Summary of key points

4 • Variables or data associated with numerical observations are called **quantitative variables** or **quantitative data**.
  • Variables or data associated with non-numerical observations are called **qualitative variables** or **qualitative data**.

5 • A variable that can take any value in a given range is a **continuous variable**.
  • A variable that can take only specific values in a given range is a **discrete variable**.

6 • When data is presented in a grouped frequency table, the specific data values are not shown. The groups are more commonly known as **classes**.
  • Class boundaries tell you the maximum and minimum values that belong in each class.
  • The midpoint is the average of the class boundaries.
  • The class width is the difference between the upper and lower class boundaries.

Summary of key points

1 A common definition of an outlier is any value that is:
- either greater than $Q_3 + k(Q_3 - Q_1)$
- or less than $Q_1 - k(Q_3 - Q_1)$

2 The process of removing anomalies from a data set is known as cleaning the data.

3 On a histogram, to calculate the height of each bar (the **frequency density**) use the formula

area of bar $= k \times$ frequency.

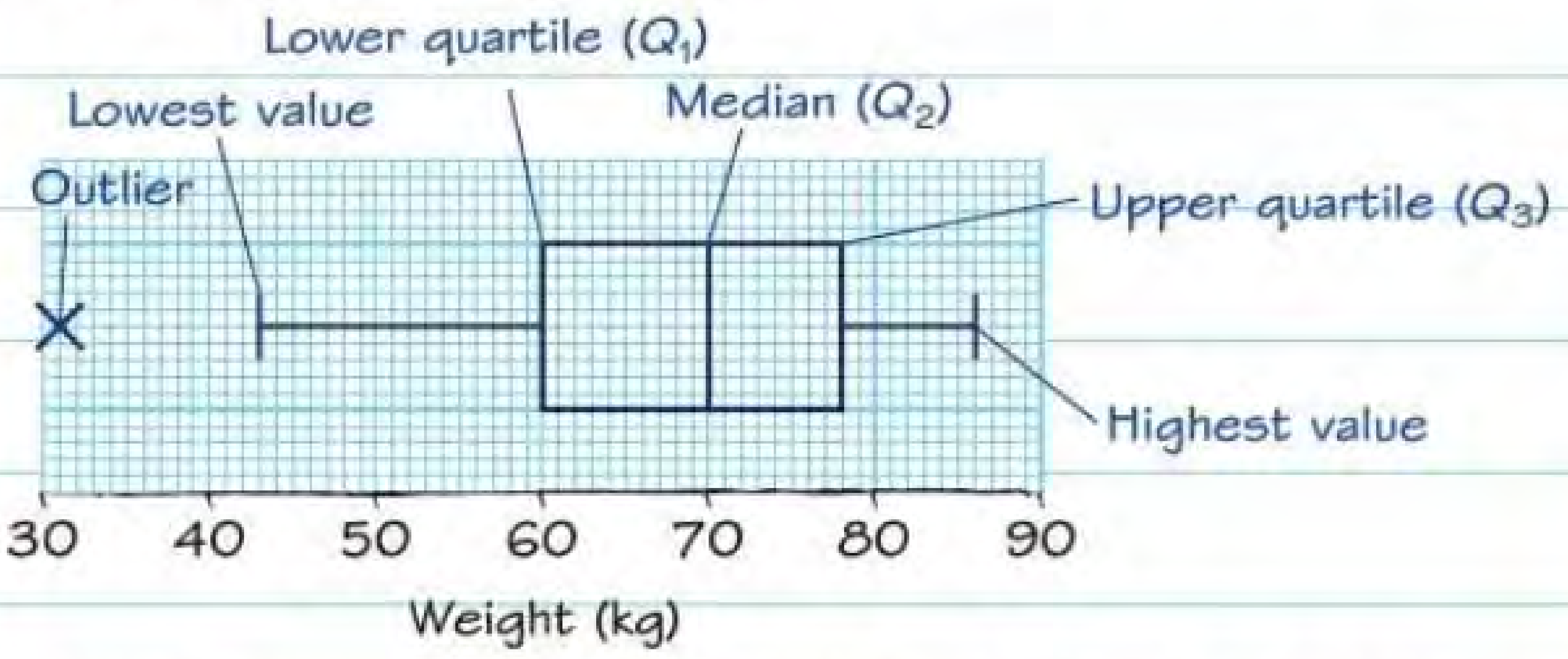
4 Joining the middle of the top of each bar in a histogram with equal class widths forms a frequency polygon.

5 When comparing data sets you can comment on:
- a measure of location
- a measure of spread

Had a look ☐ Nearly there ☐ Nailed it! ☐

Box plots and outliers

A box plot shows the **maximum** and **minimum** values and the **quartiles** of a distribution. It is usually drawn on graph paper with a scale.



Half the weights were between 60 kg and 78 kg. 25% of the weights were less than 60 kg.

Outliers

A value which doesn't fall within the main body of the data is called an **outlier**.

One common definition of an outlier is:

- values less than $Q_1 - 1.5 \times \text{IQR}$
- values greater than $Q_3 + 1.5 \times \text{IQR}$

On the diagram on the left, the IQR is $78 - 60 = 18$. The data value 31 is an outlier because it is less than $60 - 1.5 \times 18 = 33$.

Outliers are sometimes removed, or **cleaned**, from a data set. There is more about this on page 67.

Worked example

Here is a table showing information about the test scores for all the students in a school.

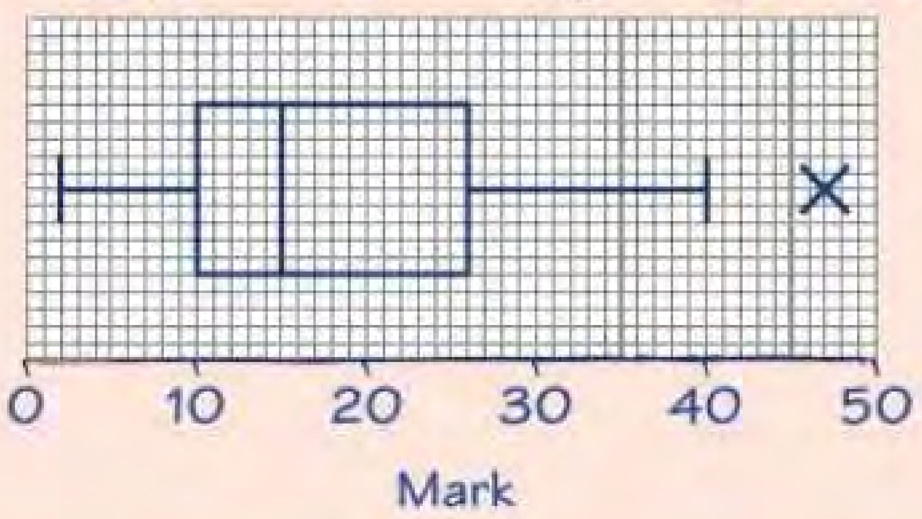
Two lowest values	2, 5
Lower quartile	10
Median	15
Upper quartile	26
Two highest values	40, 47

An outlier is a value that is greater than Q_3 plus 1.0 times the interquartile range or less than Q_1 minus 1.0 times the interquartile range.

Draw a box plot to represent these data, indicating clearly any outliers. **(5 marks)**

$\text{IQR} = 26 - 10 = 16$

$Q_1 - 1.0 \times \text{IQR} = -6$ $\quad Q_3 + 1.0 \times \text{IQR} = 42$



Read the question carefully – any rule used to specify outliers will **always** be given in the question. You need to work out the lower and upper boundaries for the outliers before deciding which data values (if any) are outliers. 47 is greater than $Q_3 + 1.0 \times \text{IQR}$ so using this definition it is an outlier – you need to mark it on your box plot with a cross.

Whiskers and outliers

When there is an outlier, you can use either of these as your 'highest' or 'lowest' value:

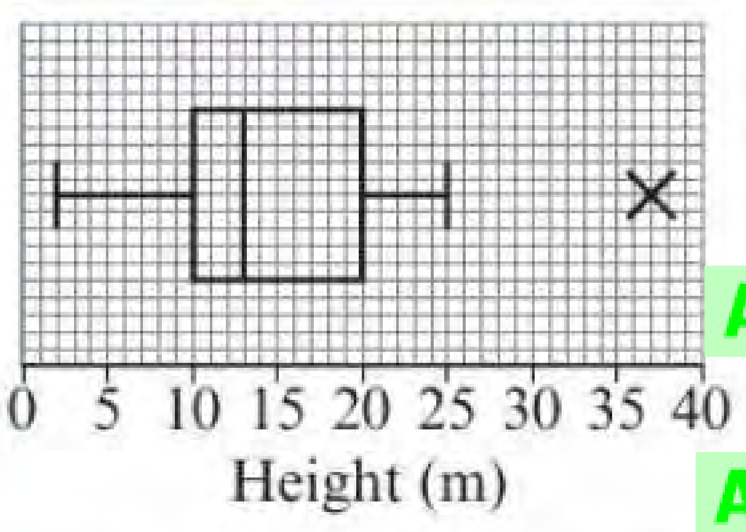
1. the highest or lowest data value which is **NOT** an outlier

This is shown in the Worked example on the left.

2. the boundary of the normal data values that are not outliers.

In the Worked example, you could also draw the right-hand whisker up to 42.

Now try this



The box plot shows a summary of the heights, in metres, of the trees in a park.

(a) (i) Write down the height that 75% of the trees are shorter than.

Answer

(ii) State the name given to this value.

Answer

(2 marks)

In (b), write down the name, what it means **and** one way of determining it.

(b) Explain what you understand by the × on the box plot.

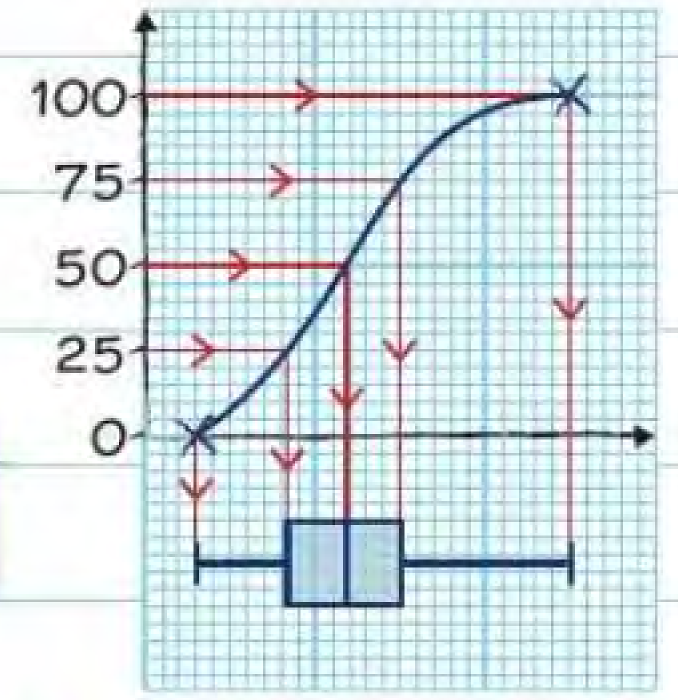
Answer **(2 marks)**

Cumulative frequency diagrams

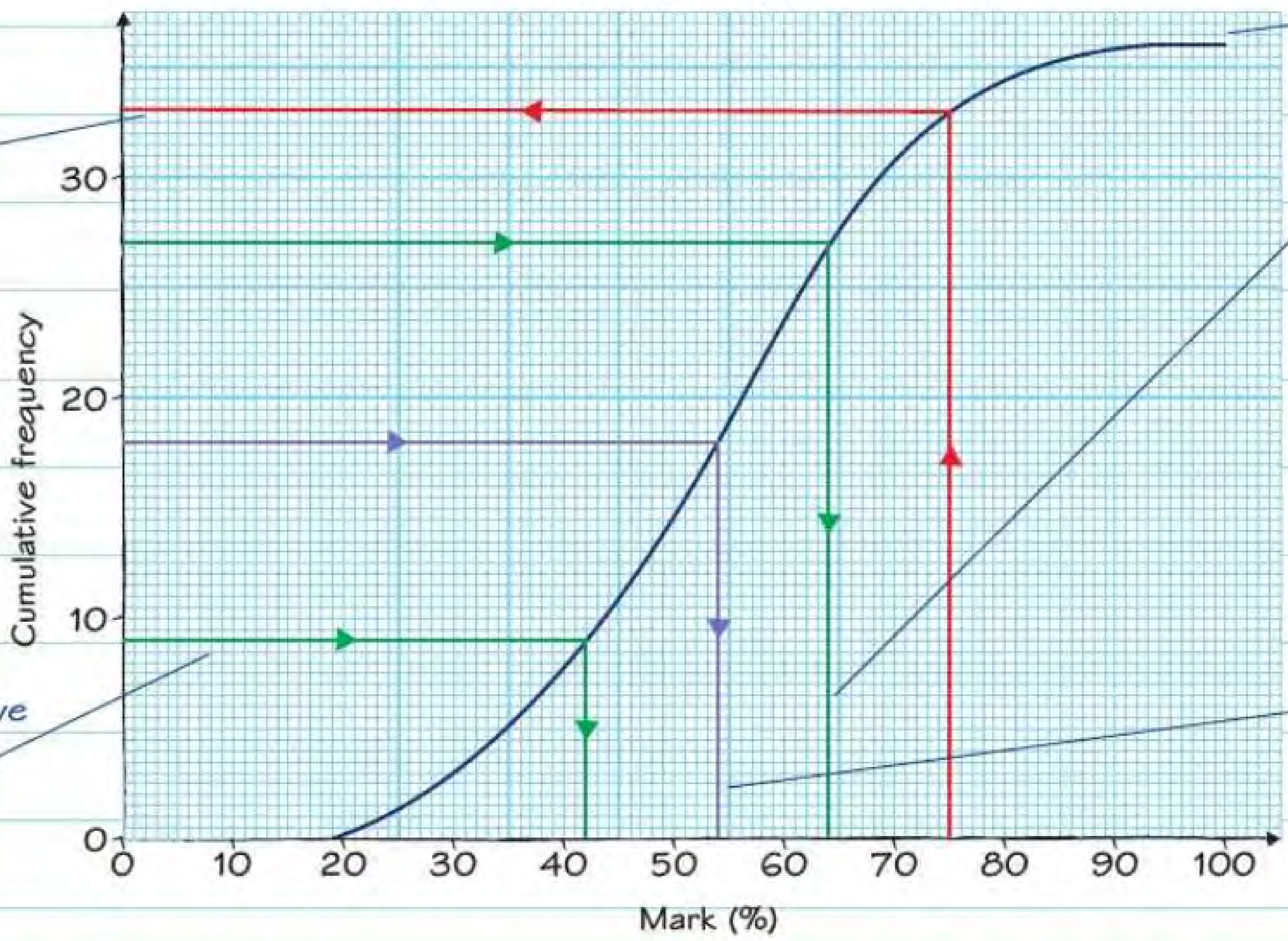
You can use a cumulative frequency diagram to find the median, quartiles and percentiles of grouped continuous data. Cumulative frequency diagrams are also a good way to **compare** continuous frequency distributions.

There is more about comparing distributions on page 66.

You can use a cumulative frequency diagram to draw a box plot.



Cumulative frequency diagram of test results



Now try this

This frequency table summarises the daily mean wind speeds in Perth between May and October 1987.

Wind speed, w (kn)	Frequency
$w < 2$	12
$2 \leqslant w < 4$	31
$4 \leqslant w < 6$	49
$6 \leqslant w < 8$	39
$8 \leqslant w < 10$	29
$10 \leqslant w < 12$	17
$w \geqslant 12$	7

Given that the minimum wind speed recorded was 0.2 kn and the maximum was 21.8 kn,

(a) draw a cumulative frequency diagram to represent these data. Answer **(3 marks)**

(b) Use your cumulative frequency diagram to estimate the 10th to 90th percentile range. Answer **(3 marks)**

To find the 10th percentile, read across from $\frac{n}{10}$ on the vertical axis to the curve, then down to the horizontal axis.

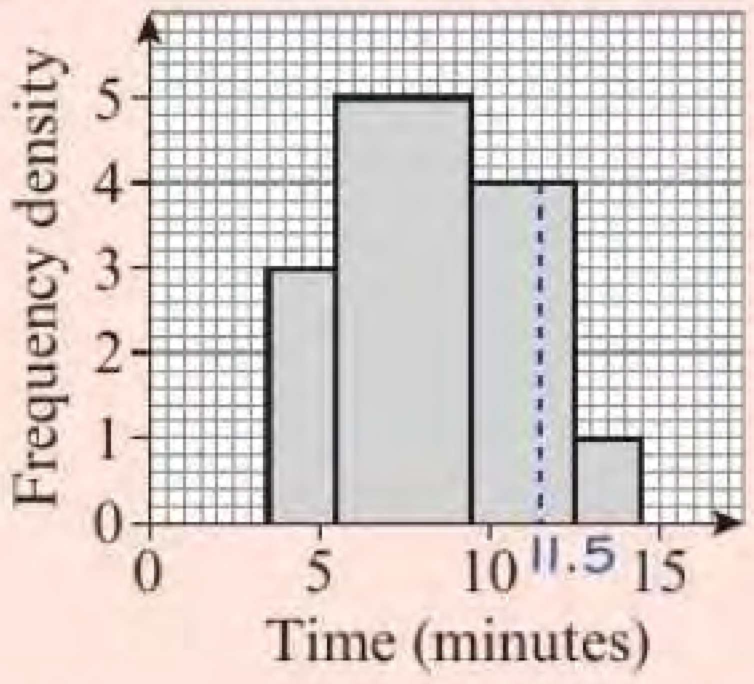
Had a look ☐ Nearly there ☐ Nailed it! ☐

Histograms

Histograms are usually used to represent **grouped continuous data**. You probably won't be asked to draw a whole histogram in your exam, but you might have to **interpret** one, or make calculations about **widths** and **heights** of bars.

Worked example

The histogram shows the finishing times, to the nearest minute, of 40 runners in a 1500 m race.



(a) Complete the table. **(2 marks)**

Finishing time (min)	Frequency
4–5	6
6–9	20
10–12	12
13–14	2

(b) Estimate the number of runners who finished the race in between 5.5 and 11.5 minutes. **(2 marks)**

$20 + \frac{2}{3} \times 12 = 28$

Your answer is only an estimate because you don't know how the data are distributed within each class.

Histogram facts

- No gaps between the bars.
- **area** of each bar is proportional to frequency.
- Vertical axis is labelled 'Frequency density'.
- Bars can be different widths.
- Frequency density = $\frac{\text{frequency}}{\text{class width}}$
- Bars are drawn between **class boundaries**.

Have a look at page 59 for a reminder about upper and lower class boundaries.

The 6–9 bar is plotted from 5.5 to 9.5, so it has width 4. Its frequency density is 5.

Frequency = frequency density × class width

= 5 × 4 = 20

You can work out the frequency in the 10–12 class using the total number of runners (40 − 20 − 6 − 2 = 12), or using the formula for frequency density.

Problem solved!

Be careful – 11.5 isn't one of the class boundaries. Frequency is proportional to **area** in a histogram. $\frac{2}{3}$ of the 10–12 bar is below 11.5, so you should include $\frac{2}{3}$ of these runners.

You will need to use problem-solving skills throughout your exam – **be prepared!**

Now try this

Alison weighed 50 apples, to the nearest gram. This table shows her results.

Weight (g)	55–57	58–63	64–68
Frequency	12	30	8

A histogram was drawn and the bar representing the 58–63 class was 4 cm wide and 6 cm high. For the 55–57 class, find

(a) the width Answer

(b) the height Answer

of the bar representing this class. **(3 marks)**

(a) When calculating the width, remember that the data has been rounded, so the class boundaries are half a unit above or below the values given in the table. The bar for the 58–63 class would be drawn from 57.5 to 63.5.

(b) Use the fact that the **area** of each bar is **proportional** to the frequency.

Comparing distributions

If you need to compare two distributions, you can use **two** things.

1. Measures of **location** like the mean, median, mode and quartiles
Revise these on pages 57–59.

2. Measures of **spread**, like the range, interquartile range, variance, or standard deviation
Revise these on pages 58, 60 and 61.

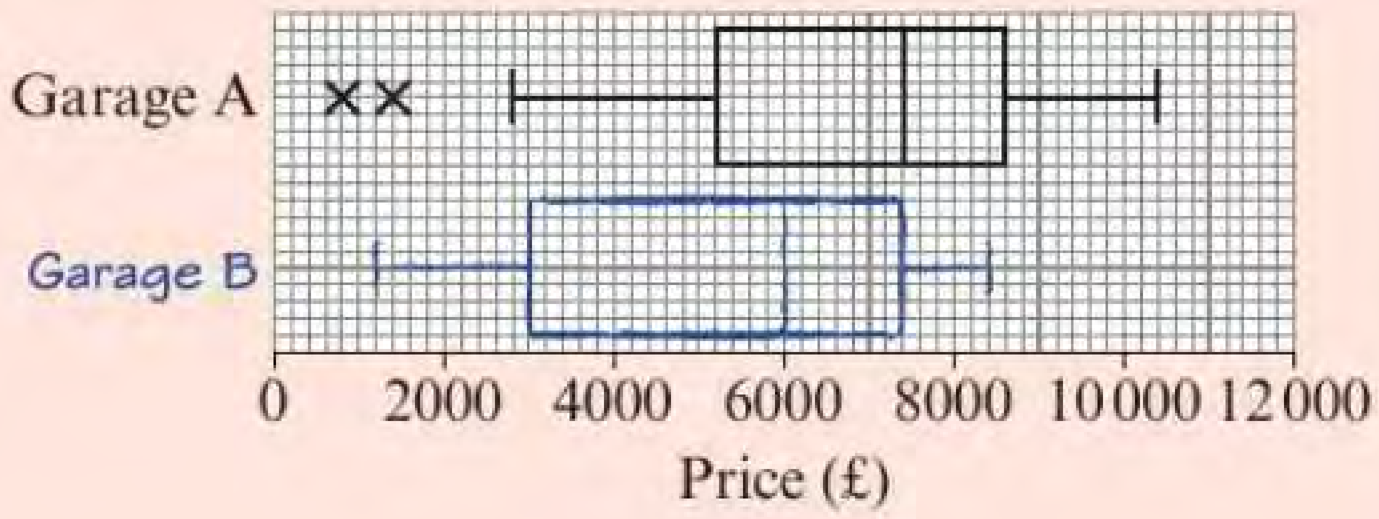
Worked example

This box plot shows the distribution of the prices of cars at Garage A.

At Garage B, the cheapest car was £1200 and the most expensive was £8400. The three quartiles were £3000, £6000 and £7400 respectively.

(a) On the same axes, draw a box plot to represent the data from Garage B. **(3 marks)**

(b) Compare and contrast the two box plots. **(3 marks)**



Median for Garage A (£7400) > Median for Garage B (£6000)

IQR for Garage B (£4400) > IQR for Garage A (£3400)

At Garage A 50% of the cars cost less than £7400, but at Garage B 75% of the cars cost less than £7400.

Try to make three different observations for part (b).
1. Compare one measure of **location**, such as the median.
2. Compare a measure of **spread**, such as the interquartile range.
3. Describe the distributions in the **context** given in the question.

Now try this

The daily mean air temperature in °C is measured in Jacksonville in May 1987 and May 2015 and recorded in the large data set. The data are summarised as follows:

	Minimum	Q_1	Q_2	Q_3	Maximum	$\sum x$	$\sum x^2$
1987	19.5	22.2	23.2	23.9	25.8	710.9	16 364.77
2015	17.5	22.7	24.5	25.0	26.5	727.9	17 255.61

(a) Calculate the mean and standard deviation of the data for each of the two years. Answer **(4 marks)**

(b) Compare the distributions for the two years. How far does the data support the conclusion that average air temperatures in Jacksonville have increased between 1987 and 2015? Answer **(4 marks)**

The mean and standard deviation use all of the data values, whereas the median and IQR ignore extreme data values. If you are comparing data you should use either the mean together with the standard deviation, or the median together with the IQR. You could also consider the sample size when commenting on the conclusion.

Removing obviously incorrect data values is sometimes called **cleaning data**. You might want to **keep** an outlier if:

- it appears to be genuine
- the data is from a reliable source, or has already been cleaned.

You might want to **remove** an outlier if:

- it is obviously a mistake
- you want to intentionally omit extreme data values