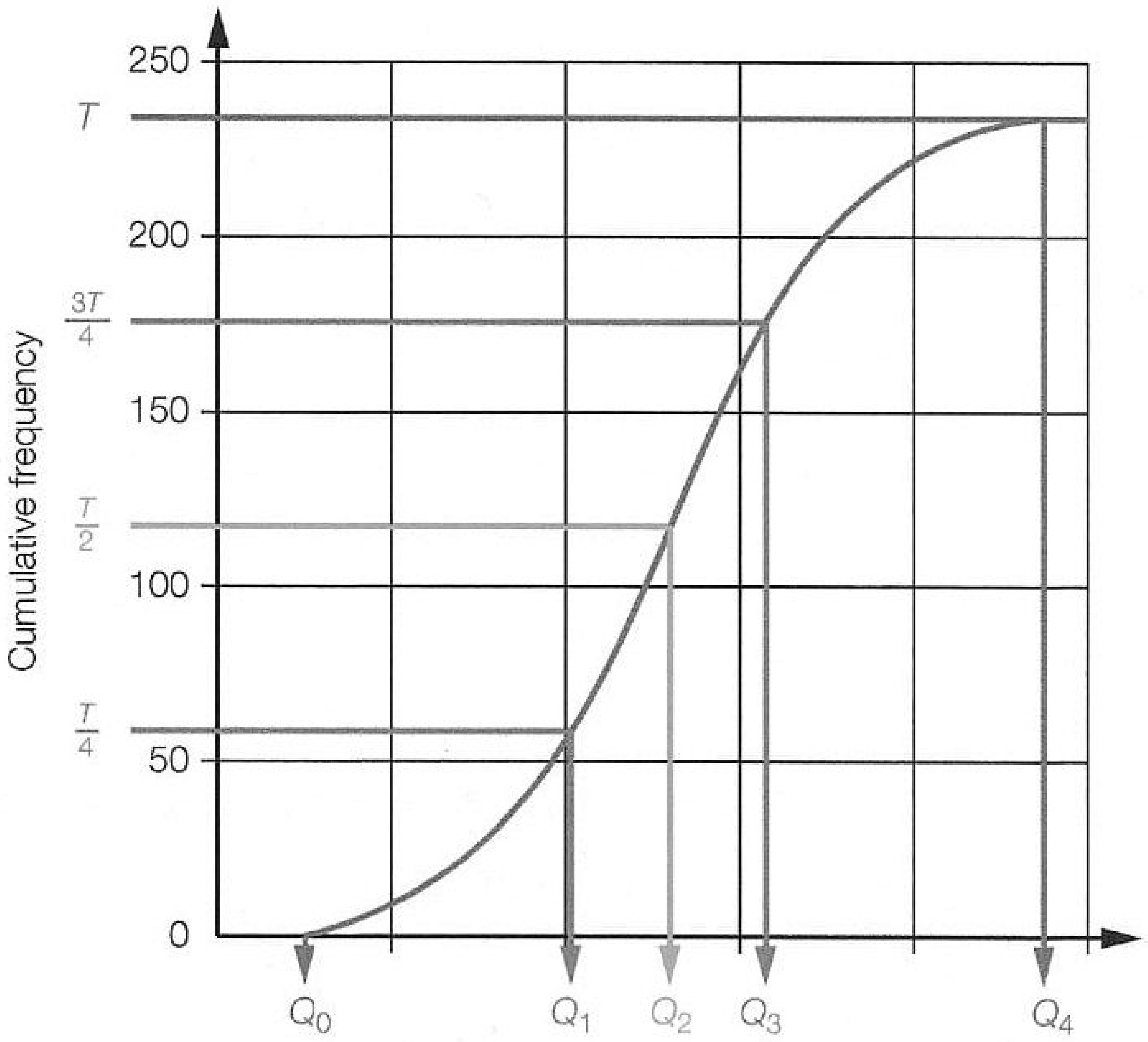
3 DESCRIPTIVE STATISTICS

WHAT YOU NEED TO KNOW

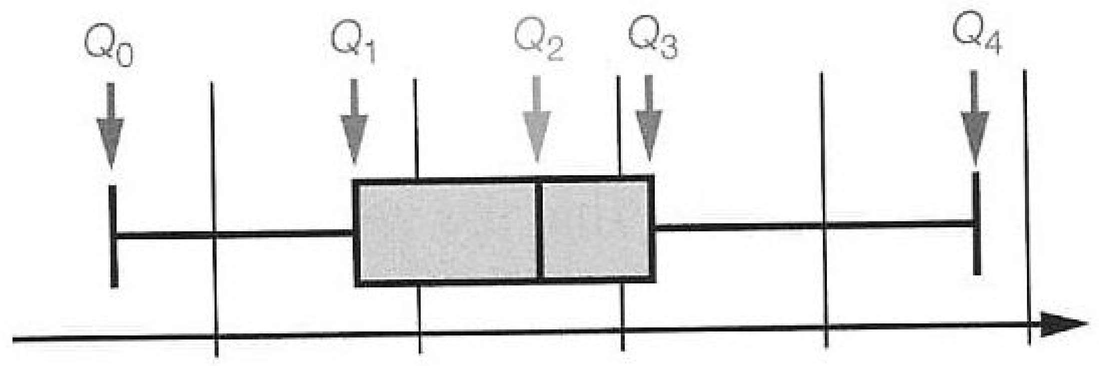
- Grouping a large sample of data into groups (or classes) makes it easier to summarise.
    - The upper and lower class boundaries are the largest and smallest data values that would be included in that group.
    - The mid-interval value is the value that is half-way between the upper and lower class boundaries.
    - If the data are discrete, the upper and lower class boundaries are as shown in the grouped frequency table. For example, a group 12–15 includes the values 12, 13, 14 and 15, and the mid-interval value is 13.5.
    - If the data are continuous but have been rounded, the class boundaries need to be adjusted. For example, if lengths have been rounded to the nearest metre, then 12–15 means that $11.5 \le \text{length} < 15.5$, and the mid-interval value is 13.5.
- Three measures of central tendency are the mean, median and mode.
    - The mean is the sum of all the data values divided by the total number of items in the data set. The formula for finding the mean is:

      $\bar{x} = \dfrac{\sum_{i=1}^{k} f_i x_i}{n}$, where $n = \sum_{i=1}^{k} f_i$ is the total frequency, x_i is the ith distinct data value and f_i is the frequency of that value.
        - To find the mean for grouped data, assume that every item is at the centre of its group; that is, use the mid-interval value of the group for x_i.
    - The median is the middle of a data set whose values have been arranged in order of size. If there are an odd number of items, the middle is the $\left(\dfrac{n+1}{2}\right)$th value. If there are an even number of items, find the mean of the two middle values, the $\left(\dfrac{n}{2}\right)$th and $\left(\dfrac{n}{2}+1\right)$th values.
    - The mode is the most common data value. If the data is grouped, the group with the largest frequency is called the modal group or modal class.
- There are three ways of measuring how spread out a data set is: range, interquartile range and standard deviation.
    - The range is the difference between the highest value and the lowest value.
    - The interquartile range (IQR) is the difference between the upper quartile (Q_3) and the lower quartile (Q_1): $\text{IQR} = Q_3 - Q_1$.
        - To find the quartiles, first divide the data set (with numbers arranged in order) into two halves. The lower quartile is the middle value of the bottom half. The upper quartile is the middle value of the top half.

- The standard deviation is a measure of the average distance of the data values from the mean. It can be calculated using a GDC.
- In a histogram, the height of each bar indicates the frequency of that group. The horizontal scale should be continuous, with each bar covering the group it represents.
- Cumulative frequency is the total frequency up to a certain data value.
  - To draw a cumulative frequency curve from a grouped frequency table, mark the values of the upper class boundaries along the x-axis and plot the cumulative frequency values along the y-axis. Join up the points with a smooth increasing curve.
  - The cumulative frequency curve can be used to find the median and quartiles. For the median, the value on the y-axis will be at $\frac{\text{total frequency}}{2}$; for the lower quartile, the value will be at $\frac{\text{total frequency}}{4}$, and for the upper quartile it will be at $\frac{3 \times \text{total frequency}}{4}$.



- A box and whisker diagram summarises five important values from a set of data: the smallest value (Q_0), the lower quartile (Q_1), the median (Q_2), the upper quartile (Q_3) and the largest value (Q_4).



  - Each 'whisker' and each half of the 'box' contains a quarter (25%) of all the data.
  - Box and whisker diagrams can be used to compare two or more sets of data.

- To compare two sets of data, use the mean and median to compare 'averages', and use the standard deviation and IQR to compare the spread of the data. The data set with the smaller spread can be described as 'less varied' or 'more consistent'.

EXAM TIPS AND COMMON ERRORS

- When using a calculator to find statistics, make sure you show the numbers that you are using.

- A common error when drawing a cumulative frequency curve is forgetting to plot the first point, which corresponds to zero frequency at the lower boundary of the first group.

- In a box and whisker diagram, the width of the box represents the interquartile range, **not** the number of data values in that range; it is a common mistake to say that a wider box 'contains more people'. When comparing two box and whisker diagrams, always make it clear whether it is the width of the box or its position that is being referred to. 'Higher IQR' should mean that the box is wider, not that the quartiles are larger.

3.1 MEDIAN AND QUARTILES

WORKED EXAMPLE 3.1

Consider the following set of data:

14, 12, 18, 20, 17, 18, 12, 16, 15, x

(a) Given that the median of the data set is 16, find the value of x.

(b) Find the interquartile range of the data.

(a) 12, 12, 14, 15, 16, 17, 18, 18, 20

Put the nine known numbers in order of size. Including x, there are 10 numbers, so the median is the mean of the 5th and 6th values.

The median is the mean of 16 and x.

Since the median is 16, $x = 16$.

The known number 16 is either the 5th or the 6th value, so the median is the mean of 16 and another number. Since we are given that the median is 16, the other number must also be 16.

(b) Lower half: 12, 12, **14**, 15, 16 $\Rightarrow Q_1 = 14$

Upper half: 16, 17, **18**, 18, 20 $\Rightarrow Q_3 = 18$

IQR $= 18 - 14 = 4$

To find the quartiles Q_1 and Q_3, divide the data set (with numbers in order) into two halves and find the median for each half. Then use the formula IQR $= Q_3 - Q_1$.

Practice questions 3.1

1. The following data set contains 13 numbers:

2, 4, 5, 10, 12, 14, 14, 16, 18, 21, 23, 23, 25

(a) Explain why the lower quartile is 7.5.

(b) Find the interquartile range of the data.

2. The median of the numbers $x - 2, x + 1, x + 3, x + 6, x + 7, x + 10$ is 6.5. Find the value of x.

3. Find the median and the interquartile range of the grades summarised in the table:

Grade	2	3	4	5	6	7
Frequency	5	12	36	42	27	19

When entering frequency table data into your GDC, check that the value of n is what you expect – in this case 141.

3.2 MEAN, STANDARD DEVIATION AND FREQUENCY HISTOGRAMS

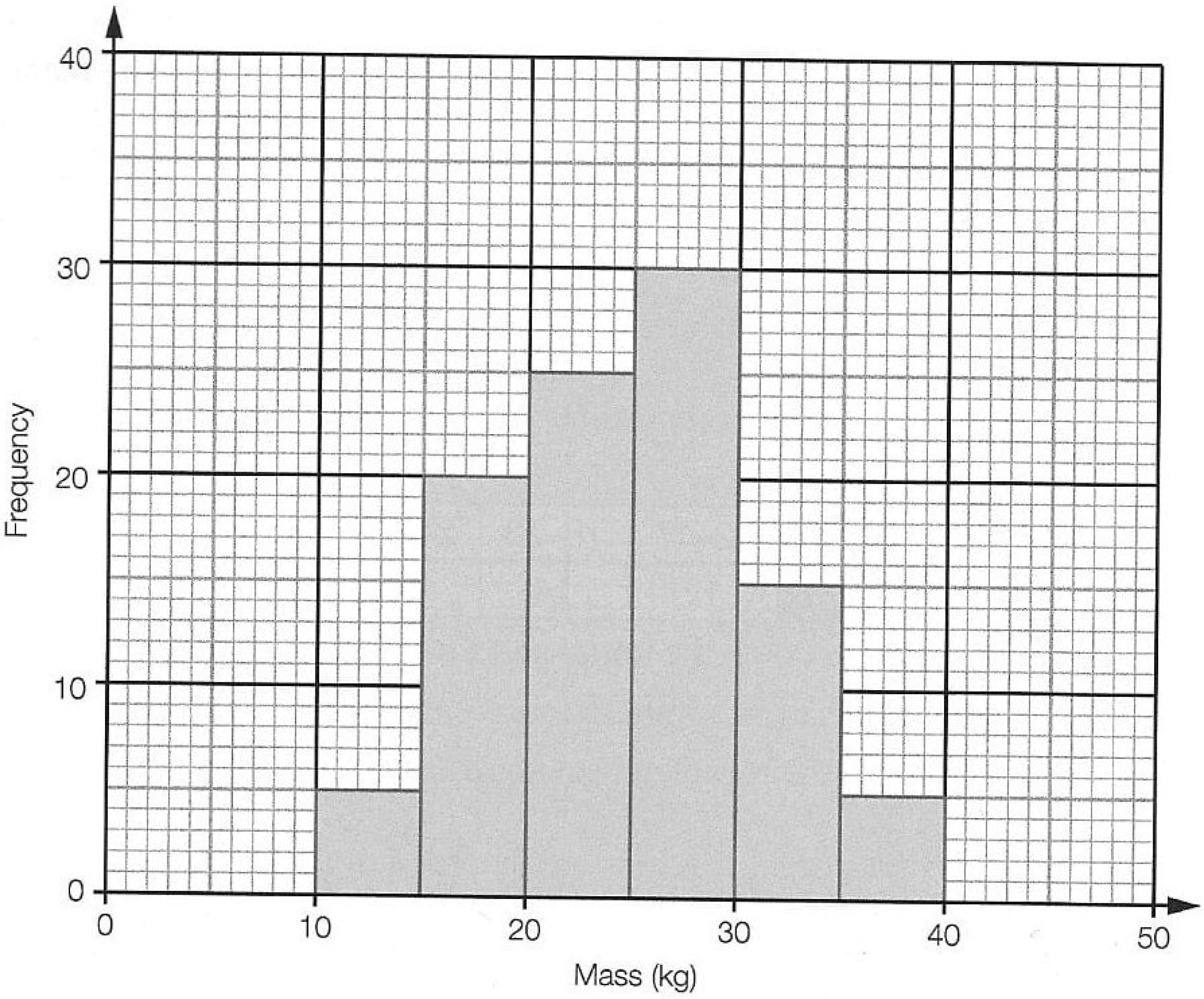
WORKED EXAMPLE 3.2

The histogram below shows the masses of some dogs.

(a) How many dogs were included in the sample?

(b) What is the modal group?

(c) Estimate the standard deviation of the masses.



(a) Total frequency
$=5+20+25+30+15+5=100$
There were 100 dogs in the sample.

The number of dogs in each group is represented by the height of the histogram bar.

(b) The modal group is 25–30 kg.

The modal group is the data class with the highest bar.

(c)

Mid-interval value	12.5	17.5	22.5	27.5	32.5	37.5
Frequency	5	20	25	30	15	5

(From GDC) Standard deviation = 6.22 (3 SF)

To estimate the standard deviation, we need to use the mid-interval value of each group, which is the mean of the lower and upper class boundaries of that group. For the first group, the mid-interval value is $\frac{10+15}{2}=12.5$.

A GDC gives two standard deviation values. You need to use the smaller one.

Practice questions 3.2

4. The mean of the numbers 4, 5, 3, 3, 5 and k is 3.5.

(a) Find the value of k.

(b) Find the standard deviation of the numbers.

5. The frequency table shows the heights of 26 trees. The mean height is 6.5 m.

Height (m)	3	5	y	10
Frequency	4	x	11	5

(a) Find the values of x and y.

(b) Calculate the standard deviation of the heights.

6. The heights of 50 buildings, rounded to the nearest metre, are summarised in the following table:

Height (m)	12–17	18–23	24–29	30–35
Frequency	12	14	16	8

(a) Write down the upper and lower boundaries of the 24–29 class.

(b) Draw a histogram to represent the data.

(c) Find the mean and standard deviation of the heights.

7. The ages of 35 teachers are recorded in the table:

Age (years)	21–30	31–40	41–50	51–60	61–70
Frequency	5	9	11	7	3

(a) Write down the upper and lower boundaries of the 21–30 group and find its mid-interval value.

(b) Draw a histogram to represent the data.

(c) Find the mean and standard deviation of the ages.

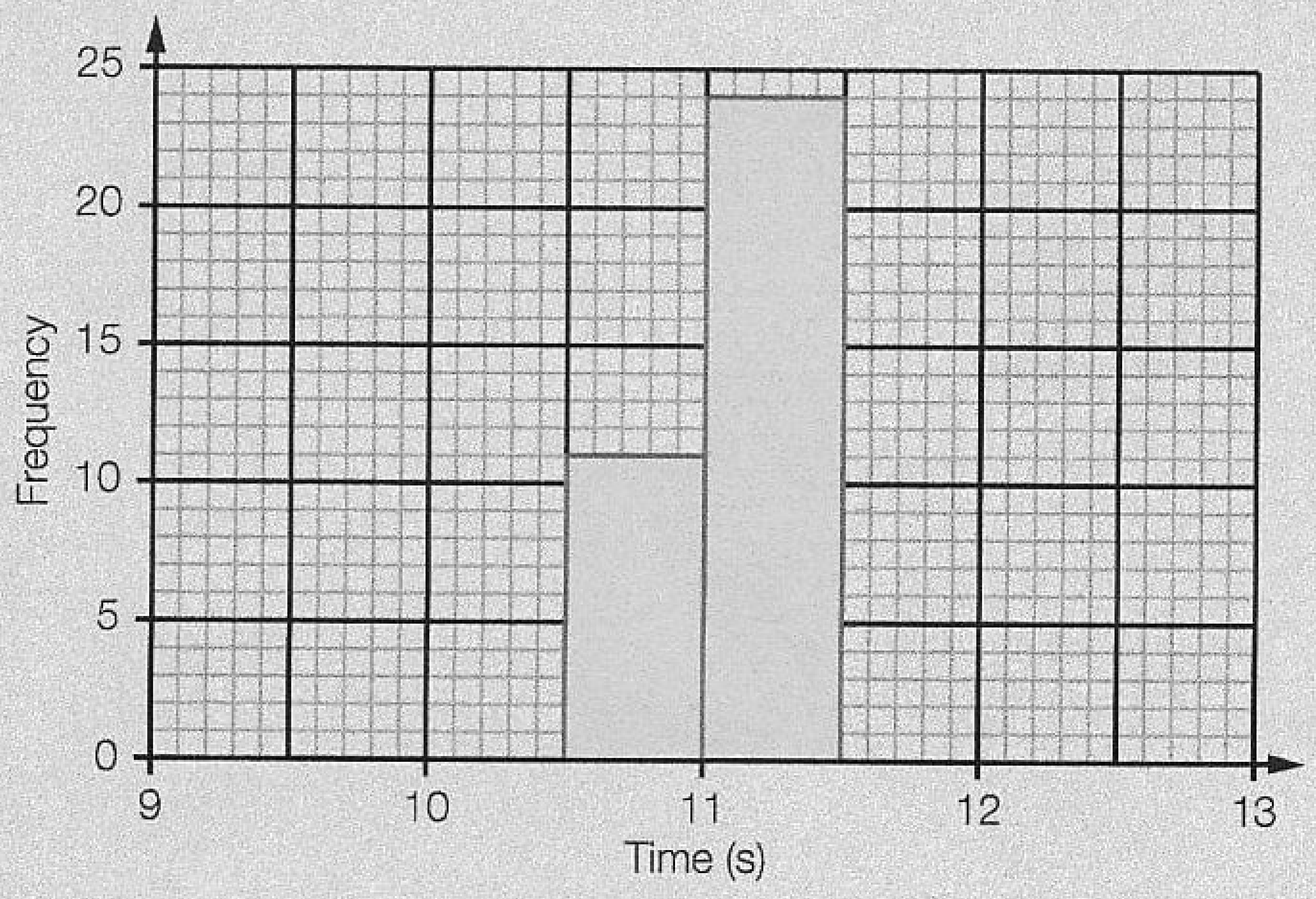
8. The times taken by a group of athletes to run 100 m were recorded in the table and histogram.

Time, t, in seconds	Frequency
$10.0 < t \le 10.5$	8
$< t \le$	11
$11.0 < t \le 11.5$	
$11.5 < t \le 12.0$	16

(a) How many athletes took part?

(b) Use the table to complete the histogram.

(c) Use the histogram to complete the table.

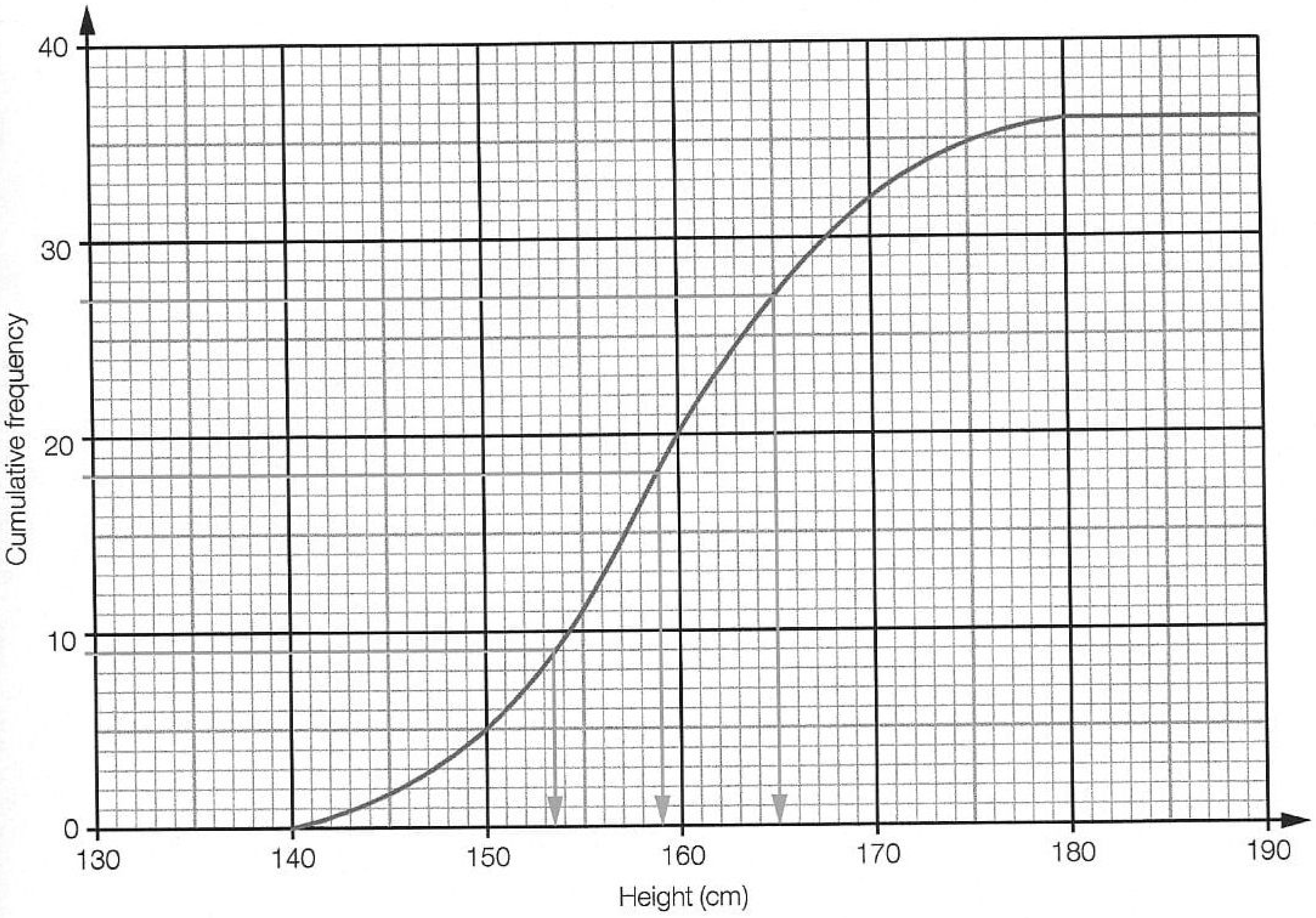


3.3 CUMULATIVE FREQUENCY

WORKED EXAMPLE 3.3

The cumulative frequency curve below represents the heights of 36 children.

Estimate the median and the interquartile range of the heights.



Median ≈ 159 cm

$Q_1 \approx 153$ cm

$Q_3 \approx 165$ cm

IQR ≈ 165 − 153 = 12 cm

$\frac{1}{2}$ of 36 = 18; $\frac{1}{4}$ of 36 = 9; $\frac{3}{4}$ of 36 = 27

To estimate the median, draw a line horizontally across from 18 on the vertical axis to the curve and then vertically down until it meets the horizontal axis. Similarly find Q_1 and Q_3, and hence calculate the IQR.

Practice questions 3.3

9. The ages of students at a school are summarised in the table.

Age (years)	Frequency
4–7	20
8–11	40
12–15	80
16–19	60

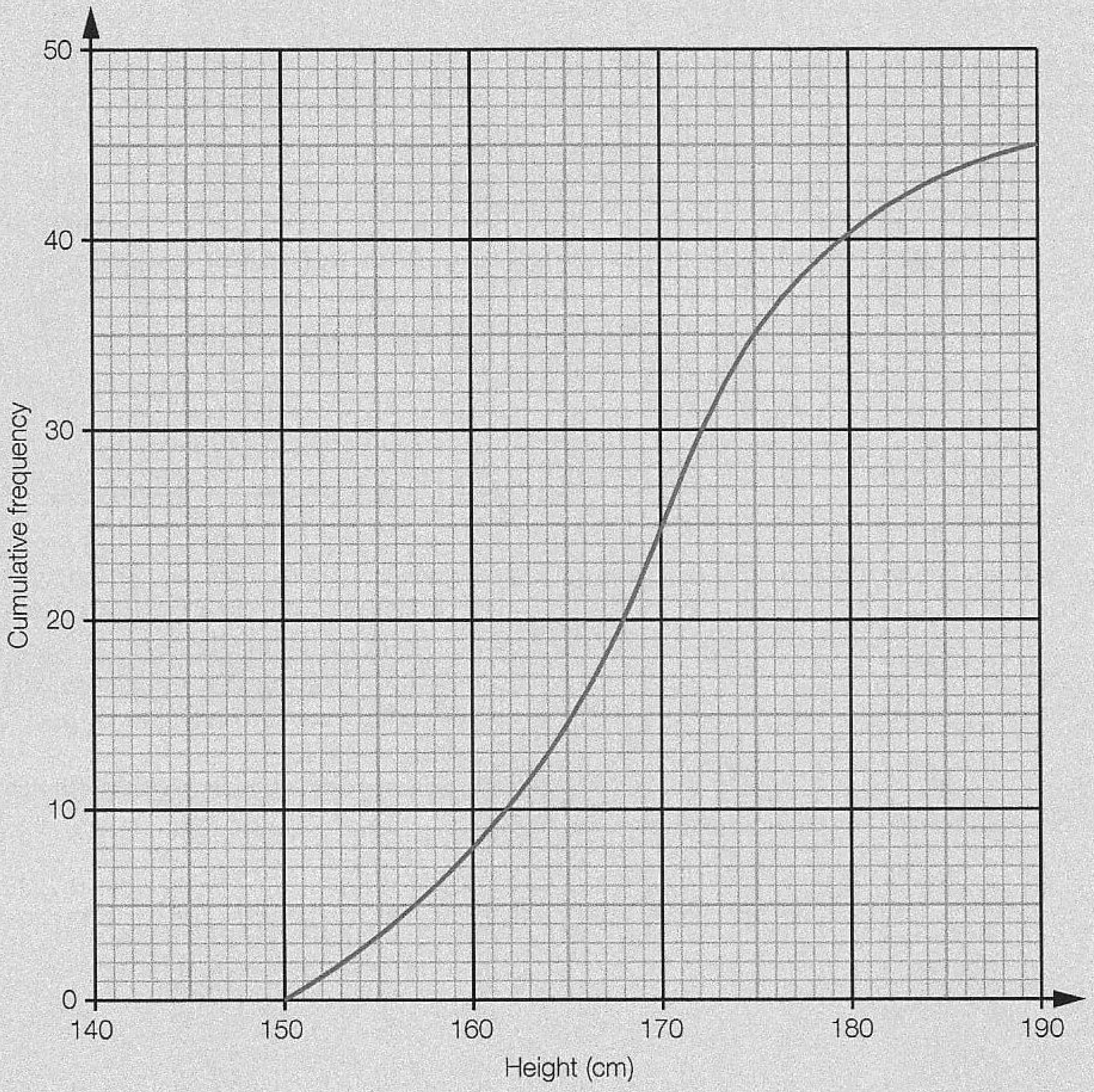
(a) Draw a cumulative frequency curve to represent the information.

(b) Find the median age.

(c) The oldest 10% of the students are older than m years. Find the value of m.

10. Use the cumulative frequency curve to complete the frequency table.

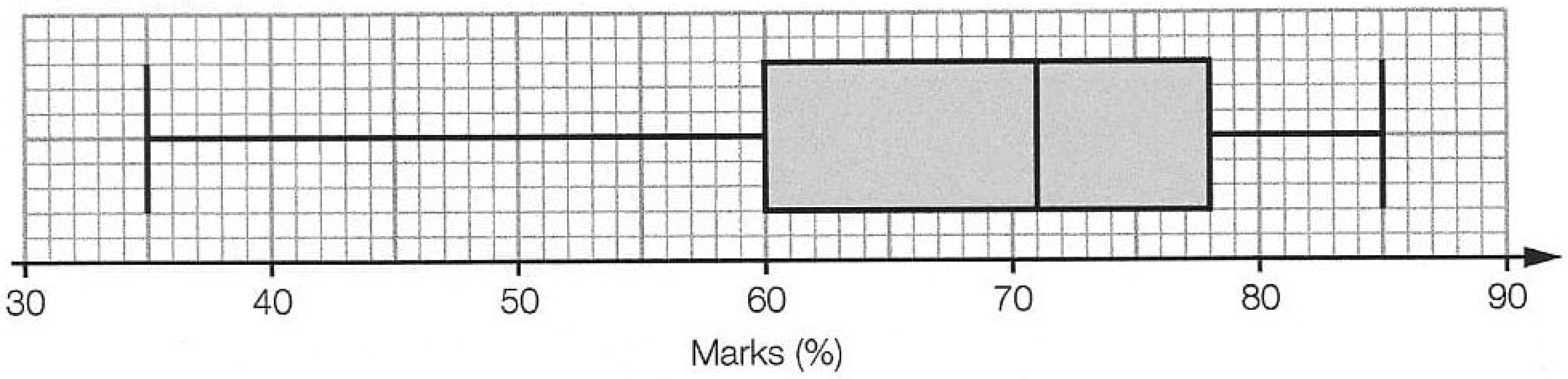
Height, h (cm)	Frequency
$150 \le h < 160$	8
$160 \le h < 168$	
$168 \le h < 175$	
$175 \le h < 190$	



3.4 BOX AND WHISKER DIAGRAMS AND COMPARING DATA SETS

WORKED EXAMPLE 3.4

The English test marks (in %) of a group of students are summarised in the following diagram.



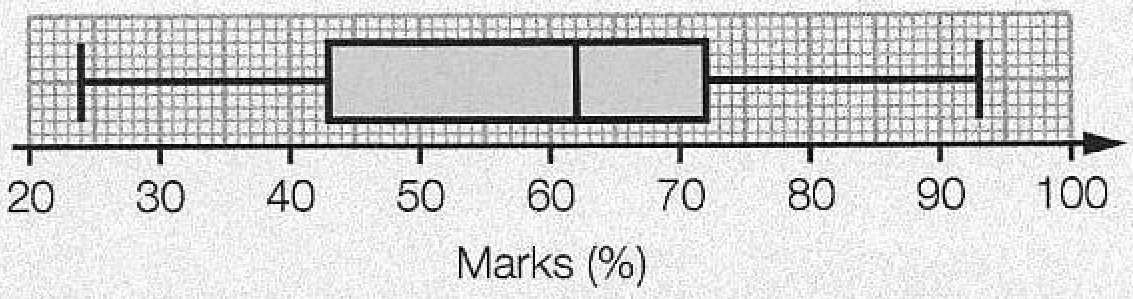
The marks for the same group of students on a Mathematics test had the following features:

lowest mark 24%, highest mark 93%, lower quartile 43%, median 62%, upper quartile 72%.

(a) Draw a box and whisker diagram for the Mathematics marks.

(b) Compare the performance of the group in the two tests.

(a)



Use the five figures given to draw the box and whisker diagram.

(b) The English marks are higher on average (median of 71% compared to 62%).

The IQR is 18 for English and 29 for Maths. So the English marks are more consistent.

The IQR indicates the spread of the data.

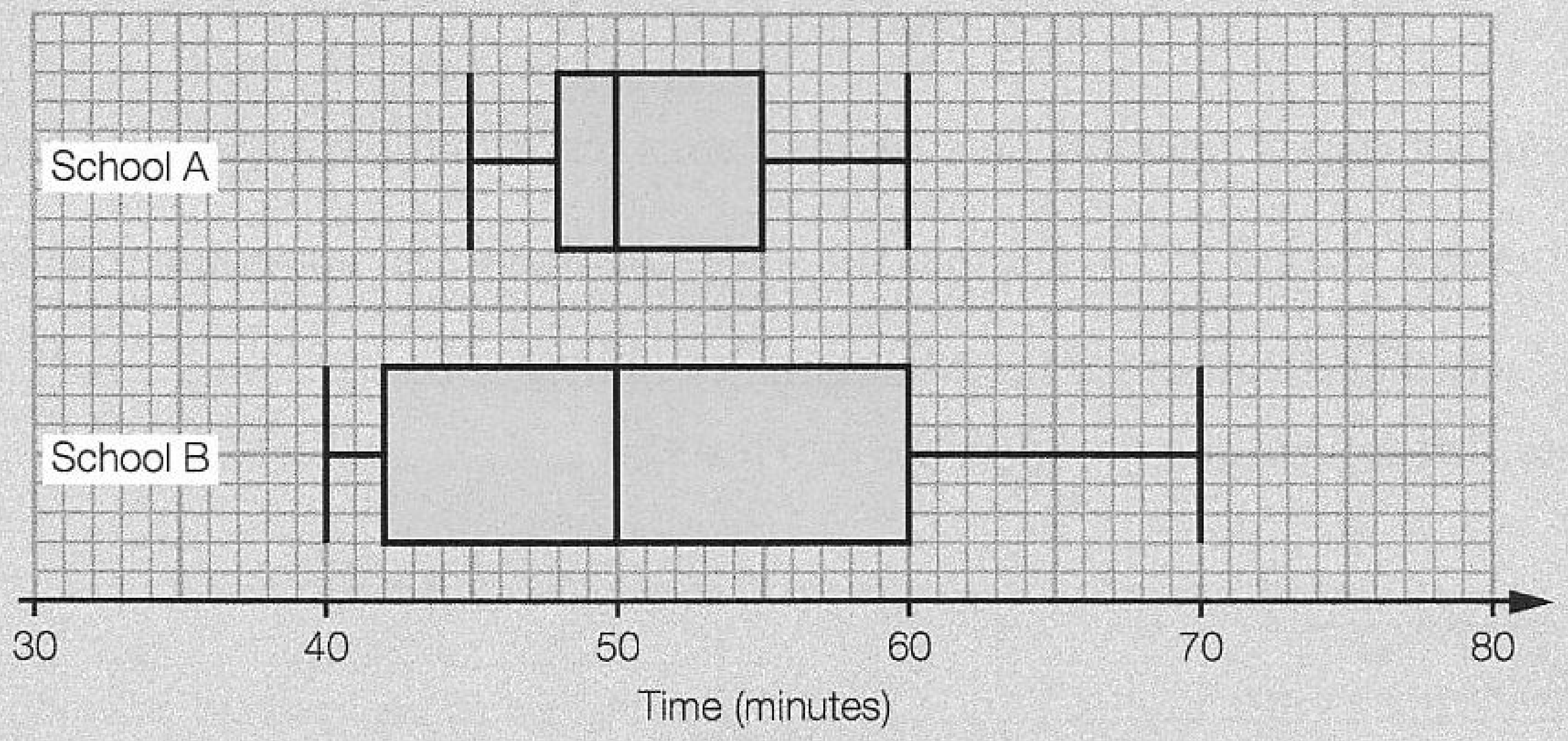
⚠ When asked to compare data you should comment on both the centre and the spread.

Practice questions 3.4

11. Draw a box and whisker diagram to represent the following data:

Smallest	Largest	Mean	Standard deviation	Median	Q_1	Q_3
12	45	31.6	5.5	35	23	42

12. Two schools took part in a cross-country race. The times are summarised in the diagram below. Write three comments comparing the times of the students from the two schools.



When comparing two sets of data, always refer to the context of the question. For example, instead of 'the men have a higher mean' you should say 'the men are older on average'.

Mixed practice 3

1. The mode of the following list of numbers is 5:

   1, 2, 2, 5, 4, 5, 6, 3, 6, x.

   (a) Find the value of x.

   (b) Find the median of the numbers.

2. The results of a Physics exam for two different schools are summarised in the table below:

	Lowest mark	Highest mark	Median	Lower quartile	Upper quartile
School 1	20	52	32	26	41
School 2	31	60	39	34	48

   (a) Calculate the interquartile ranges of the marks for the two schools.

   (b) Draw two box and whisker diagrams to represent the results.

   (c) Describe one similarity and one difference between the two schools' results.

3. The table shows the History grades of IB students at a college:

Grade	3	4	5	6	7
Number of students	4	12	x	17	9

   (a) Given that the mean grade is 5.23 (to three significant figures), find the value of x.

   (b) Find the median grade.

4. The maximum speed (in km/h) for 130 cars is recorded in the frequency table:

Speed (km/h)	140–160	160–175	175–185	185–190	190–200
Number of cars	14	35	41	23	17

   (a) Draw a cumulative frequency curve to represent the information.

   (b) Draw a box and whisker diagram to represent the data.

   (c) Find the interquartile range of the speeds.

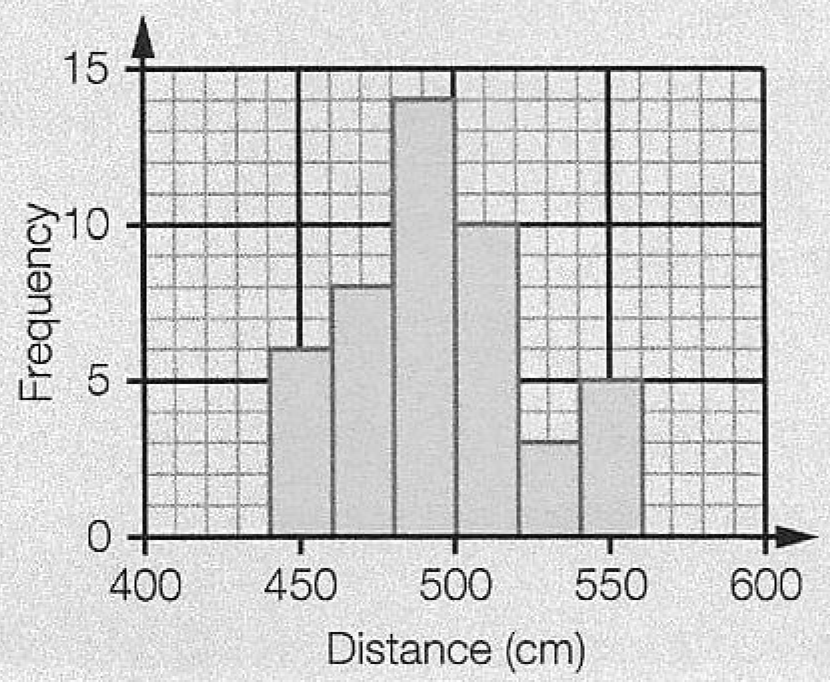
5. The ages of employees at a company are summarised in the cumulative frequency table. The youngest employee is 16 years old.

   (a) Draw a histogram to represent the data.

   (b) Estimate the mean and standard deviation of the ages.

Age (years)	Cumulative frequency
≤ 26	12
≤ 36	46
≤ 46	82
≤ 56	90

6. All athletes in a club competed in a long jump competition. Their results are shown in the histogram:



(a) Use the histogram to complete the frequency table, where the distances have been rounded to the nearest centimetre.

Distance (cm)	441–460	461–480	481–500	501–520	521–540	541–560
Frequency						

(b) Estimate the mean and the standard deviation of the distances.

(c) Draw a cumulative frequency curve to represent the information.

(d) (i) Find the percentage of athletes who jumped further than 4.80 m.

(ii) Two athletes are selected at random. What is the probability that they both jumped further than 4.80 m?

(e) The top 20% of athletes will qualify for a regional competition. Estimate the minimum distance required for qualification.

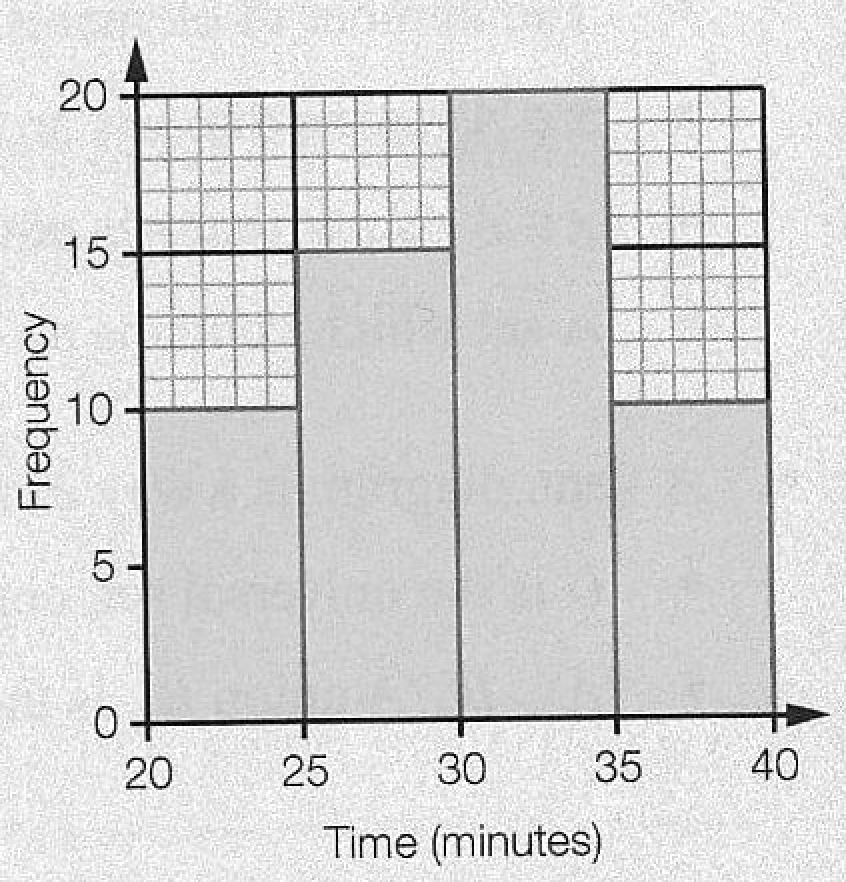
Going for the top 3

1. The frequency table summarises 36 pieces of data with mean $\frac{47}{9}$. Find the values of x and y.

Value	4	5	6	7
Frequency	9	13	x	y

2. The set of numbers 3, 2, 3, 7, 10, 5, 7, 12, x, y, z has mode 3, median 6 and mean 6. Find the values of x, y and z.

3. The histogram shows the times a group of 55 students took to complete their homework.



(a) Estimate the number of students who took:

(i) more than 25 minutes

(ii) more than 37 minutes.

(b) Given that a student took more than 25 minutes to complete their homework, find the probability that they took more than 37 minutes.