

**Chapter**

**13**

# Statistics

**Contents:**

- A** Measuring the centre of data
- B** Choosing the appropriate measure
- C** Using frequency tables
- D** Grouped data
- E** Measuring the spread of data
- F** Box and whisker diagrams
- G** Outliers
- H** Parallel box and whisker diagrams
- I** Cumulative frequency graphs
- J** Variance and standard deviation



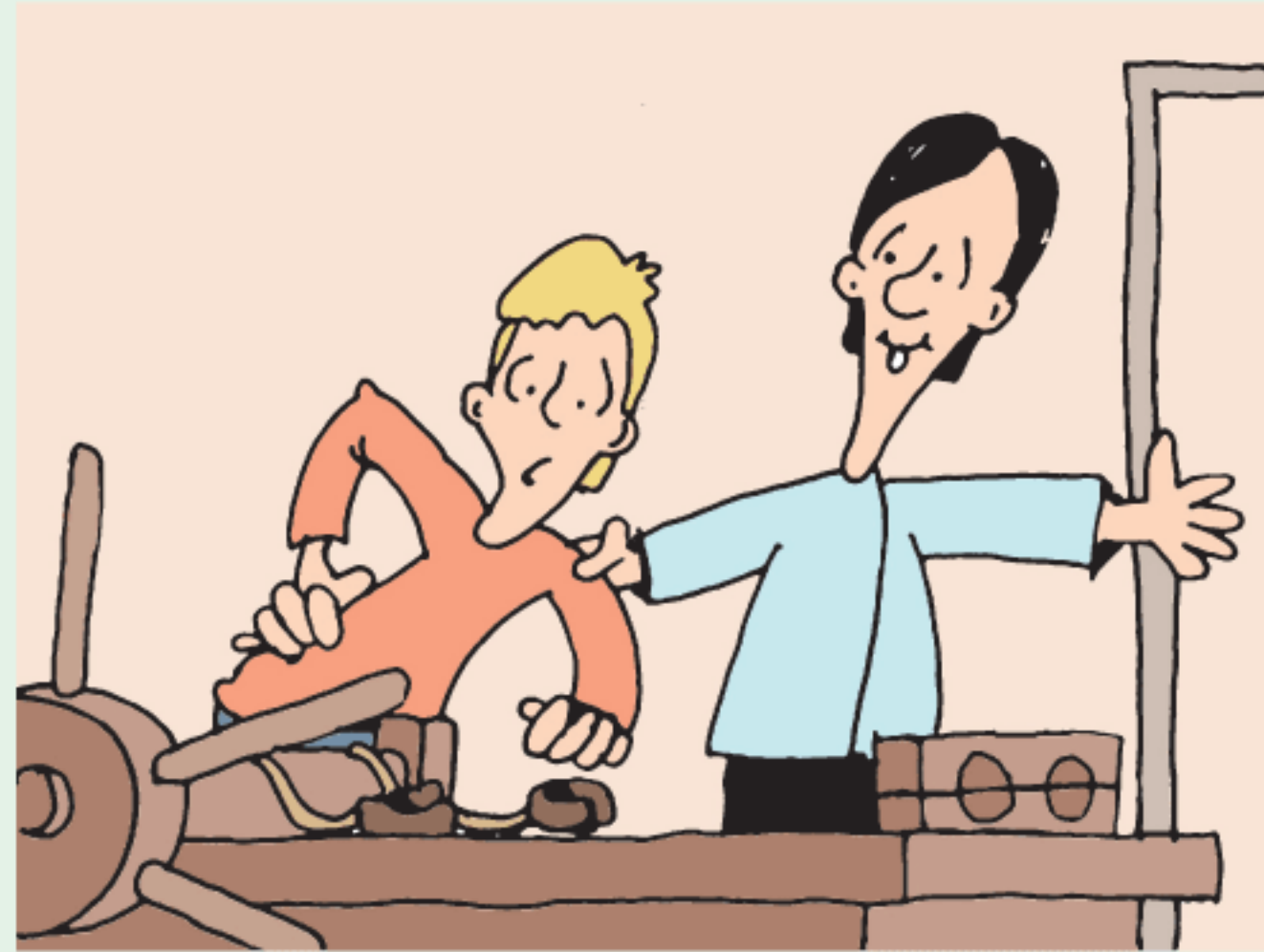


## OPENING PROBLEM

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:

<i>Before:</i>	7	9	5	6	9	7	10	6	8	9
	8	7	9	8	10	4	8	6	7	8
<i>After:</i>	4	7	6	3	8	5	9	4	5	8
	7	6	5	4	7	2	5	3	4	6



### Things to think about:

- What statistics can we calculate to measure the *centre* of each data set?
- How can we use a graph to make a visual comparison between the data sets?
- 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  $\{x_1, x_2, x_3, \dots, x_n\}$ ,

$$\begin{aligned} \text{mean} &= \frac{\text{sum of all data values}}{\text{the number of data values}} \\ &= \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \\ &= \frac{\sum_{i=1}^n x_i}{n} \end{aligned}$$

We use  $\bar{x}$  to represent the mean of a **sample**, and  $\mu$  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  $\mu$  is unknown. Instead we collect data from a sample of the population, and use the mean of the sample  $\bar{x}$  as an approximation for  $\mu$ .

$\mu$  is the Greek letter “mu” which we pronounce as “mew”.



## 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** mean =  $\frac{3 + 4 + 4 + \dots + 3 + 7 + 1}{21}$  ← sum of all the data values  
 ← 21 data values

$$= \frac{113}{21}$$

$$\approx 5.38 \text{ faulty products}$$

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

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

∴ 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 =  $\frac{113 + 9}{22} = \frac{122}{22} \approx 5.55$  faulty products.

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

The new ordered data set is:

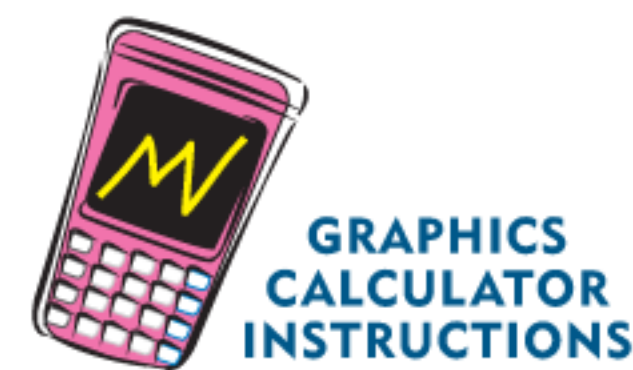
~~1 1 3 3 3 3 4 4 4 5 5 6 6 7 7 8 8 8 9 9 9~~

two middle data values

∴ the new median =  $\frac{5+6}{2} = 5.5$  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.

**EXERCISE 13A**

- 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:

<b>Motichoor ladoo</b>								<b>Malai jamun</b>							
62	76	55	65	49	78	71	82	37	52	71	59	63	47	56	68
79	47	60	72	58	82	76	67	43	67	38	73	54	55	61	49
50	61	70	85	77	69	48	74	50	48	53	39	45	60	46	51
63	56	81	75	63	74	54		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.

<b>Bus</b>							<b>Tram</b>						
30	43	40	53	70	50	63	58	68	43	45	70	79	
41	38	21	28	23	43	48	38	23	30	22	63	73	
20	26	35	48	41	33		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.

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

$$\therefore \text{sum of masses} = 53.7 \times 6$$

$$\therefore \text{the total mass} = 322.2 \text{ kg}$$

- 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 Given  $\bar{x} = 11.6$  and  $n = 10$ , calculate  $\sum_{i=1}^{10} x_i$ .
- 11 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.
- 12 Find  $x$  if 5, 9, 11, 12, 13, 14, 17, and  $x$  have a mean of 12.
- 13 Find  $a$  if 3, 0,  $a$ ,  $a$ , 4,  $a$ , 6,  $a$ , and 3 have a mean of 4.
- 14 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?
- 15 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.
- 16 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:**

- Consider the set of data: 4, 5, 6, 6, 6, 7, 7, 8, 9, 10. Calculate:
  - the mean
  - the mode
  - the median.
- 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:
  - the mean
  - the mode
  - the median.
- Comment on the effect that the extreme value has on:
  - the mean
  - the mode
  - the median.
- Which of the three measures of centre is most affected by the inclusion of an outlier?
- 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.

<i>Statistic</i>	<i>Properties</i>
<b>Mode</b>	<ul style="list-style-type: none"> <li>• gives the most usual value</li> <li>• only takes common values into account</li> <li>• not affected by extreme values</li> </ul>
<b>Mean</b>	<ul style="list-style-type: none"> <li>• commonly used and easy to understand</li> <li>• takes all values into account</li> <li>• affected by extreme values</li> </ul>
<b>Median</b>	<ul style="list-style-type: none"> <li>• gives the halfway point of the data</li> <li>• only takes middle values into account</li> <li>• not affected by extreme values</li> </ul>

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 13B**

- 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

- Calculate the mean and median selling prices. Comment on your results.
- Which measure would you use if you were:
  - a vendor wanting to sell your house
  - 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

- Find the mode, mean, and median salaries of this group.
- Explain why the mode is an unsatisfactory measure of the centre in this case.
- 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

- Use technology to find the mean, median, and mode of the data.
  - Explain why the median is not the most suitable measure of centre for this set of data.
  - Explain why the mode is not the most suitable measure of centre for this set of data.
  - Identify the outliers in this data set.
  - 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

- Calculate the mean, median, and modal number of children per family.
- Is the mode a useful statistic in this case?
- 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  $\{x_1, x_2, \dots, x_n\}$  include the:

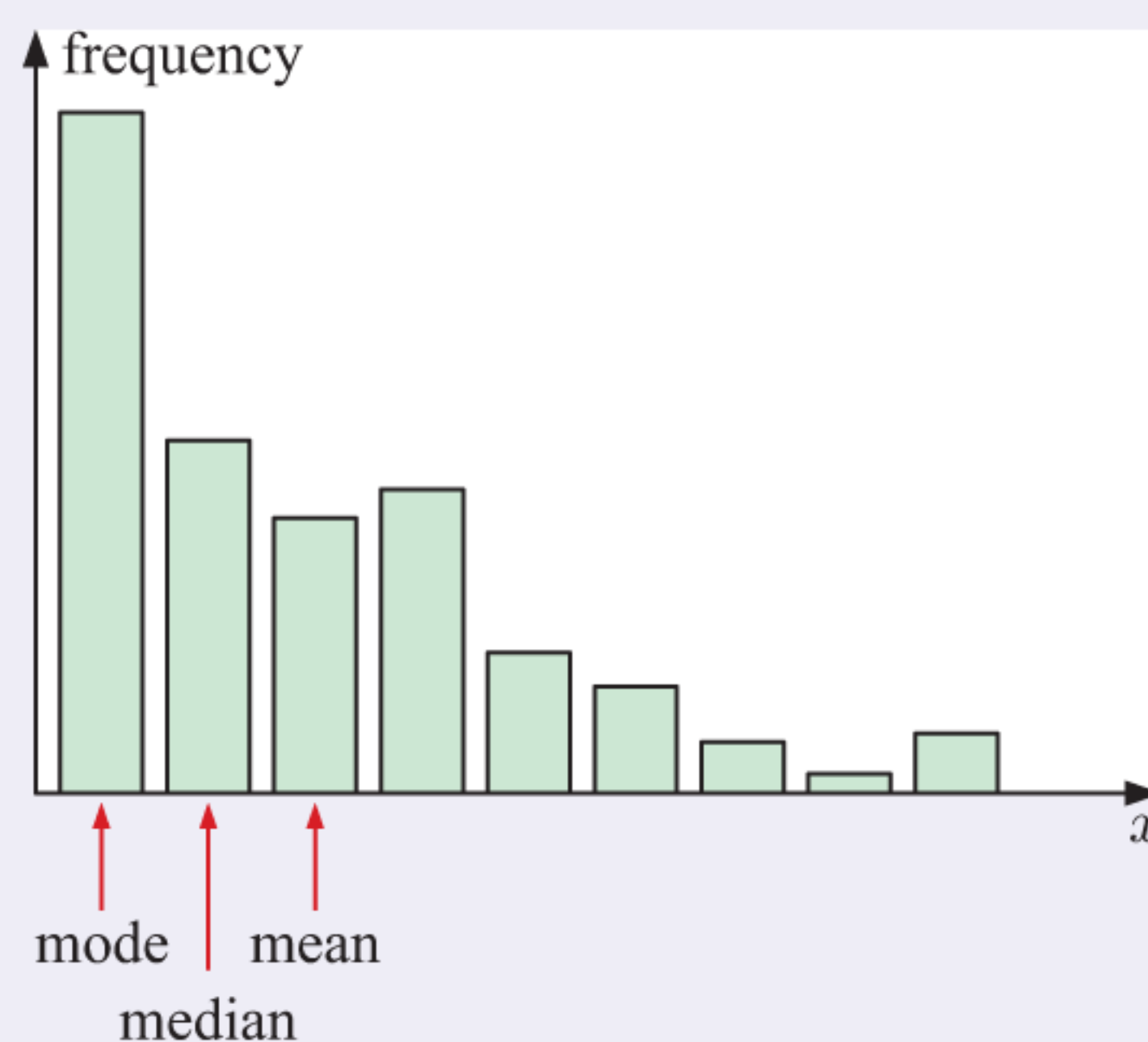
- geometric mean** =  $\sqrt[n]{x_1 \times x_2 \times \dots \times x_n}$
- mid-range value** =  $\frac{\text{maximum} + \text{minimum}}{2}$ .

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.

- Does the mode give an accurate indication of a data set’s centre in these cases?
- 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	$9 \times 5 = 45$
<i>Total</i>	$\sum f = 40$	$\sum xf = 278$

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

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_k f_k}{f_1 + f_2 + f_3 + \dots + f_k}$$

where  $k$  is the number of *different* values in the data.

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

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



## THE MEDIAN

Since  $\frac{n+1}{2} = \frac{41}{2} = 20.5$ , 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.

$$\therefore \text{the median} = \frac{7+7}{2} = 7$$

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	$\sum f = 55$	$\sum xf = 179$	

$$\begin{aligned} \mathbf{a} \quad \bar{x} &= \frac{\sum xf}{\sum f} \\ &= \frac{179}{55} \\ &\approx 3.25 \text{ aces} \end{aligned}$$

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



- b** There are 55 data values, so  $n = 55$ .  $\frac{n+1}{2} = 28$ , 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.



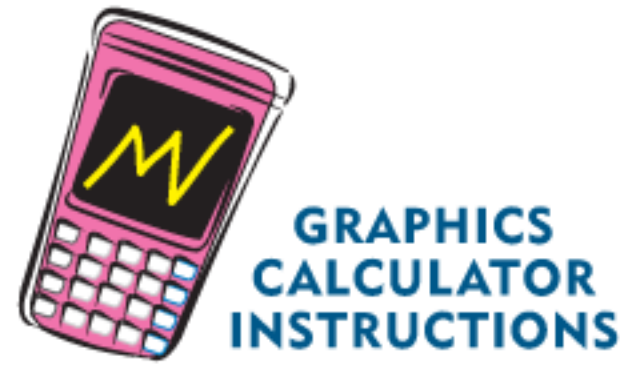
### EXERCISE 13C

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

Calculate the:

- a mode                      b median                      c mean.

Check your answers using your graphics calculator.



Number of people	Frequency
1	13
2	8
3	4
4	5
<i>Total</i>	30

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

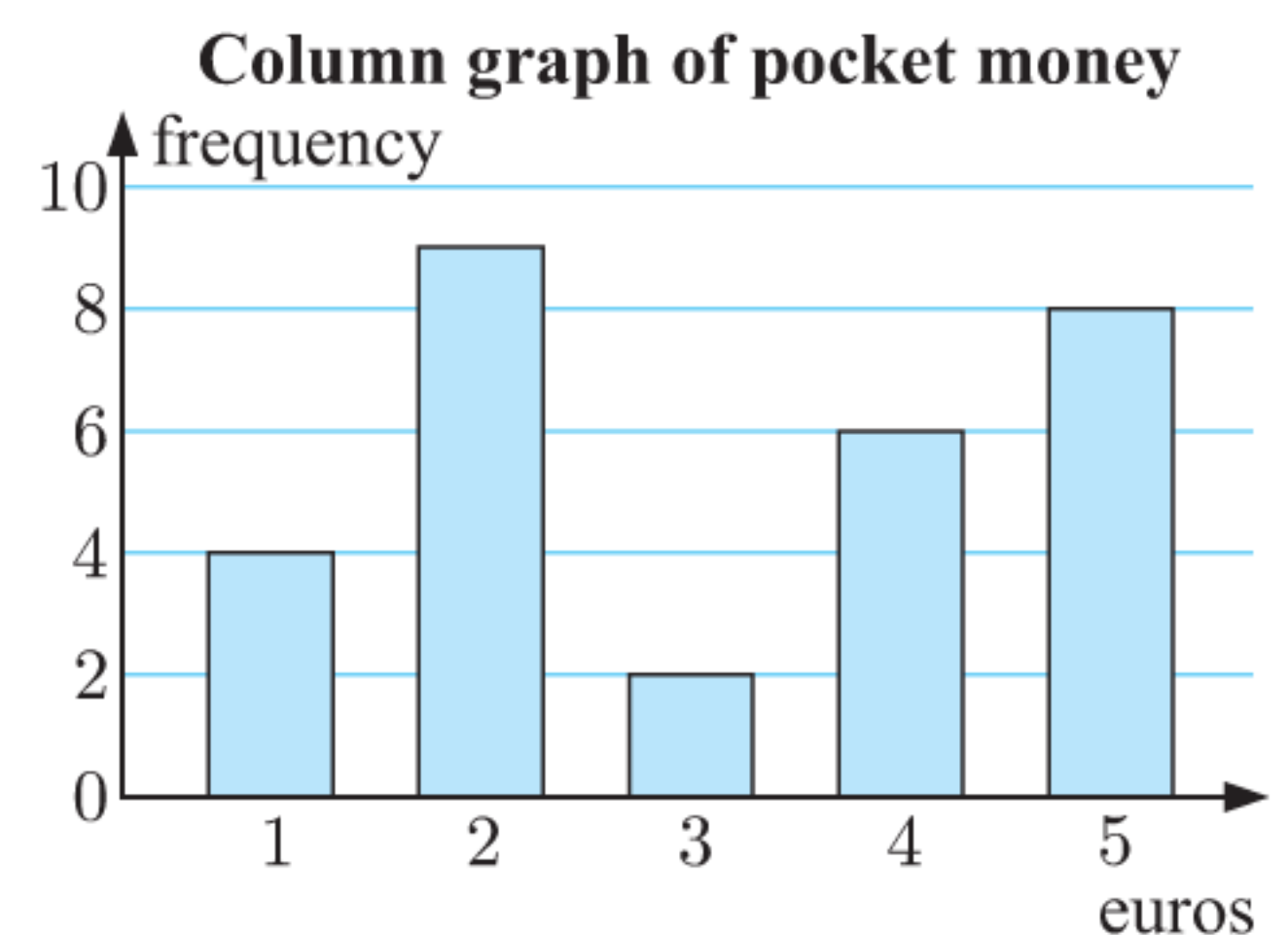
3 Families at a school in Manchester were surveyed, and the number of children in each family was recorded. The results of the survey are shown alongside.

- 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?

Number of children	Frequency
1	5
2	28
3	15
4	8
5	2
6	1
<i>Total</i>	59

4 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?



5 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.



- 6 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:
- the median salary
  - the modal salary
  - 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?



- 7 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?

This table summarises the marks out of 50 received by students in 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:

- 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* .....”
- 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* .....”
- 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.
  - Calculate the mean of these results.
  - How does this result compare with lower and upper limits found in **1** and **2**?
  - Copy and complete: “The mean Physics examination mark was approximately .....”
- 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**



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	$\sum f = 137$		$\sum xf = 5161$

$$\begin{aligned} \bar{x} &= \frac{\sum xf}{\sum f} \\ &= \frac{5161}{137} \\ &\approx 37.7 \end{aligned}$$

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

**EXERCISE 13D**

- Simone recorded the lengths of her phone calls for one week. The results are shown in the table alongside.
  - How many phone calls did she make during the week?
  - Estimate the mean length of the calls.

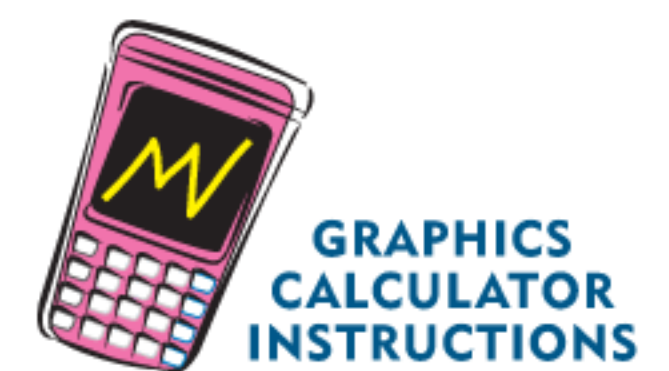
Time ( $t$ min)	Frequency
$0 \leq t < 10$	17
$10 \leq t < 20$	10
$20 \leq t < 30$	9
$30 \leq t < 40$	4

The midpoint of an interval is the average of its endpoints.



- 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.

- The table shows the petrol sales in one day by a number of city service stations.
  - How many service stations were involved in the survey?
  - Estimate the total amount of petrol sold for the day by the service stations.
  - Estimate the mean amount of petrol sold for the day.
  - Find the modal class for this distribution. Explain your answer.

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



4 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?



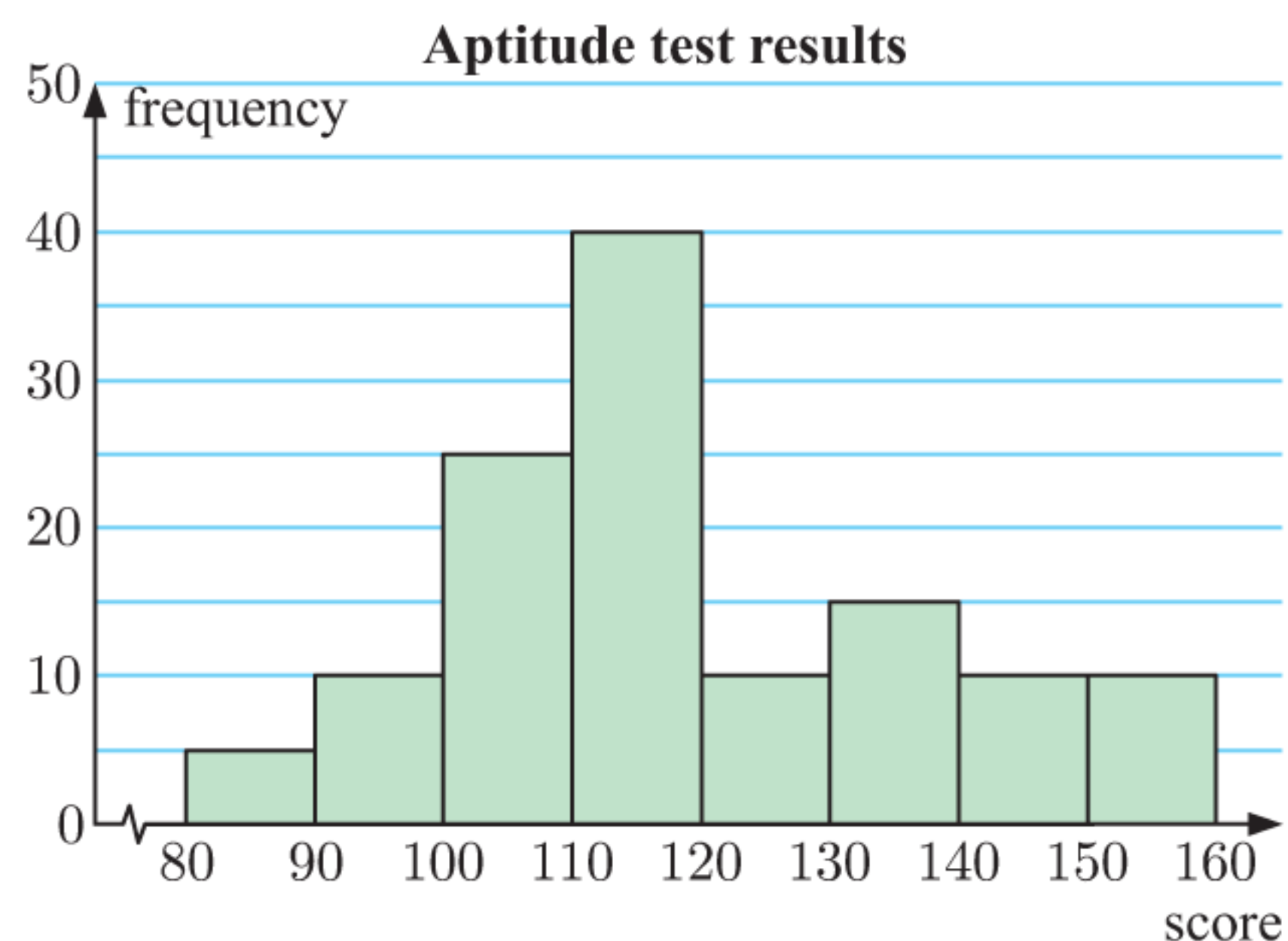
5 The manager of a bank decides to investigate the time customers wait to be served. The results for 300 customers are shown in the table alongside.

- 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
$0 \leq t < 1$	$p$
$1 \leq t < 2$	42
$2 \leq t < 3$	50
$3 \leq t < 4$	78
$4 \leq t < 5$	60
$5 \leq t < 6$	30
$6 \leq t < 7$	16

6 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 **interquartile range**
- the **variance**
- 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.

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

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.

$$\begin{aligned} \text{Range} &= \text{maximum} - \text{minimum} \\ &= 5.1 - 2.3 \\ &= 2.8 \text{ kg} \end{aligned}$$

## 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.

$$\begin{aligned} \text{interquartile range} &= \text{upper quartile} - \text{lower quartile} \\ \text{IQR} &= Q_3 - Q_1 \end{aligned}$$

The median is sometimes referred to as  $Q_2$  because it is the 2nd quartile.



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

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

$\therefore$  median = 5

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

$\underbrace{2 \ 3 \ 3 \ 4 \ 4 \ 5}_{\text{lower half}} \quad \underbrace{5 \ 6 \ 6 \ 7 \ 7 \ 8}_{\text{upper half}}$

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

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

- c  $\text{IQR} = Q_3 - Q_1 = 6.5 - 3.5 = 3$











## F

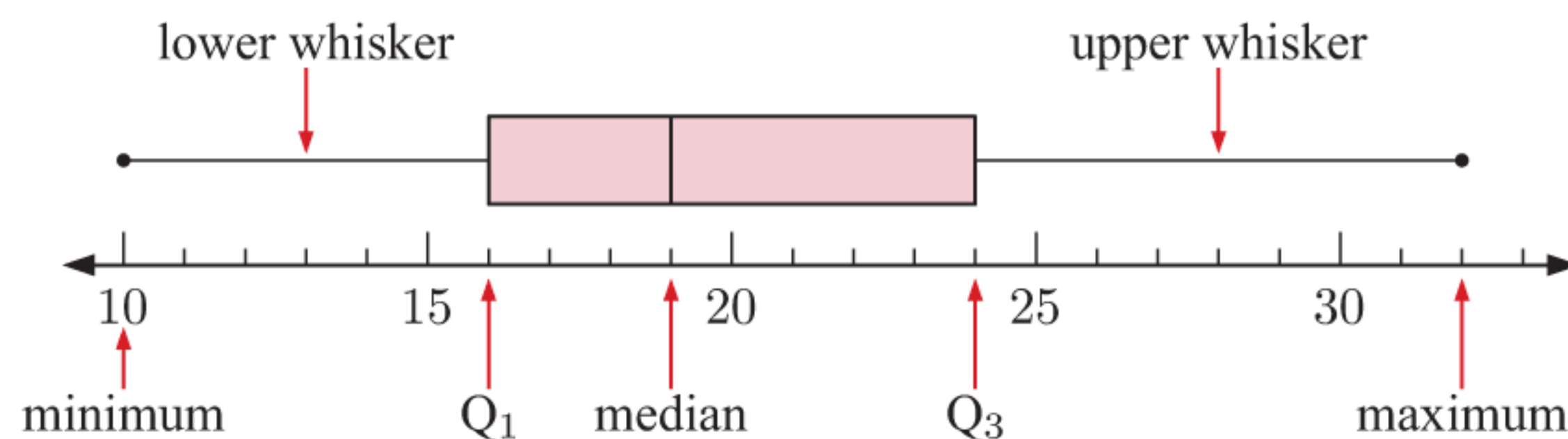
## BOX AND WHISKER DIAGRAMS

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 330, the five-number summary and box plot are:

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

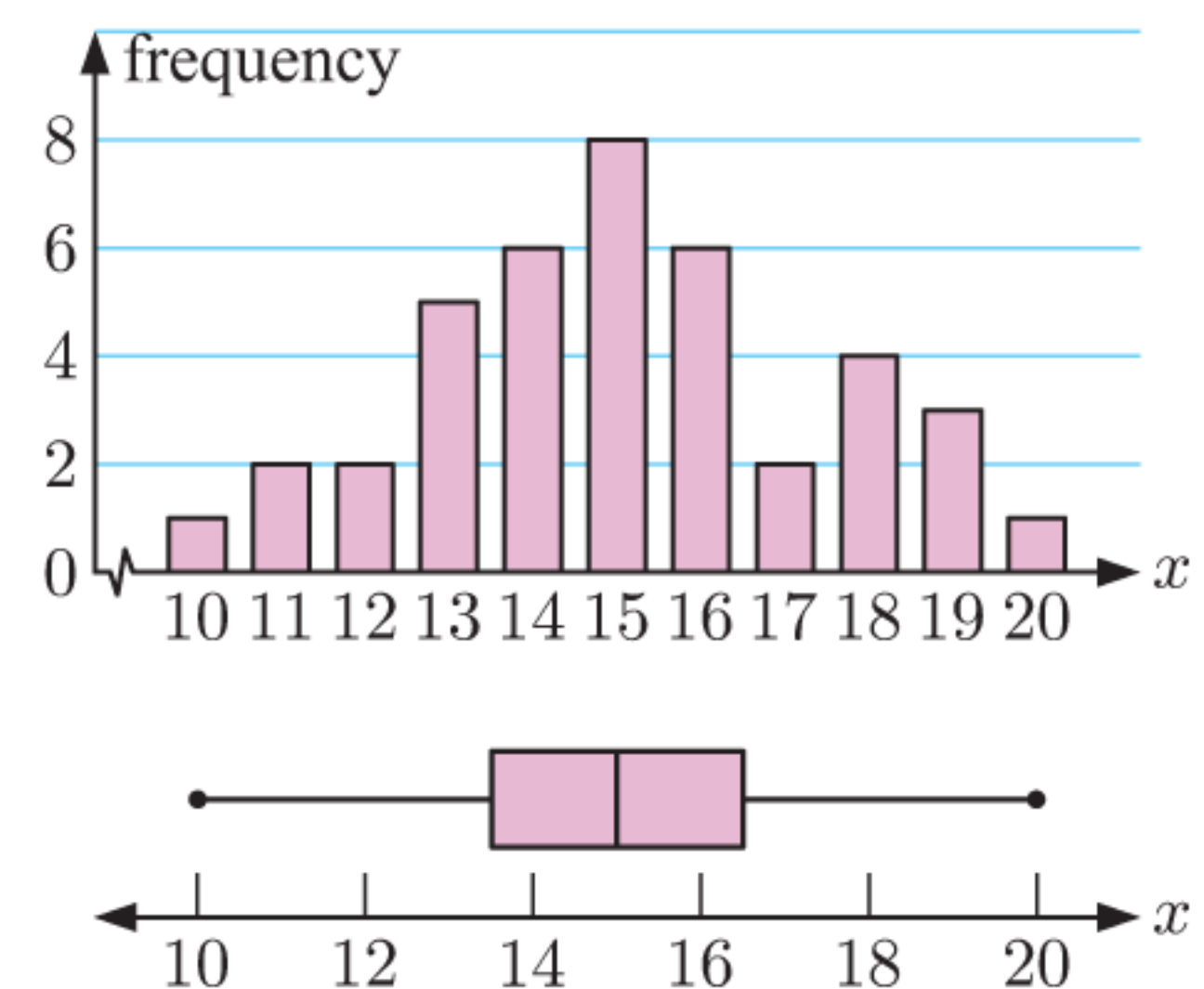


You should notice that:

- 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.

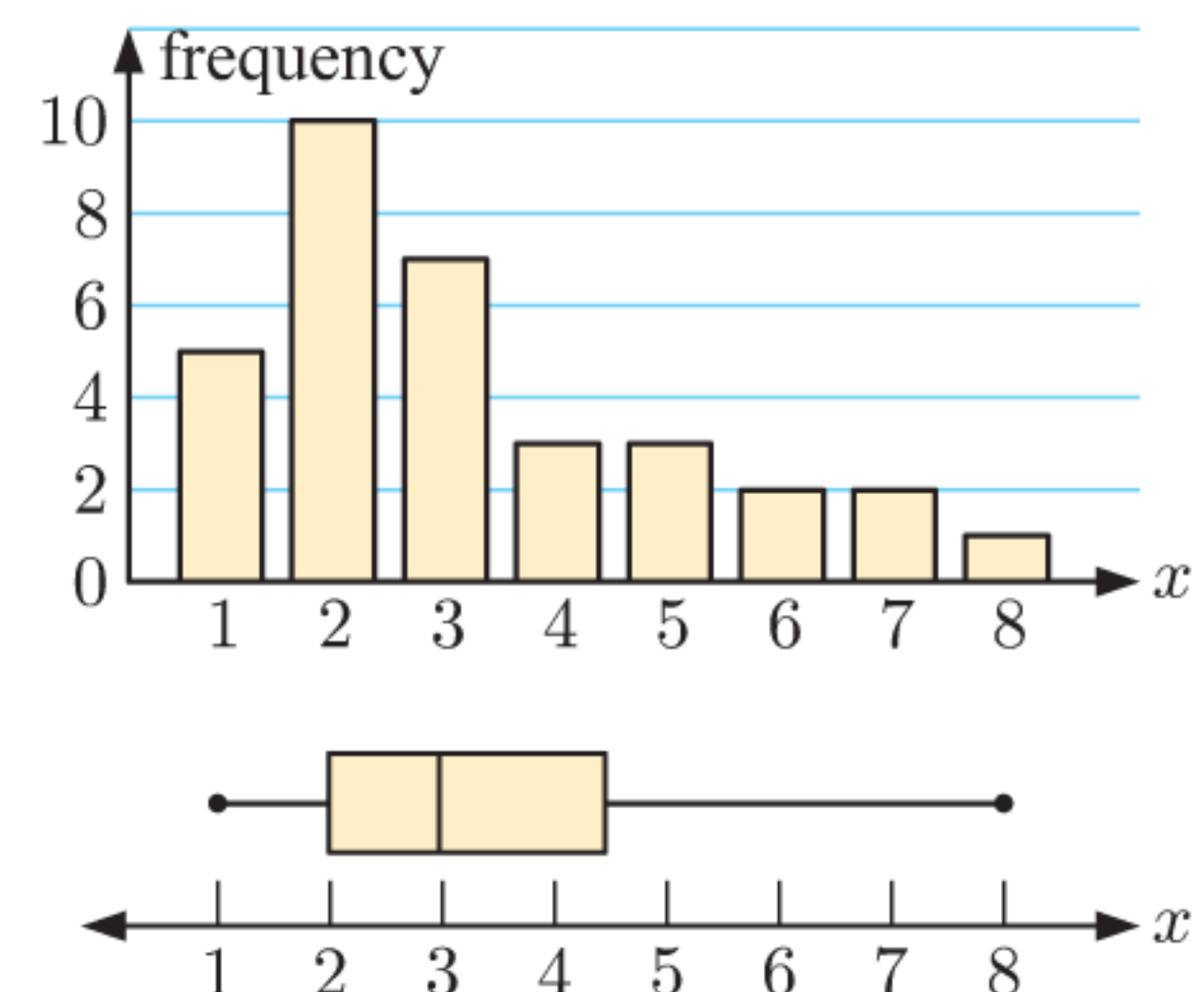
## INTERPRETING A BOX PLOT

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.



## F

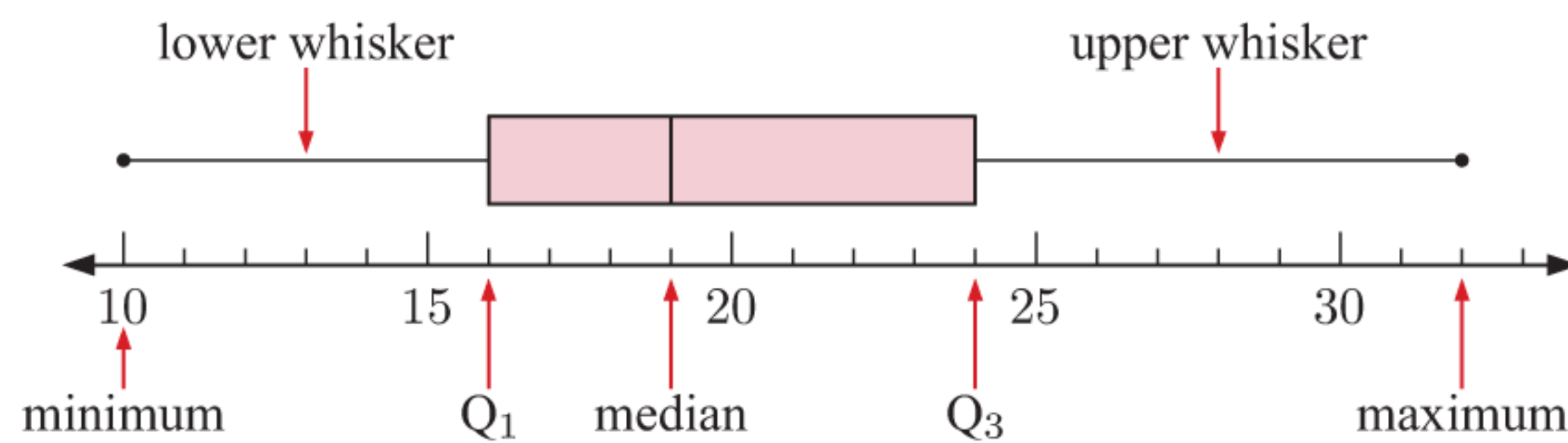
## BOX AND WHISKER DIAGRAMS

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 330, the five-number summary and box plot are:

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

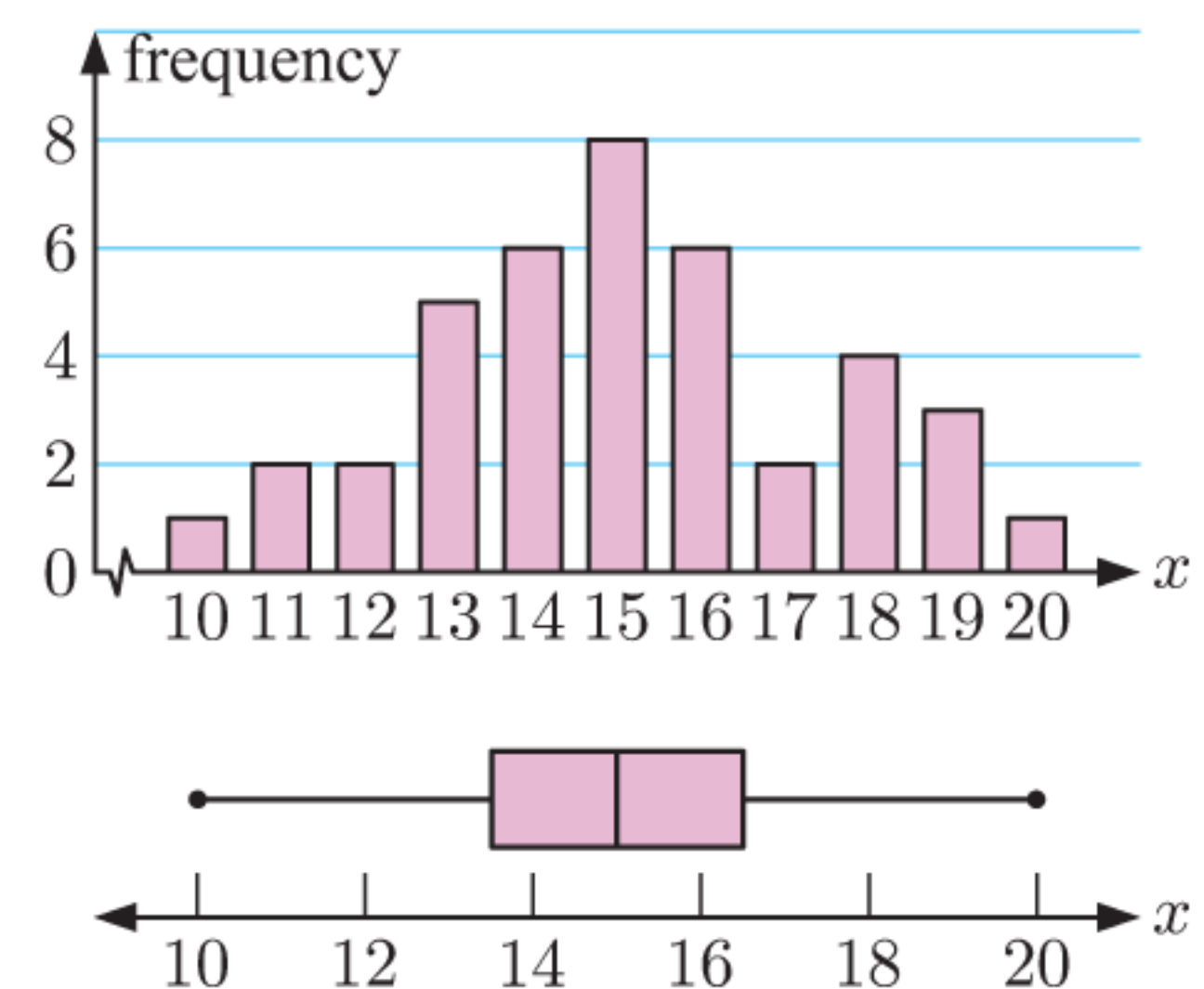


You should notice that:

- 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.

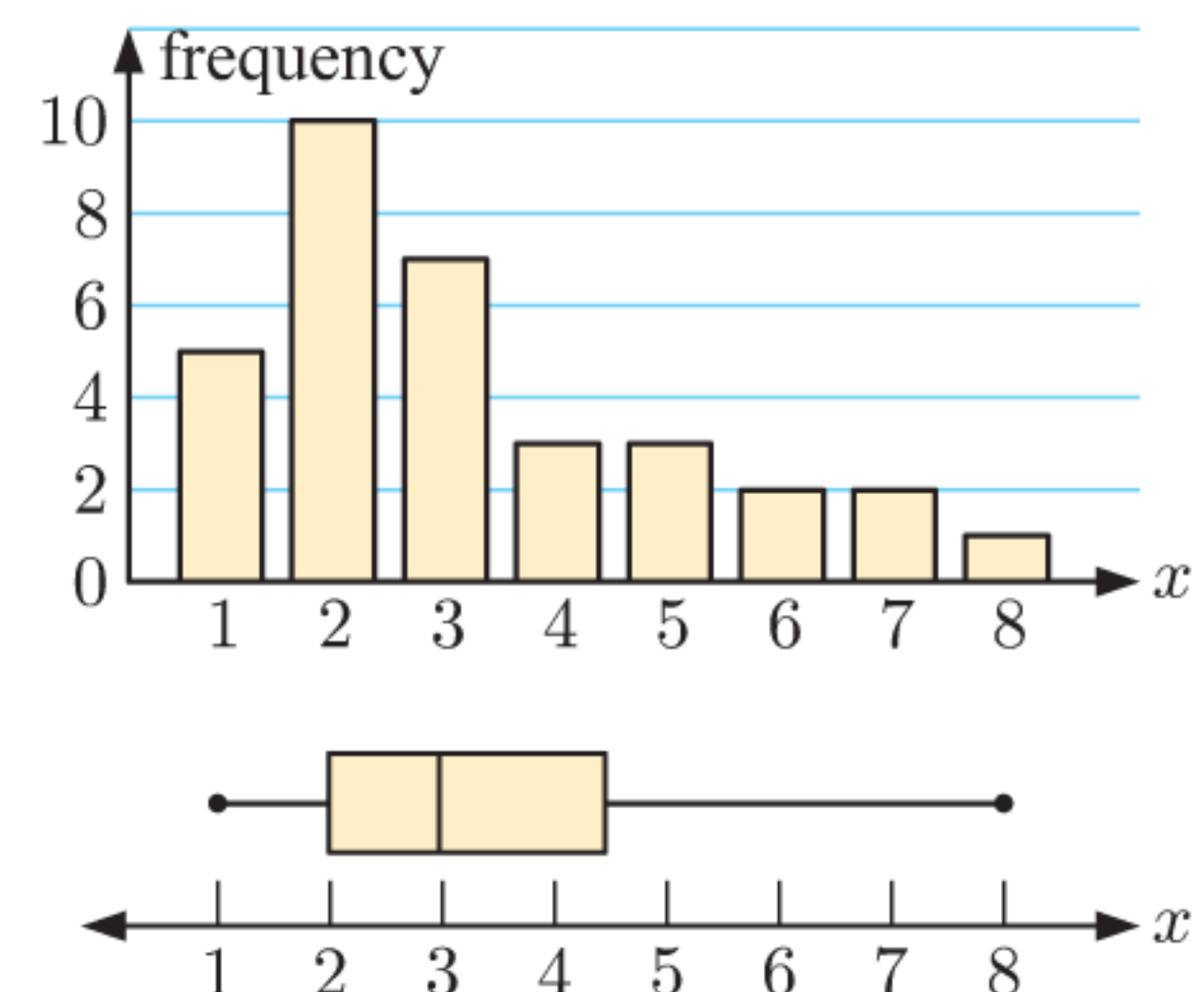
## INTERPRETING A BOX PLOT

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.







# 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. There may be 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

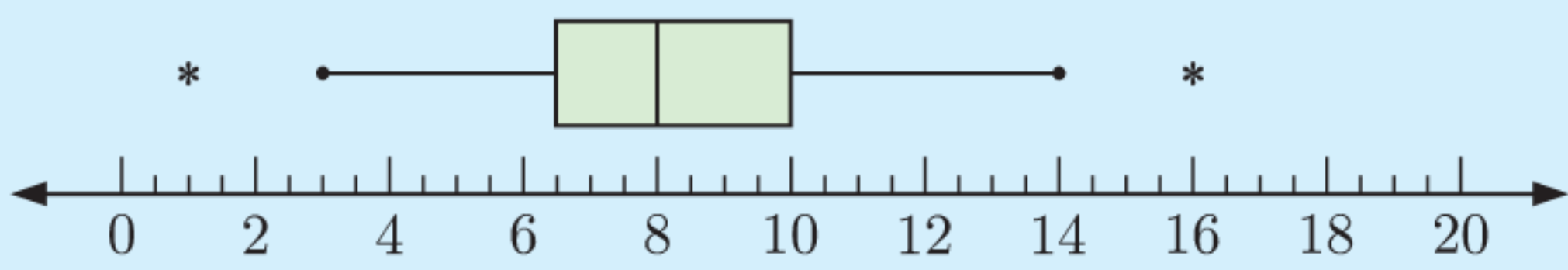
The ordered data set is:

1	3	3	5	6	7	7	7	8	8	8	8	9	9	10	10	12	13	14	16	{n = 20}
↓				↓				↓	↓					↓					↓	
min = 1				Q <sub>1</sub> = 6.5				median = 8						Q <sub>3</sub> = 10					max = 16	


IQR = Q<sub>3</sub> – Q<sub>1</sub> = 3.5

<i>Test for outliers:</i>	upper boundary = upper quartile + 1.5 × IQR = 10 + 1.5 × 3.5 = 15.25	and	lower boundary = lower quartile – 1.5 × IQR = 6.5 – 1.5 × 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.



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



## EXERCISE 13G

- 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.



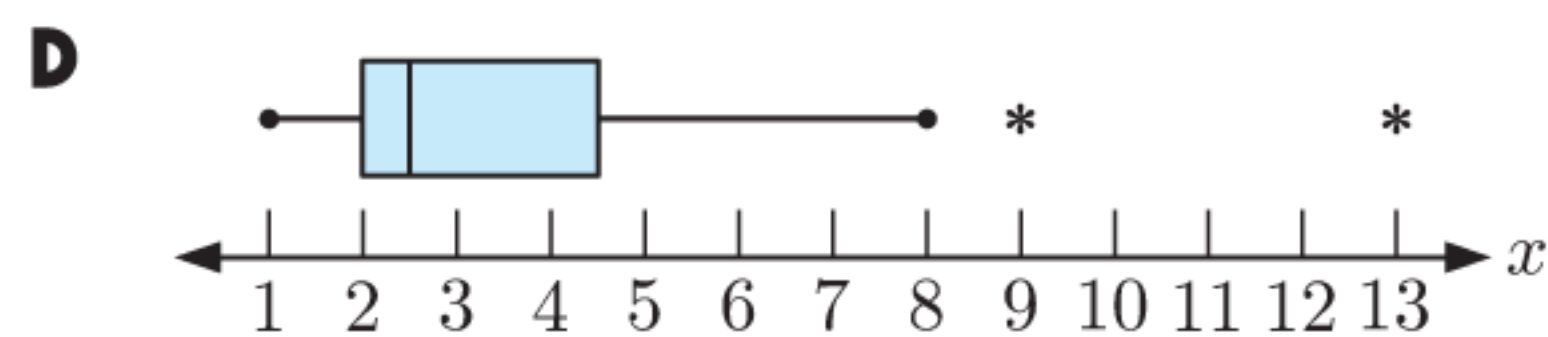
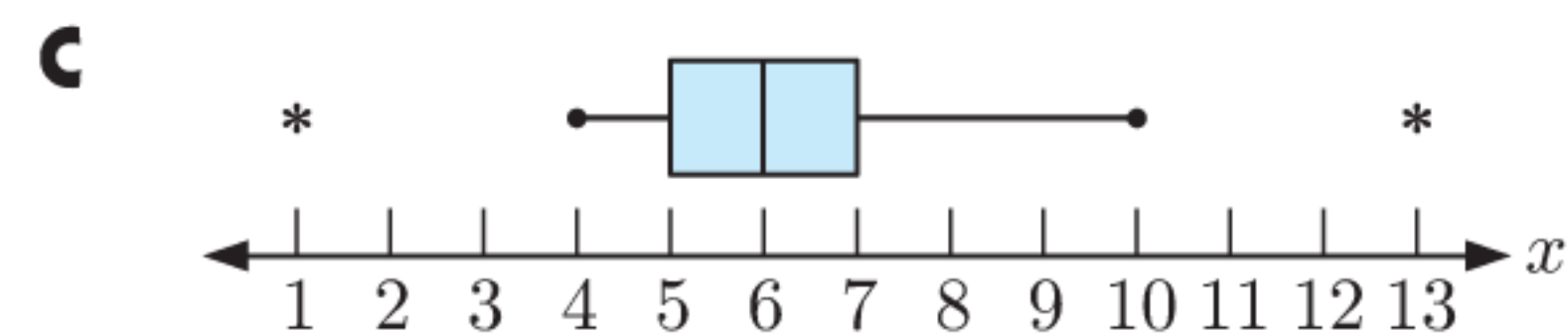
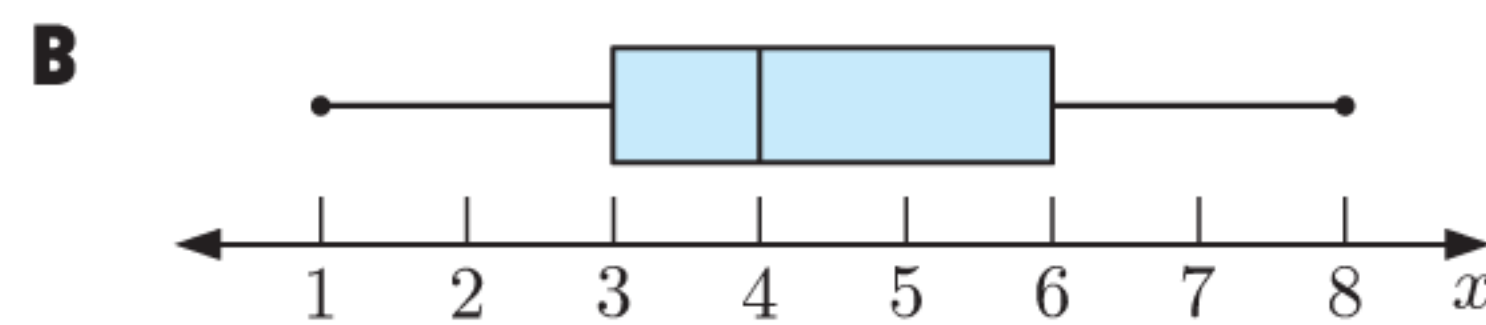
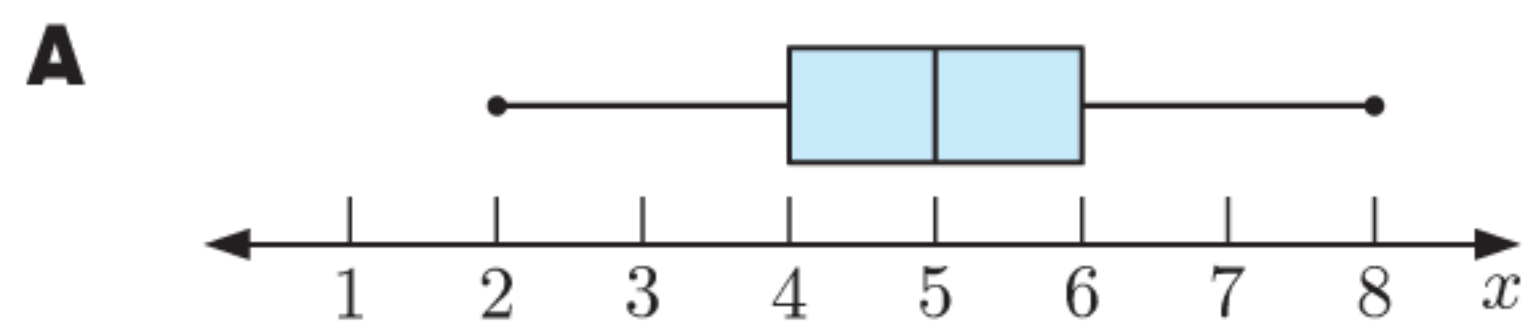
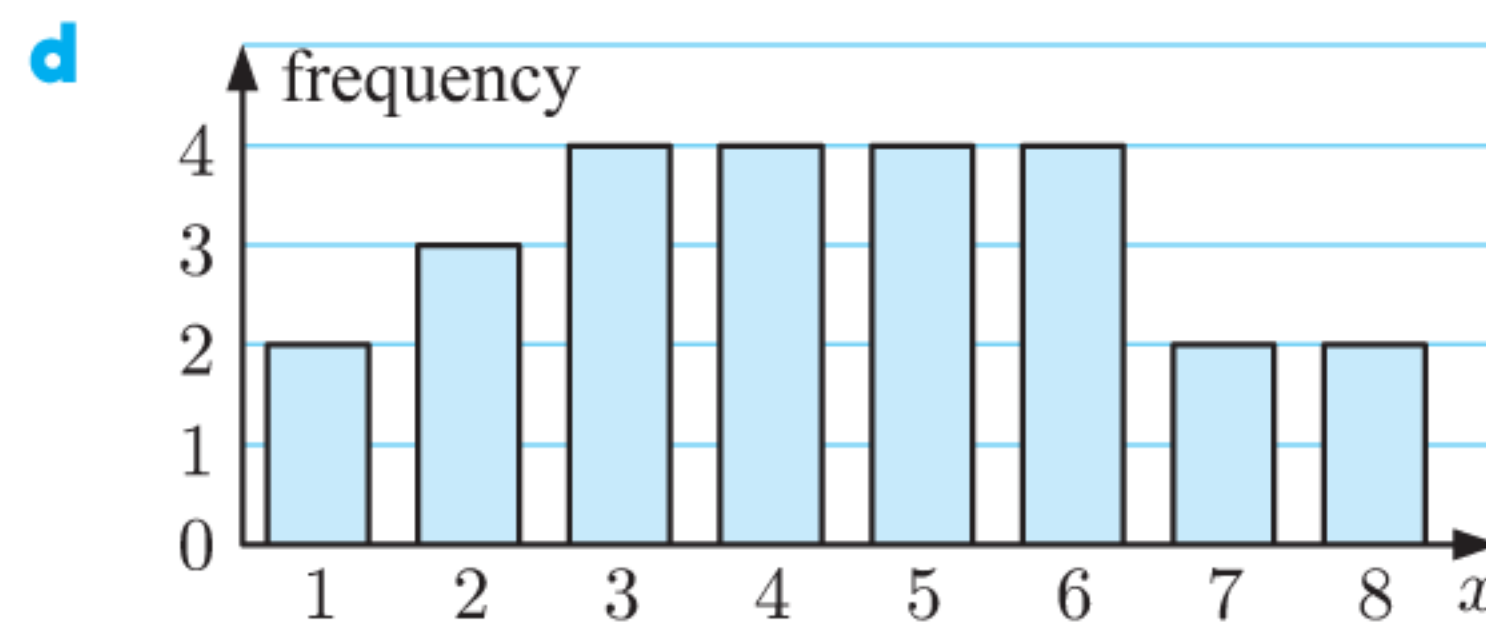
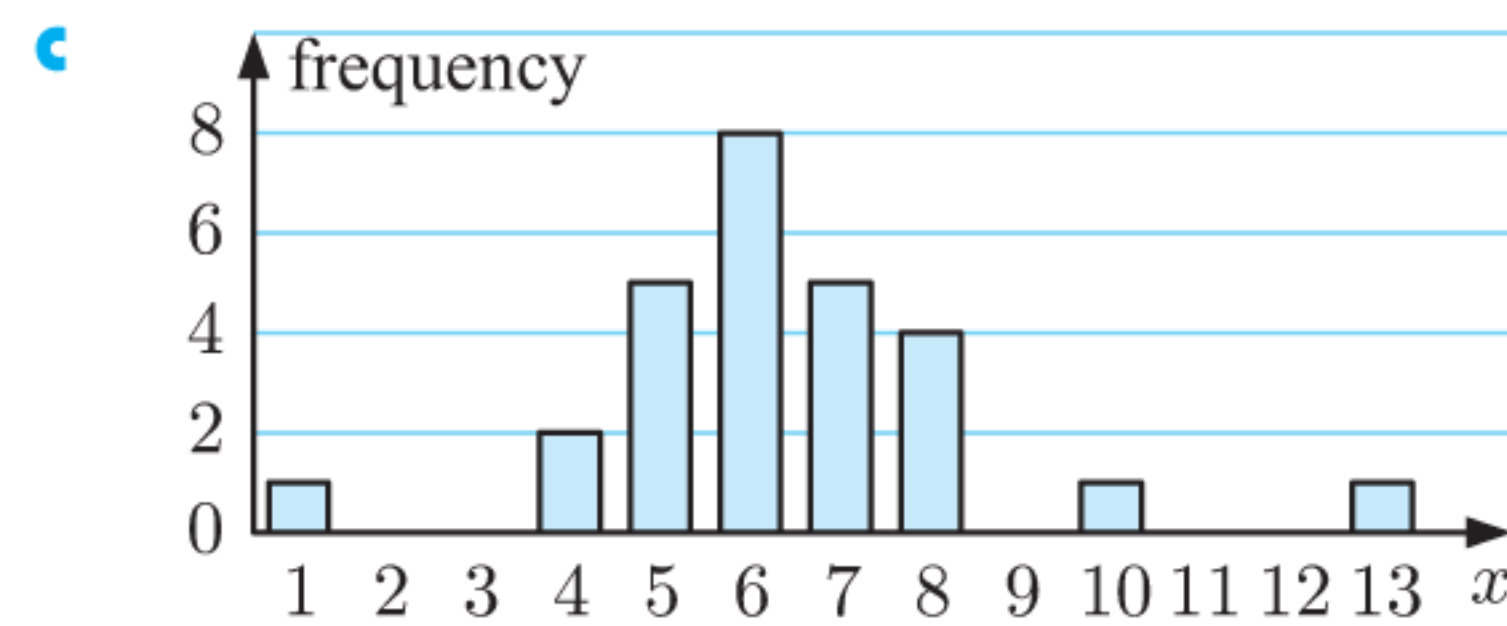
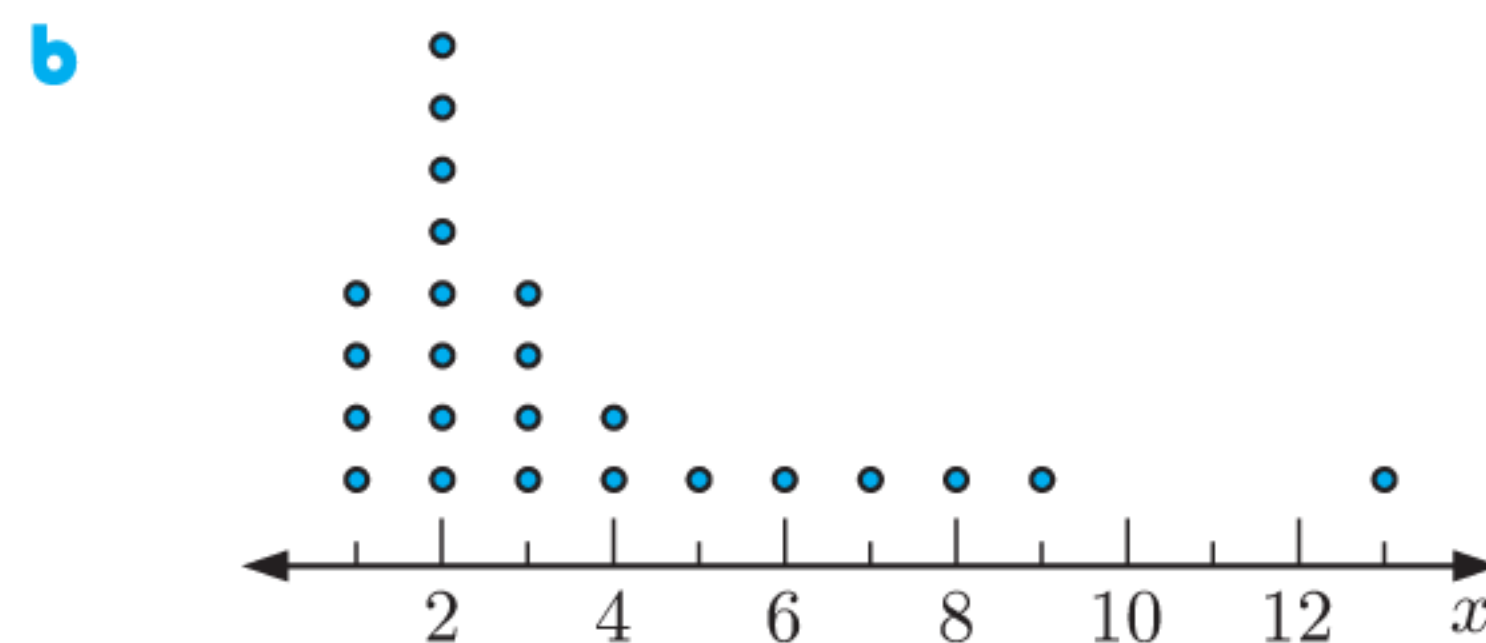
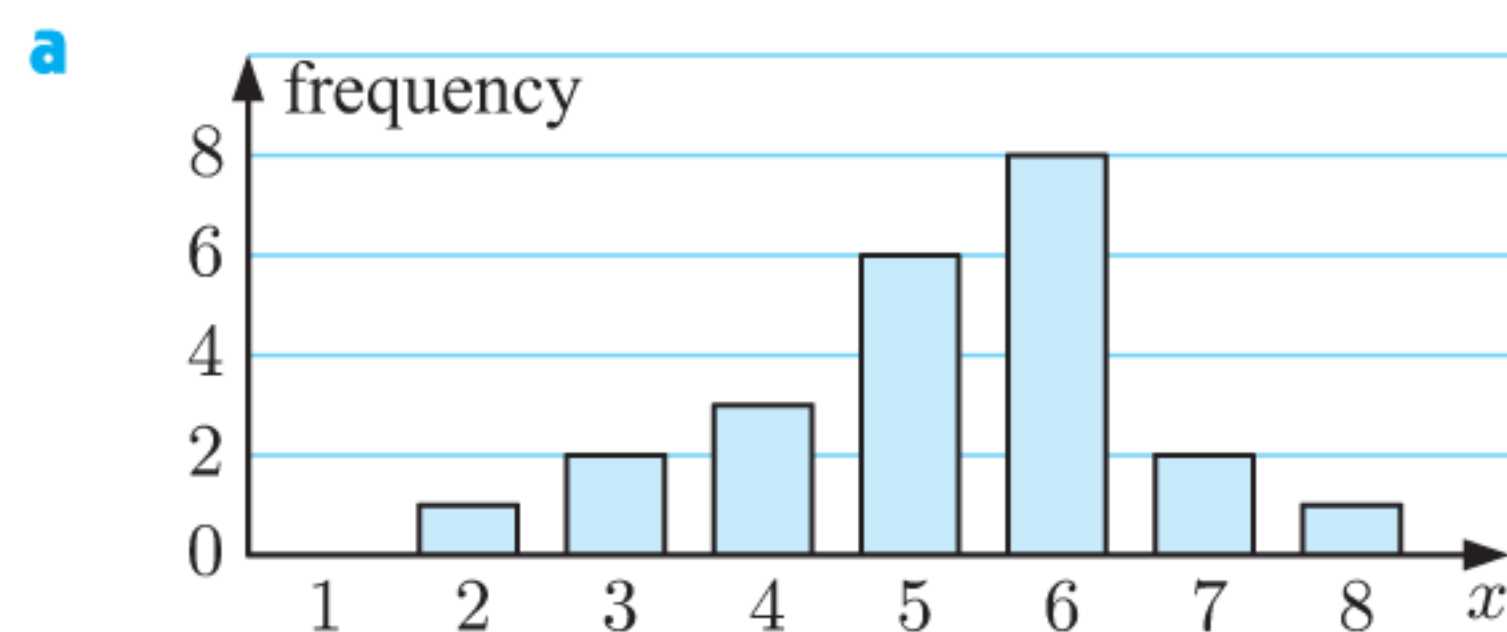
2 James goes bird watching for 25 days. The number of birds he sees each day are:

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:



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.



# 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.

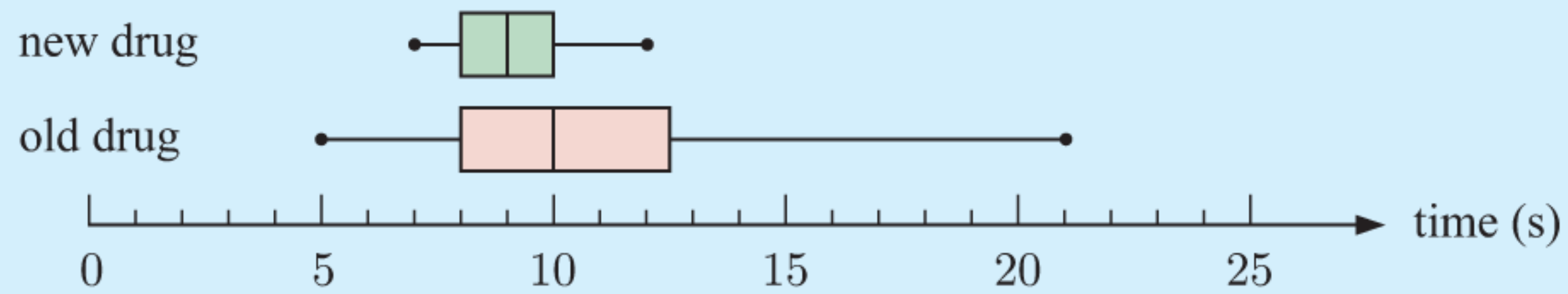
*Old drug times (s):* 8, 12, 9, 8, 16, 10, 14, 7, 5, 21,  
13, 10, 8, 10, 11, 8, 11, 9, 11, 14

*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	For the new drug:	min = 7
	Q <sub>1</sub> = 8		Q <sub>1</sub> = 8
	median = 10		median = 9
	Q <sub>3</sub> = 12.5		Q <sub>3</sub> = 10
	max = 21		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:

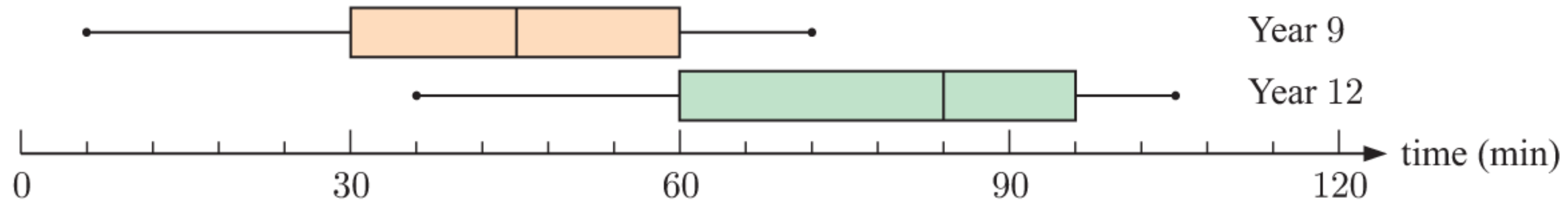
range for old drug = 21 – 5	range for new drug = 12 – 7
= 16	= 5
IQR for old drug = Q <sub>3</sub> – Q <sub>1</sub>	IQR for new drug = Q <sub>3</sub> – Q <sub>1</sub>
= 12.5 – 8	= 10 – 8
= 4.5	= 2

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



### EXERCISE 13H

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 <sub>1</sub>		
median		
Q <sub>3</sub>		
maximum		

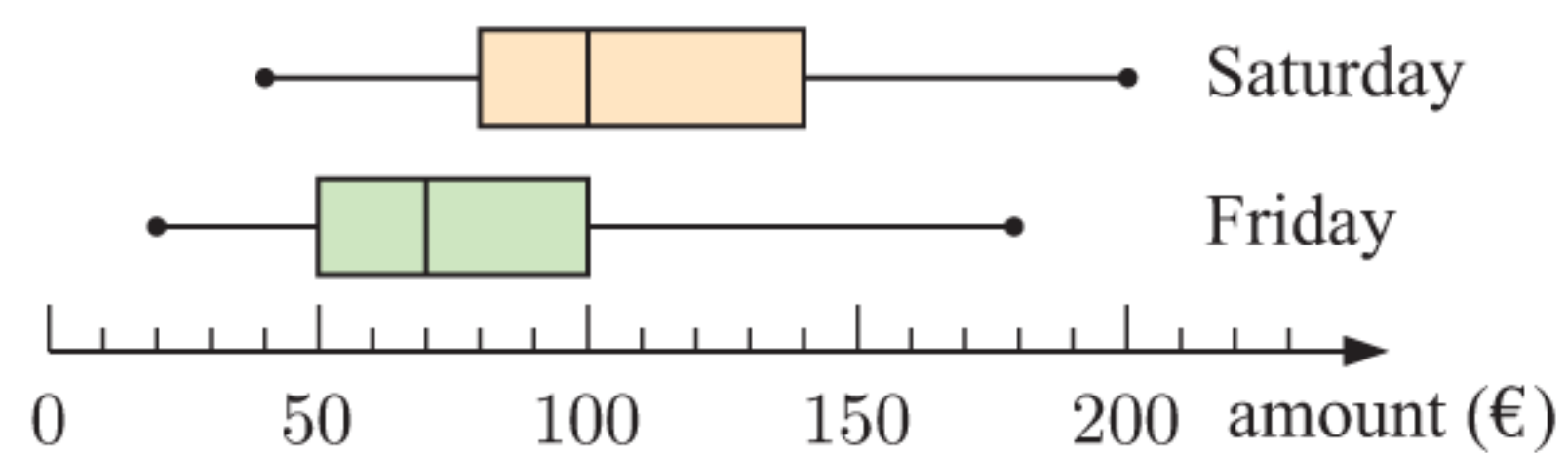
b For each group, determine the:

- 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.

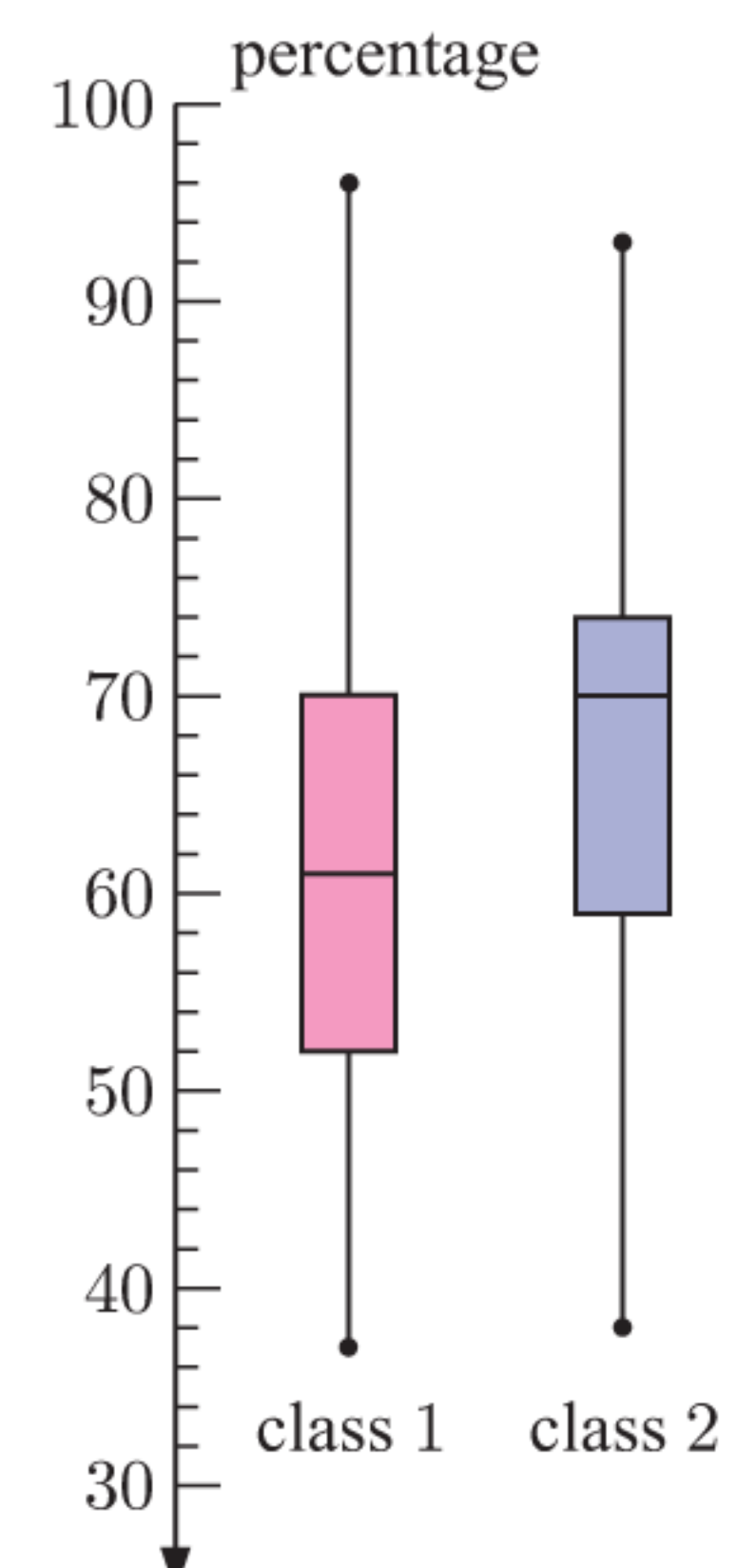
e Describe the distribution of marks in:

- 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

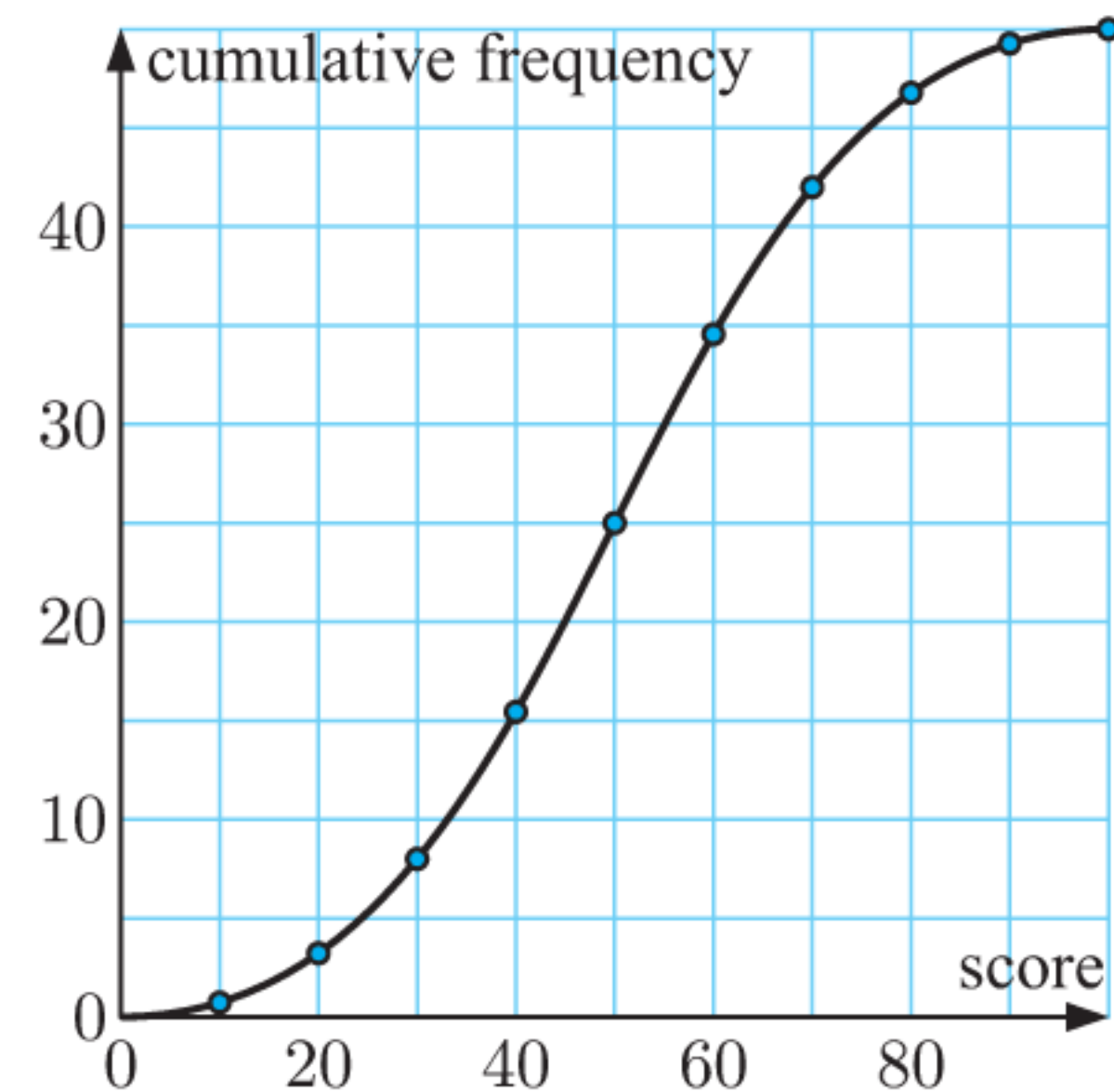
- a Is the variable discrete or continuous?
- b Enter the data into a graphics calculator or statistics package. Compare the measures of centre and spread.
- c Draw a parallel box plot.
- d Describe the shape of each distribution.
- e What conclusions, if any, can be drawn from the data?



# 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** ( $Q_1$ ) is the 25th percentile
- the **median** ( $Q_2$ ) is the 50th percentile
- the **upper quartile** ( $Q_3$ ) 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.

- Add a cumulative frequency column to the table.
- Represent the data on a cumulative frequency graph.
- Use your graph to estimate the:
  - median finishing time
  - number of competitors who finished in less than 155 minutes
  - percentage of competitors who took more than 159 minutes to finish
  - time taken by a competitor who finished in the top 20% of runners completing the marathon.

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

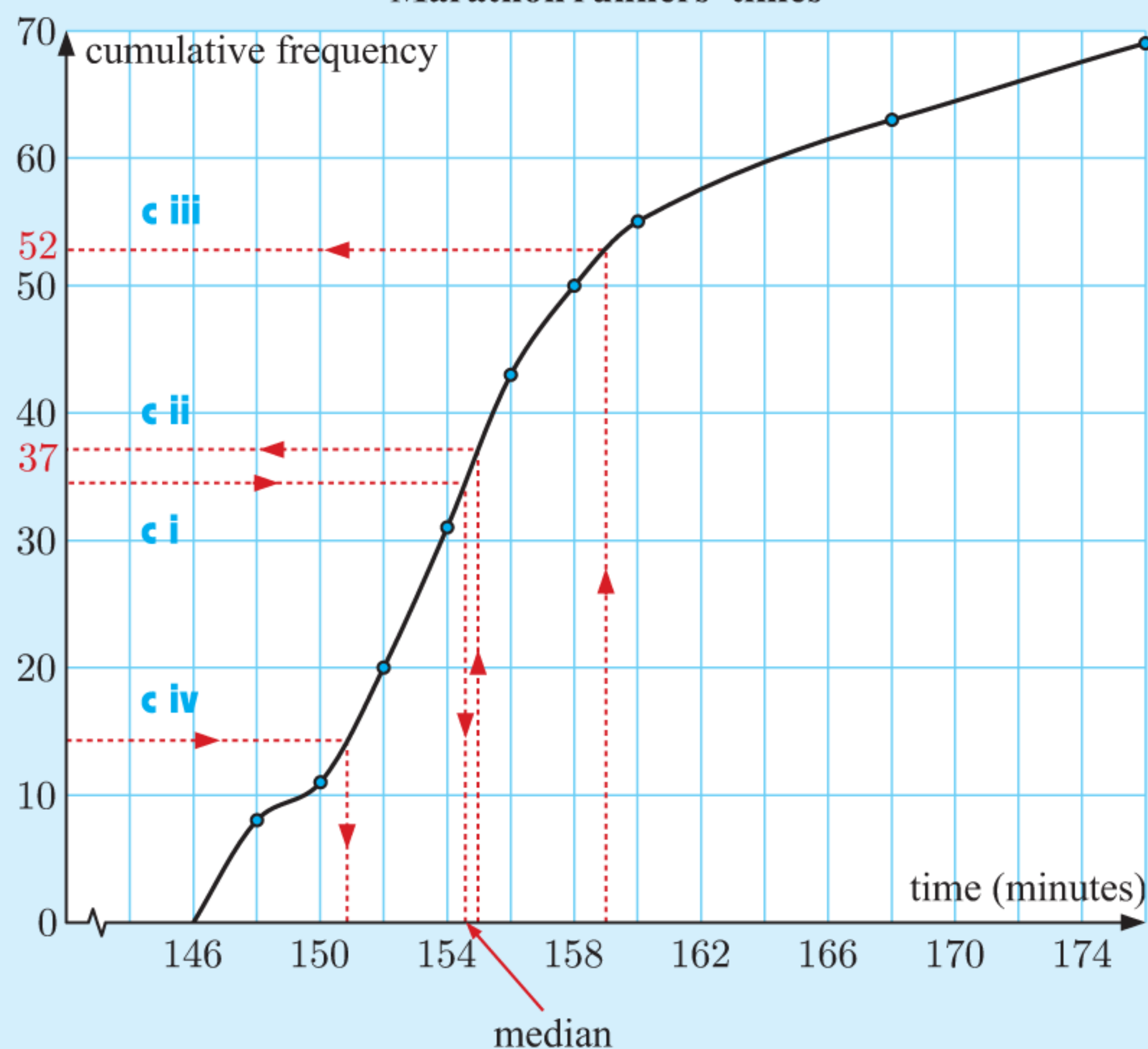


**a**

Time ( $t$ min)	Frequency	Cumulative frequency
$146 \leq t < 148$	8	8
$148 \leq t < 150$	3	11
$150 \leq t < 152$	9	20
$152 \leq t < 154$	11	31
$154 \leq t < 156$	12	43
$156 \leq t < 158$	7	50
$158 \leq t < 160$	5	55
$160 \leq t < 168$	8	63
$168 \leq t < 176$	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**
**Marathon runners' times**


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

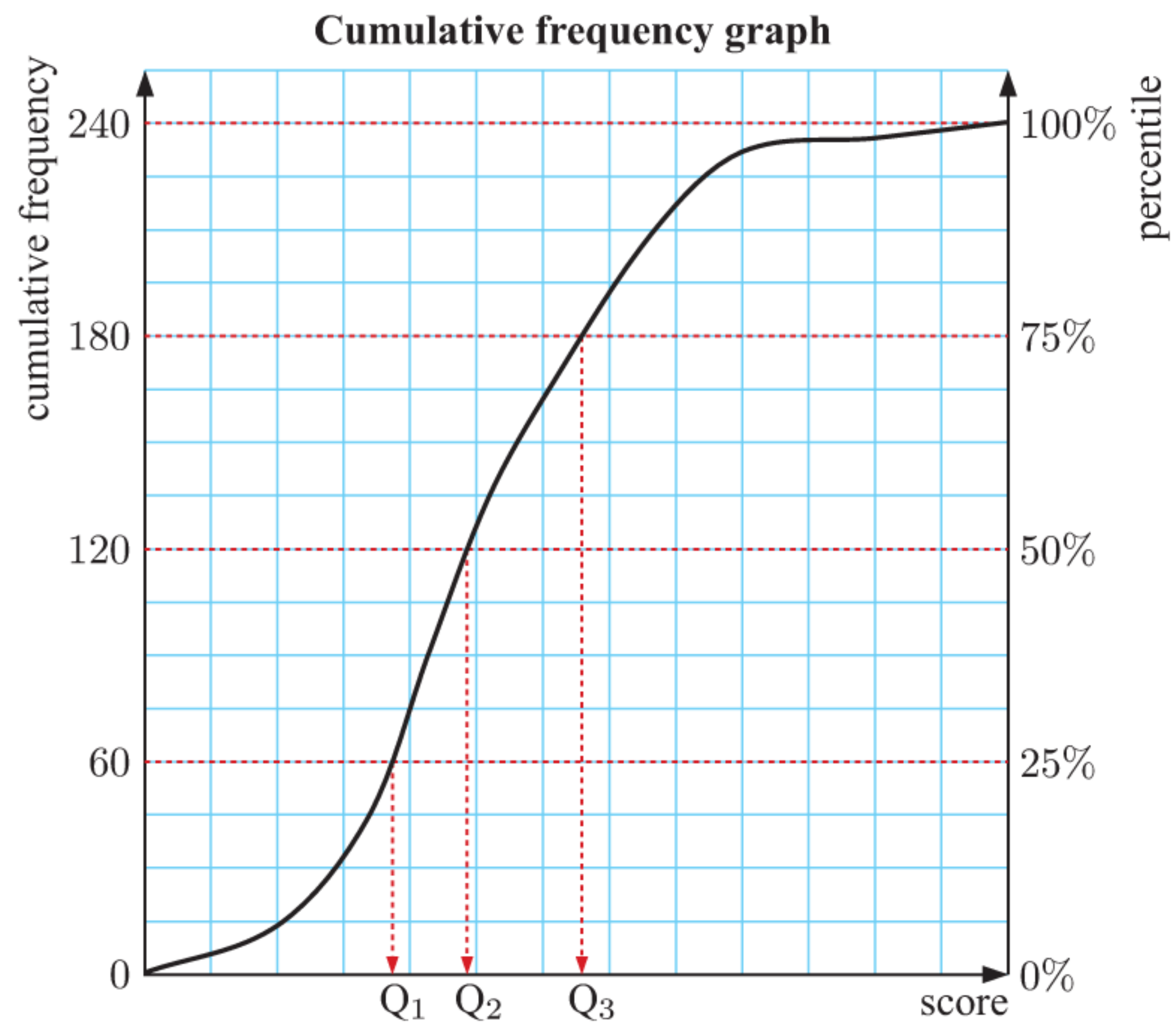


- c**
- i** The median is the 50th percentile. As 50% of 69 is 34.5, we start with the cumulative frequency 34.5 and find the corresponding time.  
The median  $\approx 154.5$  min.
  - ii** Approximately 37 competitors took less than 155 min to complete the race.
  - iii**  $69 - 52 = 17$  competitors took more than 159 min.  
 $\therefore \frac{17}{69} \approx 24.6\%$  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.

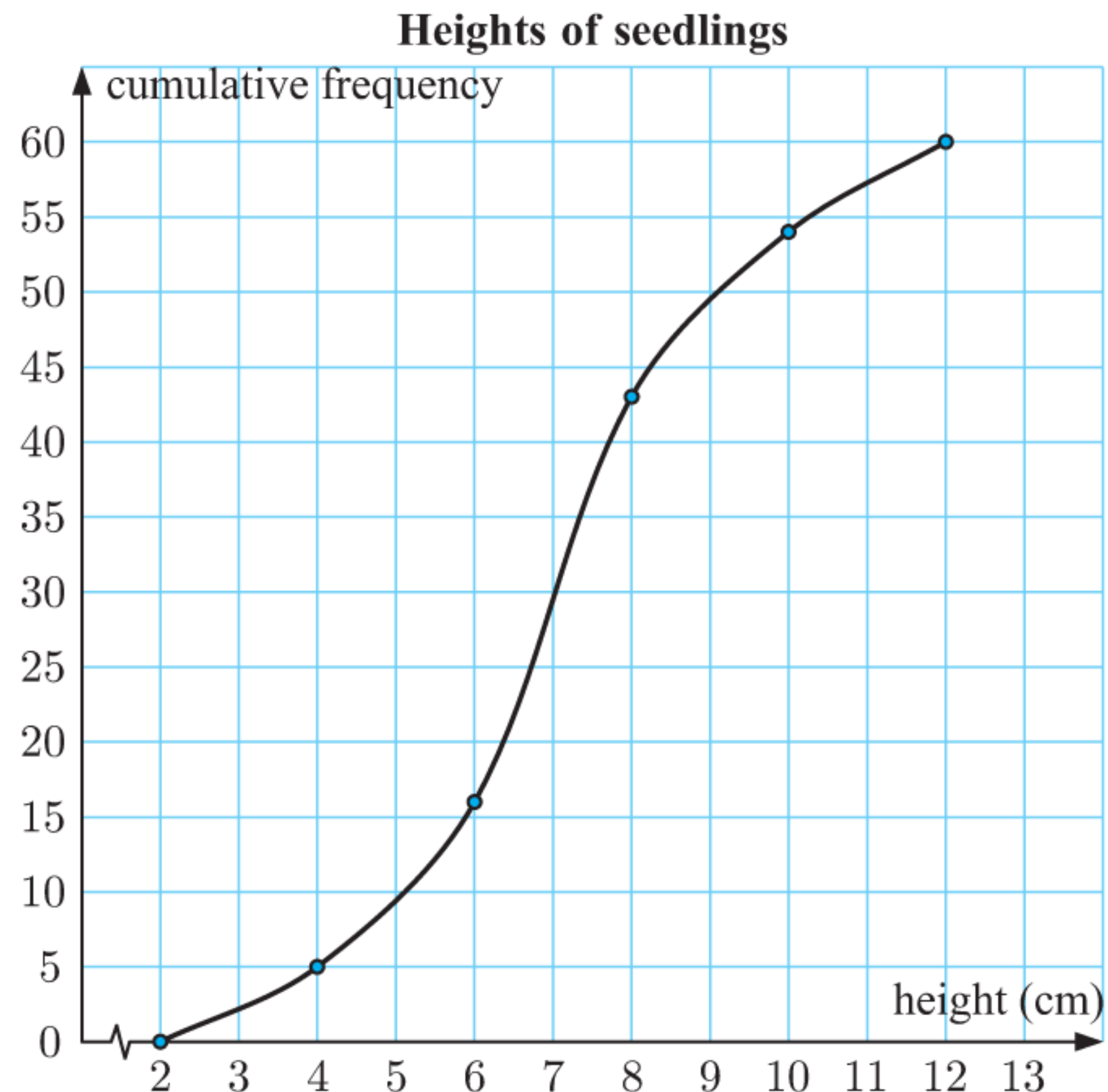


**EXERCISE 13I**

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

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

- 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.





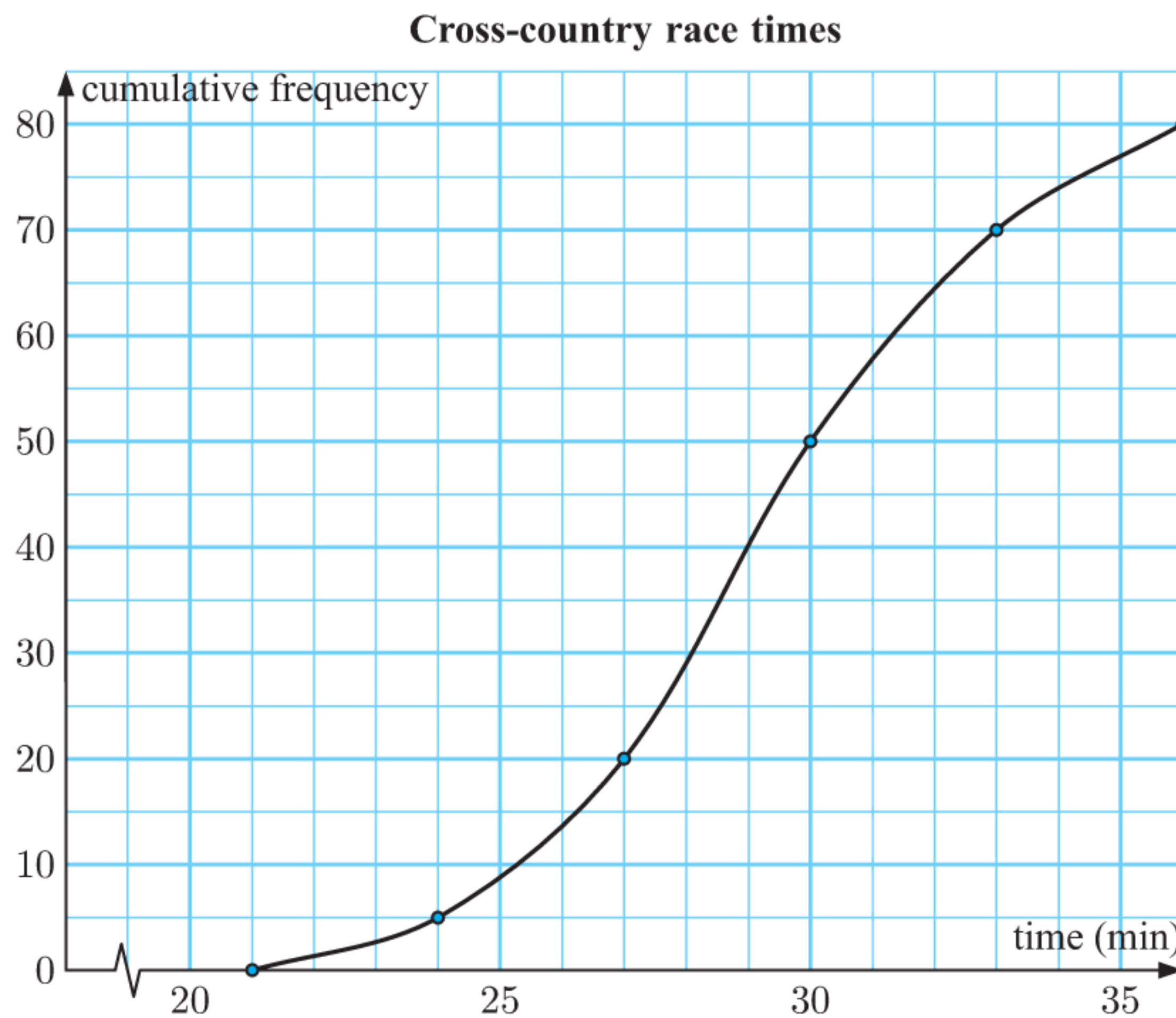
- 3** The following table summarises the age groups of car drivers involved in accidents in a city for a given year.
- a** Draw a cumulative frequency graph for the data.
  - b** Estimate the median age of the drivers involved in accidents.
  - c** Estimate the percentage of drivers involved in accidents 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.

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

- 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  $24 \leq x < 27$ ,  $27 \leq x < 30$ , 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)	$21 \leq t < 24$	$24 \leq t < 27$	$27 \leq t < 30$	$30 \leq t < 33$	$33 \leq t < 36$
Number of competitors					



- 6 The table shows the lifetimes of a sample of electric light globes.
- Draw a cumulative frequency graph for the data.
  - Estimate the median life of a globe.
  - Estimate the percentage of globes which had a life of 2700 hours or less.
  - Estimate the number of globes which had a life between 1500 and 2500 hours.

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

- 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

- Between what limits are lengths rounded to 20 cm?
- Rewrite the frequency table to show the data in the class intervals you have just described.
- Hence draw a cumulative frequency graph for the data.
- Estimate:
  - the median foot length
  - 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  $\{x_1, x_2, x_3, \dots, x_n\}$  is

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

where  $\mu$  is the population mean  
and  $n$  is the number of data values.

The variance is the average of the squares of the distances from the mean.



We observe that if the data values  $x_i$  are situated close together around the mean  $\mu$ , 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  $\{x_1, x_2, x_3, \dots, x_n\}$  is

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

The standard deviation measures the degree to which the data *deviates* from the mean.



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  $\sigma^2$  would be in  $\text{kg}^2$ , and  $\sigma$  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.

**Example 12**

**Self Tutor**

Find the population variance and standard deviation for the data set: 3 12 8 15 7

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

The population variance  $\sigma^2 = \frac{\sum (x - \mu)^2}{n}$   
 $= \frac{86}{5}$   
 $= 17.2$

The population standard deviation  $\sigma = \sqrt{17.2}$   
 $\approx 4.15$

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

**SAMPLE VARIANCE AND STANDARD DEVIATION**

If we are only given a *sample* of data from a larger population, we calculate a statistic called the **sample variance**. This statistic is used to *estimate* the variance of the population.

For a sample of  $n$  data values  $\{x_1, x_2, x_3, \dots, x_n\}$  with sample mean  $\bar{x}$ :

- The **sample variance** is  $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$ .

- The **sample standard deviation** is  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$ .

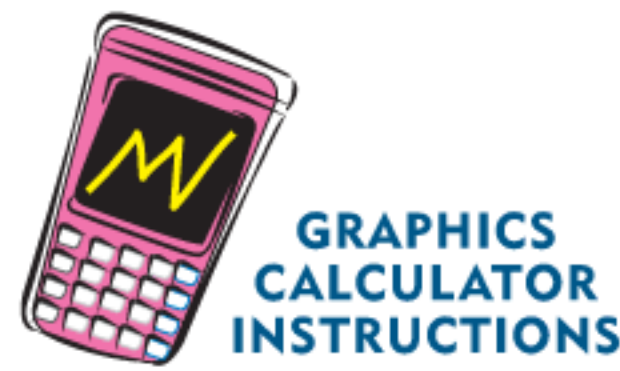
The sample variance  $s^2$  is an **unbiased estimate** of the population variance  $\sigma^2$ .





## CALCULATING VARIANCE AND STANDARD DEVIATION

We commonly use technology to find the standard deviation. If we are given the whole population we use  $\sigma_x$ , and if we are given a sample we use  $s_x$ . We can then square these values to find the population and sample variance,  $\sigma_x^2$  and  $s_x^2$  respectively.



	Deg	Norm1	d/c	Real
<b>1-Variable</b>				
$\bar{x}$	=	30.0333333		
$\Sigma x$	=	901		
$\Sigma x^2$	=	33211		
$\sigma x$	=	14.3189462		
$s x$	=	14.5637323		
n	=	30		

### Example 13

### Self Tutor

Kylie is interested in the ages of spectators at a rugby match. She selects a sample of 30 spectators. Their ages are shown below:

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

Use technology to find the sample standard deviation.

#### Casio fx-CG50

	Deg	Norm1	d/c	Real
<b>1-Variable</b>				
$\bar{x}$	=	30.0333333		
$\Sigma x$	=	901		
$\Sigma x^2$	=	33211		
$\sigma x$	=	14.3189462		
$s x$	=	14.5637323		
n	=	30		

#### TI-84 Plus CE

NORMAL FLOAT AUTO REAL RADIAN MP	
<b>1-Var Stats</b>	
$\bar{x}$	=30.03333333
$\Sigma x$	=901
$\Sigma x^2$	=33211
$s x$	=14.56373231
$\sigma x$	=14.31894627
n	=30
minX	=7
↓Q1	=19

#### HP Prime

Statistics 1Var Numeric View	
	H1
Min	7
Q1	19
Med	28.5
Q3	42
Max	59
$\Sigma x$	901
$\Sigma x^2$	33,211
$\bar{x}$	30.0333333333
$s x$	14.5637323118
$\sigma x$	14.3189462679
Sample standard deviation of X	
More	OK

The sample standard deviation  $s \approx 14.6$  years.

The reason there are two formulae for standard deviation is that if we have data which is sampled from a large population, the sample standard deviation  $s$  provides a better estimate for the actual population standard deviation  $\sigma$  than if we use the formula for  $\sigma$  on the sample.

However, in the **Mathematics: Analysis and Approaches HL** course you are expected to calculate all standard deviations as though they were populations. For this reason, two answers are given for some questions in the following Exercise.

## EXERCISE 13J

- 1 Consider the following data sets:

Data set A: 10 7 5 8 10

Data set B: 4 12 11 14 1 6

- Show that each data set has mean 8.
- Which data set appears to have the greater spread? Explain your answer.
- 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 Describe the population in this case.
  - b Use technology to find the population standard deviation of the data.
  - c 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 standard deviation of the data.

Mathematics: Analysis and Approaches students should use the population standard deviation.



- 5 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 standard deviation for each data set.
  - d Which person studies more consistently?
- 6 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
<i>Mean <math>\bar{x}</math></i>		
<i>Median</i>		
<i>Standard deviation</i>		
<i>Range</i>		



- 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?





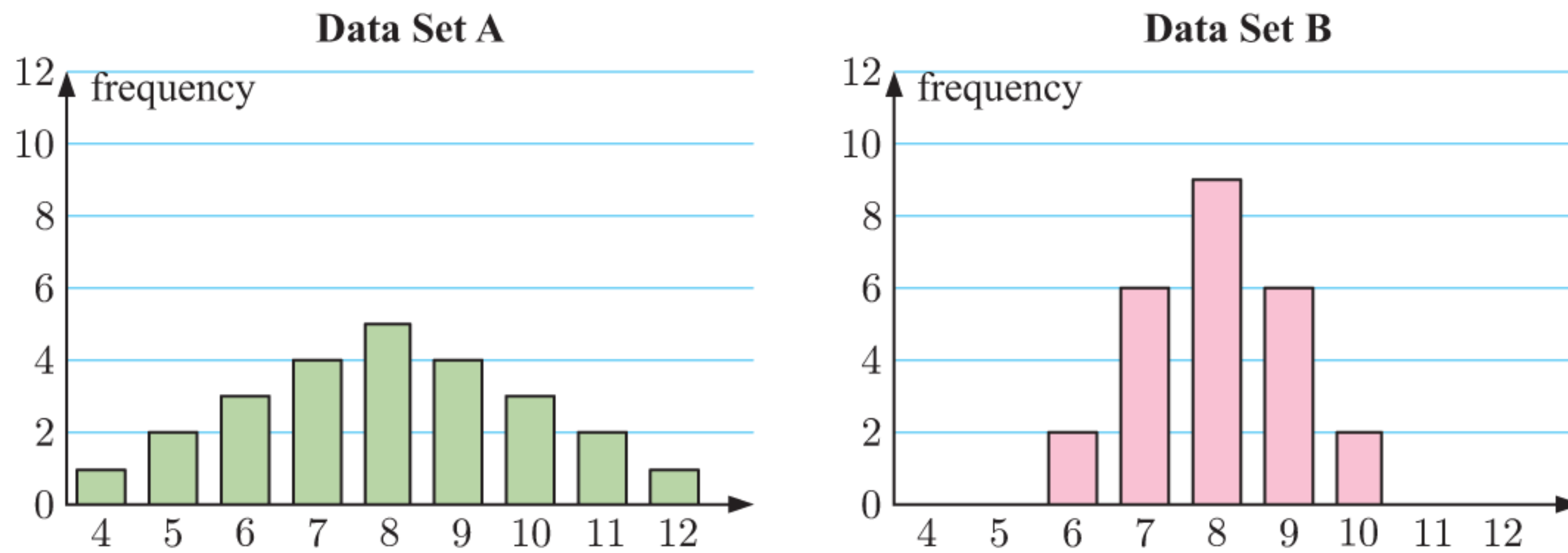


14 The table shows the ages of squash players at the Junior National Squash Championship.

<i>Age</i>	11	12	13	14	15	16	17	18
<i>Frequency</i>	2	1	4	5	6	4	2	1

Find the mean and population standard deviation of the ages.

15 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.

<i>Data set</i>	<i>Range</i>	<i>IQR</i>
A	8	3
B	4	2

In what way does the standard deviation give a better description of how the data is distributed?

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

- 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.

<i>Score</i>	<i>Females</i>	<i>Males</i>
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

17 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 14****Self Tutor**

The table alongside summarises the examination scores for 80 randomly selected students. Estimate the standard deviation for the data.

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

For continuous data or data grouped in classes, use the mid-interval value to represent all data in that interval.

**Casio fx-CG50**

1-Variable	
$\bar{x}$	=59.75
$\Sigma x$	=4780
$\Sigma x^2$	=308200
$\sigma x$	=16.8058769
$sx$	=16.9119087
$n$	=80

**TI-84 Plus CE**

1-Var Stats	
$\bar{x}$	=59.75
$\Sigma x$	=4780
$\Sigma x^2$	=308200
$Sx$	=16.91190877
$\sigma x$	=16.80587695
$n$	=80
minX	=4.5
$\downarrow Q_1$	=54.5

**TI-nspire**

OneVar data:freq: stat.results	
"Title"	"One-Variable Statistics"
" $\bar{x}$ "	59.75
" $\Sigma x$ "	4780.
" $\Sigma x^2$ "	308200.
" $Sx := S_{n-1}x$ "	16.9119
" $\sigma x := \sigma_{n-1}x$ "	16.8059
" $n$ "	80.
"MinX"	4.5

The sample standard deviation  $\approx 16.9$ .

The population standard deviation  $\approx 16.8$ .

- 18** 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
$40 \leq L < 42$	1
$42 \leq L < 44$	1
$44 \leq L < 46$	3
$46 \leq L < 48$	7
$48 \leq L < 50$	11
$50 \leq L < 52$	5
$52 \leq L < 54$	2







**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  $\mu$  and standard deviation  $\sigma$ . 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:

<i>Sample 1:</i>	10	14	16	9	16	15	15	21	9	21
<i>Sample 2:</i>	11	9	11	16	16	13	10	12	21	16
<i>Sample 3:</i>	12	10	14	7	13	11	21	20	15	9
<i>Sample 4:</i>	20	19	19	19	13	19	22	15	10	19
<i>Sample 5:</i>	19	13	23	11	17	4	14	21	13	11
<i>Sample 6:</i>	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  $\sigma$  and  $\sigma^2$  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  $s$ ,  $s^2$ ,  $\sigma$ , and  $\sigma^2$  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	<b>Actual values</b>				<b>Estimator</b>	<b>Average estimate</b>
2	$\mu$	15		Variance	$\sigma^2$	23.794
3	$\sigma$	5			$s^2$	25.046
4	$\sigma^2$	25		Standard deviation	$\sigma$	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  $\mu$  and  $\sigma$  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  $\mu$  or  $\sigma$  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 13A

- 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

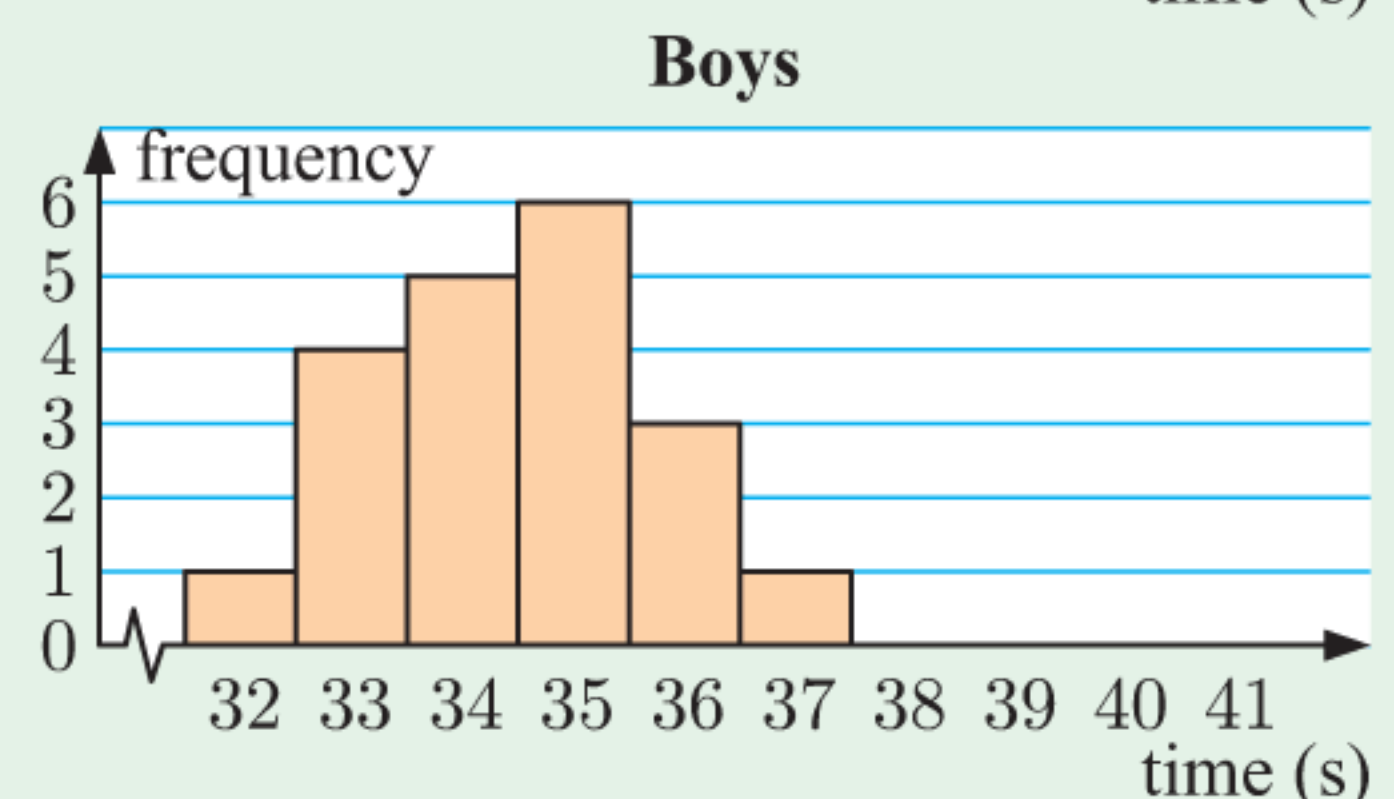
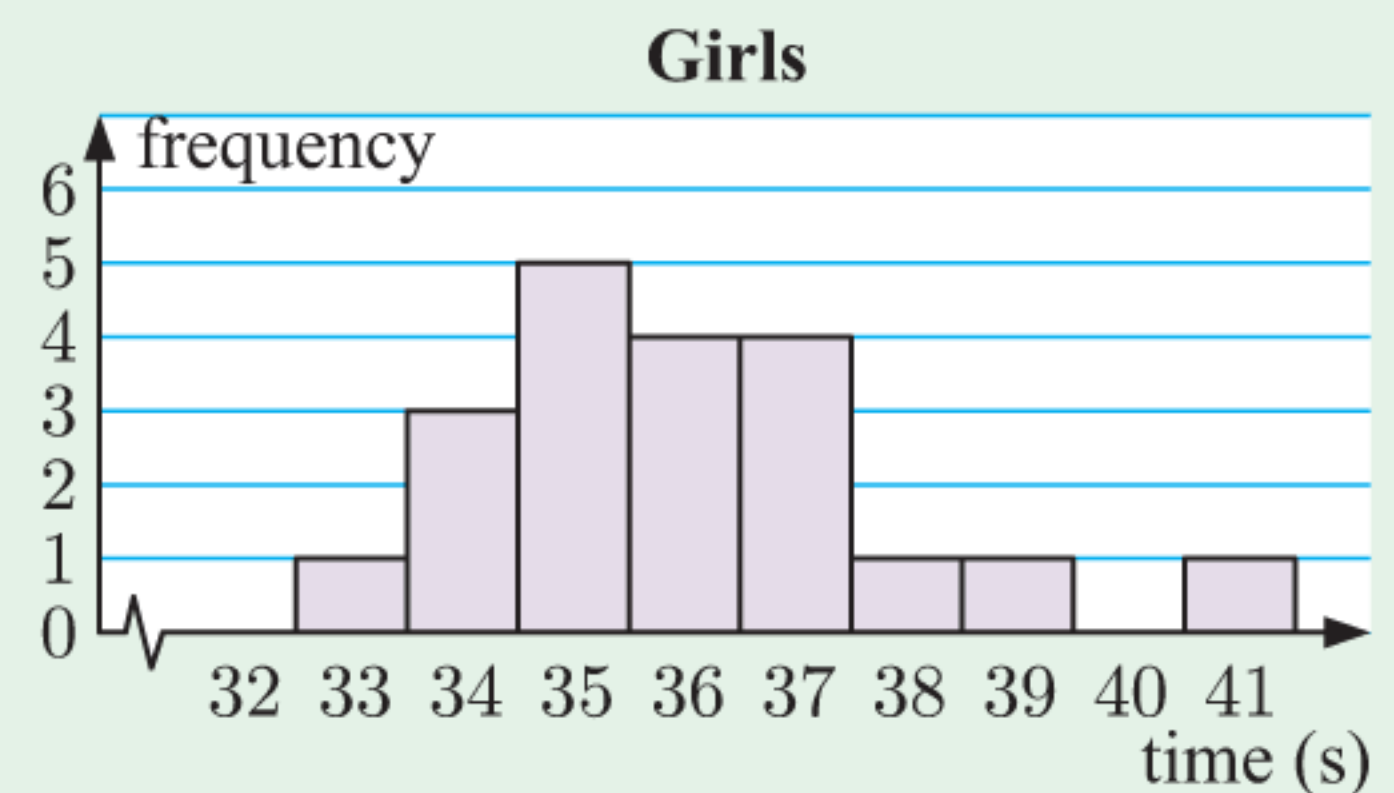
- a Draw a graph to display this data.
  - b Describe the distribution.
  - c Find the:
    - i** mode
    - ii** mean
    - iii** median.
  - d Which of the measures of centre is most appropriate for this data? Explain your answer.
- 3 The histograms alongside show the times for the 50 metre freestyle 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:  $k - 2$ ,  $k$ ,  $k + 3$ ,  $k + 3$ .
- Show that the mean of the data set is equal to  $k + 1$ .
  - 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.

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

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

- 7 Consider this data set:

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

- Find the five-number summary for the data.
- Find the range and IQR.
- Draw a box plot of the data set.

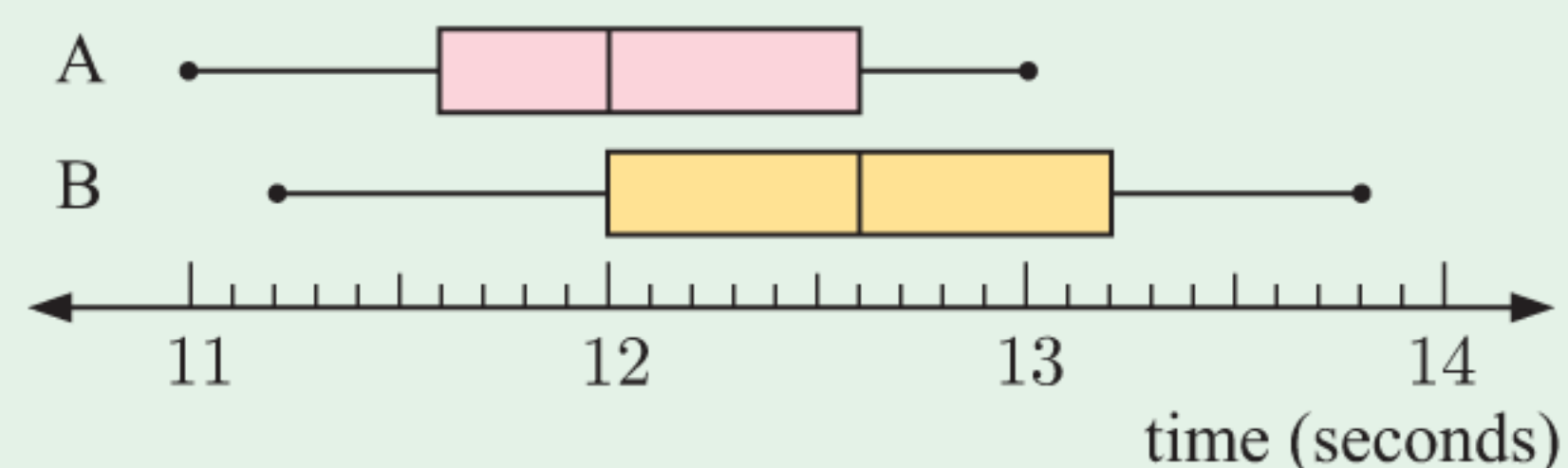
- 8 Katja's golf scores for her last 20 rounds were:

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:

- median
- interquartile range
- mean
- standard deviation.

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

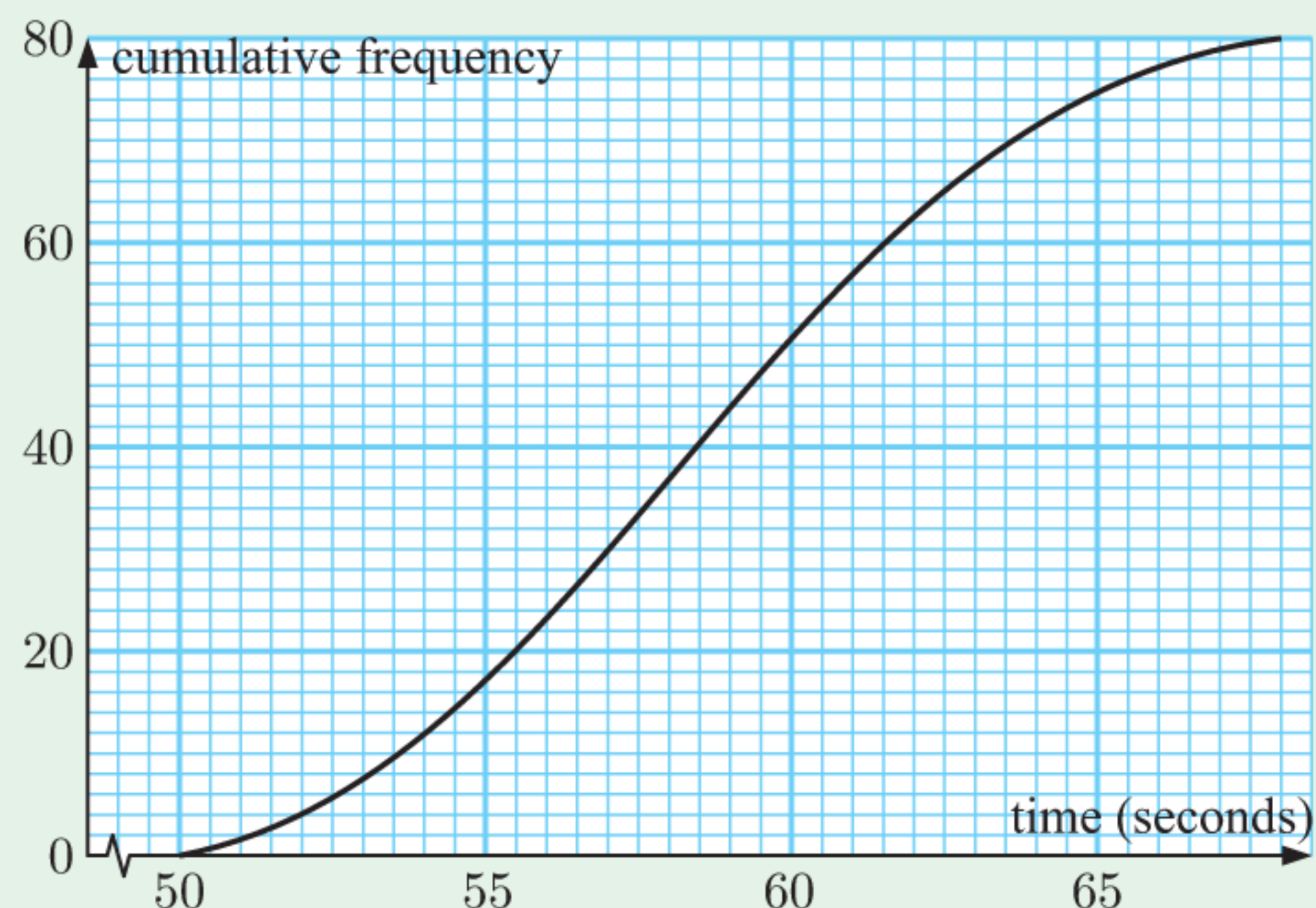


- Determine the five-number summaries for both A and B.
- For each group, calculate the range and interquartile range.
- Copy and complete:
  - The members of squad ..... generally ran faster because .....
  - The times in squad ..... are more varied because .....

- 10 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:

- the median
- the interquartile range
- the time corresponding to the top 10% of runners.

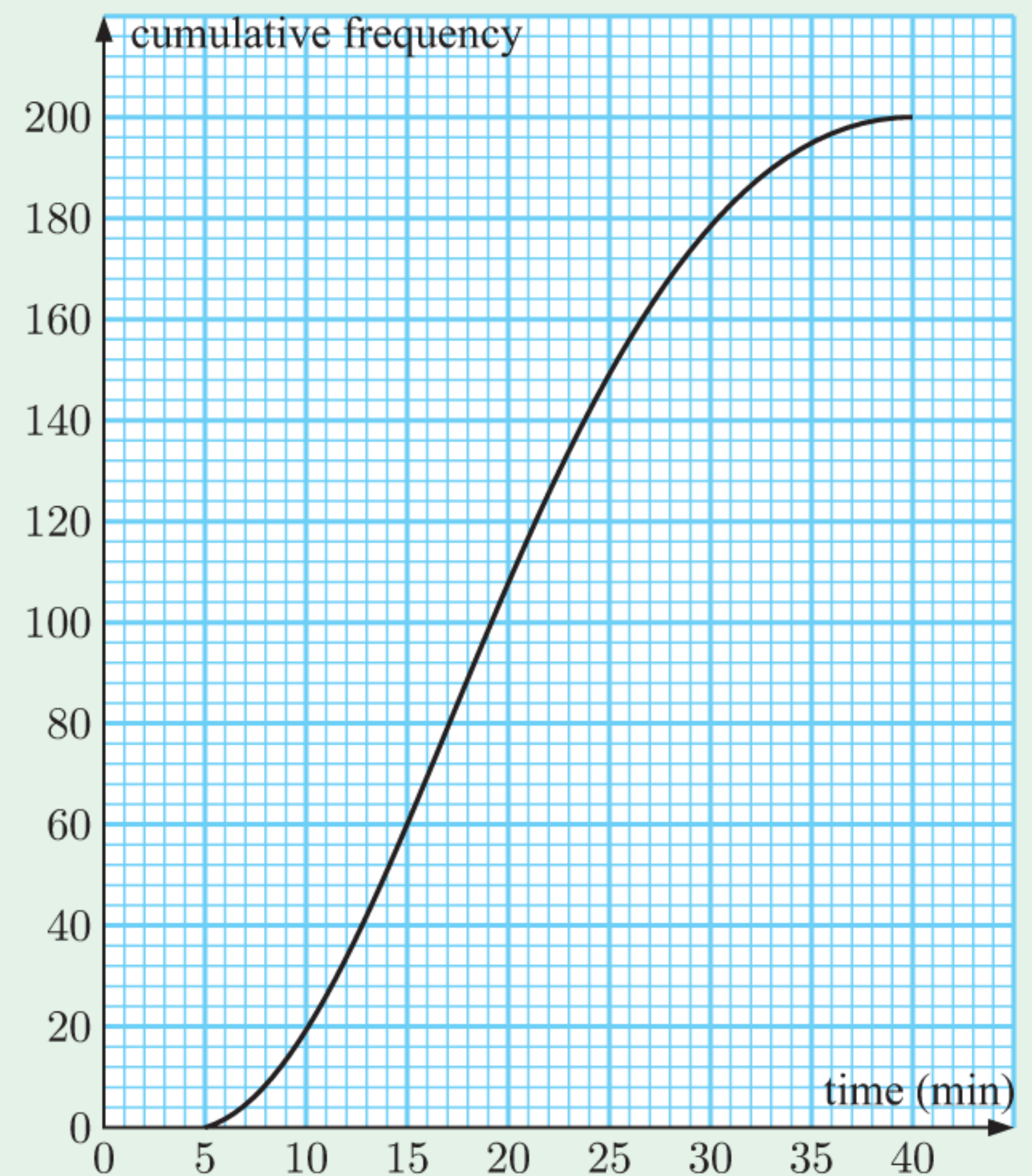




**11** This cumulative frequency curve shows the times taken for 200 students to travel to school by bus.

- 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
$5 \leq t < 10$	
$10 \leq t < 15$	
⋮	
$35 \leq t < 40$	



**12** 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

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

<i>Number</i>	47	48	49	50	51	52
<i>Frequency</i>	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** The number of litres of petrol purchased by a random sample of motor vehicle drivers is shown alongside. For the given data, estimate the:

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

- a** mean
- b** standard deviation.

**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 13B

- 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:

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

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

- 3 The data in the table alongside has mean 5.7.

<i>Value</i>	2	5	$x$	$x + 6$
<i>Frequency</i>	3	2	4	1

- a** Find the value of  $x$ .  
**b** Find the median of the distribution.

- 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.

<i>Brand X</i>						<i>Brand Y</i>					
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 the table alongside.  
**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.

	<i>Brand X</i>	<i>Brand Y</i>
min		
Q <sub>1</sub>		
median		
Q <sub>3</sub>		
max		
IQR		

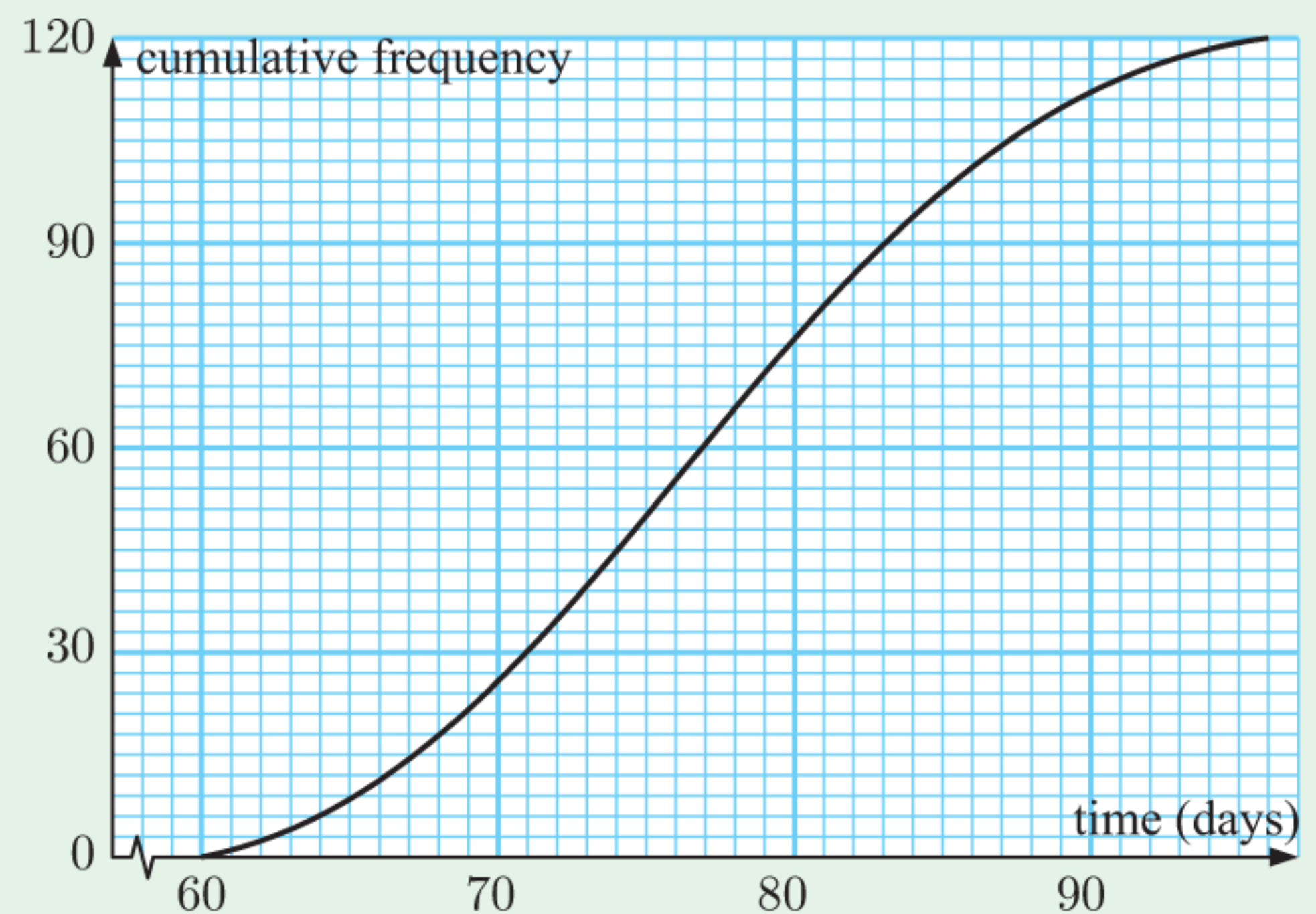
- 9** Consider the frequency table alongside:

- 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  $\bar{x}$  as a fraction.

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

- 10** 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.



- 11** Consider the data in the table below:

Scores ( $x$ )	0 ≤ $x$ < 10	10 ≤ $x$ < 20	20 ≤ $x$ < 30	30 ≤ $x$ < 40	40 ≤ $x$ < 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.



**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
$0 \leq t < 30$	1
$30 \leq t < 60$	4
$60 \leq t < 90$	12
$90 \leq t < 120$	18
$120 \leq t < 150$	7
$150 \leq t < 180$	2

- How many players were surveyed?
- Write down the modal class.
- Draw a cumulative frequency graph for the data.
- The game is considered too easy if either the mean or median completion time is below 90 minutes.
  - Estimate the median completion time using your cumulative frequency graph.
  - Estimate the mean completion time.
  - 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** A random sample of weekly supermarket bills was recorded in the table alongside.

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

For the given data, estimate the:

- mean
- standard deviation.

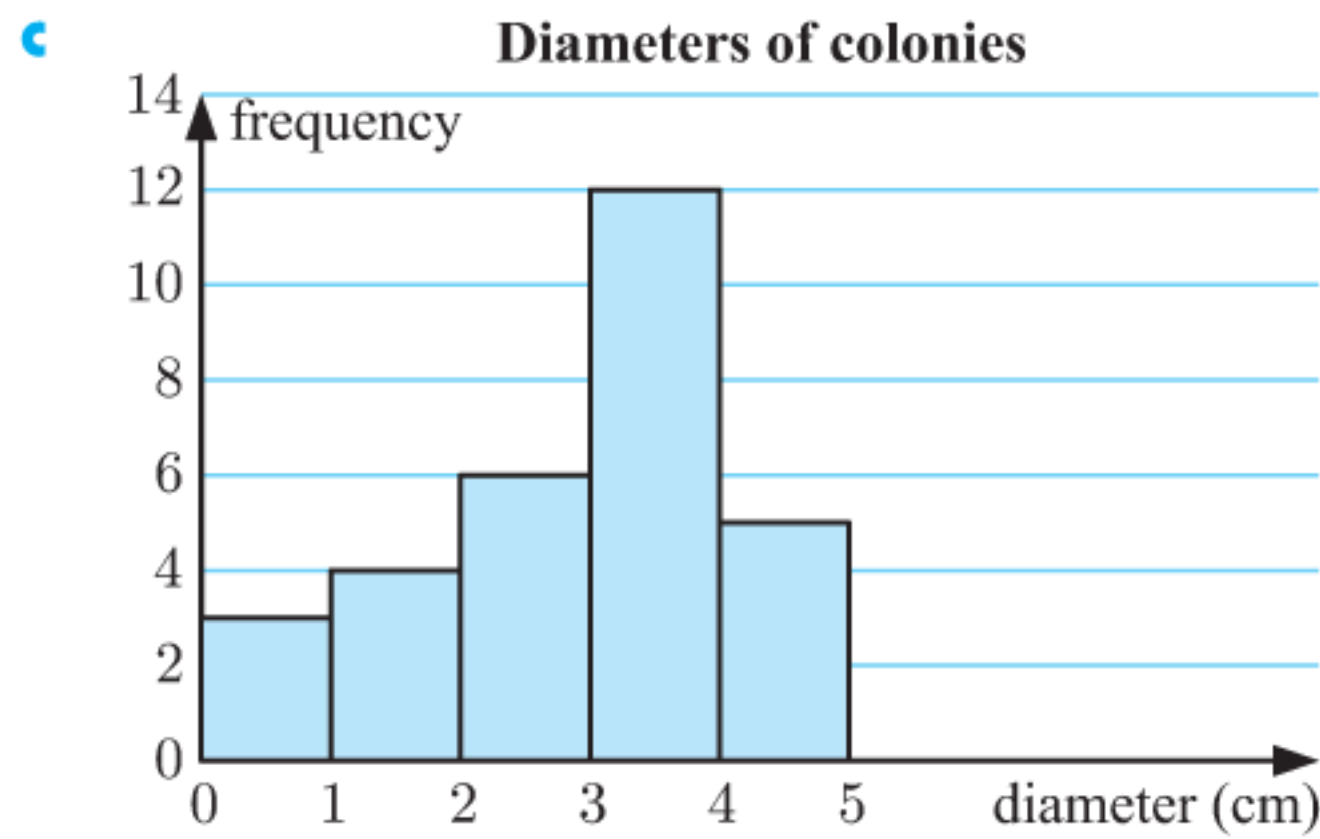
**14** Friends Kevin and Felicity each selected a sample of 20 crossword puzzles. The times they took, in minutes, to complete each puzzle were:

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

- Find the mean of each data set.
- Find the standard deviation for each person.
- Who generally solves crossword puzzles faster?
- Who is more consistent in their time taken to solve the puzzles?

**15** A data set has  $s^2 = 4.1$  and  $\sigma^2 = 3.69$ . How many data values does the data set have?

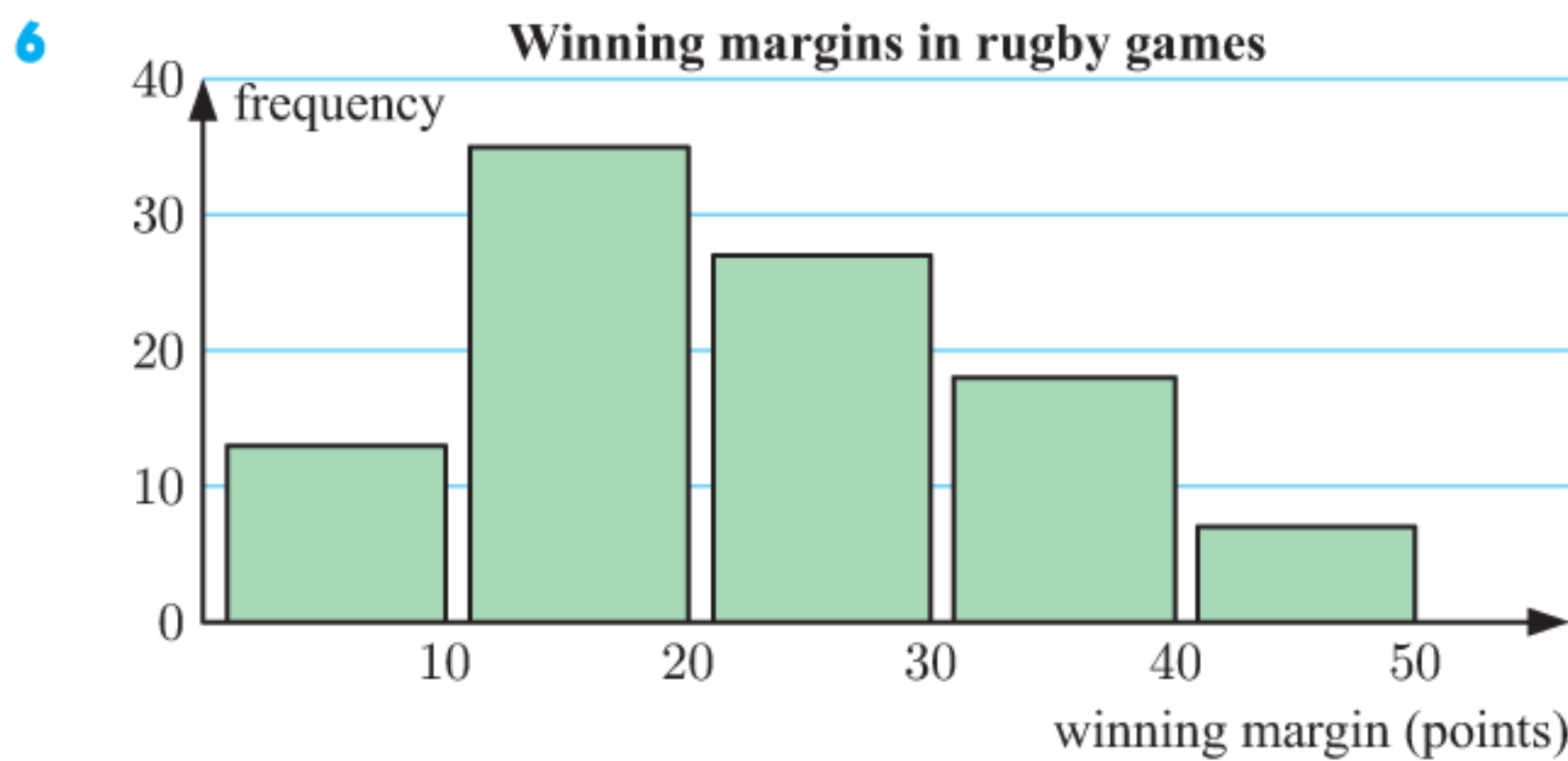




- d**  $3 \leq d < 4$  cm      **e** slightly negatively skewed

**REVIEW SET 12B**

- 1** **a** discrete      **b** continuous      **c** discrete
- 2** **a** systematic sampling
- b** A house will be visited if the last digit in its number is equal to the random number chosen by the promoter, with the random number 10 corresponding to the digit 0. Each house therefore has a 1 in 10 chance of being visited.
- c** Once the first house number has been chosen, the remaining houses chosen must all have the same second digit in their house number, that is, they are not randomly chosen. For example, it is impossible for two consecutively numbered houses to be selected for the sample.
- 3** **a** Petra's teacher colleagues are quite likely to ignore the emailed questionnaire as emails are easy to ignore.
- b** It is likely that the teachers who have responded will have strong opinions either for or against the general student behaviour. These responses may therefore not be representative of all teachers' views.
- 4** Did you learn about our services via:
- friends/family      • the internet      • newspaper
  - television      • elsewhere?
- 5** **a** The tone is not neutral and it is a structured question. The only responses possible are yes or no.
- b** How would you describe your general behaviour when you were a child?

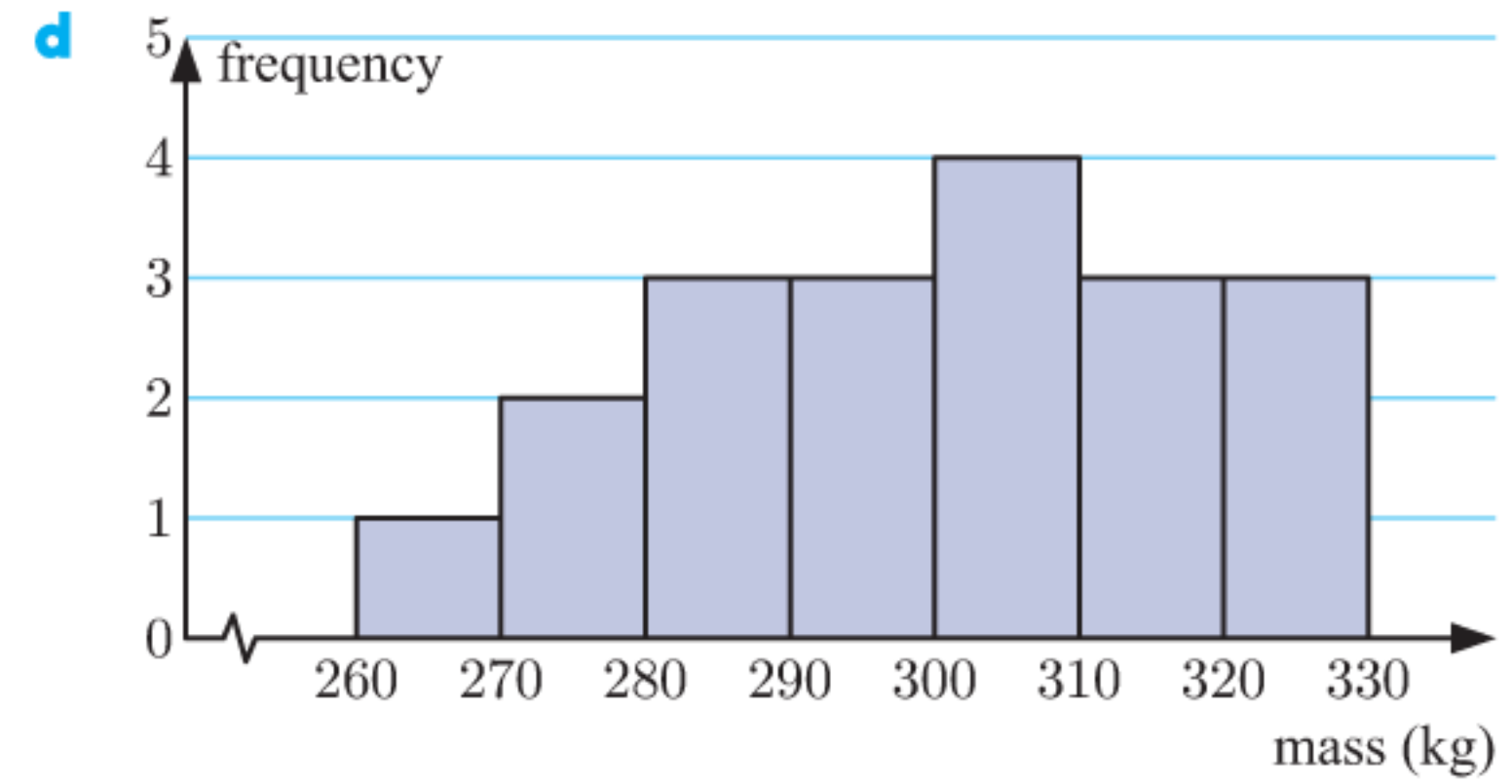


- 7** **a** Mass is measured on a continuous scale.

**b**

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

- c**  $300 \leq m < 310$  kg

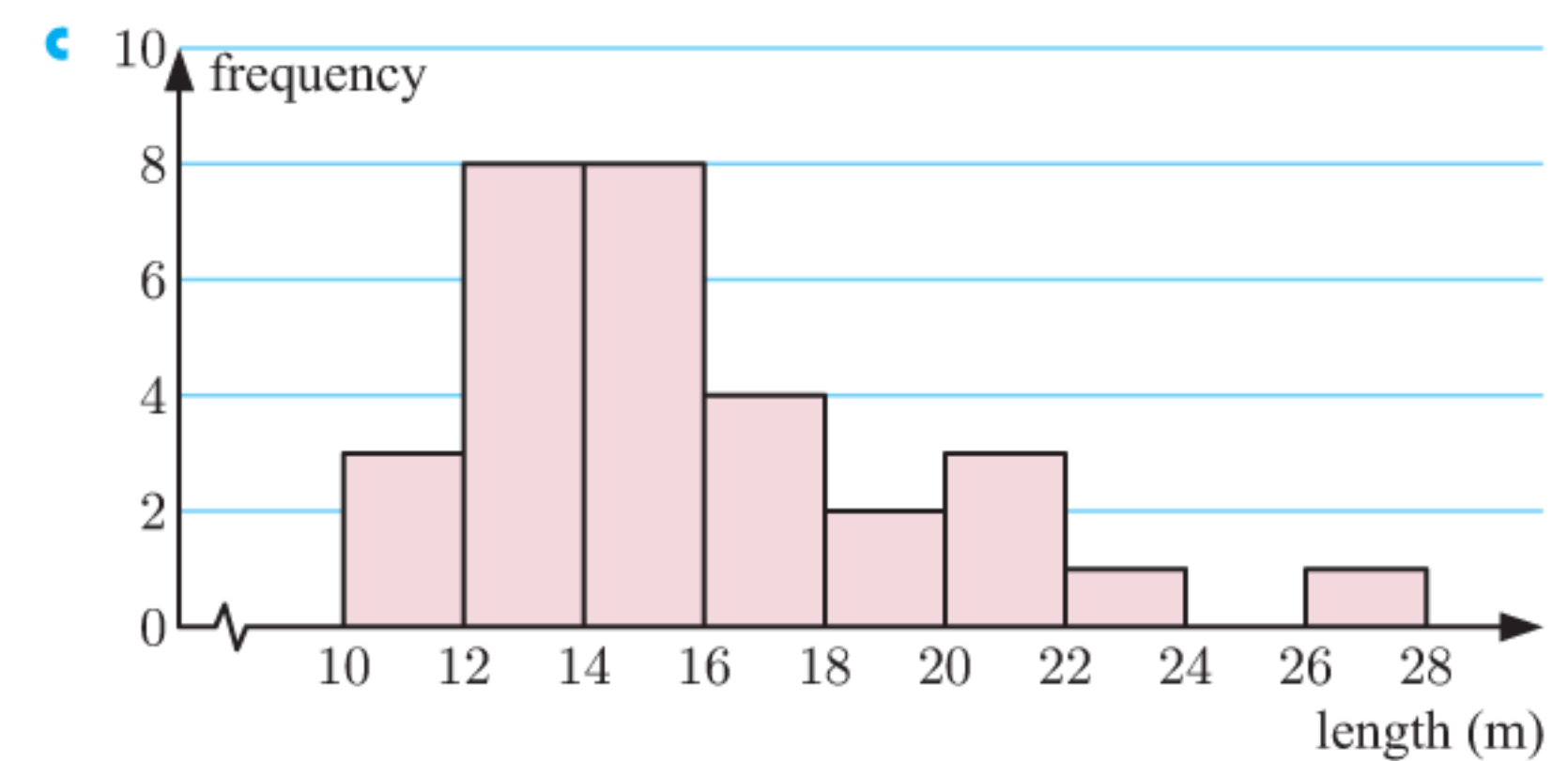


- e** slightly negatively skewed

- 8** **a** continuous

**b**

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



- d** positively skewed, one outlier (27.4 m)

**EXERCISE 13A**

- 1** **a** 1 cup      **b** 2 cups      **c** 1.8 cups
- 2** **a** **i**  $\approx 5.61$       **ii** 6      **iii** 6
- b** **i**  $\approx 16.3$       **ii** 17      **iii** 18
- c** **i**  $\approx 24.8$       **ii** 24.9      **iii** 23.5
- 3** 9      **4** Ruth
- 5** **a** data set A:  $\approx 6.46$ , data set B:  $\approx 6.85$
- 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:  $\approx 67.1$ , malai jamun:  $\approx 53.6$
- 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** 116      **11** 17.25 goals per game
- 12**  $x = 15$       **13**  $a = 5$       **14** 37 marks
- 15**  $\approx 14.8$       **16** 6 and 12



**EXERCISE 13B**

- 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
- 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.
- 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
- a** mean  $\approx$  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 13C**

- a** 1 person      **b** 2 people      **c**  $\approx$  2.03 people
- a** **i** 2.96 phone calls      **ii** 2 phone calls  
**iii** 2 phone calls

**b**

**Phone calls in a day**

frequency

mode, median (2)      mean (2.96)

number of phone calls

**c** positively skewed

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

**e** the mean
- a** **i**  $\approx$  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.

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

- 10.1 cm
- a** **i** \$63 000      **ii** \$56 000      **iii** \$66 600      **b** the mean
- a**  $x = 5$       **b** 75%

**EXERCISE 13D**

- a** 40 phone calls      **b**  $\approx$  15 minutes      **2**  $\approx$  31.7
- a** 70 service stations      **b**  $\approx$  411 000 litres ( $\approx$  411 kL)

**c**  $\approx$  5870 L

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

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

**b**  $\approx$  14.8 runs

**c**  $\approx$  14.9 runs; the estimate in **b** was very accurate.
- a**  $p = 24$       **b**  $\approx$  3.37 minutes      **c**  $\approx$  15.3%
- a** 125 people      **b**  $\approx$  119 marks      **c**  $\frac{3}{25}$       **d** 28%

**EXERCISE 13E**

- 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
- 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
- a** range = 60, IQR = 8.5      **b** ‘67’ is an outlier.

**c** range = 18, IQR = 8      **d** the range
- 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.
- a**  $g$       **b** **i**  $m - a$       **ii**  $\left(\frac{j+k}{2}\right) - \left(\frac{c+d}{2}\right)$

Measure	median	mode	range	interquartile range
<b>a</b>	11	9	13	6
<b>b</b>	18	14	26	12

**EXERCISE 13F**

- a** 35 points      **b** 78 points      **c** 13 points      **d** 53 points

**e** 26 points      **f** 65 points      **g** 27 points
- a** **i** 98, 25 marks      **ii** 70 marks      **iii** 85 marks  
**iv** 55, 85 marks

**b** range = 73, IQR = 30
- a** **i** min = 3,  $Q_1 = 5$ , med = 6,  $Q_3 = 8$ , max = 10

**ii**

**iii** range = 7, IQR = 3

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

**ii**

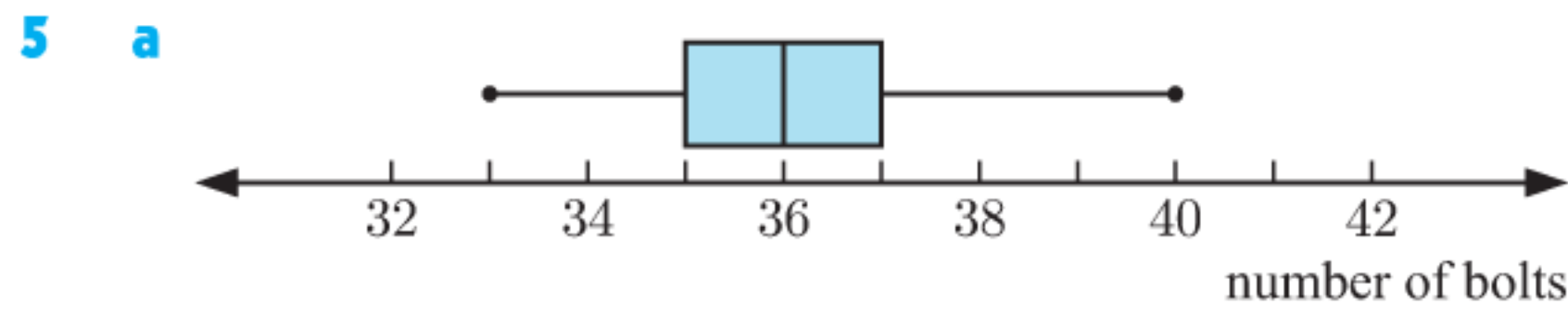
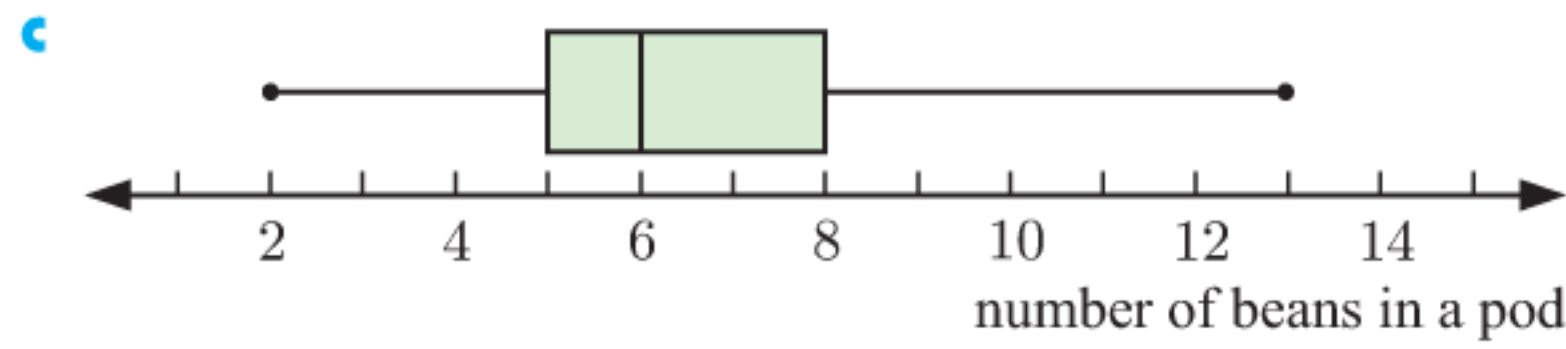
**iii** range = 9, IQR = 4

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

**ii**

**iii** range = 34, IQR = 21
- a** median = 6,  $Q_1 = 5, Q_3 = 8$       **b** IQR = 3

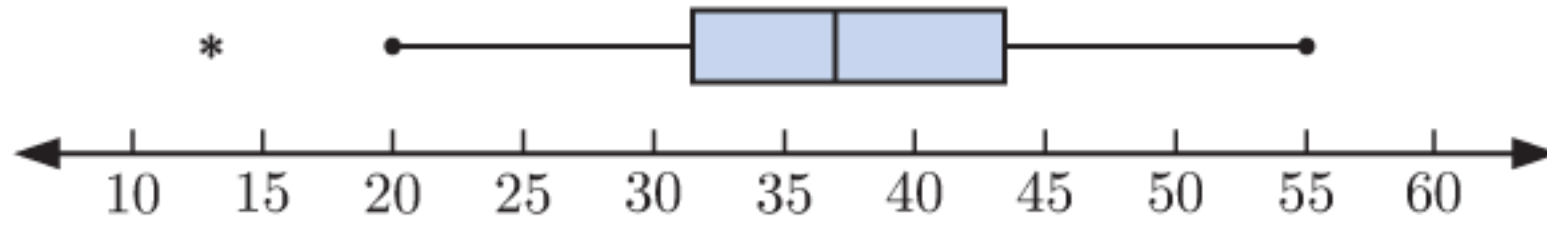




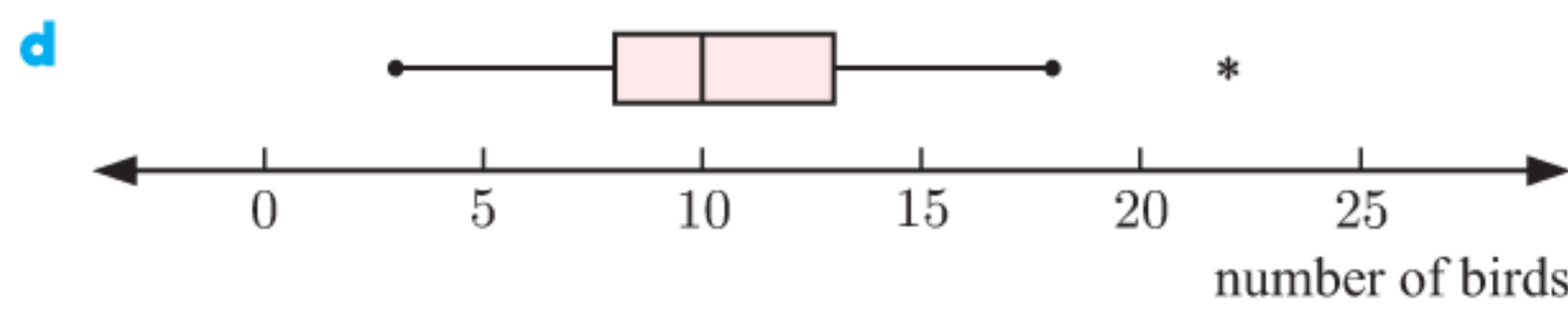
**b** range = 7, IQR = 2

**EXERCISE 13G**

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

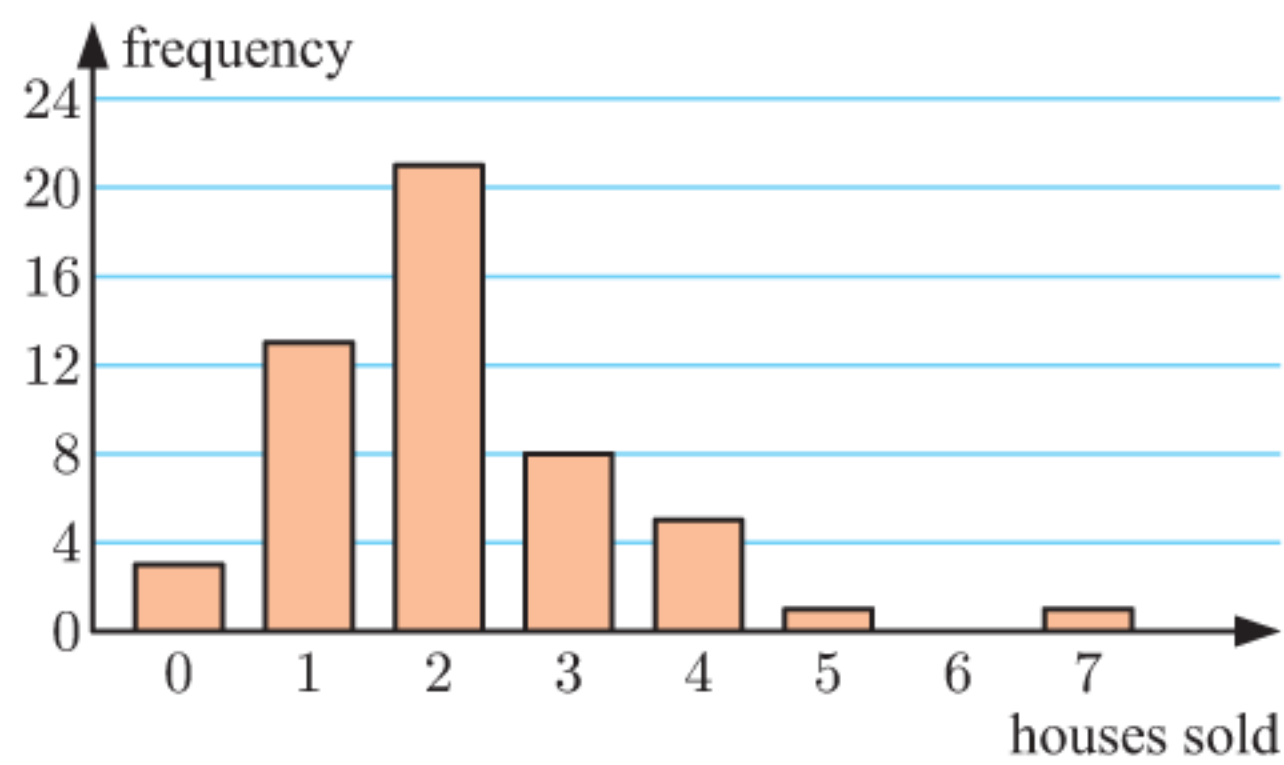


**2 a** median = 10,  $Q_1 = 8$ ,  $Q_3 = 13$     **b** IQR = 5  
**c** lower = 0.5, upper = 20.5, 22 is an outlier

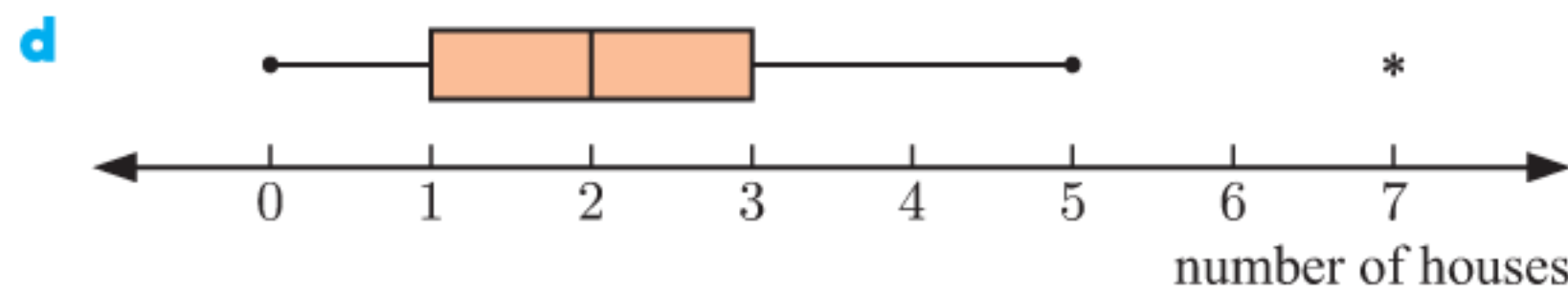


**3 a** **A**    **b** **D**    **c** **C**    **d** **B**

**4 a** Houses sold by a real estate agent



**b** 7 houses appears to be an outlier.  
**c** lower boundary = -2, upper boundary = 6  
 7 houses is an outlier



**EXERCISE 13H**

**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

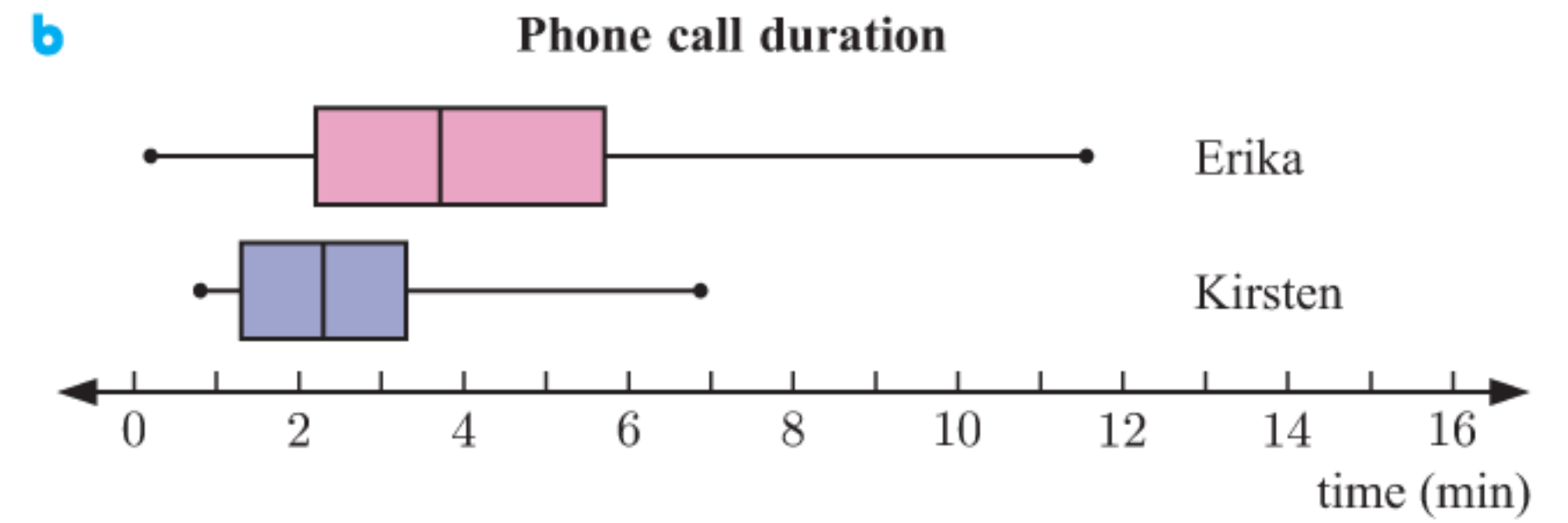
**c i** cannot tell    **ii** true, since Year 9  $Q_1 <$  Year 12 min

**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

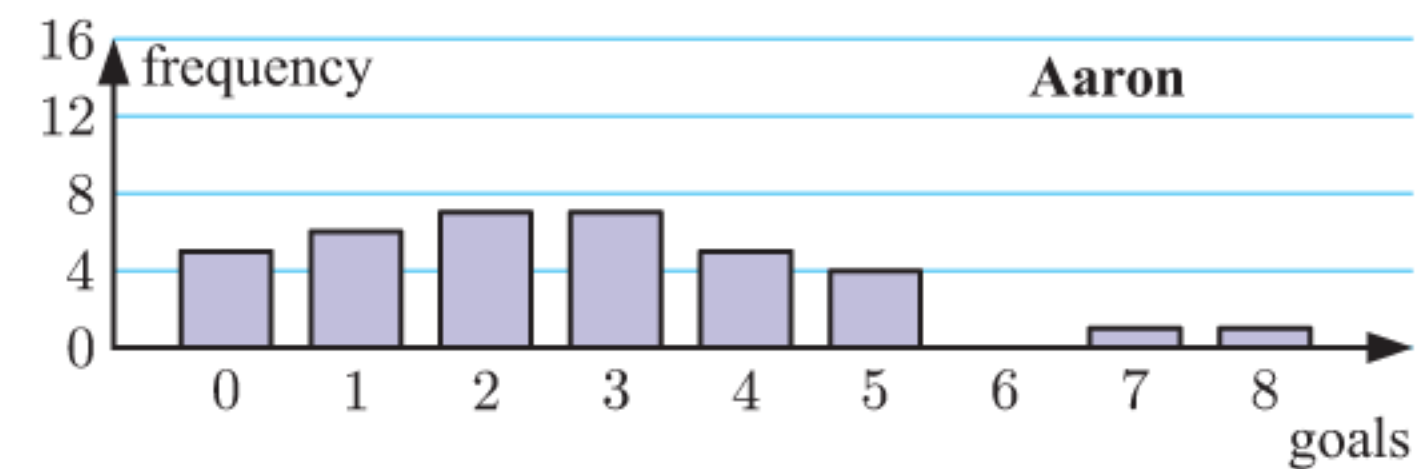
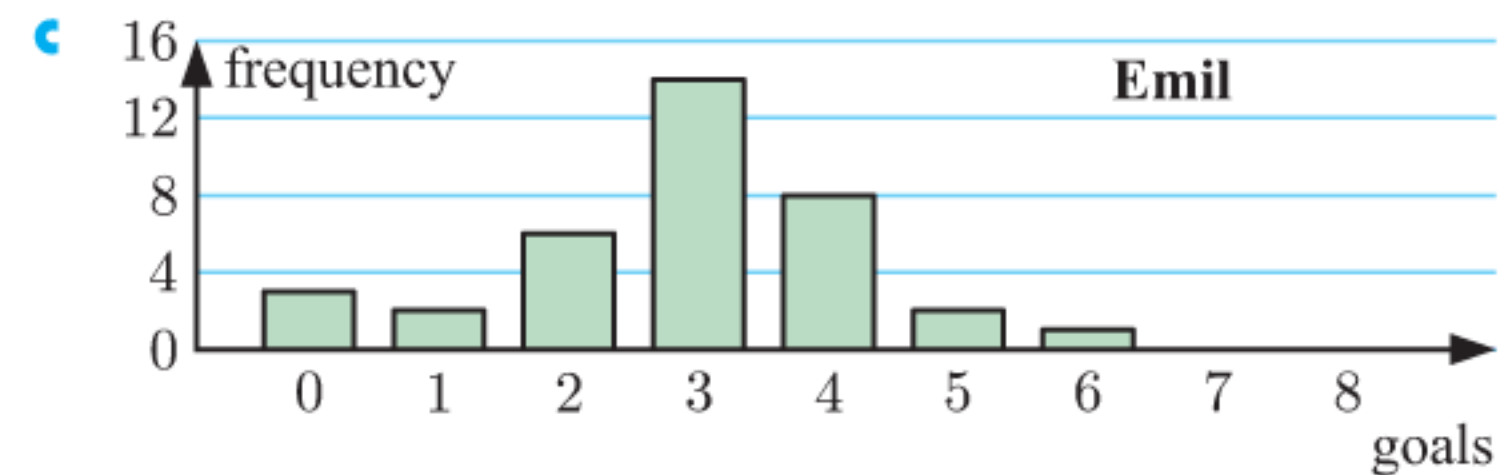
**3 a i** class 1 (96%)    **ii** class 1 (37%)    **iii** class 1  
**b** 18%    **c** 55%    **d i** 25%    **ii** 50%  
**e i** slightly positively skewed    **ii** negatively skewed  
**f** .... class 2, .... class 1

**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



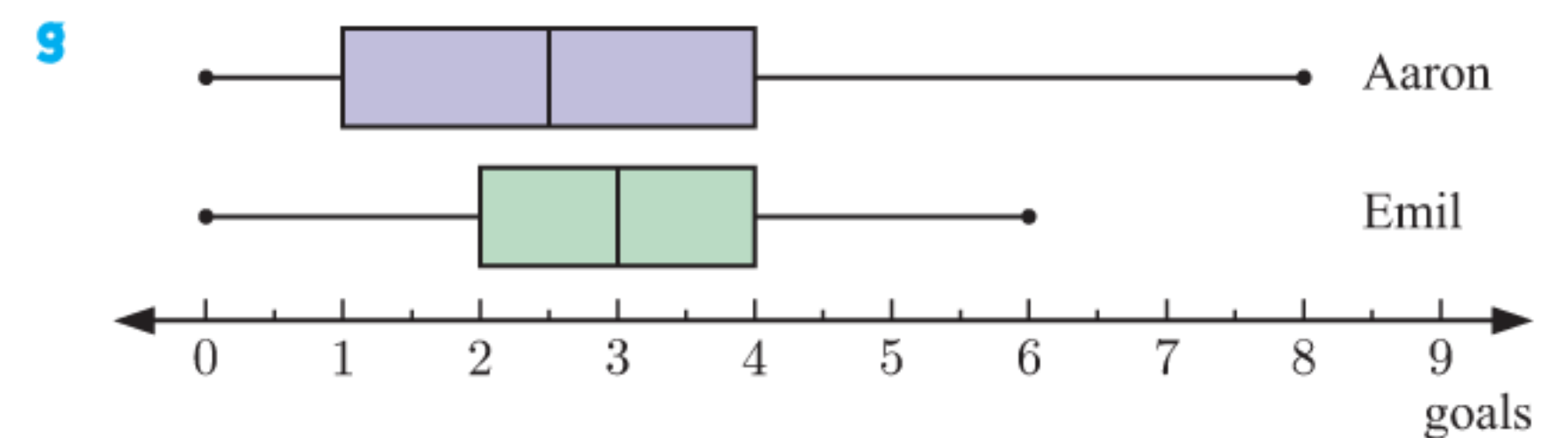
**c** Both are positively skewed (Erika's more so than Kirsten's).  
 Erika's phone calls were more varied in duration.

**5 a** discrete



**d** Emil: approximately symmetrical  
 Aaron: positively skewed  
**e** Emil: mean  $\approx 2.89$ , median = 3, mode = 3  
 Aaron: mean  $\approx 2.67$ , 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.



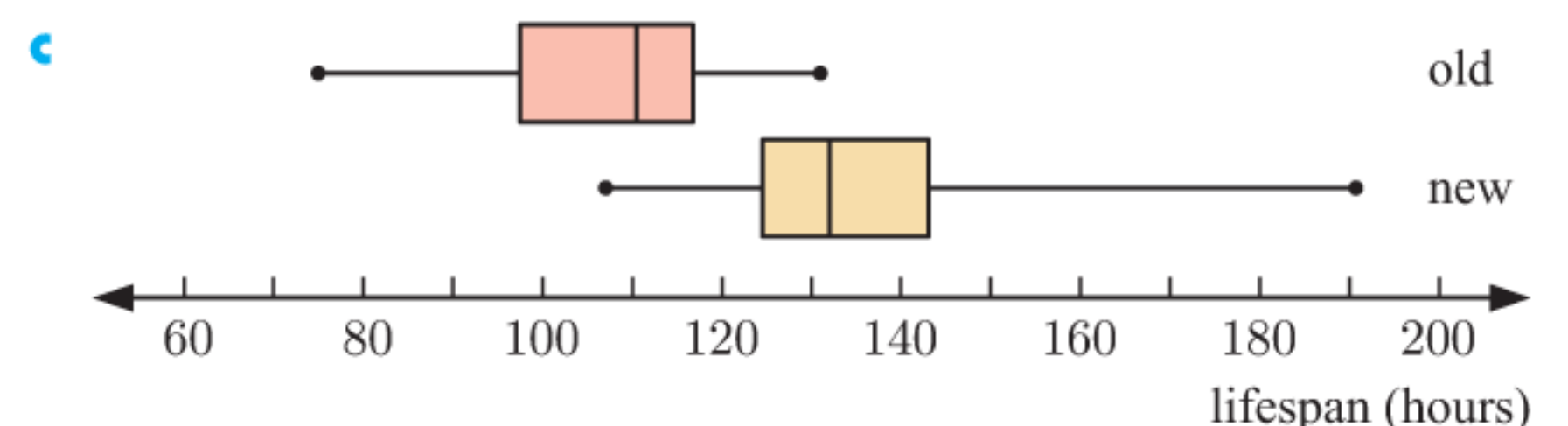
**h** Emil is more consistent with his scoring (in terms of goals) than Aaron.

**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.

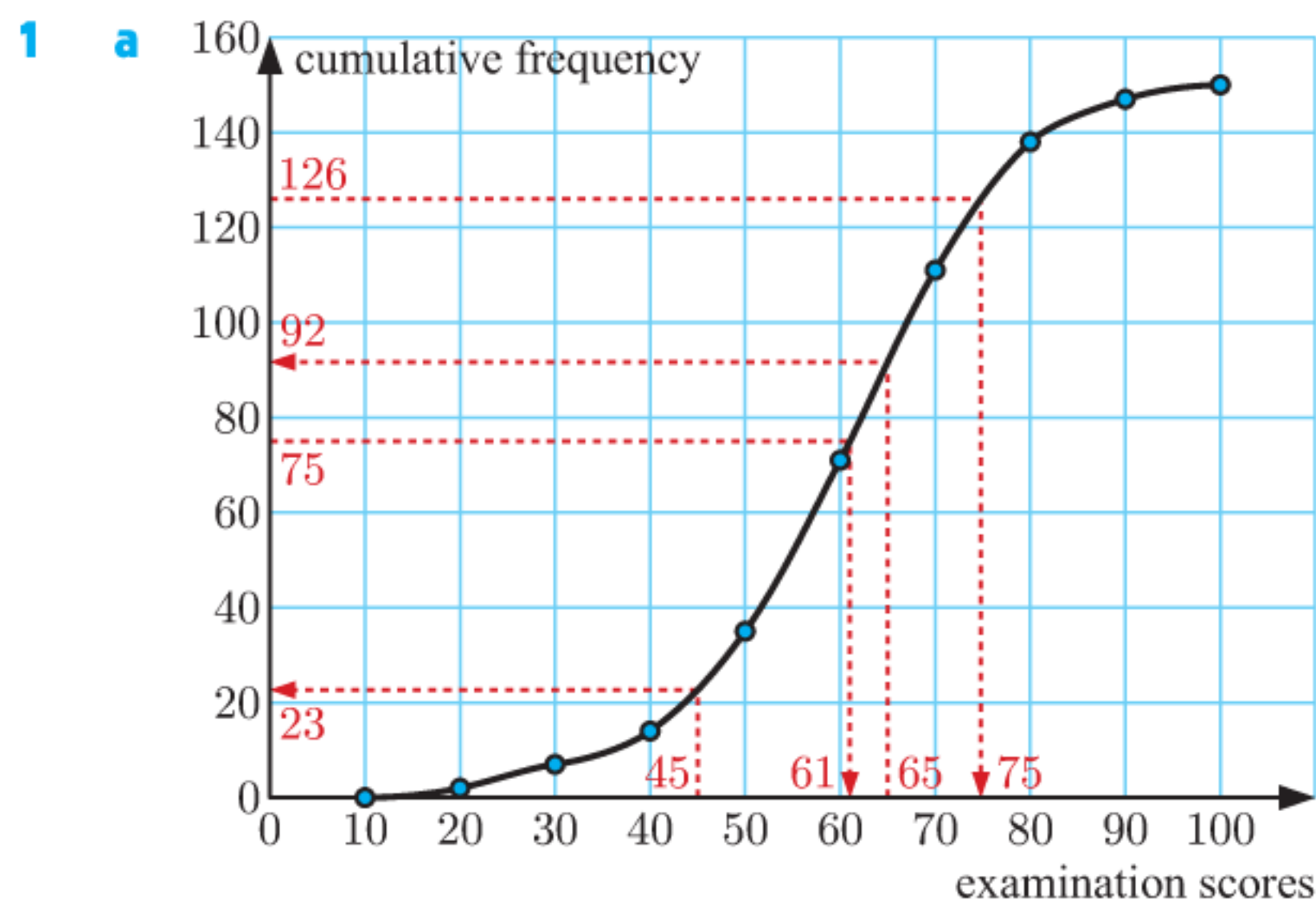


**d** Old type: negatively skewed  
 New type: positively skewed



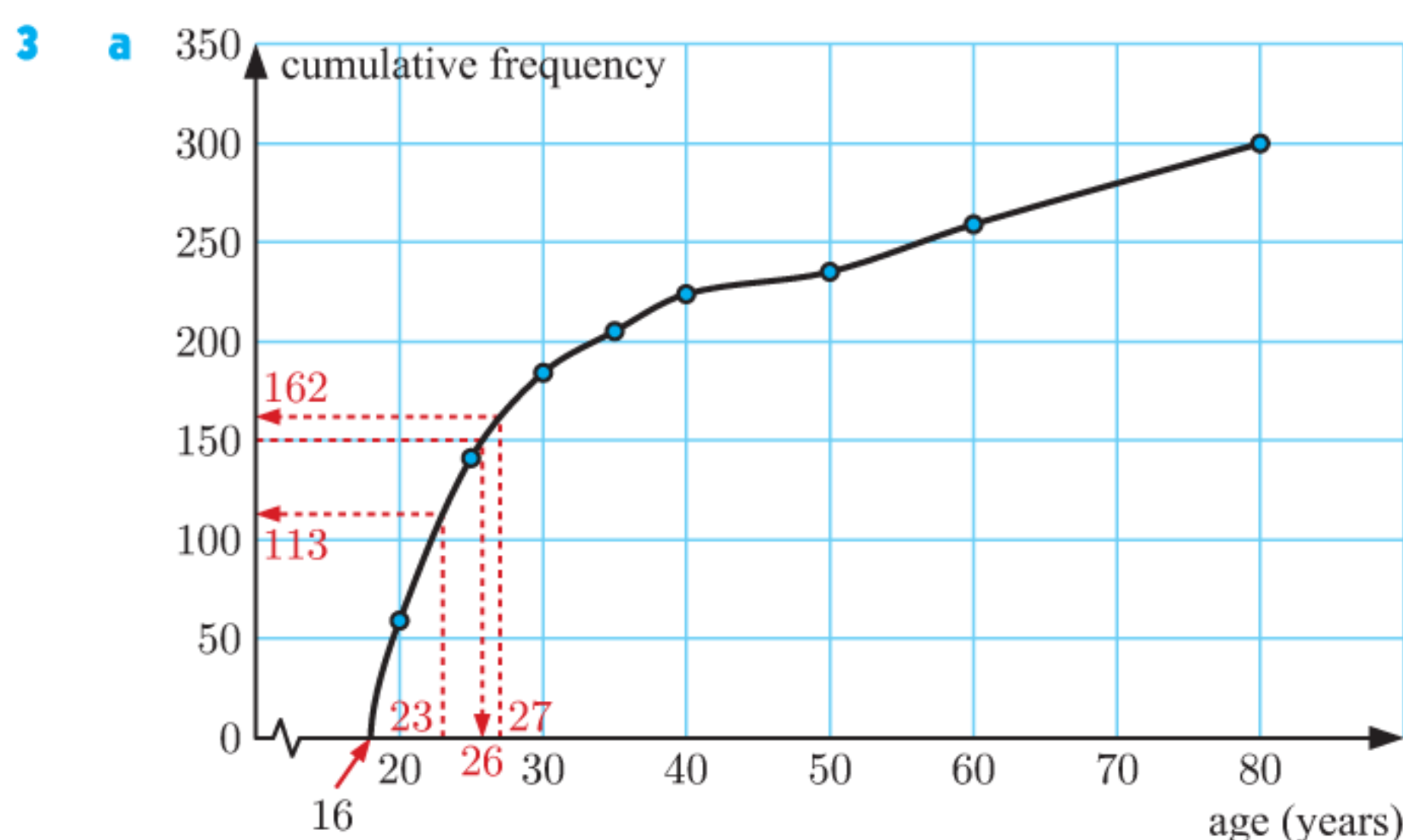
e The “new” type of light globes do last longer than the “old” type. From c, both the mean and median for the “new” type are close to 20% greater than that of the “old” type. The manufacturer’s claim appears to be valid.

**EXERCISE 13I**



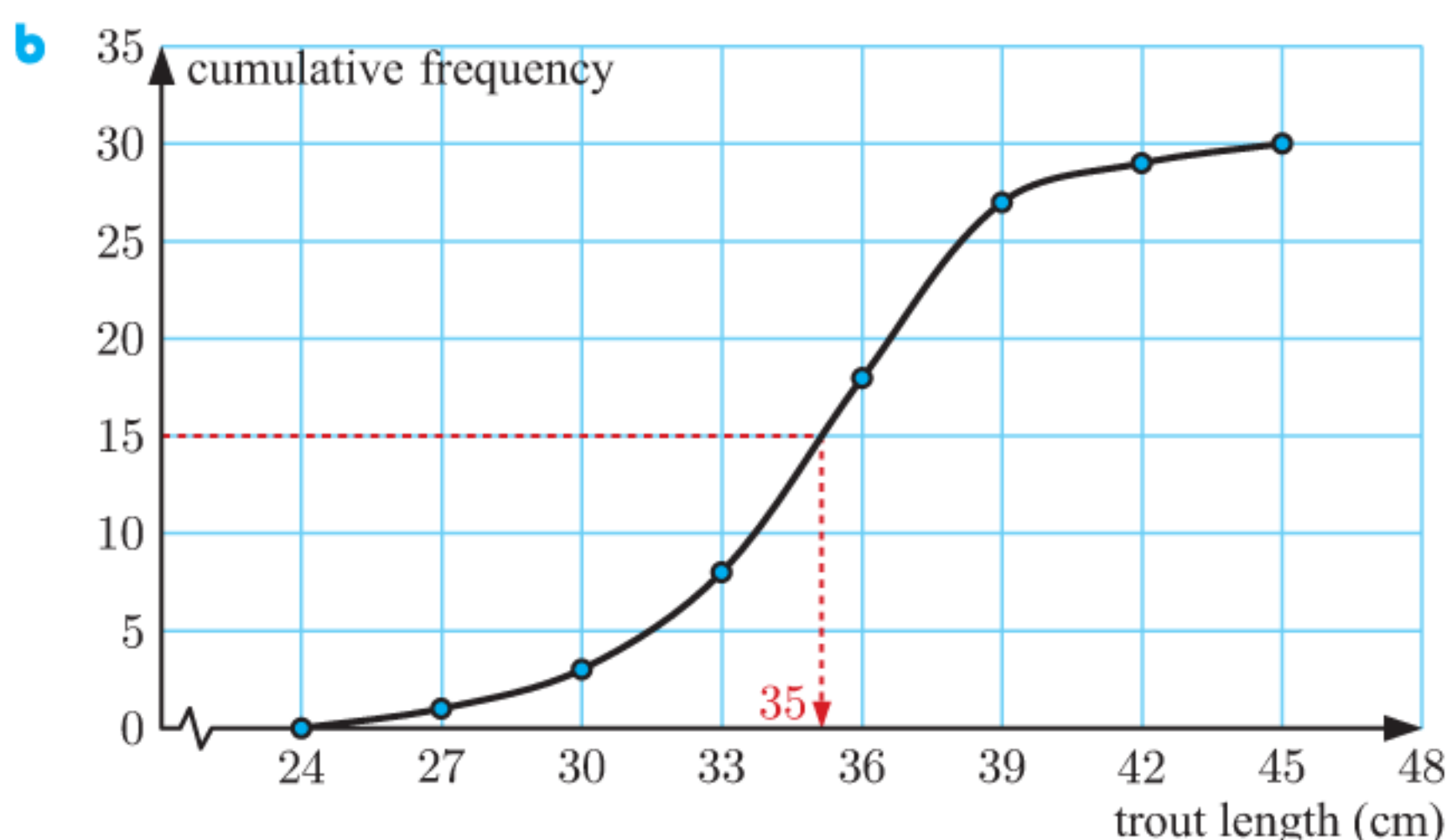
b  $\approx 61$  marks      c  $\approx 92$  students      d 76 students  
 e  $\approx 23$  students      f  $\approx 75$  marks

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.



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

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



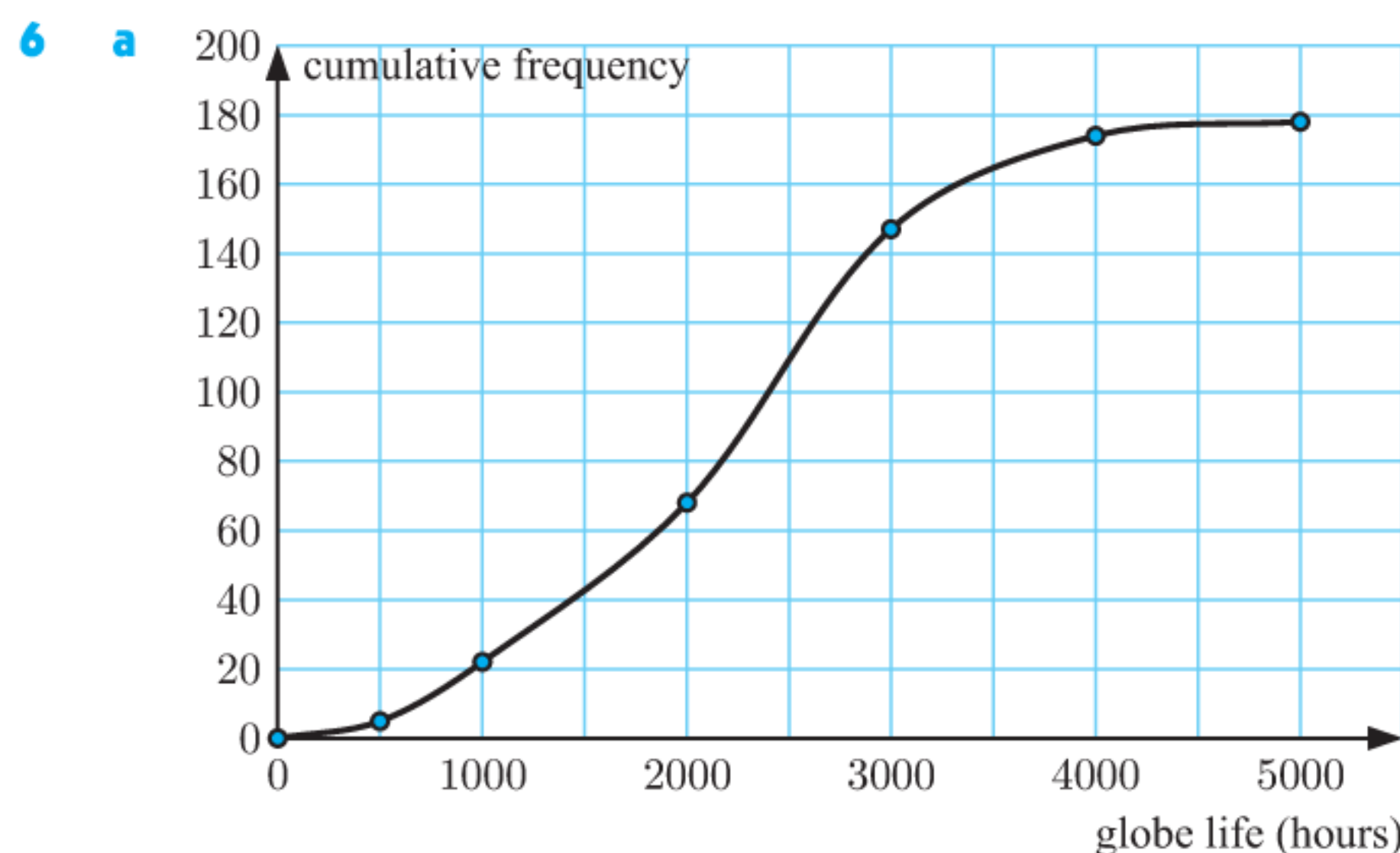
c median  $\approx 35$  cm  
 d median = 34.5 cm; the median found from the graph is a good approximation.

5 a  $\approx 27$  min      b  $\approx 29$  min      c  $\approx 31.3$  min  
 d  $\approx 4.3$  min      e  $\approx 28.2$  min

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

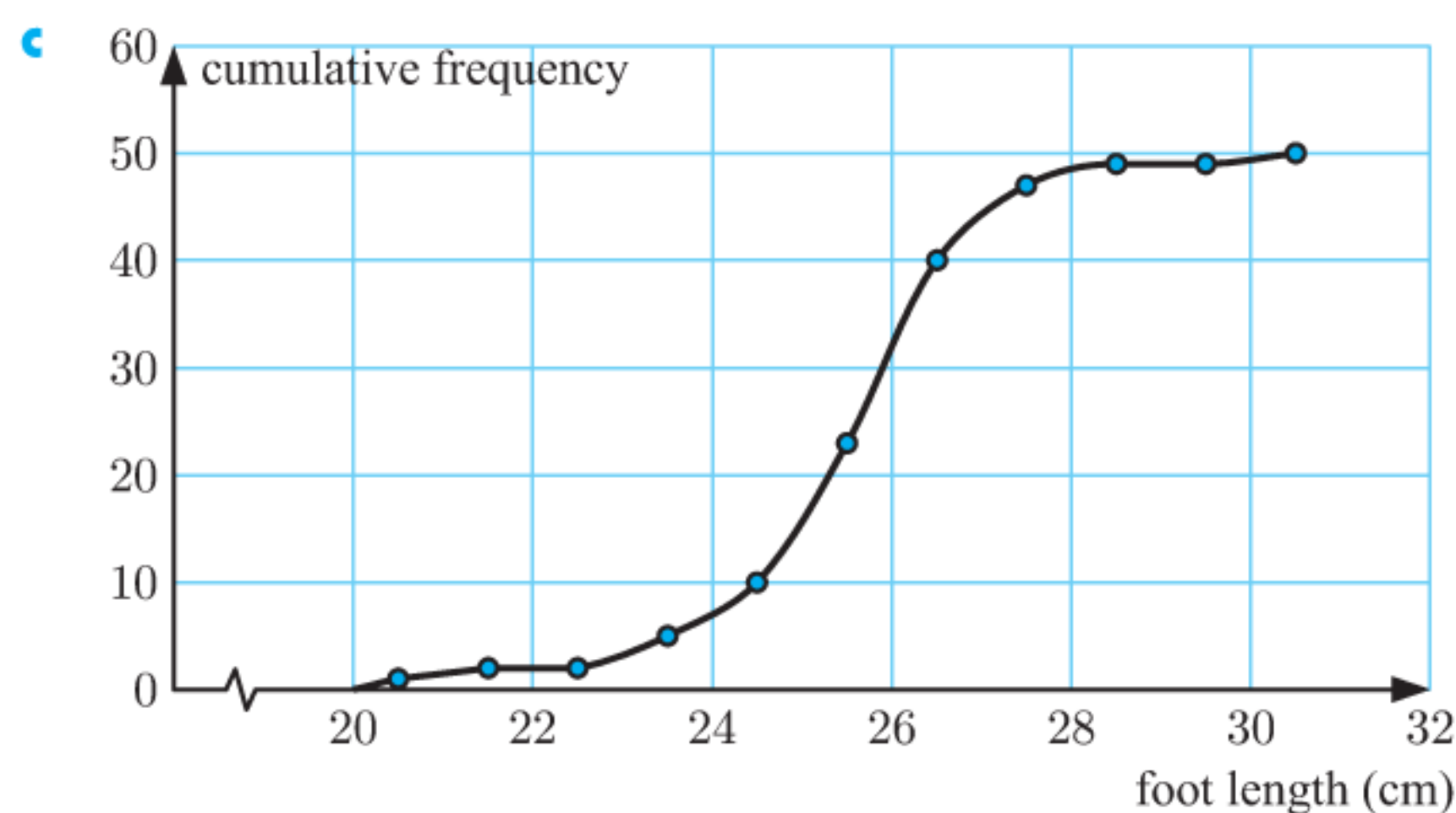
Time ( $t$ min)	$30 \leq t < 33$	$33 \leq t < 36$
Number of competitors	20	10



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

7 a  $19.5 \leq l < 20.5$  cm

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



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

**EXERCISE 13J**

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

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

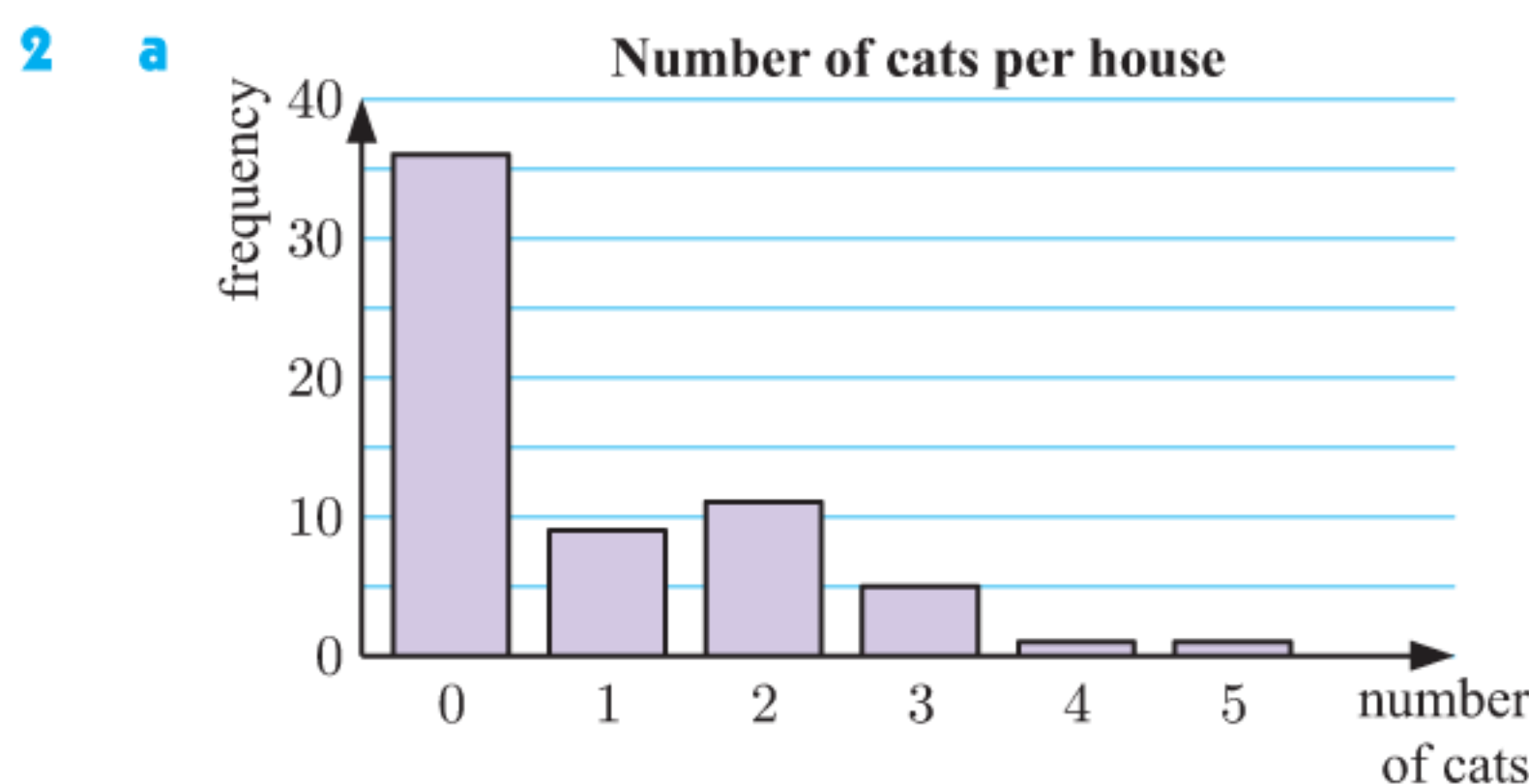


- b** Data set B appears to have a greater spread than data set A, as data set B has more values that are a long way from the mean, such as 1 and 14.
- c** Data set A:  $\sigma^2 = 3.6$ ,  $\sigma \approx 1.90$   
Data set B:  $\sigma^2 \approx 21.7$ ,  $\sigma \approx 4.65$
- 2 a** The data is positively skewed.      **b**  $\sigma \approx 1.59$   
**c**  $\sigma^2 \approx 2.54$
- 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**  $\sigma \approx 2.64$ ,  $s \approx 2.71$
- 5 a** Danny:  $\approx 3.21$  hours; Jennifer: 2 hours  
**b** Danny  
**c** Danny:  $\sigma \approx 0.700$  hours,  $s \approx 0.726$  hours;  
Jennifer:  $\sigma \approx 0.423$  hours,  $s \approx 0.439$  hours  
**d** Jennifer
- 6 a**
- |       | Mean $\bar{x}$ | Median | Standard deviation |                | Range |
|-------|----------------|--------|--------------------|----------------|-------|
|       |                |        | $\sigma$           | $s$            |       |
| Boys  | 32.02          | 31.05  | $\approx 4.52$     | $\approx 4.77$ | 13.8  |
| Girls | 34.77          | 35.85  | $\approx 3.76$     | $\approx 3.96$ | 11.7  |
- b i** boys      **ii** boys  
**c** Tyson could increase his sample size.
- 7 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:  $\sigma = 3.9$ ,  $s \approx 4.11$  ← greater variability  
Bullets:  $\sigma \approx 3.29$ ,  $s \approx 3.47$   
**c** standard deviation
- 8 a i** Museum:  $\approx 934$  visitors; Art gallery:  $\approx 1230$  visitors  
**ii** Museum:  $\sigma \approx 208$  visitors,  $s \approx 211$  visitors;  
Art gallery:  $\sigma \approx 84.6$  visitors,  $s \approx 86.0$  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  $\approx 965$  visitors,  $\sigma \approx 121$  visitors,  
 $s \approx 123$  visitors  
**v** The outlier had greatly increased the population standard deviation.
- 9**  $s_A > s_B$  does not imply that  $\sigma_A > \sigma_B$ .  
**Hint:** Find a counter example.
- 10**  $p = 6$ ,  $q = 9$       **11**  $a = 8$ ,  $b = 6$       **12 b**  $\mu = \pm 8.7$
- 13**  $\sigma \approx 0.775$       **14**  $\mu = 14.48$  years,  $\sigma \approx 1.75$  years
- 15 a** Data set A      **b** Data set A: 8, Data set B: 8  
**c** Data set A: 2, Data set B:  $\approx 1.06$   
Data set A does have a wider spread.  
**d** The standard deviation takes all of the data values into account, not just two.
- 16 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:  $\mu \approx 15.5$ ,  $\sigma \approx 1.65$

- 17** 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.
- 18 a**  $\bar{x} \approx 48.3$  cm      **b**  $\sigma \approx 2.66$  cm,  $s \approx 2.70$  cm
- 19 a**  $\bar{x} \approx 17.45$       **b**  $\sigma \approx 7.87$ ,  $s \approx 7.91$
- 20 a**  $\bar{x} \approx \$780.60$       **b**  $\sigma \approx \$31.74$ ,  $s \approx \$31.82$
- 21 a**  $\bar{x} = 40.35$  hours,  $\sigma \approx 4.23$  hours,  $s \approx 4.28$  hours  
**b**  $\bar{x} = 40.6$  hours,  $\sigma \approx 4.10$  hours,  $s \approx 4.15$  hours  
The mean increases slightly; the standard deviation decreases slightly. These are good approximations.

## REVIEW SET 13A

- 1 a i**  $\approx 4.67$       **ii** 5      **b i** 3.99      **ii** 3.9



- b** positively skewed  
**c i** 0 cats      **ii**  $\approx 0.873$  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**  $\approx 22.6$  points
- 7 a** min = 3,  $Q_1 = 12$ , med = 15,  $Q_3 = 19$ , max = 31  
**b** range = 28, IQR = 7  
**c**
- 
- 8 a** 101.5      **b** 7.5      **c** 100.2      **d**  $\approx 7.59$
- 9 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.
- 10 a**  $\approx 58.5$  s      **b**  $\approx 6$  s      **c**  $\approx 53$  s



11 a  $\approx 88$  students

b  $m \approx 24$

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

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

13 a  $\bar{x} \approx 49.6$  matches,  $\sigma \approx 1.60$  matches,  $s \approx 1.60$  matches

b The claim is not justified, but a larger sample is needed.

14 a  $\bar{x} \approx 33.6$  L      b  $\sigma \approx 7.63$  L,  $s \approx 7.66$  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.

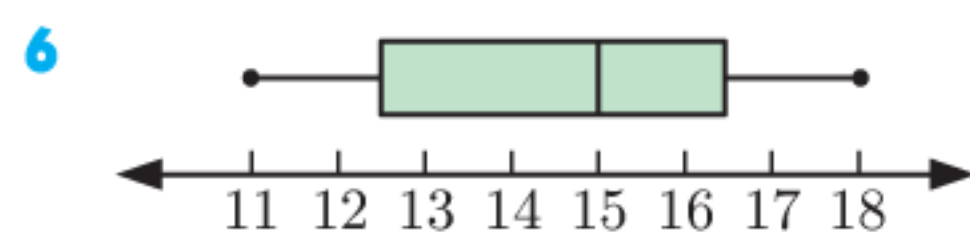
**REVIEW SET 13B**

	mean (seconds)	median (seconds)
Week 1	$\approx 16.0$	16.3
Week 2	$\approx 15.1$	15.1
Week 3	$\approx 14.4$	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.

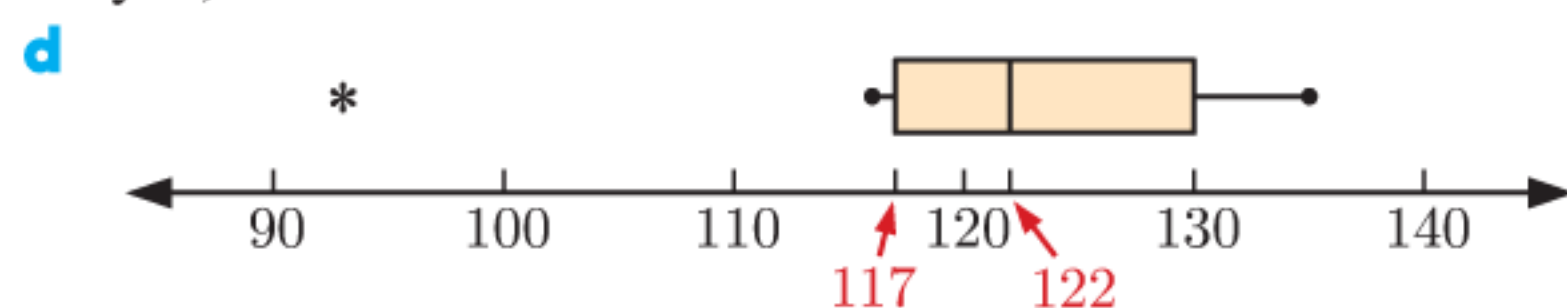
2 a 5      b 3.52      c 3.5      3 a  $x = 7$       b 6

4  $p = 7, q = 9$  (or  $p = 9, q = 7$ )      5  $\approx 414$  patrons

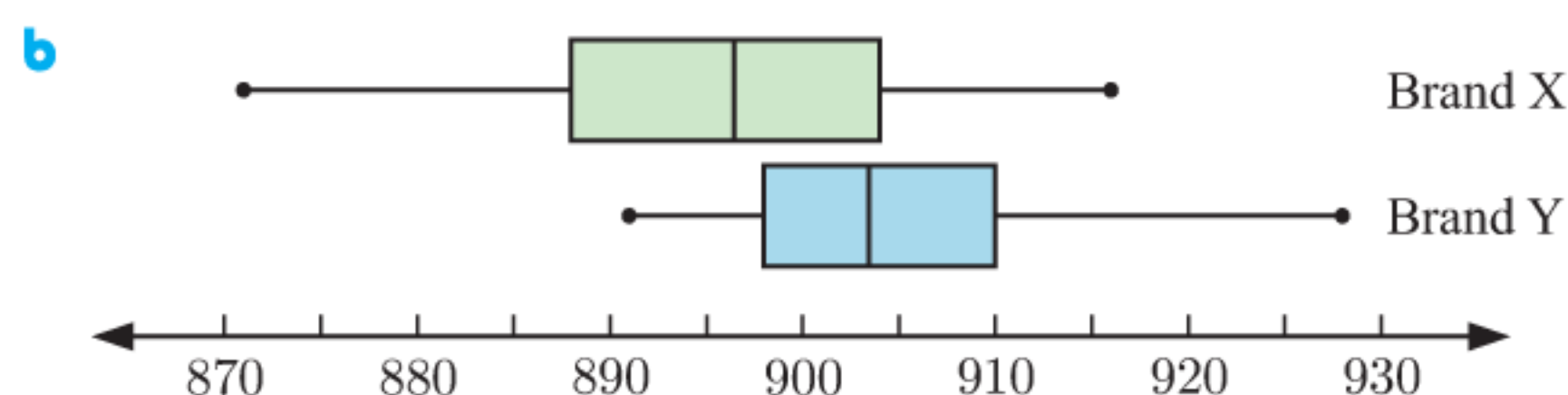


7 a  $\sigma \approx 11.7, s \approx 12.4$       b  $Q_1 = 117, Q_3 = 130$

c yes, 93



	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



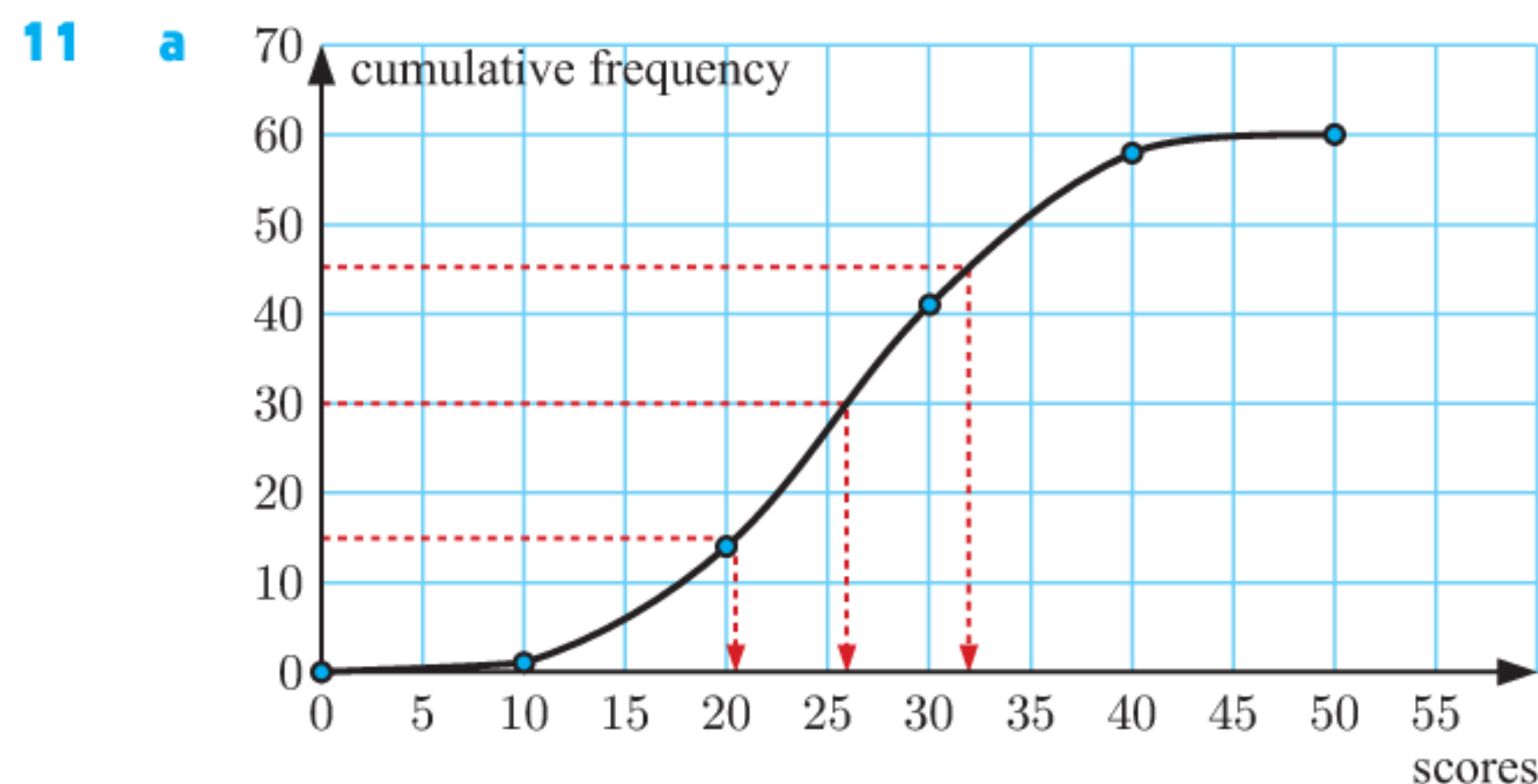
c i Brand Y, as the median is higher.  
ii Brand Y, as the IQR is lower, so less variation.

9 a  $p = 12, m = 6$

c  $\bar{x} = \frac{254}{30} = \frac{127}{15}$

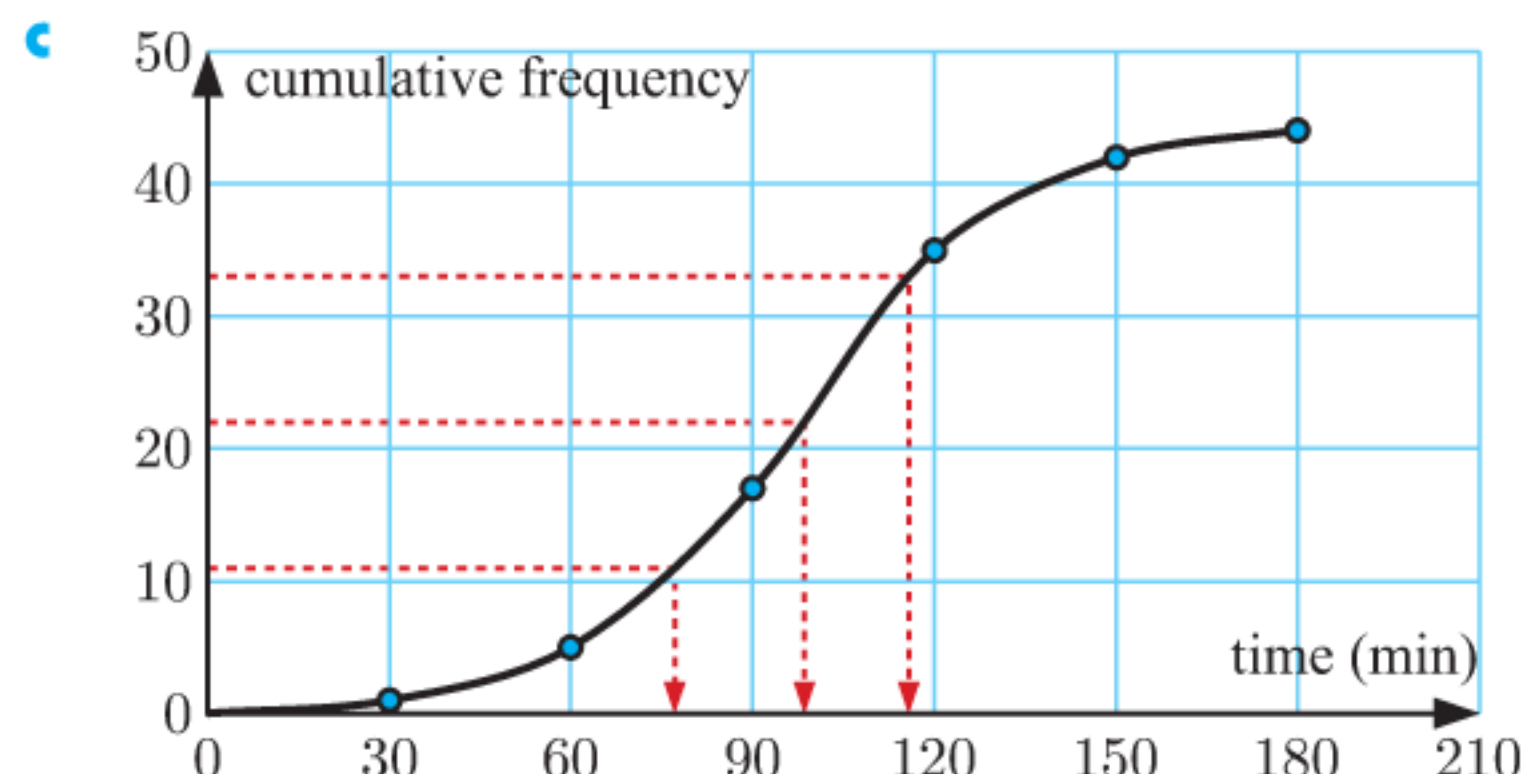
Measure	Value
mode	9
median	9
range	4

10 a  $\approx 77$  days      b  $\approx 12$  days



b i median  $\approx 26$       ii IQR  $\approx 11.5$   
iii  $\bar{x} \approx 26.0$       iv  $\sigma \approx 8.31$

12 a 44 players      b  $90 \leq t < 120$  min



d i  $\approx 98.6$  min      ii  $\approx 96.8$  min      iii no  
e "... between 77.2 and 115.7 minutes."

13 a  $\bar{x} \approx \text{€}207.02$       b  $\sigma = \text{€}38.80, s \approx \text{€}38.89$

14 a Kevin:  $\bar{x} = 41.2$  min; Felicity:  $\bar{x} = 39.5$  min

b Kevin:  $\sigma \approx 7.61$  min,  $s \approx 7.81$  min;  
Felicity:  $\sigma \approx 9.22$  min,  $s \approx 9.46$  min

c Felicity      d Kevin

15 10 data values

**EXERCISE 14A**

1 a  $y = x^2 - 3x + 1$

$x$	-2	-1	0	1	2
$y$	11	5	1	-1	-1

b  $y = x^2 + 2x - 5$

$x$	-2	-1	0	1	2
$y$	-5	-6	-5	-2	3

c  $y = 2x^2 - x + 3$

$x$	-4	-2	0	2	4
$y$	39	13	3	9	31

d  $y = -3x^2 + 2x + 4$

$x$	-4	-2	0	2	4
$y$	-52	-12	4	-4	-36

2 a no      b yes      c yes      d yes      e no      f yes

3 a  $x = -1$  or  $-2$       b  $x = 2$       c  $x = 1$  or  $5$

d  $x = -3$  or  $\frac{1}{2}$       e  $x = -6$  or  $1$       f no real solutions

**EXERCISE 14B.1**

