Chapter 12: Averages and measures of spread

Key words

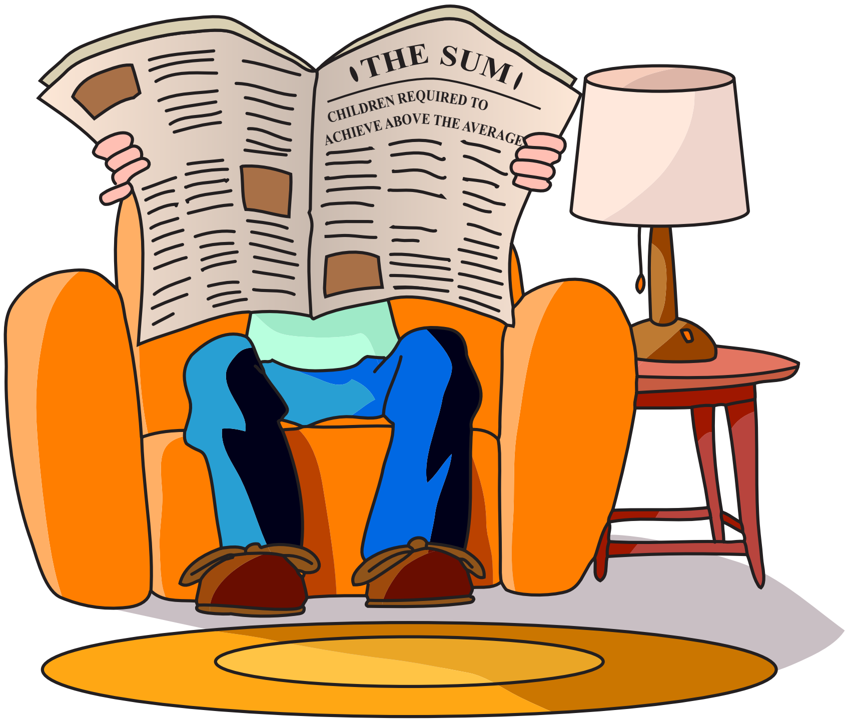
- Average
- Mode
- Mean
- Spread
- Median
- Range
- Discrete
- Continuous
- Grouped data
- Estimated mean
- Modal class
- Percentiles
- Upper quartile
- Lower quartile
- Interquartile range
- Box-and-whisker plot

In this chapter you will learn how to:

- calculate the mean, median and mode of sets of data
- calculate and interpret the range as a measure of spread
- interpret the meaning of each result and compare data using these measures

EXTENDED

- construct and use frequency distribution tables for grouped data
- identify the class that contains the median of grouped data
- calculate and work with quartiles
- divide data into quartiles and calculate the interquartile range
- identify the modal class from a grouped frequency distribution.
- Construct and interpret box-and-whisker plots.



The newspaper headline is just one example of a situation in which statistics have been badly misunderstood. It is important to make sure that you fully understand the statistics before you use it to make any kind of statement!

When you are asked to interpret data and draw conclusions you need to think carefully and to look at more than one element of the data. For example, if a student has a mean mark of 70% overall, you could conclude that the student is doing well. However if the student is getting 90% for three subjects and 40% for the other two, then that conclusion is not sound. Similarly, if the number of bullying incidents in a school goes down after a talk about bullying, it could mean the talk was effective, but it could also mean that the reporting of incidents went down (perhaps because the bullies threatened worse bullying if they were reported).

It is important to remember the following:

Correlation is not the same as causation. For example, if a company's social media account suddenly gets lots of followers and at the same time their sales in a mall increase, they may (mistakenly) think the one event caused the other.

Sometimes we only use the data that confirms our own biases (this is called confirmation bias). For example, if you were asked whether a marketing campaign to get more followers was successful and the data showed that more people followed you on Instagram, but there was no increase in your Facebook following, you could use the one data set to argue that the campaign was successful, especially if you believed it was.

Sometimes you need to summarise data to make sense of it. You do not always need to draw a diagram; instead you can calculate numerical summaries of average and spread. Numerical summaries can be very useful for comparing different sets of data but, as with all statistics, you must be careful when interpreting the results.

RECAP

You should already be familiar with the following data handling work:

Averages (Year 9 Mathematics)

Mode – value that appears most often

Median – middle value when data is arranged in ascending order

Mean – $\frac{\text{sum of values}}{\text{number of values}}$

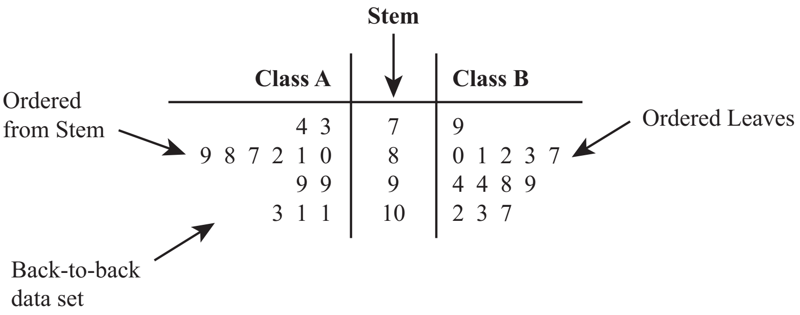
For the data set: 3, 4, 5, 6, 6, 10, 11, 12, 12, 12, 18

Mode = 12

Median = 10

Mean = $\frac{3+4+5+6+6+10+11+12+12+12+18}{11} = \frac{99}{11} = 9$

Stem and leaf diagram (Chapter 4)



Class A	Stem	Class B
4 3	7	9
9 8 7 2 1 0	8	0 1 2 3 7
9 9	9	4 4 8 9
3 1 1	10	2 3 7

	Class A	Class B
Key	3 | 7 = 73	7 | 9 = 79

12.1 Different types of average

Tip

There is more than one ‘average’, so you should never refer to *the* average. Always specify *which* average you are talking about: the *mean, median* or *mode.*

If you take the mean of *n* items and multiply it by *n*, you get the total of all *n* values.

LINK

Geographers use averages to summarise numerical results for large populations. This saves them from having to show every piece of numerical data that has been collected!

An **average** is a single value used to represent a set of data. There are three types of average used in statistics and the following shows how each can be calculated.

The shoe sizes of 19 students in a class are shown below:

4 7 6 6 7 4 8 3 8 11 6 8 6 3 5 6 7 6 4

How would you describe the shoe sizes in this class?

If you count how many size fours, how many size fives and so on, you will find that the most common (most frequent) shoe size in the class is six. This average is called the **mode**.

What most people think of as the average is the value you get when you add up all the shoe sizes and divide your answer by the number of students:

$$\frac{\text{total of shoe sizes}}{\text{number of students}} = \frac{115}{19} = 6.05\,(2\,\text{d.p.})$$

This average is called the **mean**. The mean value tells you that the shoe sizes appear to be **spread** in some way around the value 6.05. It also gives you a good impression of the general ‘size’ of the data. Notice that the value of the mean, in this case, is not a possible shoe size.

The mean is sometimes referred to as the measure of ‘central tendency’ of the data. Another measure of central tendency is the middle value when the shoe sizes are arranged in ascending order.

3 3 4 4 4 5 6 6 6 6 6 6 7 7 7 8 8 8 11

If you now think of the first and last values as one pair, the second and second to last as another pair, and so on, you can cross these numbers off and you will be left with a single value in the middle.

$$\cancel{3}\ \ \cancel{3}\ \ \cancel{4}\ \ \cancel{4}\ \ \cancel{4}\ \ \cancel{5}\ \ \cancel{6}\ \ \cancel{6}\ \ \cancel{6}\ \ \boxed{6}\ \ \cancel{6}\ \ \cancel{6}\ \ \cancel{7}\ \ \cancel{7}\ \ \cancel{7}\ \ \cancel{8}\ \ \cancel{8}\ \ \cancel{8}\ \ \cancel{11}$$

This middle value, (in this case six), is known as the **median**.

Crossing off the numbers from each end can be cumbersome if you have a lot of data. You may have noticed that, counting from the left, the median is the 10th value. Adding one to the number of students and dividing the result by two, $\frac{(19+1)}{2}$, also gives 10 as the median position.

What if there had been 20 students in the class? For example, add an extra student with a shoe size of 11. Crossing off pairs gives this result:

$$\cancel{3}\ \ \cancel{3}\ \ \cancel{4}\ \ \cancel{4}\ \ \cancel{4}\ \ \cancel{5}\ \ \cancel{6}\ \ \cancel{6}\ \ \cancel{6}\ \ \boxed{6\ \ 6}\ \ \cancel{6}\ \ \cancel{7}\ \ \cancel{7}\ \ \cancel{7}\ \ \cancel{8}\ \ \cancel{8}\ \ \cancel{8}\ \ \cancel{11}\ \ \cancel{11}$$

You are left with a middle pair rather than a single value. If this happens then you simply find the mean of this middle pair: $\frac{(6+6)}{2}=6$.

Notice that the position of the first value in this middle pair is $\frac{20}{2}=10$.

Adding an extra size 11 has not changed the median or mode in this example, but what will have happened to the mean?

In summary:

Mode The value that appears in your list more than any other. There can be more than one mode but if there are no values that occur more often than any other then there is no mode.

Mean $\frac{\text{total of all data}}{\text{number of values}}$. The mean may not be one of the actual data values.

Median

1. Arrange the data into ascending numerical order.
2. If the number of data is n and n is odd, find $\frac{n+1}{2}$ and this will give you the position of the median.
3. If n is even, then calculate $\frac{n}{2}$ and this will give you the position of the first of the middle pair. Find the mean of this pair.

Dealing with extreme values

Sometimes you may find that your collection of data contains values that are extreme in some way. For example, if you were to measure the speeds of cars as they pass a certain point you may find that some cars are moving unusually slowly or unusually quickly. It is also possible that you may have made a mistake and measured a speed incorrectly, or just written the wrong numbers down!

Suppose the following are speeds of cars passing a particular house over a five minute period (measured in kilometres per hour):

67.2 58.3 128.9 65.0 49.0 55.7

One particular value will catch your eye immediately. 128.9 km/h seems somewhat faster than any other car. How does this extreme value affect your averages?

You can check yourself that the mean of the above data *including* the extreme value is 70.7 km/h.

This is *larger* than all but one of the values and is not representative. Under these circumstances the mean can be a poor choice of average. If you discover that the highest speed was a mistake, you can exclude it from the calculation and get the much more realistic value of 59.0 km/h (try the calculation for yourself).

Tip

You could be asked to give reasons for choosing the mean or median as your average.

If the extreme value is genuine and cannot be excluded, then the median will give you a better impression of the main body of data. Writing the data in rank order:

49.0 55.7 58.3 65.0 67.2 128.9

The median is the mean of 58.3 and 65.0, which is 61.7. Notice that the median reduces to 58.3 if you remove the highest value, so this doesn't change things a great deal.

As there is an even number of speeds', the median is the mean of the 3rd and 4th data points.

There is no mode for these data.

Worked example 1

After six tests, Graham has a mean average score of 48. He takes a seventh test and scores 83 for that test.

a What is Graham's total score after six tests?
b What is Graham's mean average score after seven tests?

a Since $\text{mean} = \dfrac{\text{total of all data}}{\text{number of values}}$

then,
$$\begin{aligned}\text{total of all data} &= \text{mean} \times \text{number of values}\\ &= 48 \times 6\\ &= 288\end{aligned}$$

b
$$\begin{aligned}\text{Total of all seven scores} &= \text{total of first six plus seventh}\\ &= 288 + 83\\ &= 371\end{aligned}$$

$$\text{mean} = \frac{371}{7} = 53$$

Tip

This is good example of where you need to think before you conclude that Graham is an average student (scoring 53%). He may have had extra tuition and will get above 80% for all future tests.

Exercise 12.1

1 For each of the following data sets calculate:

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

a	12	2	5	6	9	3	12	13	10	
b	5	9	7	3	8	2	5	8	8	2
c	2.1	3.8	2.4	7.6	8.2	3.4	5.6	8.2	4.5	2.1
d	12	2	5	6	9	3	12	13	43	

2 Look carefully at the lists of values in parts (a) and (d) above. What is different? How did your mean, median and mode change?

3 Andrew and Barbara decide to investigate their television watching patterns and record the number of minutes that they watch the television for 8 days:

Andrew:	38	10	65	43	125	225	128	40
Barbara:	25	15	10	65	90	300	254	32

a Find the median number of minutes spent watching television for each of Andrew and Barbara.
b Find the mean number of minutes spent watching television for each of Andrew and Barbara.

4 Find a list of five numbers with a mean that is larger than all but one of the numbers.

Tip

Think about confirmation bias and how the player might ignore data that doesn't support his claim to be a good player.

5 A keen ten pin bowler plays five rounds in one evening. His scores are 98, 64, 103, 108 and 109. Which average (mode, mean or median) will he choose to report to his friends at the end of the evening? Explain your answer carefully, showing all your calculations clearly.

6 If the mean height of 31 children is 143.6 cm, calculate the sum of the heights used to calculate the mean.

7 The mean mass of 12 bags of potatoes is 2.4 kg. If a 13th bag containing 2.2 kg of potatoes is included, what would the new mean mass of the 13 bags be?

8 The mean temperature of 10 cups of coffee is 89.6 °C. The mean temperature of a different collection of 20 cups of coffee is 92.1 °C. What is the mean temperature of all 30 cups of coffee?

9 Find a set of five numbers with mean five, median four and mode four.

10 Find a set of five *different* whole numbers with mean five and median four.

11 The mean mass of a group of m boys is X kg and the mean mass of a group of n girls is Y kg. Find the mean mass of all of the children combined.

12.2 Making comparisons using averages and ranges

Having found a value to represent your data (an average) you can now compare two or more sets of data. However, just comparing the averages can sometimes be misleading.

It can be helpful to know how *consistent* the data is and you do this by thinking about how spread out the values are. A simple measure of spread is the **range**.

Range = largest value − smallest value

The larger the range, the more spread out the data is and the less consistent the values are with one another.

Worked example 2

Two groups of athletes want to compare their 100 m sprint times. Each person runs once and records his or her time as shown (in seconds).

Team Pythagoras	14.3	16.6	14.3	17.9	14.1	15.7
Team Socrates	13.2	16.8	14.7	14.7	13.6	16.2

a Calculate the mean 100 m time for each team.
b Which is the smaller mean?
c What does this tell you about the 100 m times for Team Pythagoras in comparison with those for Team Socrates?
d Calculate the range for each team.
e What does this tell you about the performance of each team?

a Team Pythagoras:

$$\text{Mean} = \frac{14.3 + 16.6 + 14.3 + 17.9 + 14.1 + 15.7}{6} = \frac{92.9}{6} = 15.48 \text{ seconds}$$

Team Socrates:

$$\text{Mean} = \frac{13.2 + 16.8 + 14.7 + 14.7 + 13.6 + 16.2}{6} = \frac{89.2}{6} = 14.87 \text{ seconds}$$

b Team Socrates have the smaller mean 100 m time.

c The smaller time means that Team Socrates are slightly faster as a team than Team Pythagoras.

d Team Pythagoras' range = 17.9 − 14.1 = 3.8 seconds
Team Socrates' range = 16.8 − 13.2 = 3.6 seconds

e Team Socrates are slightly faster as a whole and they are slightly more consistent. This suggests that their team performance is not improved significantly by one or two fast individuals but rather all team members run at more or less similar speeds. Team Pythagoras is less consistent and so their mean is improved by individuals.

Tip

When comparing means or ranges, make sure that you refer to the original context of the question.

Exercise 12.2

1 Two friends, Ricky and Oliver, are picking berries. Each time they fill a carton its mass, in kg, is recorded. The masses are shown below:

Ricky	0.145	0.182	0.135	0.132	0.112	0.155	0.189	0.132	0.145	0.201	0.139
Oliver	0.131	0.143	0.134	0.145	0.132	0.123	0.182	0.134	0.128		

a For each boy calculate:
i the mean mass of berries collected per box **ii** the range of masses.

b Which boy collected more berries per load?

c Which boy was more consistent when collecting the berries?

2 The marks obtained by two classes in a Mathematics test are show below. The marks are out of 20.

Class Archimedes	12	13	4	19	20	12	13	13	16	18	12
Class Bernoulli	13	6	9	15	20	20	13	15	17	19	3

a Calculate the median score for each class.

b Find the range of scores for each class.

c Which class did better in the test overall?

d Which class was more consistent?

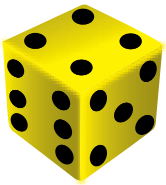
3 Three shops sell light bulbs. A sample of 100 light bulbs is taken from each shop and the working life of each is measured in hours. The following table shows the mean time and range for each shop:

Shop	Mean (hours)	Range (hours)
Brightlights	136	18
Footlights	145	36
Backlights	143	18

Which shop would you recommend to someone who is looking to buy a new light bulb and why?

12.3 Calculating averages and ranges for frequency data

So far, the lists of data that you have calculated averages for have been quite small. Once you start to get more than 20 pieces of data it is better to collect the data with the same value together and record it in a table. Such a table is known as a *frequency distribution table* or just a *frequency distribution*.



Data shown in a frequency distribution table

If you throw a single die 100 times, each of the six numbers will appear several times. You can record the number of times that each appears like this:

Number showing on the upper face	1	2	3	4	5	6
Frequency	16	13	14	17	19	21

Mean

You need to find the total of all 100 throws. Sixteen 1s appeared giving a sub-total of $1 \times 16 = 16$, thirteen 2s appeared giving a sub-total of $13 \times 2 = 26$ and so on. You can extend your table to show this:

Tip

You can add columns to a table given to you! It will help you to organise your calculations clearly.

Number showing on the upper face	Frequency	Frequency × number on the upper face
1	16	$1 \times 16 = 16$
2	13	$2 \times 13 = 26$
3	14	$3 \times 14 = 42$
4	17	$17 \times 4 = 68$
5	19	$19 \times 5 = 95$
6	21	$21 \times 6 = 126$

REWIND

Sometimes you will need to retrieve the data from a diagram like a bar chart or a pictogram and then calculate a mean. These charts were studied in chapter 4. ◀

The total of all 100 die throws is the sum of all values in this third column:

$$= 16 + 26 + 42 + 68 + 95 + 126$$
$$= 373$$

So the mean score per throw $= \dfrac{\text{total score}}{\text{total number of throws}} = \dfrac{373}{100} = 3.73$

Median

There are 100 throws, which is an even number, so the median will be the mean of a middle pair. The first of this middle pair will be found in position $\dfrac{100}{2} = 50$

The table has placed all the values in order. The first 16 are ones, the next 13 are twos and so on. Notice that adding the first three frequencies gives $16 + 13 + 14 = 43$. This means that the first 43 values are 1, 2 or 3. The next 17 values are all 4s, so the 50th and 51st values will both be 4. The mean of both 4s is 4, so this is the median.

Mode

For the mode you simply need to find the die value that has the highest frequency. The number 6 occurs most often (21 times), so 6 is the mode.

Range

The highest and lowest values are known, so the range is $6 - 1 = 5$

Data organised into a stem and leaf diagram

You can determine averages and the range from stem and leaf diagrams.

Mean

As a stem and leaf diagram shows all the data values, the mean is found by adding all the values and dividing them by the number of values in the same way you would find the mean of any data set.

Median

You can use an ordered stem and leaf diagram to determine the median. An ordered stem and leaf diagram has the leaves for each stem arranged in order from smallest to greatest.

Mode

An ordered stem and leaf diagram allows you see which values are repeated in each row. You can compare these to determine the mode.

Range

In an ordered stem and leaf diagram, the first value and the last value can be used to find the range.

Worked example 3

The ordered stem and leaf diagram shows the number of customers served at a supermarket checkout every half hour during an 8-hour shift.

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

Key
0 | 2 = 2 customers

a What is the range of customers served?
b What is the modal number of customers served?
c Determine the median number of customers served.
d How many customers were served altogether during this shift?
e Calculate the mean number of customers served every half hour.

a The lowest number is 2 and the highest number is 21. The range is 21 − 2 = 19 customers.

b 6 is the value that appears most often.

c There are 16 pieces of data, so the median is the mean of the 8th and 9th values.

$$\frac{(11+13)}{2} = \frac{24}{2} = 12$$

d To calculate this, find the sum of all the values.
Find the total for each row and then combine these to find the overall total.
Row 1: 2 + 5 + 5 + 6 + 6 + 6 + 6 = 36
Row 2: 11 + 13 + 13 + 15 + 15 + 16 + 17 + 17 = 117
Row 3: 21
36 + 117 + 21 = 174 customers in total

e

$$\text{Mean} = \frac{\text{sum of data values}}{\text{number of data values}}$$

$$= \frac{174}{16} = 10.875 \text{ customers per half hour.}$$

In summary:

- **Mode** The value that has the highest frequency will be the mode. If more than one value has the same highest frequency then there is no single mode.
- **Mean** $\frac{\text{total of all data}}{\text{number of values}} = \frac{\text{sum of frequency} \times \text{value}}{\text{total frequency}}$

  (Remember to extend the table so that you can fill in a column for calculating frequency × value in each case.)
- **Median**
  - If the number of data is n and n is odd, find $\frac{n+1}{2}$ and this will give you the position of the median.
  - If n is even, then calculate $\frac{n}{2}$ and this will give you the position of the first of the middle pair. Find the mean of this pair.
  - Add the frequencies in turn until you find the value whose frequency makes you exceed (or equal) the value from one or two above. This is the median.

Exercise 12.3

1 Construct a frequency table for the following data and calculate:

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

3	4	5	1	2	8	9	6	5	3	2	1	6	4	7	8	1
1	5	5	2	3	4	5	7	8	3	4	2	5	1	9	4	5
6	7	8	9	2	1	5	4	3	4	5	6	1	4	4	8	

2 Tickets for a circus were sold at the following prices: 180 at \$6.50 each, 215 at \$8 each and 124 at \$10 each.

a Present this information in a frequency table.
b Calculate the mean price of tickets sold (give your answer to 3 significant figures).

3 A man kept count of the number of letters he received each day over a period of 60 days. The results are shown in the table below.

Number of letters per day	0	1	2	3	4	5
Frequency	28	21	6	3	1	1

For this distribution, find:

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

4 A survey of the number of children in 100 families gave the following distribution:

Number of children in family	0	1	2	3	4	5	6	7
Number of families	4	36	27	21	5	4	2	1

For this distribution, find:

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

5 The distribution of marks obtained by the students in a class is shown in the table below.

Mark obtained	0	1	2	3	4	5	6	7	8	9	10
Number of students	1	0	3	2	2	4	3	4	6	3	2

Find:

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

d The class teacher is asked to report on her class's performance and wants to show them to be doing as well as possible. Which average should she include in her report and why?

6 The masses of 20 soccer players were measured to the nearest kilogram and this stem and leaf diagram was produced.

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

Key
4 | 6 = 46 kilograms

a Redraw the stem and leaf diagram to make an ordered data set.

b How many players have a mass of 60 kilograms or more?

c Why is the mode not a useful statistic for this data?

d What is the range of masses?

e What is the median mass of the players?

7 The number of electronic components produced by a machine every hour over a 24-period is:

143, 128, 121, 126, 134, 150, 128, 132, 140, 131, 146, 128

133, 138, 140, 125, 142, 129, 136, 130, 133, 142, 126, 129

a Using two intervals for each stem, draw an ordered stem and leaf diagram of the data.

b Determine the range of the data.

c Find the median.

E

12.4 Calculating averages and ranges for grouped continuous data

Some data is **discrete** and can only take on certain values. For example, if you throw an ordinary die then you can only get one of the numbers 1, 2, 3, 4, 5 or 6. If you count the number of red cars in a car park then the result can only be a whole number.

Some data is **continuous** and can take on *any* value in a given range. For example, heights of people, or the temperature of a liquid, are continuous measurements.

Continuous data can be difficult to process effectively unless it is summarised. For instance, if you measure the heights of 100 children you could end up with 100 different results. You can group the data into frequency tables to make the process more manageable – this is now **grouped data**. The groups (or classes) can be written using *inequality* symbols. For example, if you want to create a class for heights (h cm) between 120 cm and 130 cm you could write:

$120 \leqslant h < 130$

This means that h is greater than or equal to 120 but strictly less than 130. The next class could be:

$130 \leqslant h < 140$

Notice that 130 is not included in the first class but is included in the second. This is to avoid any confusion over where to put values at the boundaries.

E

The following worked example shows how a grouped frequency table is used to find the **estimated mean** and range, and also to find the **modal class** and the median classes (i.e. the classes in which the mode and median lie).

Tip

You may be asked to explain why your calculations only give an estimate. Remember that you don't have the exact data, only frequencies and classes.

Worked example 4

The heights of 100 children were measured in cm and the results recorded in the table below:

Height in cm (h)	Frequency (f)
$120 \leqslant h < 130$	12
$130 \leqslant h < 140$	16
$140 \leqslant h < 150$	38
$150 \leqslant h < 160$	24
$160 \leqslant h < 170$	10

Find an estimate for the mean height of the children, the modal class, the median class and an estimate for the range.

None of the children's heights are known exactly, so you use the *midpoint* of each group as a best estimate of the height of each child in a particular class. For example, the 12 children in the $120 \leqslant h < 130$ class have heights that lie between 120 cm and 130 cm, and that is all that you know. Halfway between 120 cm and 130 cm is $\frac{(120+130)}{2} = 125\,\text{cm}$.

A good estimate of the total height of the 12 children in this class is 12 × 125 (= frequency × midpoint).

So, extend your table to include midpoints and then totals for each class:

Height in cm (h)	Frequency (f)	Midpoint	Frequency × midpoint
$120 \leqslant h < 130$	12	125	12 × 125 = 1500
$130 \leqslant h < 140$	16	135	16 × 135 = 2160
$140 \leqslant h < 150$	38	145	38 × 145 = 5510
$150 \leqslant h < 160$	24	155	24 × 155 = 3720
$160 \leqslant h < 170$	10	165	10 × 165 = 1650

An estimate for the mean height of the children is then:

$$\frac{1500+2160+5510+3720+1650}{12+16+38+24+10} = \frac{14540}{100} = 145.4\,\text{cm}$$

To find the median class you need to find where the 50th and 51st tallest children would be placed. Notice that the first two frequencies add to give 28, meaning that the 28th child in an ordered list of heights would be the tallest in the $130 \leqslant h < 140$ class. The total of the first *three* frequencies is 66, meaning that the 50th child will be somewhere in the $140 \leqslant h < 150$ class. This then, makes $140 \leqslant h < 150$ the median class.

The class with the highest frequency is the modal class. In this case it is the same class as the median class: $140 \leqslant h < 150$.

The shortest child could be as small as 120 cm and the tallest could be as tall as 170 cm. The best estimate of the range is, therefore, 170 – 120 = 50 cm.

Exercise 12.4

E

The symbol Σ is the Greek letter capital 'sigma'. It is used to mean 'sum'. So, Σf simply means, 'the sum of all the frequencies'.

1 The table shows the heights of 50 sculptures in an art gallery. Find an estimate for the mean height of the sculptures.

Heights (h cm)	Frequency (f)
$130 < h \leqslant 135$	7
$135 < h \leqslant 140$	13
$140 < h \leqslant 145$	15
$145 < h \leqslant 150$	11
$150 < h \leqslant 155$	4
Total	$\Sigma f = 50$

2 The table shows the lengths of 100 telephone calls.

Time (t minutes)	Frequency (f)
$0 < t \leqslant 1$	12
$1 < t \leqslant 2$	14
$2 < t \leqslant 4$	20
$4 < t \leqslant 6$	14
$6 < t \leqslant 8$	12
$8 < t \leqslant 10$	18
$10 < t \leqslant 15$	10

a Calculate an estimate for the mean time, in minutes, of a telephone call.
b Write your answer in minutes and seconds, to the nearest second.

3 The table shows the temperatures of several test tubes during a Chemistry experiment.

Temperature (T °C)	Frequency (f)
$45 \leqslant T < 50$	3
$50 \leqslant T < 55$	8
$55 \leqslant T < 60$	17
$60 \leqslant T < 65$	6
$65 \leqslant T < 70$	2
$70 \leqslant T < 75$	1

Calculate an estimate for the mean temperature of the test tubes.

4 Two athletics teams – the *Hawks* and the *Eagles* – are about to compete in a race. The masses of the team members are shown in the table below.

Hawks

Mass (M kg)	Frequency (f)
$55 \leqslant M < 65$	2
$65 \leqslant M < 75$	8
$75 \leqslant M < 85$	12
$85 \leqslant M < 100$	3

Eagles

Mass (M kg)	Frequency (f)
$55 \leqslant M < 65$	1
$65 \leqslant M < 75$	7
$75 \leqslant M < 85$	13
$85 \leqslant M < 100$	4

E

a Calculate an estimate for the mean mass of each team.
b Calculate the range of masses of each team.
c Comment on your answers for (a) and (b).

5 The table below shows the lengths of 50 pieces of wire used in a Physics laboratory. The lengths have been measured *to the nearest centimetre*. Find an estimate for the mean.

Length	26–30	31–35	36–40	41–45	46–50
Frequency (f)	4	10	12	18	6

Be careful when calculating the midpoints here. Someone who is just a day short of 31 will still be in the 21–30 class. What difference does this make?

6 The table below shows the ages of the teachers in a secondary school to the nearest year.

Age in years	21–30	31–35	36–40	41–45	46–50	51–65
Frequency (f)	3	6	12	15	6	7

Calculate an estimate for the mean age of the teachers.

12.5 Percentiles and quartiles

Fashkiddler's accountancy firm is advertising for new staff to join the company and has set an entrance test to examine the ability of candidates to answer questions on statistics. In a statement on the application form the company states that, *'All those candidates above the 80th percentile will be offered an interview.'* What does this mean?

FAST FORWARD

You will do further work with percentiles in chapter 20 when you learn about cumulative frequency curves. You will also find the solution to *Fashkiddler's* problem. ▶

The median is a very special example of a **percentile**. It is placed exactly half way through a list of ordered data so that 50% of the data is smaller than the median. Positions other than the median can, however, also be useful.

The tenth percentile, for example, would lie such that 10% of the data was smaller than its value. The 75th percentile would lie such that 75% of the values are smaller than its value.

Quartiles

Two very important percentiles are the **upper** and **lower quartiles**. These lie 25% and 75% of the way through the data respectively.

Use the following rules to estimate the positions of each quartile within a set of ordered data:

FAST FORWARD

You will use quartiles and the interquartile range when you plot boxplots later in this chapter. ▶

Q_1 = lower quartile = value in position $\frac{1}{4}(n+1)$
Q_2 = median (as calculated earlier in the chapter)
Q_3 = upper quartile = value in position $\frac{3}{4}(n+1)$

If the position does not turn out to be a whole number, you simply find the mean of the pair of numbers on either side. For example, if the position of the lower quartile turns out to be 5.25, then you find the mean of the 5th and 6th pair.

Interquartile range

As with the *range*, the **interquartile range** gives a measure of how spread out or consistent the data is. The main difference is that the interquartile range (IQR) avoids using extreme data by finding the difference between the lower and upper quartiles. You are, effectively, measuring the spread of the central 50% of the data.

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

If one set of data has a smaller IQR than another set, then the first set is more consistent and less spread out. This can be a useful comparison tool.

FAST FORWARD

In chapter 20 you will learn about *cumulative frequency* graphs. These enable you to calculate estimates for the median when there are too many data to put into order, or when you have grouped data. ▶

E

Worked example 5

For each of the following sets of data calculate the median, upper and lower quartiles.
In each case calculate the interquartile range.

a 13 12 8 6 11 14 8 5 1 10 16 12
b 14 10 8 19 15 14 9

a First *sort the data into ascending order*.

1 5 6 8 8 10 11 12 12 13 14 16

There is an even number of items (12). So for the median, you find the value of the middle pair, the first of which is in position $\frac{12}{2} = 6$. So the median is $\frac{(10+11)}{2} = 10.5$

There are 12 items so, for the quartiles, you calculate the positions

$$\frac{1}{4}(12+1) = 3.25 \text{ and } \frac{3}{4}(12+1) = 9.75$$

Notice that these are not whole numbers, so the lower quartile will be the mean of the 3rd and 4th values, and the upper quartile will be the mean of the 9th and 10th values.

$$Q_1 = \frac{(6+8)}{2} = 7 \text{ and } Q_3 = \frac{(12+13)}{2} = 12.5$$

Thus, the $\text{IQR} = 12.5 - 7 = 5.5$

b The ordered data is:

8 9 10 14 14 15 19

The number of data is odd, so the median will be in position $\frac{(7+1)}{2} = 4$. The median is 14.

There are seven items, so calculate $\frac{1}{4}(7+1) = 2$ and $\frac{3}{4}(7+1) = 6$

These are whole numbers so the lower quartile is in position two and the upper quartile is in position six.

So $Q_1 = 9$ and $Q_3 = 15$.

$\text{IQR} = 15 - 9 = 6$

Worked example 6

Two companies sell sunflower seeds. Over the period of a year, seeds from Allbright produce flowers with a median height of 98 cm and IQR of 13 cm. In the same year seeds from Barstows produce flowers with a median height of 95 cm and IQR of 4 cm. Which seeds would you buy if you wanted to enter a competition for growing the tallest sunflower and why?

I would buy Barstows' seeds. Although Allbright sunflowers seem taller (with a higher median) they are less consistent. So, whilst there is a chance of a very big sunflower there is also a good chance of a small sunflower. Barstows' sunflowers are a bit shorter, but are more consistent in their heights so you are more likely to get flowers around the height of 95 cm.

Tip

Remember to count the data in ascending order when you work with the left hand side. The lowest values are closest to the stem in each row.

Worked example 7

The back-to-back stem and leaf diagram shows the concentration of low density lipoprotein (bad) cholesterol in the blood (milligrams per 100 ml of blood (mg/dl)) in 70 adults, half of whom are smokers and half of whom are non-smokers.

Non-smokers		Smokers
Leaf	Stem	Leaf
8 0	9	
8 8 3 1	10	
9 9 8 8 6 5 2	11	2
9 9 8 7 7 0 0	12	0 1
9 6 5 1 1 1	13	0 2 3
4 2 2	14	1 3 5 6
8 2 1	15	0 4 5 5 9
6 5	16	0 1 4 7 8 9
3	17	2 3 6 8 8
	18	0 2 4 5
	19	1 6 8
	20	1
	21	5

Key
Non-smokers 0 | 9 = 90 Smokers 11 | 2 = 112

a Determine the median for each group.

b Find the range for:

i Non-smokers **ii** Smokers

c Determine the interquartile range for:

i Non-smokers **ii** Smokers

d LDL levels of <130 are desirable, levels of 130 – 160 are considered borderline high and levels >160 are considered high risk (more so for people with medical conditions that increase risk. Using these figures, comment on what the distribution on the stem and leaf diagram suggests.

a The data is already ordered and there are 35 values in each set. $\frac{1}{2}(35 + 1)^{\text{th}} = 18$, so median is the 18th value.
Non smokers median = 128
Smokers median = 164

b **i** Range = 173 – 90 = 83 **ii** 215 – 112 = 103

c Determine the position of Q1 and Q3.

The lower quartile $= \frac{1}{4}(35 + 1)^{\text{th}} = 9^{\text{th}}$ value

The upper quartile $= \frac{3}{4}(35 + 1)^{\text{th}} = 27^{\text{th}}$ value

i IQR = Q3 – Q1 = 142 – 116 = 26 for non-smokers

ii IQR = Q3 – Q1 = 180 – 145 = 35 for smokers

d For non-smokers the data is skewed toward the lower levels on the stem and leaf diagram. More than half the values are in the desirable range, with only three in the high risk range. For smokers, the data is further spread out. Only 3 values are in the desirable range, 12 are borderline high and 20 are in the high risk category, suggesting that smokers have higher levels of bad cholesterol in general. However, without considering other risk factors or medical history, you cannot say this for certain from one set of data.

Exercise 12.5

1 Find the median, quartiles and interquartile range for each of the following. Make sure that you show your method clearly.

a	5	8	9	9	4	5	6	9	3	6	4				
b	12	14	12	17	19	21	23								
c	4	5	12	14	15	17	14	3	18	19	18	19	14	4	15
d	3.1	2.4	5.1	2.3	2.5	4.2	3.4	6.1	4.8						
e	13.2	14.8	19.6	14.5	16.7	18.9	14.5	13.7	17.0	21.8	12.0	16.5			

Applying your skills

Try to think about what the calculations in each question tell you about each situation.

2 Gideon walks to work when it is not raining. Each week for 15 weeks Gideon records the number of walks that he takes and the results are shown below:

5	7	5	8	4
2	9	9	4	7
6	4	6	12	4

Find the median, quartiles and interquartile range for this data.

3 Paavan is conducting a survey into the traffic on his road. Every Monday for eight weeks in the summer Paavan records the number of cars that pass by his house between 08.00 a.m. and 09.00 a.m. He then repeats the experiment during the winter. Both sets of results are shown below:

Summer:	18	15	19	25	19	26	17	13
Winter:	12	9	14	11	13	9	12	10

a Find the median number of cars for each period.
b Find the interquartile range for each period.
c What differences do you notice? Try to explain why this might happen.

Tip

Think carefully about possible restrictions before you answer part (c).

4 Julian and Aneesh are reading articles from different magazines. They count the number of words in a random selection of sentences from their articles and the results are recorded below:

Julian (reading the *Statistician*):	23	31	12	19	23	13	24		
Aneesh (reading the *Algebraist*):	19	12	13	16	18	15	18	21	22

a Calculate the median for each article.
b Calculate the interquartile range for each article.
c Aneesh claims that the editor of the *Algebraist* has tried to control the writing and seems to be aiming it at a particular audience. What do your answers from (a) and (b) suggest about this claim?

5 The fuel economy (km/l of petrol) of 18 new car models was tested in both city traffic and open road driving conditions and the following stem and leaf diagram was produced.

New car fuel economy (km/l)

City traffic **Leaf**	**Stem**	**Open road** **Leaf**
0	8	
4 2 1 0	9	
5 3 1 1	10	
8 3 2	11	5 5 9
7 6 4	12	1 1 2 7
1	13	3 6
5 2	14	5 6 7
	15	2 7 9
	16	0 1
	17	4

Key
0 | 8 = 8.0 km/l 11 | 5 = 11.5 km/l

Stem and leaf diagrams are useful for organising up to 50 pieces of data, beyond that they become very clumsy and time consuming. Box-and-whisker plots are far more useful for summarising large data sets.

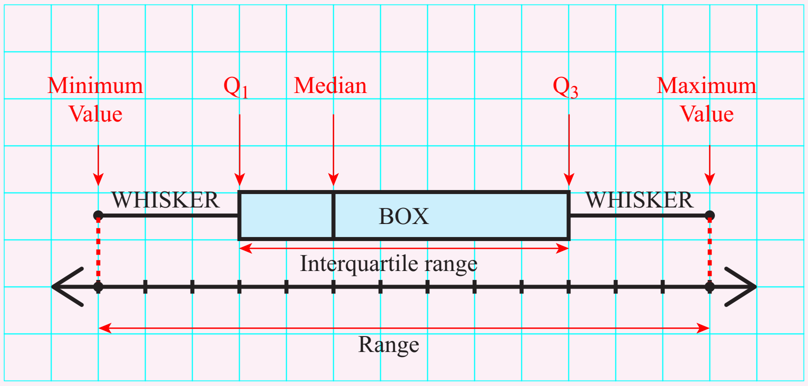
a Find the range of kilometres per litre of petrol for (i) city traffic and (ii) open road conditions.
b Find the median fuel economy for (i) city traffic and (ii) open road driving.
c Determine the interquartile range for (i) city traffic and (ii) open road driving.
d Compare and comment on the data for both city traffic and open road driving.

12.6 Box-and-whisker plots

A **box-and-whisker plot** is a diagram that shows the distribution of a set of data at a glance. They are drawn using five summary statistics: the lowest and highest values (the range), the first and third quartiles (the interquartile range) and the median.

Drawing box-and-whisker plots

All box-and-whisker plots have the same basic features. You can see these on the diagram.



Box-and-whisker plots (also called boxplots) are a standardised way of showing the range, the interquartile range and a typical value (the median). These five summary statistics are also called the 5-number summary.

Worked example 8

The masses in kilograms of 20 students were rounded to the nearest kilogram and listed in order:
48, 52, 54, 55, 55, 58, 58, 61, 62, 63, 63, 64, 65, 66, 66, 67, 69, 70, 72, 79.
Draw a box-and-whisker plot to represent this data.

The minimum and maximum values can be read from the data set.
Minimum = 48 kg
Maximum = 79 kg

Calculate the median.

There are 20 data values, so the median will lie halfway between the 10th and 11th values. In this data set they are both 63, so the median is 63 kg.

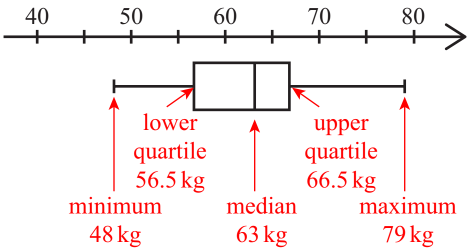
Next, calculate the lower and upper quartiles (Q_1 and Q_3). Q_1 is the mean of the 5th and 6th values and Q_3 is the mean of the 15th and 16th values.

$$Q_1 = \frac{55+58}{2} = 56.5 \text{ kg}$$

$$Q_3 = \frac{66+67}{2} = 66.5 \text{ kg}$$

To draw the box-and-whisker plot:

- Draw a scale with equal intervals that allows for the minimum and maximum values.
- Mark the position of the median and the lower and upper quartiles against the scale.
- Draw a rectangular box with Q_1 at one end and Q_3 at the other. Draw a line parallel to Q_1 and Q_3 inside the box to show the position of the median.
- Extend lines (the whiskers) from the Q_1 and Q_3 sides of the box to the lowest and highest values.



Box-and-whisker plots are very useful for comparing two or more sets of data. When you want to compare two sets of data, you plot the diagrams next to each other on the same scale.

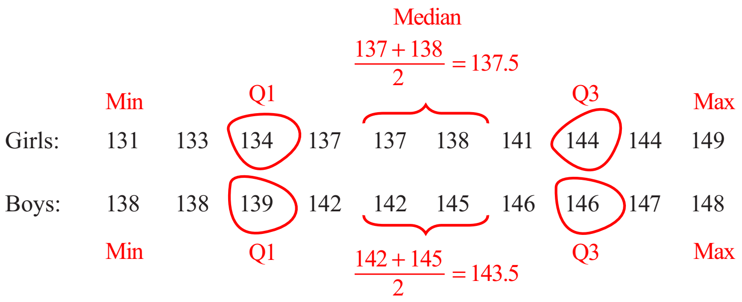
Worked example 9

The heights of ten 13-year old boys and ten 13-year old girls (to the nearest cm) are given in the table.

Girls	137	133	141	137	138	134	149	144	144	131
Boys	145	142	146	139	138	148	138	147	142	146

Draw a box-and-whisker plot for both sets of data and compare the interquartile range.

First arrange the data sets in order. Then work out the five number summary for each data set:



Median $\frac{137+138}{2} = 137.5$

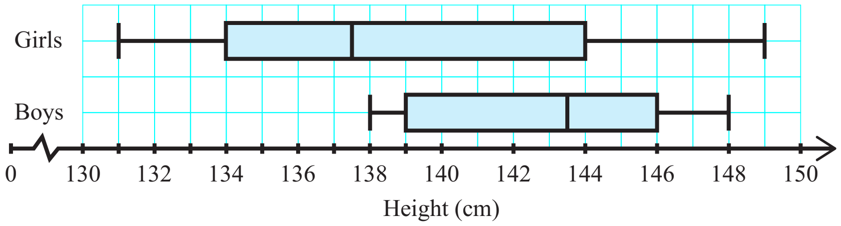
	Min		Q1					Q3		Max
Girls:	131	133	134	137	137	138	141	144	144	149
Boys:	138	138	139	142	142	145	146	146	147	148
	Min		Q1					Q3		Max

$\frac{142+145}{2} = 143.5$

E

Draw a scale that allows for the minimum and maximum values.

Plot both diagrams and label them to show which is which.

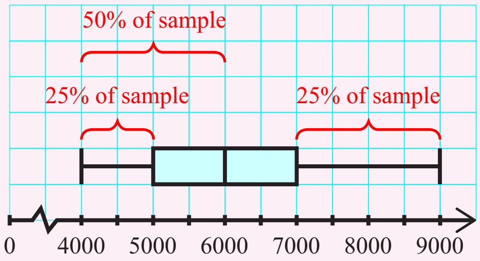


The IQR for girls (10 cm) is wider than that for boys (7 cm) showing that the data for girls is more spread out and varied.

Interpreting box-and-whisker plots

To interpret a box-and-whisker plot, you need to think about what information the diagram gives you about the dataset.

This box-and-whisker plot shows the results of a survey in which a group of teenagers wore a fitness tracker to record the number of steps they took each day.



The box-and-whisker plot shows that:

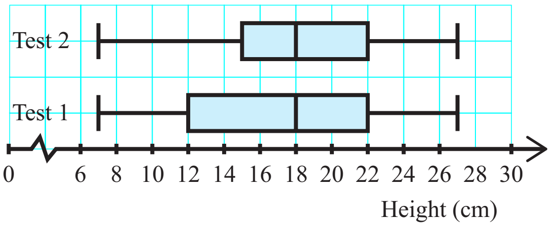
- The number of steps ranged from 4000 to 9000 per day.
- The median number of steps was 6000 steps per day.
- 50% of the teenagers took 6000 or fewer steps per day (the data below the median value)
- 25% of the teenagers took 5000 or fewer steps per day (the lower ‘whisker’ represents the lower 25% of the data)
- 25% of the teenagers took more than 7000 steps per day (the upper ‘whisker’ shows the top 25% of the data)
- The data is fairly regularly distributed because the median line is in the middle of the box (in other words, equally far from Q_1 and Q_3).

E

Worked example 10

The box-and-whisker plots below show the test results that the same group of students achieved for two tests. Test 2 was taken two weeks after Test 1.

Comment on how the students performed in the two tests.



The highest and lowest marks were the same for both tests. The marks ranged from 7 to 27, a difference of 20 marks.

Q_3 is the same for both tests. This means that 75% of the students scored $\frac{22}{30}$ or less on both tests. Only 25% of the students scored 22 or more.

For the first test, Q_1 was 12, so 75% of the students scored 12 or more marks. In the second test, Q_1 increased from 12 to 15. This means that 75% of the students scored 15 or more marks in the second test, suggesting that the group did slightly better overall in the second test.

Exercise 12.6

1 Zara weighed the contents of fifteen different bags of nuts and recorded their mass to the nearest gram.

147	150	152	150	150	148	151	146
149	151	148	146	150	145	149	

Draw a box-and-whisker plot to display the data.

2 The range and quartiles of a data set are given below. Use these figures to draw a box-and-whisker plot.

Range: 76 – 28 = 48

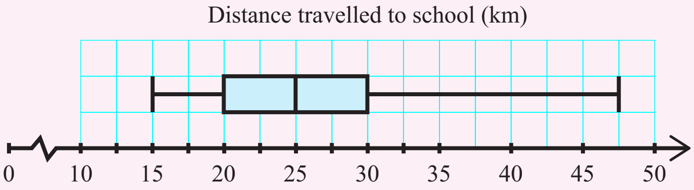
$Q_1 = 41.5$, $Q_2 = 46.5$, $Q_3 = 53.5$

3 The table shows the marks that the same group of ten students received for three consecutive assignments.

TEST 1	34	45	67	87	65	56	34	55	89	77
TEST 2	19	45	88	75	45	88	64	59	49	72
TEST 3	76	32	67	45	65	45	66	57	77	59

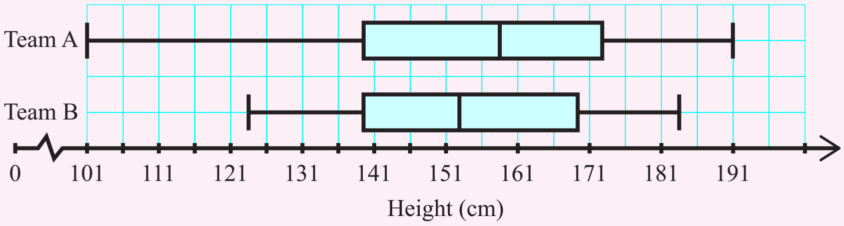
a Draw three box-and-whisker plots on the same scale to display this data.
b Use the diagrams to comment on the performance of this group of students in the three assignments.

4 The following box-and-whisker plot shows the distances in kilometres that various teachers travel to get to school each day.



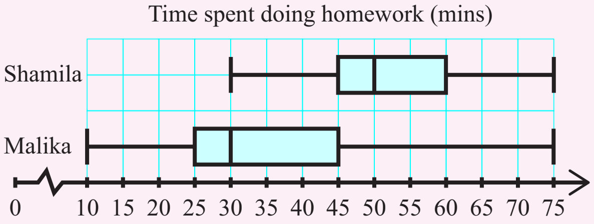
a What is the median distance travelled?
b What is the furthest that a teacher has to travel?
c What percentage of the teachers travel 30 or fewer kilometres to work?
d What percentage of the teachers travel between 15 and 25 kilometres to work?
e What is the IQR of this data set? What does it tell you?
f What does the position of the median in the box tell you about the distribution of the data?

5 Two teams of friends have recorded their scores on a game and created a pair of box-and-whisker plots.



a What is the interquartile range for Team A?
b What is the interquartile range for Team B?
c Which team has the most consistent scores?
d To stay in the game, you must score at least 120. Which team seems most likely to stay in?
e Which team gets the highest scores? Give reasons for your answer.

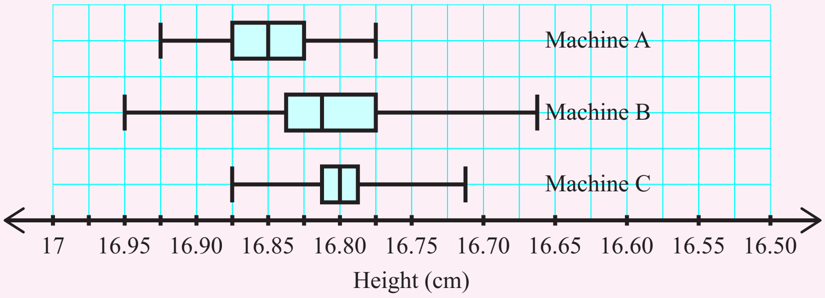
6 This diagram shows the time (in minutes) that two students spend doing homework each day for a term. What does this diagram tell you about the two students?



E

Applying your skills

7 An engineering firm has three machines that produce specialised bolts for airplanes. The bolts must have a diameter of 16.85 mm (with a tolerance of +/– 0.1 mm). During a quality inspection, a sample of 50 bolts produced by each machine is tested and the following box-and-whisker plot is produced using the test data.



Write a quality inspection report comparing the performance of the three machines.

Summary

Do you know the following?

- Averages – the mode, median and mean – are used to summarise a collection of data.
- There are two main types of numerical data – discrete and continuous.
- Discrete data can be listed or arranged in a frequency distribution.
- Continuous data can be listed or arranged into groups
- The mean is affected by extreme data.
- The median is less affected by extreme data.

E

- The median is a special example of a percentile.
- The lower quartile (Q_1) lies 25% of the way through the data.
- The upper quartile (Q_3) lies 75% of the way through the data.
- The interquartile range (IQR = $Q_3 - Q_1$) gives a measure of how spread out or consistent the data is. It is a measure of the spread of the central 50% of the data.
- A box-and-whisker plot is a diagram that shows the distribution of a data set using five values: the minimum and maximum (range); the lower and upper quartiles (IQR) and the median.

Are you able to …?

- calculate the mean, median, mode and range of data given in a list
- calculate the mean, median, mode and range of data given in a frequency distribution and a stem and leaf diagram

E

- calculate an estimate for the mean of grouped data
- find the median class for grouped data
- find the modal class for grouped data
- compare sets of data using summary averages and ranges
- find the quartiles of data arranged in ascending order
- find the interquartile range for listed data
- construct and interpret box-and-whisker plots and use them to compare and describe two or more sets of data.

Examination practice

Past paper questions

1 The time, t seconds, taken for each of 50 chefs to cook an omelette is recorded.

Time (t seconds)	$20 < t \leqslant 25$	$25 < t \leqslant 30$	$30 < t \leqslant 35$	$35 < t \leqslant 40$	$40 < t \leqslant 45$	$45 < t \leqslant 50$
Frequency	2	6	7	19	9	7

a Write down the modal time interval [1]

b Calculate an estimate of the mean time. Show all your working [4]

[Cambridge IGCSE Mathematics 0580 Paper 42 Q3 (a) & (b) October/November 2014]

2 Shahruk plays four games of golf
His four scores have a mean of 75, a mode of 78 and a median of 77
Work out his four scores [3]

[Cambridge IGCSE Mathematics 0580 Paper 22 Q11 May/June 2016]

3 **a** A farmer takes a sample of 158 potatoes from his crop.
He records the mass of each potato and the results are shown in the table.

Mass (m grams)	Frequency
$0 < m \leqslant 40$	6
$40 < m \leqslant 80$	10
$80 < m \leqslant 120$	28
$120 < m \leqslant 160$	76
$160 < m \leqslant 200$	22
$200 < m \leqslant 240$	16

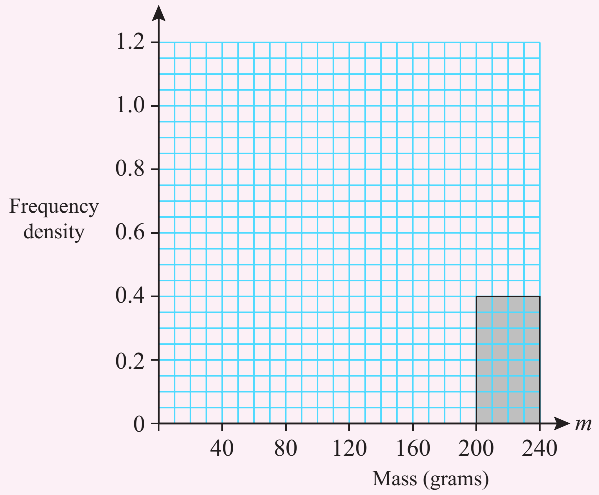
Calculate an estimate of the mean mass.
Show all your working. [4]

b A new frequency table is made from the results shown in the table in **part a**.

Mass (m grams)	Frequency
$0 < m \leqslant 80$	
$80 < m \leqslant 200$	
$200 < m \leqslant 240$	16

i Complete the table above. [2]

ii On the grid, complete the histogram to show the information in this new table.



[3]

c A bag contains 15 potatoes which have a mean mass of 136 g.
The farmer puts 3 potatoes which have a mean mass of 130 g into the bag.
Calculate the mean mass of all the potatoes in the bag. [3]

[Cambridge IGCSE Mathematics 0580 Paper 42 Q5 October/November 2012]