standard deviation.

We will use this data set as a basis: 4 2 3 3 5 2 9 7 3 5
2 1 5 3 6 6 3 3 6 7

What to do:

1 Calculate the mean and population standard deviation for the data set.

2 **a** Suppose we add 5 to each data value. Calculate the mean and population standard deviation for the new data set.

b What do you expect to happen to the mean and standard deviation if k is added to each value in a data set?

c Check your answer by:

i adding 11 to each data value **ii** subtracting 3 from each data value.

3 **a** Suppose we multiply each value in the original data set by 4. Calculate the mean and population standard deviation for the new data set.

b What do you expect to happen to the mean and standard deviation if each value in a data set is multiplied by a?

c Check your answer by:

i multiplying each value by 9 **ii** dividing each value by 4.

4 Suppose a data set $\{x_i\}$ has mean μ and standard deviation σ. Write down the mean and standard deviation for the data set:

a $\{ax_i\}$ **b** $\{x_i + k\}$ **c** $\{ax_i + k\}$

INVESTIGATION 4 ESTIMATING THE VARIANCE AND STANDARD DEVIATION OF A POPULATION

In this Investigation we consider the accuracy of using a sample to make inferences about a whole population. This will help you to see why statisticians have a subtly different formula for the standard deviation of a sample.

The Year 12 students at a school were asked to record how many minutes they spent travelling to school. The results were collected in a survey the following morning.

There are a total of 150 Year 12 students at the school, and these are split into 6 classes.

What to do:

1 Click on the icon to obtain a spreadsheet containing all of the responses to the survey.

SPREADSHEET

a Use the frequency table in the spreadsheet to draw a histogram for the data. Describe this distribution.

b The summary statistics in the spreadsheet are calculated using all of the survey responses, and hence are the *true* population values. Find the true population variance.

2 10 students were randomly selected from each class to form 6 samples. Their responses to the survey are shown below:

Sample 1:	10	14	16	9	16	15	15	21	9	21
Sample 2:	11	9	11	16	16	13	10	12	21	16
Sample 3:	12	10	14	7	13	11	21	20	15	9
Sample 4:	20	19	19	19	13	19	22	15	10	19
Sample 5:	19	13	23	11	17	4	14	21	13	11
Sample 6:	19	11	16	6	8	13	10	22	20	11

a Calculate the *sample* statistics s and s^2 for each sample.

b Calculate the *population* statistics σ and σ^2 for each sample.

c Which set of estimates from **a** and **b** are generally closer to the true population variance and standard deviation?

d Does your answer to **c** explain why we have different variance and standard deviation formulae for a sample as opposed to a population?

3 To see which set of estimators (population or sample) are better at estimating the true population variance and standard deviation, we will consider a simulation based on the survey responses from the school.

Click on the icon to obtain a spreadsheet with 1000 simulations of the survey results. The values s, s^2, σ, and σ^2 are calculated for each simulated sample. The average values for each estimator are shown in the table on the sheet labelled "Summary".

SPREADSHEET

	A	B	C	D	E	F
1	**Actual values**				**Estimator**	**Average estimate**
2	**μ**	15		*Variance*	**σ^2**	23.794
3	**σ**	5			**s^2**	25.046
4	**σ^2**	25		*Standard deviation*	**σ**	4.814
5	**n**	20			**s**	4.939

Based on the calculations in the spreadsheet, which set of estimates (population or sample) are generally closer to the true values? Does your conclusion agree with your answer to **2 c**?

4 Change the values for μ and σ in the spreadsheet. This will now effectively simulate the results for a different distribution, perhaps the travel times for the students at a different school. Does your choice of μ or σ affect your conclusion regarding the choice of estimators?

5 Why is it important to have accurate estimates of the variance and standard deviation of a population?

REVIEW SET 12A

1 For each of the following data sets, find the: **i** mean **ii** median.

a 0, 2, 3, 3, 4, 5, 5, 6, 6, 7, 7, 8 **b** 2.9, 3.1, 3.7, 3.8, 3.9, 3.9, 4.0, 4.5, 4.7, 5.4

2 Katie loves cats. She visits every house in her street to find out how many cats live there. The responses are given below:

Number of cats	0	1	2	3	4	5
Frequency	36	9	11	5	1	1

a Draw a graph to display this data.

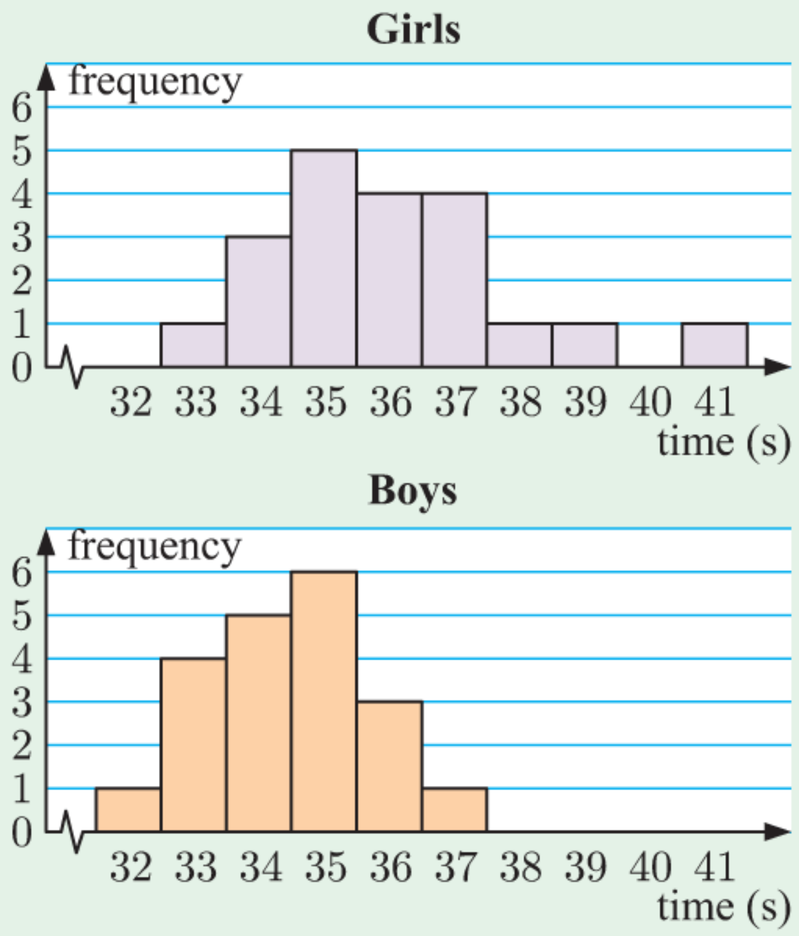
b Describe the distribution.

c Find the:

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

d Which of the measures of centre is most appropriate for this data? Explain your answer.

3 The histograms alongside show the times for the 50 metre freestyle recorded by members of a swimming squad.



a Copy and complete:

Distribution	*Girls*	*Boys*
median		
mean		
modal class		

b Discuss the distributions of times for the boys and girls. What conclusion can you make?

4 The data set $4, 6, 9, a, 3, b$ has a mean and mode of 6. Find the values of a and b given that $a > b$.

5 Consider the data set: $k-2,\ k,\ k+3,\ k+3$.

a Show that the mean of the data set is equal to $k+1$.

b Suppose each number in the data set is increased by 2. Find the new mean of the data set in terms of k.

6 The winning margins in 100 basketball games were recorded. The results are summarised alongside.

a Explain why you cannot calculate the mean winning margin from the table exactly.

b Estimate the mean winning margin.

Margin (points)	*Frequency*
1 - 10	13
11 - 20	35
21 - 30	27
31 - 40	18
41 - 50	7

7 The table alongside compares the mass of guinea pigs at birth with their mass when they are two weeks old.

a Find the mean birth mass.

b Find the mean mass after two weeks.

c Find the mean increase over the two weeks.

Guinea Pig	*Mass* (g) *at birth*	*Mass* (g) *at 2 weeks*
A	75	210
B	70	200
C	80	200
D	70	220
E	74	215
F	60	200
G	55	206
H	83	230

8 Consider this data set:

19, 7, 22, 15, 14, 10, 8, 28, 14, 18, 31, 13, 18, 19, 11, 3, 15, 16, 19, 14

a Find the five-number summary for the data. **b** Find the range and IQR.

c Draw a box plot of the data set.

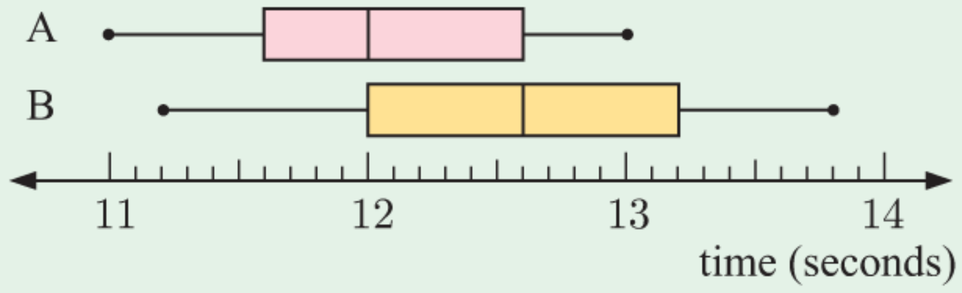
9 Katja's golf scores for her last 20 rounds were:

90	106	84	103	112	100	105	81	104	98
107	95	104	108	99	101	106	102	98	101

For this data set, find the:

a median **b** interquartile range **c** mean **d** standard deviation.

10 The parallel box plot alongside shows the 100 metre sprint times for the members of two athletics squads.



a Determine the five-number summaries for both A and B.

b For each group, calculate the range and interquartile range.

c Copy and complete:

i The members of squad ..... generally ran faster because .....

ii The times in squad ..... are more varied because .....

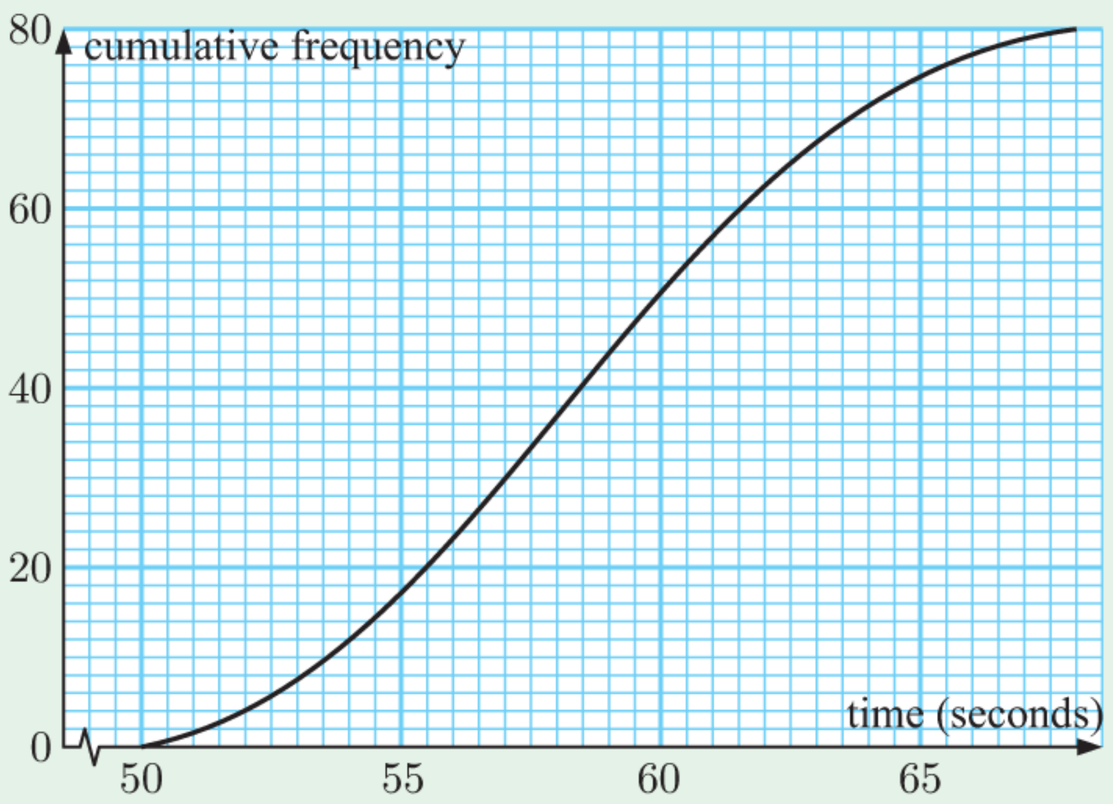
11 80 senior students ran 400 metres in a Physical Education program. Their times were recorded and the results were used to produce the cumulative frequency graph shown.

Estimate:

a the median

b the interquartile range

c the time corresponding to the top 10% of runners.



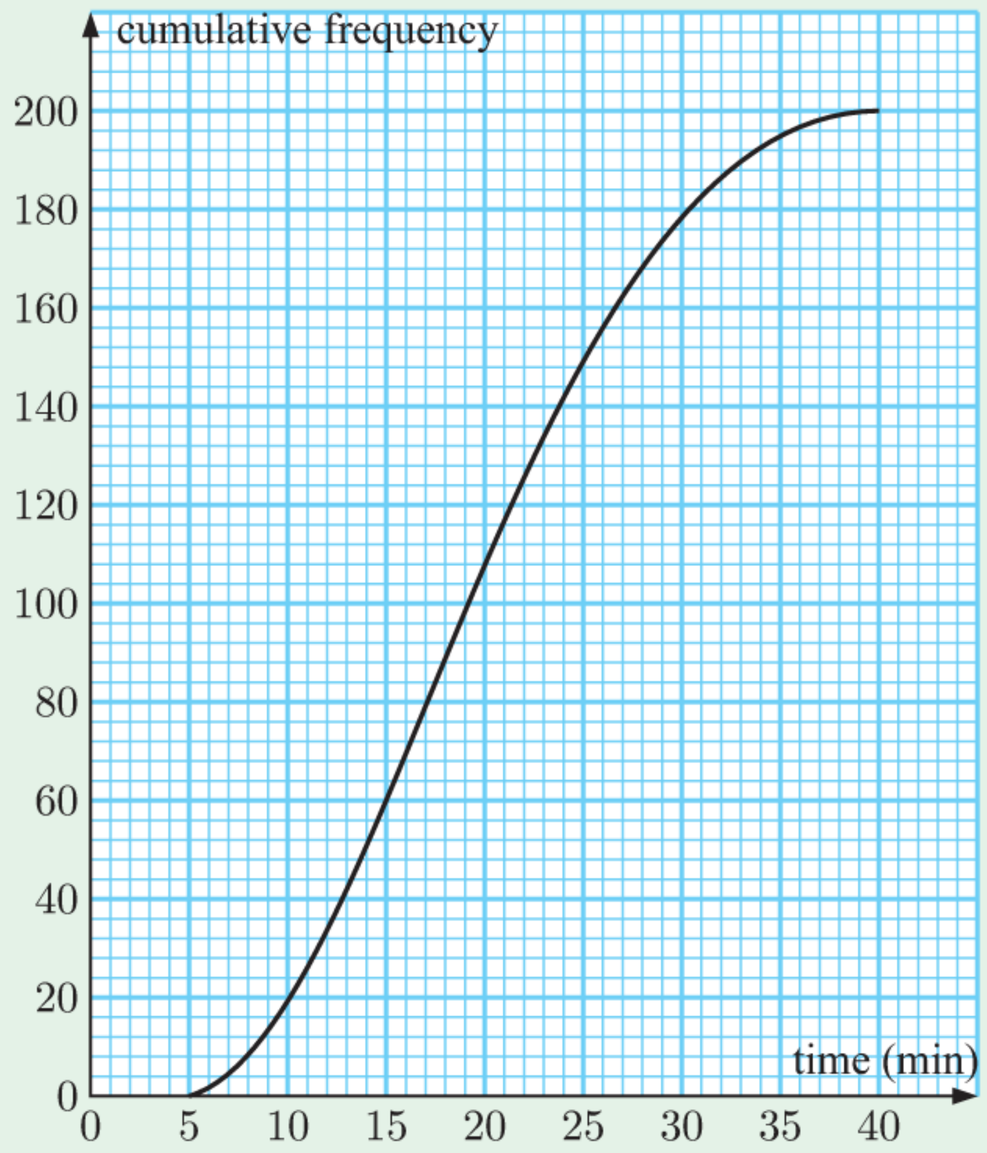
12 This cumulative frequency curve shows the times taken for 200 students to travel to school by bus.

a Estimate how many of the students spent between 10 and 20 minutes travelling to school.

b 30% of the students spent more than m minutes travelling to school. Estimate the value of m.

c Use the cumulative frequency curve to complete the following table:

Time (t min)	*Frequency*
$5 \leqslant t < 10$	
$10 \leqslant t < 15$	
$\vdots$	
$35 \leqslant t < 40$	



13 Find the population variance and standard deviation for each data set:

a 117, 129, 105, 124, 123, 128, 131, 124, 123, 125, 108

b 6.1, 5.6, 7.2, 8.3, 6.6, 8.4, 7.7, 6.2

14 The number of litres of petrol purchased by a random sample of motor vehicle drivers is shown alongside.

For the given data, estimate the:

a mean **b** standard deviation.

Litres (L)	*Number of vehicles*
$15 \leqslant L < 20$	5
$20 \leqslant L < 25$	13
$25 \leqslant L < 30$	17
$30 \leqslant L < 35$	29
$35 \leqslant L < 40$	27
$40 \leqslant L < 45$	18
$45 \leqslant L < 50$	7

15 Pratik is a quality control officer for a biscuit company. He needs to check that 250 g of biscuits go into each packet, but realises that the weight in each packet will vary slightly.

a Would you expect the standard deviation for the whole population to be the same for one day as it is for one week? Explain your answer.

b If a sample of 100 packets is measured each day, what measure would be used to check:

i that an average of 250 g of biscuits goes into each packet

ii the variability of the mass going into each packet?

c Explain the significance of a low standard deviation in this case.

REVIEW SET 12B

1 Heike is preparing for an athletics carnival. She records her times in seconds for the 100 m sprint each day for 4 weeks.

Week 1:	16.4	15.2	16.3	16.3	17.1	15.5	14.9
Week 2:	14.9	15.7	15.1	15.1	14.7	14.7	15.3
Week 3:	14.3	14.2	14.6	14.6	14.3	14.3	14.4
Week 4:	14.0	14.0	13.9	14.0	14.1	13.8	14.2

a Calculate Heike's mean and median time for each week.

b Do you think Heike's times have improved over the 4 week period? Explain your answer.

2 A die was rolled 50 times.
The results are shown in the table alongside.
Find the:

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

Number	*Frequency*
1	10
2	7
3	8
4	5
5	12
6	8

3 The data in the table alongside has mean 5.7.

a Find the value of x.

b Find the median of the distribution.

Value	2	5	x	$x+6$
Frequency	3	2	4	1

4 A set of 14 data is: 6, 8, 7, 7, 5, 7, 6, 8, 6, 9, 6, 7, p, q.
The mean and mode of the set are both 7. Find p and q.

5 The table alongside shows the number of patrons visiting an art gallery on various days.
Estimate the mean number of patrons per day.

Number of patrons	*Frequency*
250 - 299	14
300 - 349	34
350 - 399	68
400 - 449	72
450 - 499	54
500 - 549	23
550 - 599	7

6 Draw a box and whisker diagram for the following data:

11, 12, 12, 13, 14, 14, 15, 15, 15, 16, 17, 17, 18.

7 Consider the data set: 120, 118, 132, 127, 135, 116, 122, 93, 128.

a Find the standard deviation for the data.

b Find the upper and lower quartiles of the data set.

c Are there any outliers in the data set?

d Draw a box plot to display the data.

8 The number of peanuts in a jar varies slightly from jar to jar. Samples of 30 jars were taken for each of two brands X and Y, and the number of peanuts in each jar was recorded.

Brand X						*Brand Y*					
871	885	878	882	889	885	909	906	913	891	898	901
916	913	886	905	907	898	894	894	928	893	924	892
874	904	901	894	897	899	927	907	901	900	907	913
908	901	898	894	895	895	921	904	903	896	901	895
910	904	896	893	903	888	917	903	910	903	909	904

a Copy and complete this table:

	Brand X	*Brand Y*
min		
Q_1		
median		
Q_3		
max		
IQR		

b Display the data on a parallel box plot.

c Comment on which brand:

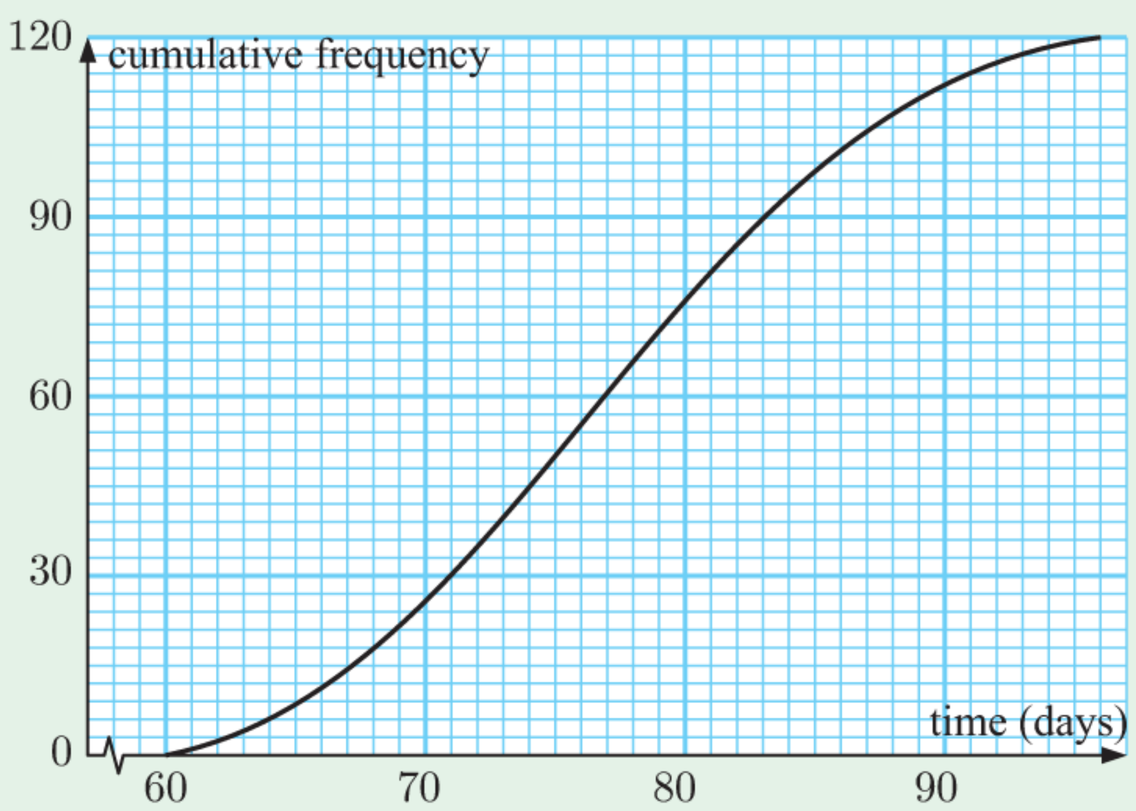
i has more peanuts per jar **ii** has a more consistent number of peanuts per jar.

9 120 people caught whooping cough in an outbreak. The times for them to recover were recorded, and the results were used to produce the cumulative frequency graph shown.

Estimate:

a the median

b the interquartile range.



10 Consider the data in the table below:

Scores (x)	$0 \leqslant x < 10$	$10 \leqslant x < 20$	$20 \leqslant x < 30$	$30 \leqslant x < 40$	$40 \leqslant x < 50$
Frequency	1	13	27	17	2

a Construct a cumulative frequency graph for the data.

b Estimate the:

i median **ii** interquartile range **iii** mean **iv** standard deviation.

11 Consider the frequency table alongside:

Score	*Frequency*	*Cumulative frequency*
6	2	2
7	4	m
8	7	13
9	p	25
10	5	30

a Find the values of p and m.

b Hence find the mode, median, and range of the data.

c Given that $\sum_{i=1}^{5} x_i f_i = 254$, write the mean $\overline{x}$ as a fraction.

12 To test the difficulty level of a new computer game, a company measures the time taken for a group of players to complete the game. Their results are displayed in the table alongside.

Completion time (t min)	*Number of players*
$0 \leqslant t < 30$	1
$30 \leqslant t < 60$	4
$60 \leqslant t < 90$	12
$90 \leqslant t < 120$	18
$120 \leqslant t < 150$	7
$150 \leqslant t < 180$	2

a How many players were surveyed?

b Write down the modal class.

c Draw a cumulative frequency graph for the data.

d The game is considered too easy if either the mean or median completion time is below 90 minutes.

i Estimate the median completion time using your cumulative frequency graph.

ii Estimate the mean completion time.

iii Hence comment on whether the game is too easy.

e Complete this sentence:

The middle 50% of players completed the game in times between ..... and ..... minutes.

13 The table below shows the number of matches in a sample of boxes.

Number	47	48	49	50	51	52
Frequency	21	29	35	42	18	31

a Find the mean and standard deviation for this data.

b Does this result justify a claim that the average number of matches per box is 50?

14 A random sample of weekly supermarket bills was recorded in the table alongside.

For the given data, estimate the:

a mean **b** standard deviation.

Bill (€b)	*Number of families*
$140 \leqslant b < 160$	27
$160 \leqslant b < 180$	32
$180 \leqslant b < 200$	48
$200 \leqslant b < 220$	25
$220 \leqslant b < 240$	37
$240 \leqslant b < 260$	21
$260 \leqslant b < 280$	18
$280 \leqslant b < 300$	7

15 Friends Kevin and Felicity each selected a sample of 20 crossword puzzles. The times they took, in minutes, to complete each puzzle were:

Kevin					Felicity				
37	53	47	33	39	33	36	41	26	52
49	37	48	32	36	38	49	57	39	44
39	42	34	29	52	48	25	34	27	53
48	33	56	39	41	38	34	35	50	31

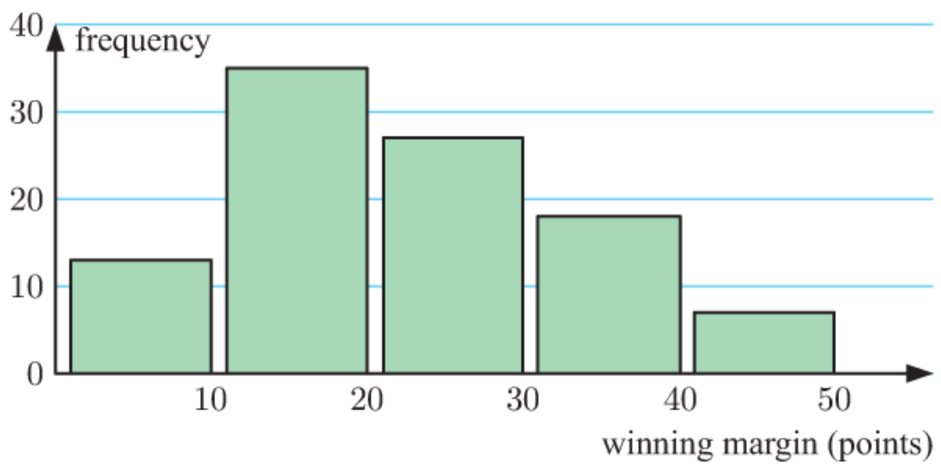
a Find the mean of each data set.

b Find the population standard deviation for each data set.

c Who generally solves crossword puzzles faster?

d Who is more consistent in their time taken to solve the puzzles?

5 **a**



b **i** 48% **ii** 25%

c No, only that it was in the interval 1 - 10.

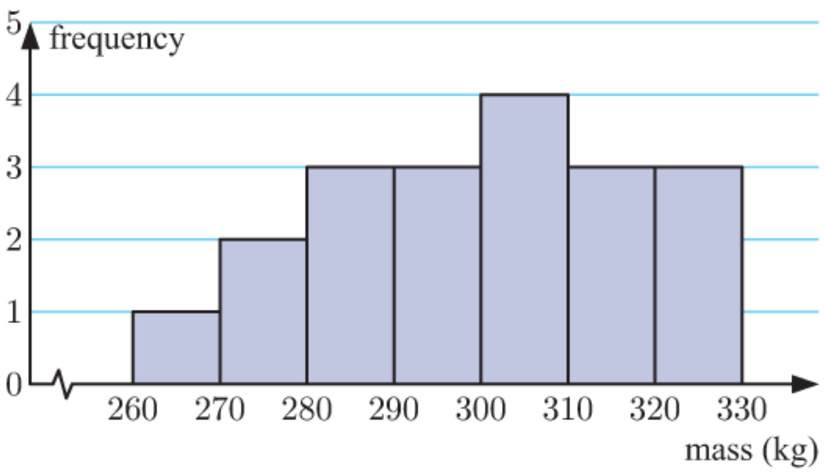
6 **a** Mass is measured on a continuous scale.

b

Mass (m kg)	*Frequency*
$260 \leqslant m < 270$	1
$270 \leqslant m < 280$	2
$280 \leqslant m < 290$	3
$290 \leqslant m < 300$	3
$300 \leqslant m < 310$	4
$310 \leqslant m < 320$	3
$320 \leqslant m < 330$	3

c $300 \leqslant m < 310$ kg

d



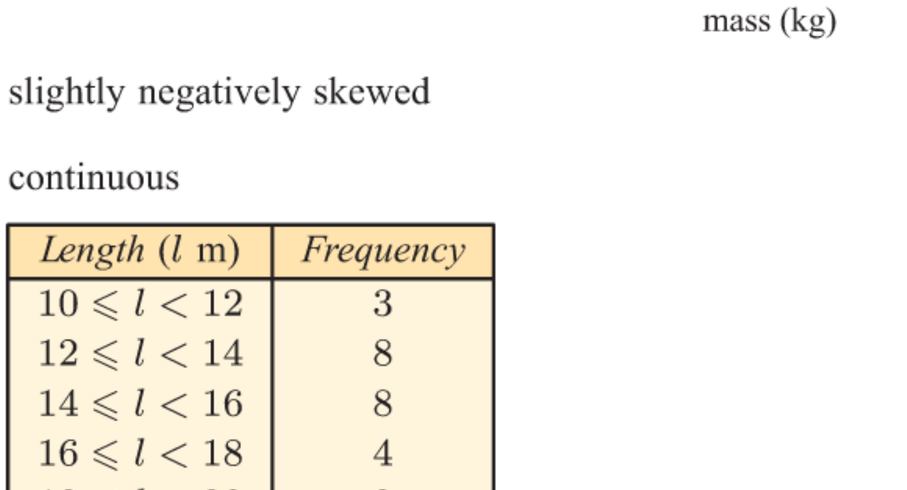
e slightly negatively skewed

7 **a** continuous

b

Length (l m)	*Frequency*
$10 \leqslant l < 12$	3
$12 \leqslant l < 14$	8
$14 \leqslant l < 16$	8
$16 \leqslant l < 18$	4
$18 \leqslant l < 20$	2
$20 \leqslant l < 22$	3
$22 \leqslant l < 24$	1
$24 \leqslant l < 26$	0
$26 \leqslant l < 28$	1

c

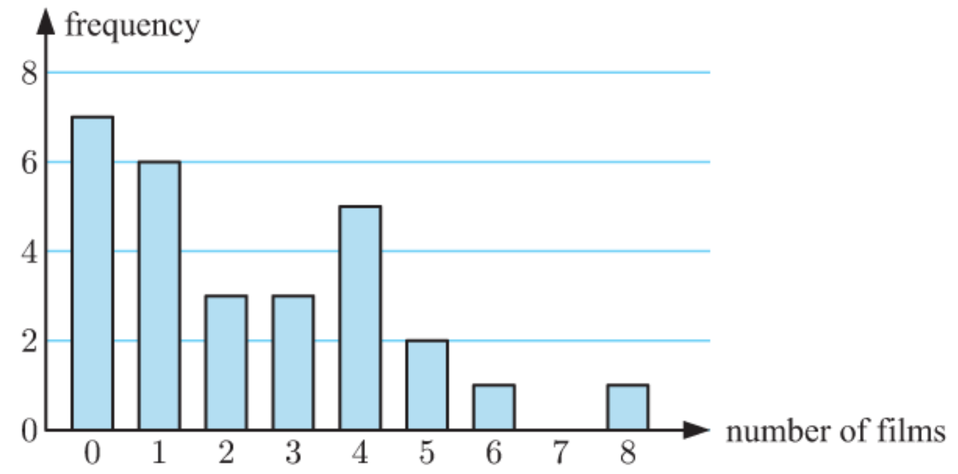


d positively skewed, one outlier (27.4 m)

8 **a**

Films watched	*Frequency*
0	7
1	6
2	3
3	3
4	5
5	2
6	1
8	1

b



c 0 **d** **i** 75% **ii** $\approx 57.1\%$

EXERCISE 12A

1 **a** 1 cup **b** 2 cups **c** 1.8 cups

2 **a** **i** ≈ 5.61 **ii** 6 **iii** 6

b **i** ≈ 16.3 **ii** 17 **iii** 18

c **i** ≈ 24.8 **ii** 24.9 **iii** 23.5

3 9 **4** Ruth

5 **a** *data set A*: ≈ 6.46, *data set B*: ≈ 6.85

b *data set A*: 7, *data set B*: 7

c Data sets A and B differ only by their last value. This affects the mean, but not the median.

6 **a** **i** *motichoor ladoo*: ≈ 67.1, *malai jamun*: ≈ 53.6

ii *motichoor ladoo*: 69, *malai jamun*: 52

b The mean and median were much higher for the motichoor ladoo, so the motichoor ladoo were more popular.

7 **a** *Bus*: mean = 39.7, median = 40.5
Tram: mean $\approx$ 49.1, median = 49

b The tram data has a higher mean and median, but since there are more bus trips per day and more people travel by bus in total, the bus is more popular.

8 **a** 44 points **b** 44 points

c **i** Decrease, since 25 is lower than the mean of 44 for the first four matches.

ii 40.2 points

9 €185 604 **10** 3144 km **11** 116

12 17.25 goals per game **13** $x = 15$ **14** $a = 5$

15 37 marks **16** ≈ 14.8 **17** 6 and 12

EXERCISE 12B

1 **a** mean = \$363 770, median = \$347 200
The mean has been affected by the extreme values (the two values greater than \$400 000).

b **i** the mean **ii** the median

2 **a** mode = \$33 000, mean = \$39 300, median = \$33 500

b The mode is the lowest value in the data set.

c No, it is too close to the lower end of the distribution.

3 **a** mean $\approx$ 3.19 mm, median = 0 mm, mode = 0 mm

b The median is not the most suitable measure of centre as the data is positively skewed.

c The mode is the lowest value.

d 42 mm and 21 mm **e** no

4 **a** mean ≈ 2.03, median $= 2$, mode $= 1$ and 2

b Yes, as Esmé can then offer a "family package" to match the most common number of children per family.

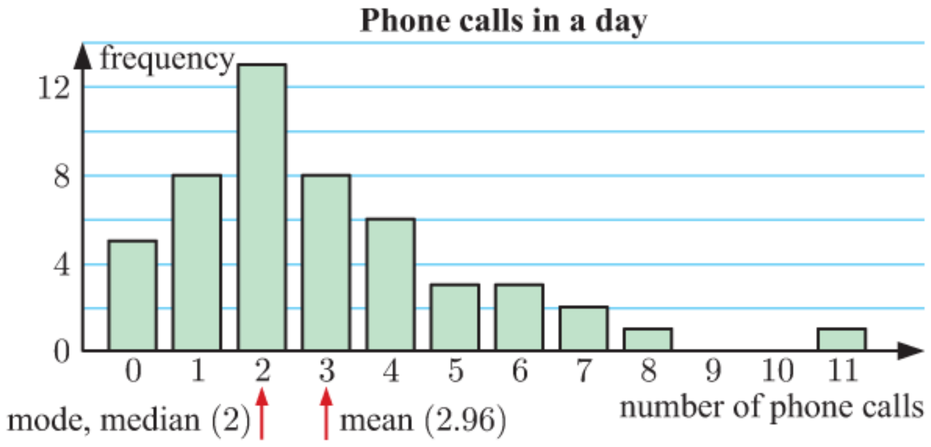
c 2 children, since this is one of the modes; it is also the median, and very close to the mean.

EXERCISE 12C

1 **a** 1 person **b** 2 people **c** ≈ 2.03 people

2 **a** **i** 2.96 phone calls **ii** 2 phone calls **iii** 2 phone calls

b



c positively skewed

d The mean takes into account the larger numbers of phone calls.

e the mean

3 **a** **i** 49 matches **ii** 49 matches **iii** ≈ 49.0 matches

b no

c The sample of only 30 is not large enough. The company could have won its case by arguing that a larger sample would have found an average of 50 matches per box.

4 **a** **i** ≈ 2.61 children **ii** 2 children **iii** 2 children

b This school has more children per family than the average British family.

c positively skewed

d The values at the higher end increase the mean more than the median and the mode.

5 **a**

Pocket money (€)	*Frequency*
1	4
2	9
3	2
4	6
5	8

b 29 children

c **i** $\approx$ €3.17 **ii** €3 **iii** €2

d the mode

6 10.1 cm

7 **a** **i** \$63 000 **ii** \$56 000 **iii** \$66 600 **b** the mean

8 **a** $x = 5$ **b** 75%

EXERCISE 12D

1 **a** 40 phone calls **b** ≈ 15 minutes **2** ≈ 31.7

3 **a** 26 days **b** 31 - 40 children **c** ≈ 41.5 children

4 **a** 70 service stations **b** $\approx 411\,000$ litres (≈ 411 kL)

c ≈ 5870 L

d $6000 < P \leqslant 7000$ L. This is the most frequently occurring amount of petrol sales at a service station in one day.

5 **a**

Runs scored	*Tally*	*Frequency*
0 - 9	𝍸 𝍸 |	11
10 - 19	𝍸 |||	8
20 - 29	𝍸 |||	8
30 - 39	||	2
	Total	29

b ≈ 14.8 runs

c ≈ 14.9 runs; the estimate in **b** was very accurate.

6 **a** $p = 24$ **b** ≈ 3.37 minutes **c** $\approx 15.3\%$

7 **a** 125 people **b** ≈ 119 marks **c** $\frac{3}{25}$ **d** 28%

EXERCISE 12E

1 **a** **i** 13 **ii** $Q_1 = 9$, $Q_3 = 18$ **iii** 16 **iv** 9

b **i** 18.5 **ii** $Q_1 = 13$, $Q_3 = 23$ **iii** 19 **iv** 10

c **i** 26.5 **ii** $Q_1 = 20$, $Q_3 = 35$ **iii** 28 **iv** 15

d **i** 37 **ii** $Q_1 = 28$, $Q_3 = 52$ **iii** 49 **iv** 24

2 **a** **i** range $= 23$ goals, IQR $= 17$ goals

ii range $= 38$ goals, IQR $= 24$ goals

b Natalie

3 **a** *Jane*: mean $= \$35.50$, median $= \$35.50$

Ashley: mean $= \$30.75$, median $= \$26.00$

b *Jane*: range $= \$18$, IQR $= \$9$

Ashley: range $= \$40$, IQR $= \$14$

c Jane **d** Ashley

4 **a** range $= 60$, IQR $= 8.5$ **b** '67' is an outlier.

c range $= 18$, IQR $= 8$ **d** the range

5 **a** *Derrick*: range $= 240$ minutes, IQR $= 30$ minutes

Gareth: range $= 170$ minutes, IQR $= 120$ minutes

b **i** Gareth's **ii** Derrick's

c The IQR is most appropriate as it is less affected by outliers.

6 **a** g **b** **i** $m - a$ **ii** $\left(\frac{j+k}{2}\right) - \left(\frac{c+d}{2}\right)$

7

Measure	median	mode	range	interquartile range
Value **a**	11	9	13	6
Value **b**	18	14	26	12

EXERCISE 12F

1 **a** **i** 35 points **ii** 78 points **iii** 13 points

iv 53 points **v** 26 points

b **i** 65 points **ii** 27 points

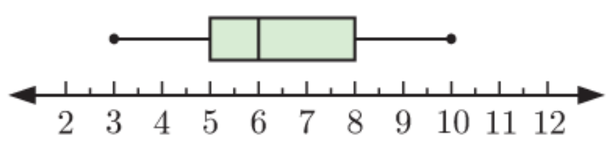
2 **a** **i** 98, 25 marks **ii** 70 marks **iii** 85 marks

iv 55, 85 marks

b 73 marks **c** 30 marks

3 **a** **i** min $= 3$, $Q_1 = 5$, med $= 6$, $Q_3 = 8$, max $= 10$

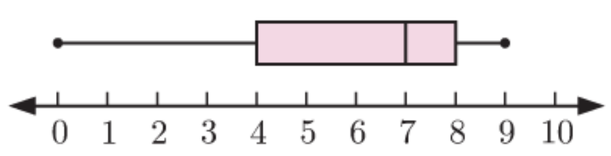
ii



iii 7 **iv** 3

b **i** min $= 0$, $Q_1 = 4$, med $= 7$, $Q_3 = 8$, max $= 9$

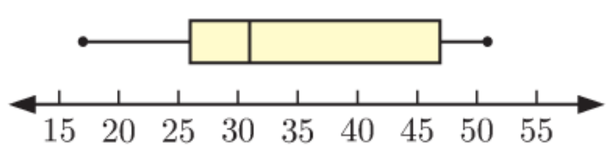
ii



iii 9 **iv** 4

c **i** min $= 17$, $Q_1 = 26$, med $= 31$, $Q_3 = 47$, max $= 51$

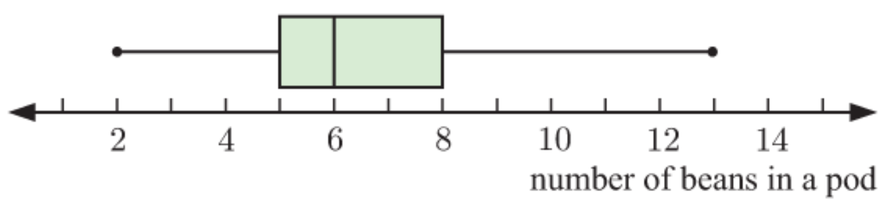
ii



iii 34 **iv** 21

4 **a** median $= 6$, $Q_1 = 5$, $Q_3 = 8$ **b** IQR $= 3$

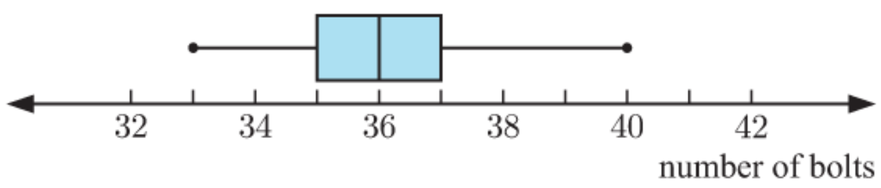
c



5 **a** min = 33, Q_1 = 35, med = 36, Q_3 = 37, max = 40

b **i** range = 7 **ii** IQR = 2

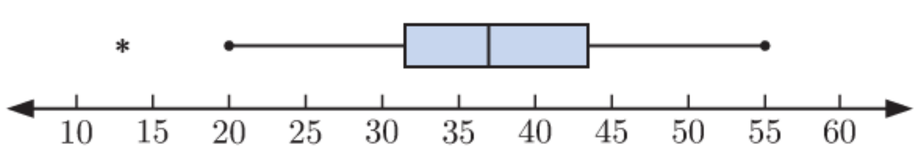
c



EXERCISE 12G

1 **a** 12 **b** lower = 13.5, upper = 61.5 **c** 13

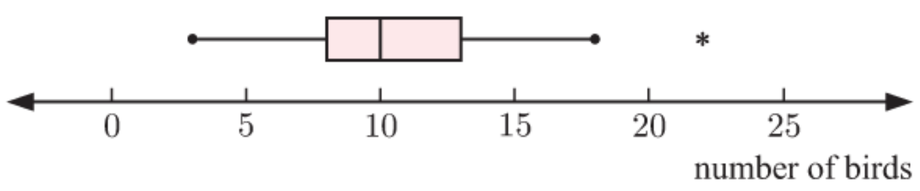
d



2 **a** median = 10, Q_1 = 8, Q_3 = 13 **b** IQR = 5

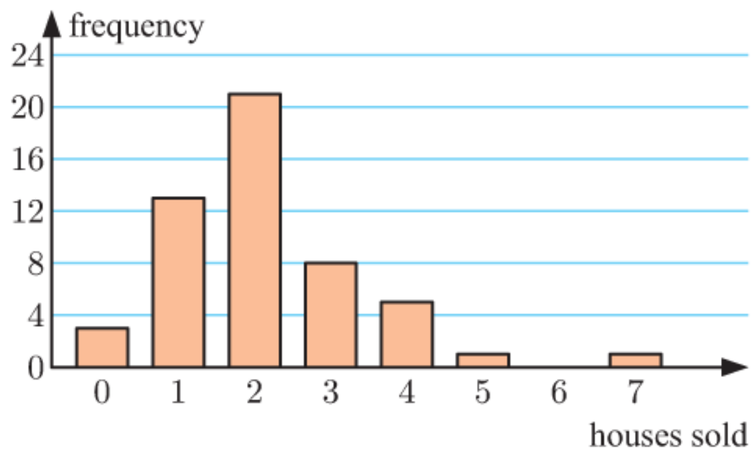
c lower = 0.5, upper = 20.5, 22 is an outlier

d



3 **a** **A** **b** **D** **c** **C** **d** **B**

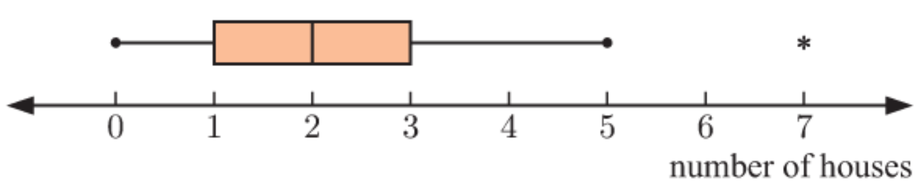
4 **a**



b 7 houses appears to be an outlier.

c lower boundary = −2, upper boundary = 6
7 houses is an outlier

d



EXERCISE 12H

1 **a**

Statistic	*Year 9*	*Year 12*
minimum	6	36
Q_1	30	60
median	45	84
Q_3	60	96
maximum	72	105

b **i** Year 9: 66 min
Year 12: 69 min

ii Year 9: 30 min
Year 12: 36 min

c **i** cannot tell **ii** true, since Year 9 Q_1 < Year 12 min

2 **a** Friday: min = €20, Q_1 = €50, med = €70, Q_3 = €100, max = €180

Saturday: min = €40, Q_1 = €80, med = €100, Q_3 = €140, max = €200

b **i** Friday: €160, Saturday: €160

ii Friday: €50, Saturday: €60

3 **a** **i** class 1 (96%) **ii** class 1 (37%) **iii** class 1

b 18% **c** 55% **d** **i** 25% **ii** 50%

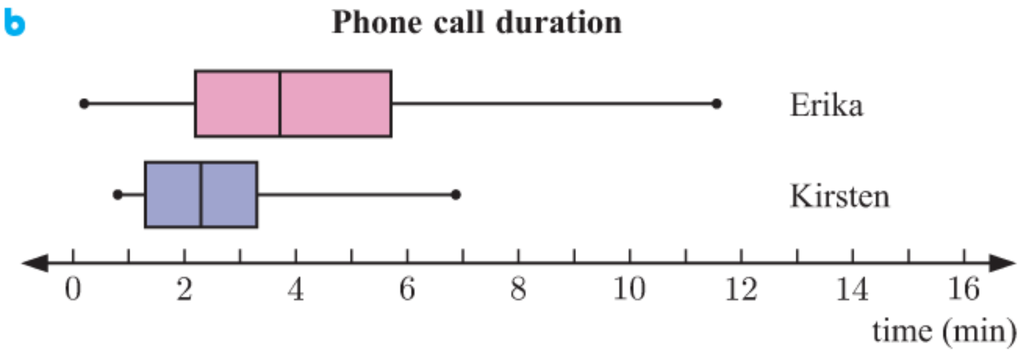
e **i** slightly positively skewed **ii** negatively skewed

f .... class 2, ... class 1

4 **a** Kirsten: min = 0.8 min, Q_1 = 1.3 min, med = 2.3 min, Q_3 = 3.3 min, max = 6.9 min

Erika: min = 0.2 min, Q_1 = 2.2 min, med = 3.7 min, Q_3 = 5.7 min, max = 11.5 min

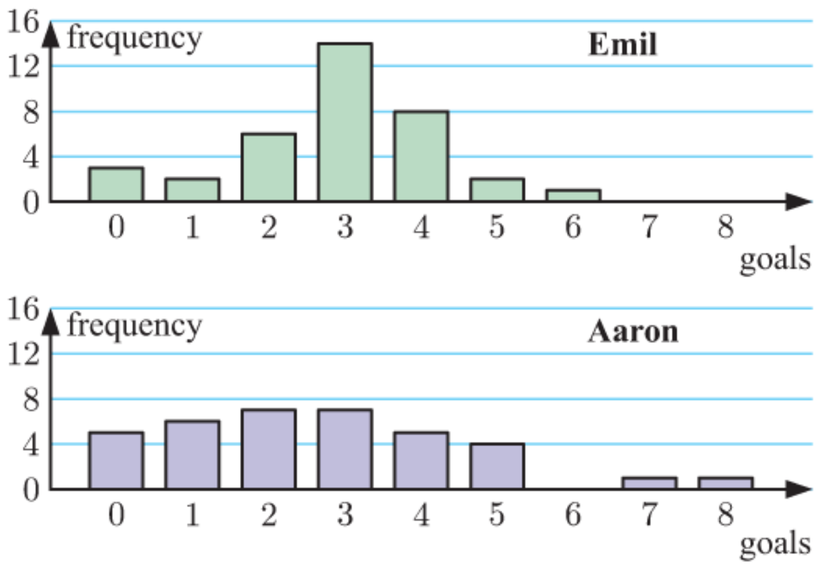
b



c Both are positively skewed (Erika's more so than Kirsten's). Erika's phone calls were more varied in duration.

5 **a** discrete

c

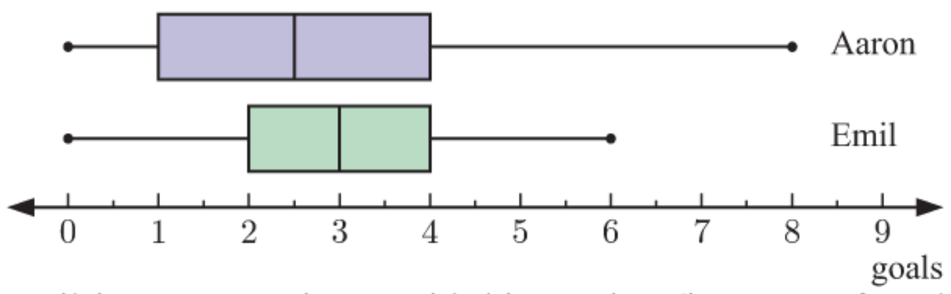


d Emil: approximately symmetrical
Aaron: positively skewed

e Emil: mean ≈ 2.89, median = 3, mode = 3
Aaron: mean ≈ 2.67, median = 2.5, mode = 2, 3
Emil's mean and median are slightly higher than Aaron's. Emil has a clear mode of 3, whereas Aaron has two modes (2 and 3).

f Emil: range = 6, IQR = 2
Aaron: range = 8, IQR = 3
Emil's data set demonstrates less variability than Aaron's.

g



h Emil is more consistent with his scoring (in terms of goals) than Aaron.

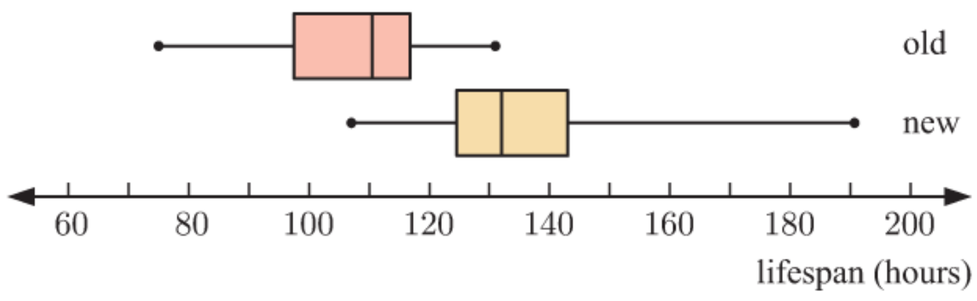
6 **a** continuous (the data is measured)

b Old type: mean = 107 hours, median = 110.5 hours, range = 56 hours, IQR = 19 hours

New type: mean = 134 hours, median = 132 hours, range = 84 hours, IQR = 18.5 hours

The "new" type of light globe has a higher mean and median than the "old" type.
The IQR is relatively unchanged going from "old" to "new", however, the range of the "new" type is greater, suggesting greater variability.

c

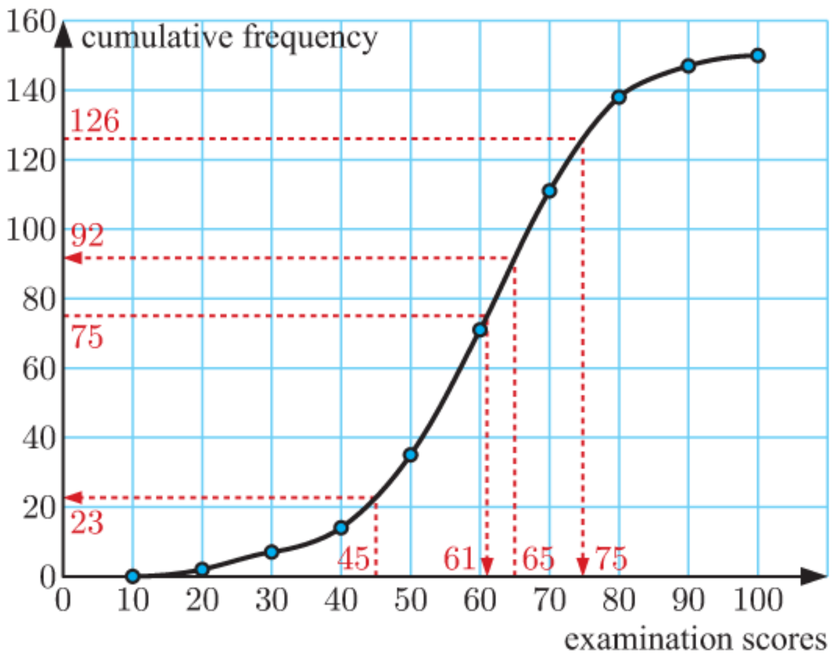


d Old type: negatively skewed
New type: positively skewed

e The "new" type of light globes do last longer than the "old" type. From **c**, both the mean and median for the "new" type are close to 20% greater than that of the "old" type. The manufacturer's claim appears to be valid.

EXERCISE 12I

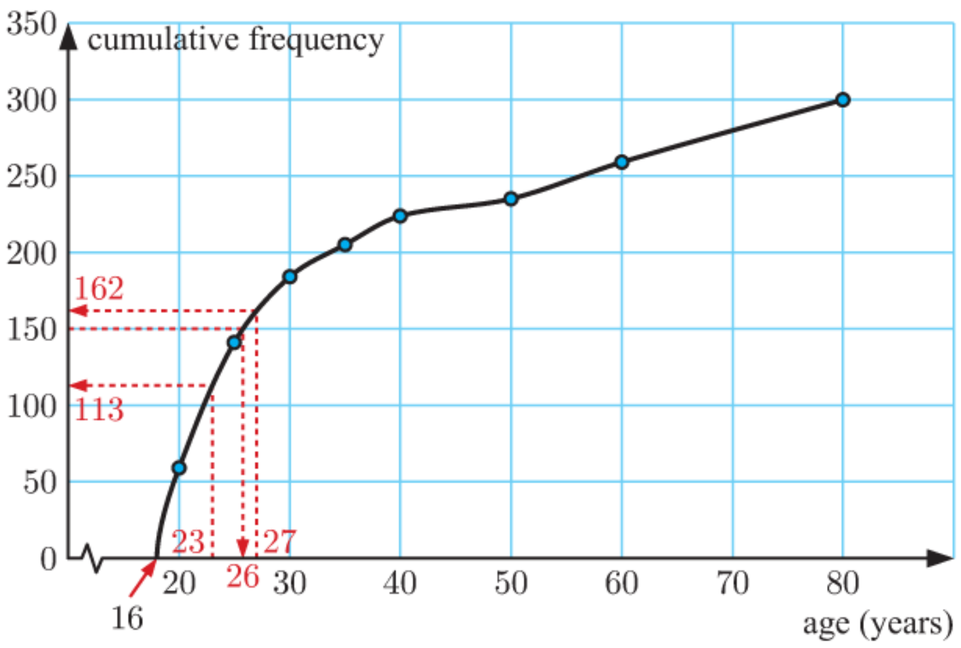
1 **a**



b $\approx$ 61 marks **c** $\approx$ 92 students **d** 76 students
e $\approx$ 23 students **f** $\approx$ 75 marks

2 **a** $\approx$ 9 seedlings **b** $\approx$ 28.3% **c** $\approx$ 7.1 cm
d $\approx$ 2.4 cm
e 10 cm, which means that 90% of the seedlings are shorter than 10 cm.

3 **a**

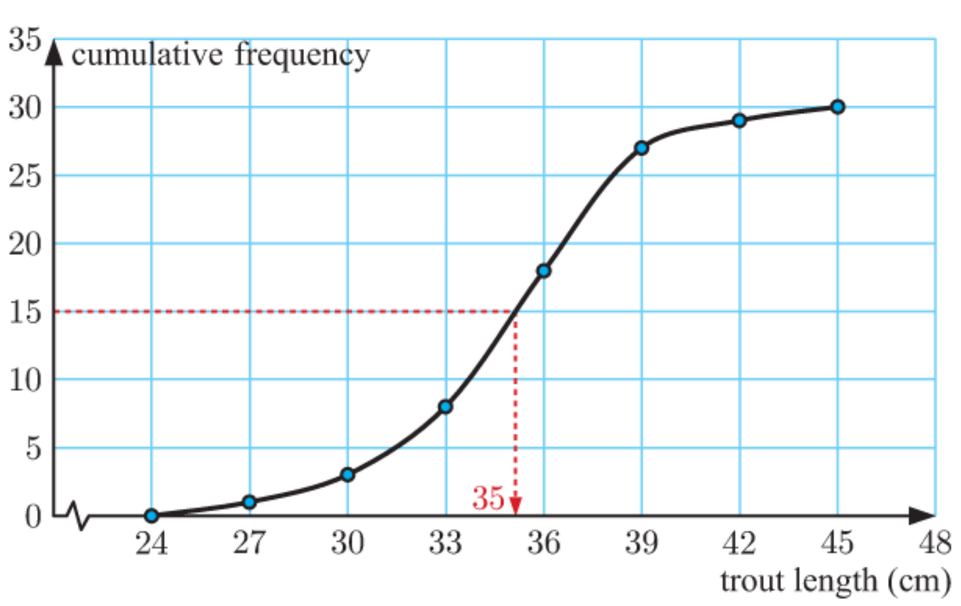


b $\approx$ 26 years **c** $\approx$ 37.7%
d **i** $\approx$ 0.54 **ii** $\approx$ 0.04

4 **a**

Length (cm)	*Frequency*	*Cumulative frequency*
$24 \leqslant x < 27$	1	1
$27 \leqslant x < 30$	2	3
$30 \leqslant x < 33$	5	8
$33 \leqslant x < 36$	10	18
$36 \leqslant x < 39$	9	27
$39 \leqslant x < 42$	2	29
$42 \leqslant x < 45$	1	30

b



c median $\approx$ 35 cm
d median = 34.5 cm; the median found from the graph is a good approximation.

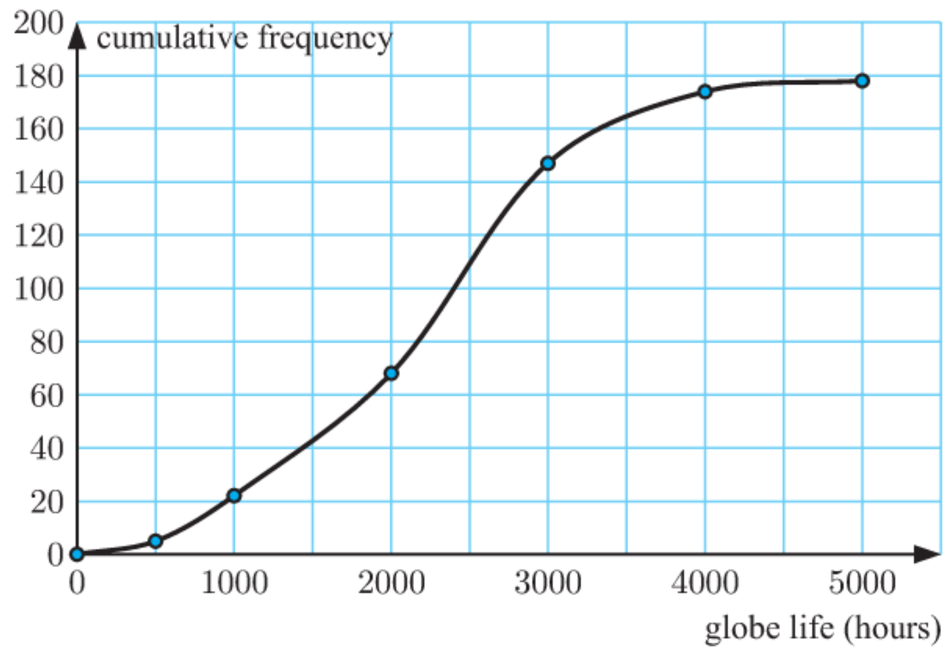
5 **a** $\approx$ 27 min **b** $\approx$ 29 min **c** $\approx$ 31.3 min
d $\approx$ 4.3 min **e** $\approx$ 28 min

f

Time (t min)	$21 \leqslant t < 24$	$24 \leqslant t < 27$	$27 \leqslant t < 30$
Number of competitors	5	15	30

Time (t min)	$30 \leqslant t < 33$	$33 \leqslant t < 36$
Number of competitors	20	10

6 **a**



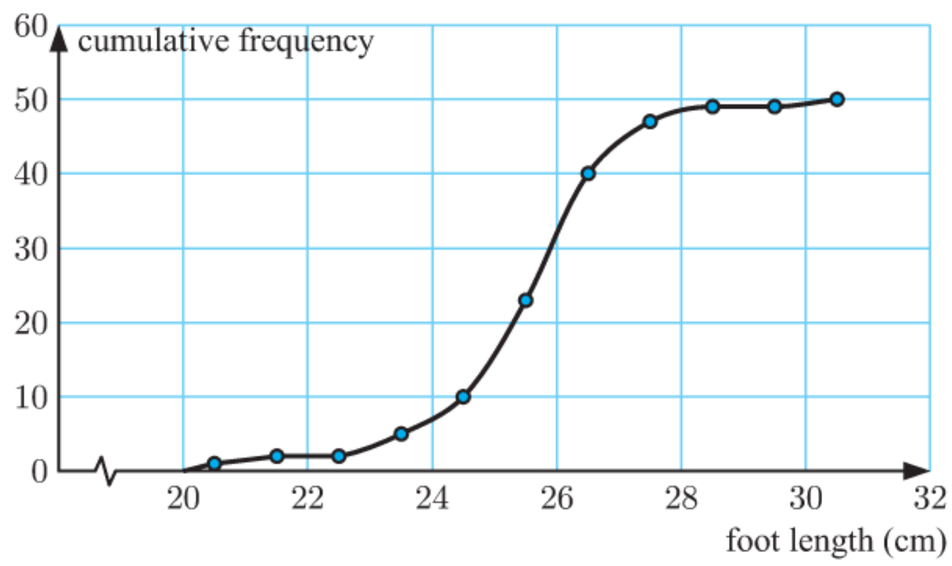
b $\approx$ 2280 hours **c** $\approx$ 71% **d** $\approx$ 67

7 **a** $19.5 \leqslant l < 20.5$ cm

b

Foot length (cm)	*Frequency*	*Cumulative frequency*
$19.5 \leqslant l < 20.5$	1	1
$20.5 \leqslant l < 21.5$	1	2
$21.5 \leqslant l < 22.5$	0	2
$22.5 \leqslant l < 23.5$	3	5
$23.5 \leqslant l < 24.5$	5	10
$24.5 \leqslant l < 25.5$	13	23
$25.5 \leqslant l < 26.5$	17	40
$26.5 \leqslant l < 27.5$	7	47
$27.5 \leqslant l < 28.5$	2	49
$28.5 \leqslant l < 29.5$	0	49
$29.5 \leqslant l < 30.5$	1	50

c



d **i** $\approx$ 25.2 cm **ii** $\approx$ 18 people

EXERCISE 12J

1 **a** *Data set A*: $\text{mean} = \dfrac{10+7+5+8+10}{5} = 8$

Data set B: $\text{mean} = \dfrac{4+12+11+14+1+6}{6} = 8$

b Data set B appears to have a greater spread than data set A, as data set B has more values that are a long way from the mean, such as 1 and 14.

c *Data set A*: $\sigma^2 = 3.6$, $\sigma \approx 1.90$
Data set B: $\sigma^2 \approx 21.7$, $\sigma \approx 4.65$

2 **a** $\sigma \approx 1.59$ **b** $\sigma^2 \approx 2.54$

3 **a** $\mu = 24.25$, $\sigma \approx 3.07$ **b** $\mu = 28.25$, $\sigma \approx 3.07$

c If each data value is increased or decreased by the same amount, then the mean will also be increased or decreased by that amount, however the population standard deviation will be unchanged.

4 $\sigma \approx 2.64$ **5** mean ≈ 30.0, $\sigma \approx 14.3$

6 **a** *Danny*: ≈ 3.21 hours; *Jennifer*: 2 hours

b Danny

c *Danny*: $\sigma \approx 0.700$ hours; *Jennifer*: $\sigma \approx 0.423$ hours

d Jennifer

7 **a**

	Mean $\overline{x}$	*Median*	*Standard deviation* σ	*Range*
Boys	32.02	31.05	≈ 4.52	13.8
Girls	34.77	35.85	≈ 3.76	11.7

b **i** boys **ii** boys

c Tyson could increase his sample size.

8 **a** *Rockets*: mean $= 5.7$, range $= 11$
Bullets: mean $= 5.7$, range $= 11$

b We suspect the Rockets, since they twice scored zero runs.
Rockets: $\sigma = 3.9$ ← greater variability
Bullets: $\sigma \approx 3.29$

c standard deviation

9 **a** **i** *Museum*: ≈ 934 visitors; *Art gallery*: ≈ 1230 visitors

ii *Museum*: ≈ 208 visitors; *Art gallery*: ≈ 84.6 visitors

b the museum

c **i** '0' is an outlier.

ii This outlier corresponded to Christmas Day, so the museum was probably closed which meant there were no visitors on that day.

iii Yes, although the outlier is not an error, it is not a true reflection of a visitor count for a particular day.

iv *Museum*: mean ≈ 965 visitors, $\sigma \approx 121$ visitors

v The outlier had greatly increased the population standard deviation.

10 $\sigma \approx 0.775$ **11** $\mu = 14.48$ years, $\sigma \approx 1.75$ years

12 **a** Data set A **b** *Data set A*: 8, *Data set B*: 8

c *Data set A*: 2, *Data set B*: ≈ 1.06
Data set A does have a wider spread.

d The standard deviation takes all of the data values into account, not just two.

13 **a** The female students' marks are in the range 16 to 20 whereas the male students' marks are in the range 12 to 19.

i the females **ii** the males

b *Females*: $\mu \approx 17.5$, $\sigma \approx 1.02$
Males: $\mu \approx 15.5$, $\sigma \approx 1.65$

14 The results for the mean will differ by 1, but the results for the standard deviation will be the same. Jess' question is worded so that the respondent will not include themselves.

15 **a** ≈ 48.3 cm **b** ≈ 2.66 cm

16 **a** ≈ 17.45 **b** ≈ 7.87

17 **a** $\approx \$780.60$ **b** $\approx \$31.74$

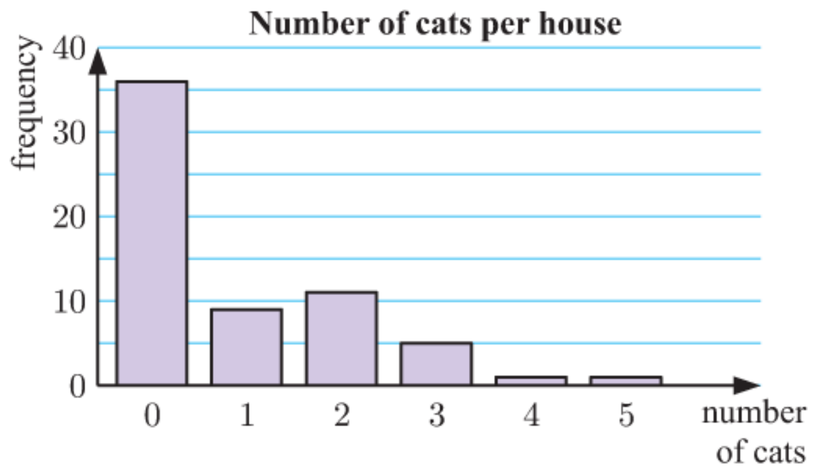
18 **a** $\overline{x} = 40.35$ hours, $\sigma \approx 4.23$ hours

b $\overline{x} = 40.6$ hours, $\sigma \approx 4.10$ hours
The mean increases slightly; the standard deviation decreases slightly. These are good approximations.

REVIEW SET 12A

1 **a** **i** ≈ 4.67 **ii** 5 **b** **i** 3.99 **ii** 3.9

2 **a**



b positively skewed

c **i** 0 cats **ii** ≈ 0.873 cats **iii** 0 cats

d The mean, as it suggests that some people have cats. (The mode and median are both 0.)

3 **a**

Distribution	*Girls*	*Boys*
median	36 s	34.5 s
mean	36 s	34.45 s
modal class	34.5 - 35.5 s	34.5 - 35.5 s

b The girls' distribution is positively skewed and the boys' distribution is approximately symmetrical. The median and mean swim times for boys are both about 1.5 seconds lower than for girls. Despite this, the distributions have the same modal class because of the skewness in the girls' distribution. The analysis supports the conjecture that boys generally swim faster than girls with less spread of times.

4 $a = 8$, $b = 6$

5 **b** $k + 3$

6 **a** We do not know each individual data value, only the intervals they fall in, so we cannot calculate the mean winning margin exactly.

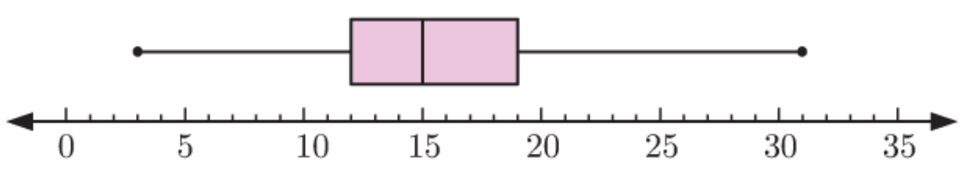
b ≈ 22.6 points

7 **a** ≈ 70.9 g **b** ≈ 210 g **c** ≈ 139.1 g

8 **a** min $= 3$, $Q_1 = 12$, med $= 15$, $Q_3 = 19$, max $= 31$

b range $= 28$, IQR $= 7$

c



9 **a** 101.5 **b** 7.5 **c** 100.2 **d** ≈ 7.59

10 **a** A: min $= 11$ s, $Q_1 = 11.6$ s, med $= 12$ s, $Q_3 = 12.6$ s, max $= 13$ s
B: min $= 11.2$ s, $Q_1 = 12$ s, med $= 12.6$ s, $Q_3 = 13.2$ s, max $= 13.8$ s

b A: range $= 2.0$ s, IQR $= 1.0$ s
B: range $= 2.6$ s, IQR $= 1.2$ s

c **i** A, the median time is lower.

ii B, the range and IQR are higher.

11 **a** ≈ 58.5 s **b** ≈ 6 s **c** ≈ 53 s

12 **a** ≈ 88 students **b** $m \approx 24$

c

Time (t min)	*Frequency*
$5 \leqslant t < 10$	20
$10 \leqslant t < 15$	40
$15 \leqslant t < 20$	48
$20 \leqslant t < 25$	42
$25 \leqslant t < 30$	28
$30 \leqslant t < 35$	17
$35 \leqslant t < 40$	5

13 **a** $\sigma^2 \approx 63.0$, $\sigma \approx 7.94$ **b** $\sigma^2 \approx 0.969$, $\sigma \approx 0.984$

14 **a** ≈ 33.6 L **b** ≈ 7.63 L

15 **a** No, extreme values have less effect on the standard deviation of a larger population.

b **i** mean **ii** standard deviation

c A low standard deviation means that the weight of biscuits in each packet is, on average, close to 250 g.

REVIEW SET 12B

1 **a**

	mean (seconds)	median (seconds)
Week 1	≈ 16.0	16.3
Week 2	≈ 15.1	15.1
Week 3	≈ 14.4	14.3
Week 4	14.0	14.0

b Yes, Heike's mean and median times have gradually decreased each week which indicates that her speed has improved over the 4 week period.

2 **a** 5 **b** 3.52 **c** 3.5

3 **a** $x = 7$ **b** 6

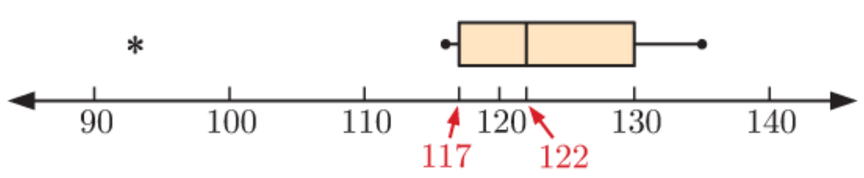
4 $p = 7$, $q = 9$ (or $p = 9$, $q = 7$)

5 ≈ 414 patrons

6

7 **a** $\sigma \approx 11.7$ **b** $Q_1 = 117$, $Q_3 = 130$ **c** yes, 93

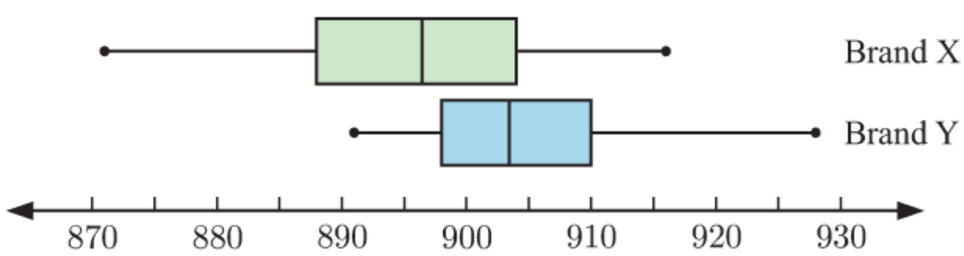
d



8 **a**

	Brand X	*Brand Y*
min	871	891
Q_1	888	898
median	896.5	903.5
Q_3	904	910
max	916	928
IQR	16	12

b

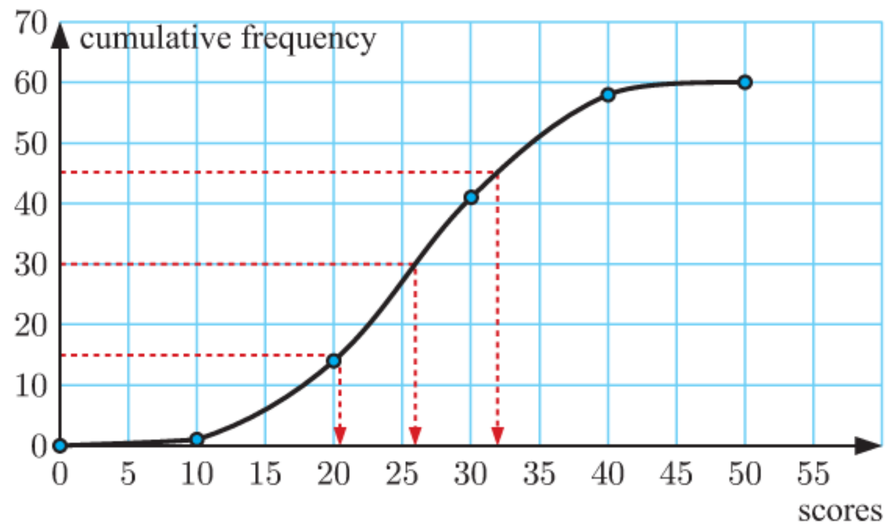


c **i** Brand Y, as the median is higher.

ii Brand Y, as the IQR is lower, so less variations.

9 **a** ≈ 77 days **b** ≈ 12 days

10 **a**



b **i** median ≈ 26.0 **ii** IQR ≈ 12

iii $\overline{x} \approx 26.0$ **iv** $\sigma \approx 8.31$

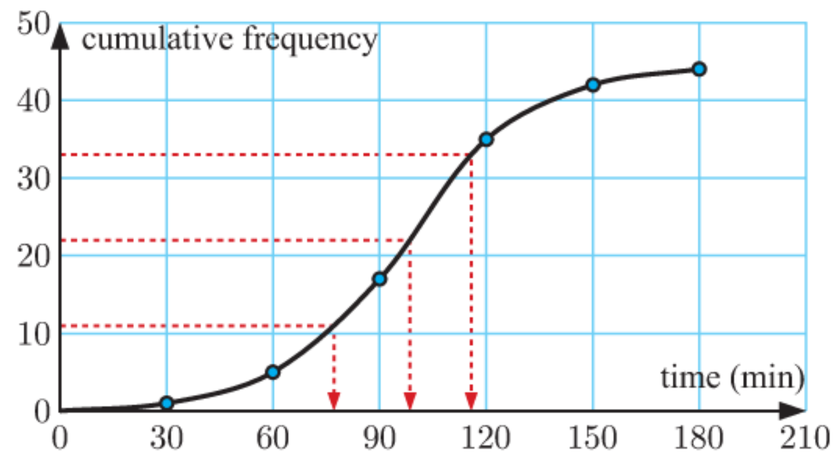
11 **a** $p = 12$, $m = 6$ **b**

Measure	*Value*
mode	9
median	9
range	4

c $\frac{254}{30} = \frac{127}{15}$

12 **a** 44 players **b** $90 \leqslant t < 120$ min

c



d **i** ≈ 98.6 min **ii** 96.8 min **iii** no

e ".... between 77.2 and 115.7 minutes."

13 **a** $\overline{x} \approx 49.6$ matches, $\sigma \approx 1.60$ matches

b The claim is not justified, but a larger sample is needed.

14 **a** $\approx$ €207.02 **b** $\approx$ €38.80

15 **a** Kevin: $\overline{x} = 41.2$ min; Felicity: $\overline{x} = 39.5$ min

b Kevin: $\sigma \approx 7.61$ min; Felicity: $\sigma \approx 9.22$ min

c Felicity **d** Kevin