# 2 Measures of location and spread

## Objectives

After completing this chapter you should be able to:

- Calculate measures of central tendency such as the mean, median and mode → pages 21–25
- Calculate measures of location such as percentiles and deciles → pages 25–28
- Calculate measures of spread such as range, interquartile range and interpercentile range → pages 28–29
- Calculate variance and standard deviation → pages 30–33
- Understand and use coding → pages 33–36

## Prior knowledge check

**1** State whether each of these variables is qualitative or quantitative:

**a** Colour of car
**b** Miles travelled by a cyclist
**c** Favourite type of pet
**d** Number of brothers and sisters.

← Section 1.4

**2** State whether each of these variables is discrete or continuous:

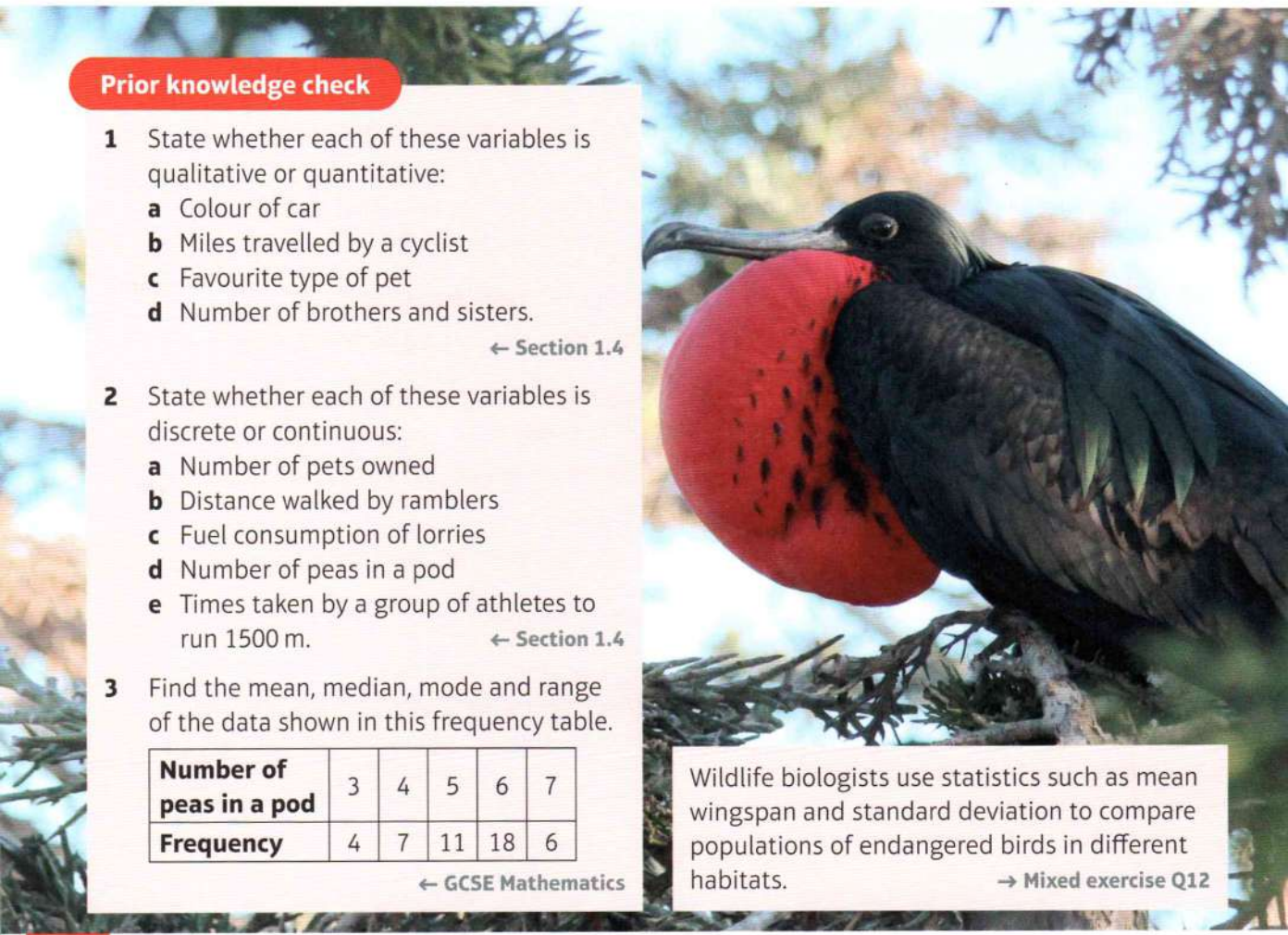
**a** Number of pets owned
**b** Distance walked by ramblers
**c** Fuel consumption of lorries
**d** Number of peas in a pod
**e** Times taken by a group of athletes to run 1500 m.

← Section 1.4

**3** Find the mean, median, mode and range of the data shown in this frequency table.

| **Number of peas in a pod** | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| **Frequency** | 4 | 7 | 11 | 18 | 6 |

← GCSE Mathematics



Wildlife biologists use statistics such as mean wingspan and standard deviation to compare populations of endangered birds in different habitats. → Mixed exercise Q12

## 2.1 Measures of central tendency

A **measure of location** is a single value which describes a position in a data set. If the single value describes the centre of the data, it is called a **measure of central tendency**. You should already know how to work out the **mean**, **median** and **mode** of a set of ungrouped data and from ungrouped frequency tables.

- **The mode or modal class is the value or class that occurs most often.**
- **The median is the middle value when the data values are put in order.**
- **The mean can be calculated using the formula $\bar{x} = \frac{\Sigma x}{n}$.**

**Notation**

- $\bar{x}$ represents the **mean** of the data. You say '$x$ bar'.
- $\Sigma x$ represents the sum of the data values.
- $n$ is the number of data values.

**Example 1**

The mean of a sample of 25 observations is 6.4. The mean of a second sample of 30 observations is 7.2. Calculate the mean of all 55 observations.

For the first set of observations:

$\bar{x} = \frac{\Sigma x}{n}$ so $6.4 = \frac{\Sigma x}{25}$

$\Sigma x = 6.4 \times 25 = 160$

Sum of data values = mean × number of data values.

For the second set of observations:

$\bar{y} = \frac{\Sigma y}{m}$ so $7.2 = \frac{\Sigma y}{30}$

$\Sigma y = 7.2 \times 30 = 216$

$\text{Mean} = \frac{160 + 216}{25 + 30} = 6.84$ (2 d.p.)

**Notation** You can use $x$ and $y$ to represent two different data sets. You need to use different letters for the number of observations in each data set.

You need to decide on the best measure to use in particular situations.

- Mode — This is used when data is qualitative, or quantitative with either a single mode or two modes (bimodal). It is not very informative if each value occurs only once.
- Median — This is used for quantitative data. It is usually used when there are extreme values, as they do not affect it.
- Mean — This is used for quantitative data and uses all the pieces of data. It therefore gives a true measure of the data. However, it is affected by extreme values.

You can calculate the mean, median and mode for discrete data presented in a frequency table.

- **For data given in a frequency table, the mean can be calculated using the formula $\bar{x} = \frac{\Sigma xf}{\Sigma f}$.**

**Notation**

- $\Sigma xf$ is the sum of the products of the data values and their frequencies.
- $\Sigma f$ is the sum of the frequencies.

## Example 2

Rebecca records the shirt collar size, $x$, of the male students in her year. The results are shown in the table.

| Shirt collar size | 15 | 15.5 | 16 | 16.5 | 17 |
|---|---|---|---|---|---|
| Number of students | 3 | 17 | 29 | 34 | 12 |

Find for this data:

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

**d** Explain why a shirt manufacturer might use the mode when planning production numbers.

**a** Mode = 16.5

16.5 is the collar size with the highest frequency.

**b** There are 95 observations

so the median is the $\frac{95+1}{2} = 48$th.

There are 20 observations up to 15.5 and 49 observations up to 16.

The 48th observation is therefore 16.

Median = 16

**c** $\bar{x} = \frac{15 \times 3 + 15.5 \times 17 + 16 \times 29 + 16.5 \times 34 + 17 \times 12}{95}$

$= \frac{45 + 263.5 + 464 + 561 + 204}{95} = \frac{1537.5}{95} = 16.2$

**Online** You can input a frequency table into your calculator, and calculate the mean and median without having to enter the whole calculation.

**d** The mode is an actual data value and gives the manufacturer information on the most common size worn/purchased.

The mean is not one of the data values and the median is not necessarily indicative of the most popular collar size.

## Exercise 2A

**1** Meryl collected wild mushrooms every day for a week. When she got home each day she weighed them to the nearest 50 g. The weights are shown below:

500 700 400 300 900 700 700

**a** Write down the mode for this data.

**b** Calculate the mean for this data.

**c** Find the median for this data.

On the next day, Meryl collects 650 g of wild mushrooms.

**d** Write down the effect this will have on the mean, the mode and the median.

**Hint** Try to answer part **d** without recalculating the averages. You could recalculate to check your answer.

**2** Joe collects six pieces of data, $x_1$, $x_2$, $x_3$, $x_4$, $x_5$ and $x_6$. He works out that $\Sigma x$ is 256.2.

**a** Calculate the mean for this data.

He collects another piece of data. It is 52.

**b** Write down the effect this piece of data will have on the mean.

**3** From the large data set, the daily mean visibility, $v$ metres, for Leeming in May and June 2015 was recorded each day. The data is summarised as follows:

May: $n = 31$, $\Sigma v = 724\,000$

June: $n = 30$, $\Sigma v = 632\,000$

**a** Calculate the mean visibility in each month.

**b** Calculate the mean visibility for the total recording period.

**Hint** You don't need to refer to the actual large data set. All the data you need is given with the question.

**4** A small workshop records how long it takes, in minutes, for each of their workers to make a certain item. The times are shown in the table.

| Worker | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Time in minutes | 7 | 12 | 10 | 8 | 6 | 8 | 5 | 26 | 11 | 9 |

**a** Write down the mode for this data.

**b** Calculate the mean for this data.

**c** Find the median for this data.

**d** The manager wants to give the workers an idea of the average time they took. Write down, with a reason, which of the answers to **a**, **b** and **c** she should use.

**5** The frequency table shows the number of breakdowns, $b$, per month recorded by a road haulage firm over a certain period of time.

| Breakdowns | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 8 | 11 | 12 | 3 | 1 | 1 |

**a** Write down the modal number of breakdowns.

**b** Find the median number of breakdowns.

**c** Calculate the mean number of breakdowns.

**d** In a brochure about how many loads reach their destination on time, the firm quotes one of the answers to **a**, **b** or **c** as the number of breakdowns per month for its vehicles. Write down which of the three answers the firm should quote in the brochure.

**6** The table shows the frequency distribution for the number of petals in the flowers of a group of celandines.

| Number of petals | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|
| Frequency | 8 | 57 | 29 | 3 | 1 |

Calculate the mean number of petals.

**7** A naturalist is investigating how many eggs the endangered kakapo bird lays in each brood cycle. The results are given in this frequency table.

| Number of eggs | 1 | 2 | 3 |
|---|---|---|---|
| Frequency | 7 | $p$ | 2 |

If the mean number of eggs is 1.5, find the value of $p$.

**Problem-solving**

Use the formula for the mean of an ungrouped frequency table to write an equation involving $p$.

You can calculate the mean, the class containing the median and the modal class for continuous data presented in a grouped frequency table by finding the midpoint of each class interval.

## Example 3

The length $x$ mm, to the nearest mm, of a random sample of pine cones is measured. The data is shown in the table.

| Length of pine cone (mm) | 30–31 | 32–33 | 34–36 | 37–39 |
|---|---|---|---|---|
| Frequency | 2 | 25 | 30 | 13 |

**a** Write down the modal class. **b** Estimate the mean. **c** Find the median class.

**a** Modal class = 34–36

**b** Mean $= \dfrac{30.5 \times 2 + 32.5 \times 25 + 35 \times 30 + 38 \times 13}{70}$

$= 34.54$

**c** There are 70 observations so the median is the 35.5th. The 35.5th observation will lie in the class 34–36.

The modal class is the class with the highest frequency.

Use $\bar{x} = \dfrac{\Sigma xf}{\Sigma f}$, taking the midpoint of each class interval as the value of $x$. The answer is an estimate because you don't know the exact data values.

← Section 1.4

## Exercise 2B

**1** The weekly wages (to the nearest £) of the production line workers in a small factory is shown in the table.

| Weekly wage (£) | Frequency |
|---|---|
| 175–225 | 4 |
| 226–300 | 8 |
| 301–350 | 18 |
| 351–400 | 28 |
| 401–500 | 7 |

**a** Write down the modal class.

**b** Calculate an estimate of the mean wage.

**c** Write down the interval containing the median.

(E) **2** The noise levels at 30 locations near an outdoor concert venue were measured to the nearest decibel. The data collected is shown in the grouped frequency table.

| Noise (decibels) | 65–69 | 70–74 | 75–79 | 80–84 | 85–89 | 90–94 | 95–99 |
|---|---|---|---|---|---|---|---|
| Frequency | 1 | 4 | 6 | 6 | 8 | 4 | 1 |

**a** Calculate an estimate of the mean noise level. **(1 mark)**

**b** Explain why your answer to part **a** is an estimate. **(1 mark)**

(E) **3** The table shows the daily mean temperature at Heathrow in October 1987 from the large data set.

| Temp (°C) | $6 \leqslant t < 8$ | $8 \leqslant t < 10$ | $10 \leqslant t < 12$ | $12 \leqslant t < 14$ | $14 \leqslant t < 16$ | $16 \leqslant t < 18$ |
|---|---|---|---|---|---|---|
| Frequency | 3 | 7 | 9 | 7 | 3 | 2 |



**a** Write down the modal class. **(1 mark)**

**b** Calculate an estimate for the mean daily mean temperature. **(1 mark)**

(P) **4** Two DIY shops (A and B) recorded the ages of their workers.

| Age of worker | 16–25 | 26–35 | 36–45 | 46–55 | 56–65 | 66–75 |
|---|---|---|---|---|---|---|
| Frequency A | 5 | 16 | 14 | 22 | 26 | 14 |
| Frequency B | 4 | 12 | 10 | 28 | 25 | 13 |

By comparing estimated means for each shop, determine which shop is better at employing older workers.

**Problem-solving**

Since age is always rounded **down**, the class boundaries for the 16–25 group are 16 and 26. This means that the midpoint of the class is 21.

## 2.2 Other measures of location

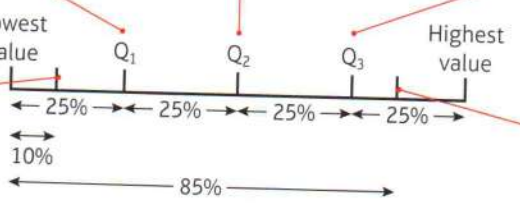
The median describes the middle of the data set. It splits the data set into two equal (50%) halves.

You can calculate other measures of location such as **quartiles** and **percentiles**.

The **lower quartile** is one-quarter of the way through the data set.

This is the median value.

The **upper quartile** is three-quarters of the way through the data set.




Percentiles split the data set into 100 parts. The 10th percentile lies one-tenth of the way through the data.

85% of the data values are less than the 85th percentile, and 15% are greater.

Use these rules to find the upper and lower quartiles for **discrete data**.

- **To find the lower quartile for discrete data, divide $n$ by 4. If this is a whole number, the lower quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.**
- **To find the upper quartile for discrete data, find $\frac{3}{4}$ of $n$. If this is a whole number, the upper quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.**

**Notation** $Q_1$ is the lower quartile, $Q_2$ is the median and $Q_3$ is the upper quartile.

**Watch out** There are different conventions for calculating quartiles. If your calculator gives a different answer, either one is valid.

**Example 4**

From the large data set, the daily maximum gust (knots) during the first 20 days of June 2015 is recorded in Hurn. The data is shown below:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 15 | 17 | 17 | 18 | 18 | 19 | 19 | 22 | 22 |
| 23 | 23 | 23 | 24 | 25 | 26 | 27 | 28 | 36 | 39 |



Find the median and quartiles for this data.

$Q_2 = \frac{20 + 1}{2}$th value = 10.5th value

$Q_2 = \frac{22 + 23}{2} = 22.5$ knots

$Q_1 = 5.5$th value

$Q_1 = 18$ knots

$Q_3 = 15.5$th value

$Q_3 = 25.5$ knots

$Q_2$ is the median. It lies halfway between the 10th and 11th data values (which are 22 knots and 23 knots respectively).

$\frac{20}{4} = 5$ so the lower quartile is halfway between the 5th and 6th data values.

$\frac{3 \times 20}{4} = 15$ so the upper quartile is halfway between the 15th and 16th data values.

When data are presented in a grouped frequency table you can use a technique called **interpolation** to estimate the median, quartiles and percentiles. When you use interpolation, you are assuming that the data values are **evenly distributed** within each class.

**Watch out** For **grouped continuous** data, or data presented in a cumulative frequency table:

$Q_1 = \frac{n}{4}$th data value

$Q_2 = \frac{n}{2}$th data value

$Q_3 = \frac{3n}{4}$th data value

## Example 5

The length of time (to the nearest minute) spent on the internet each evening by a group of students is shown in the table.

| Length of time spent on internet (minutes) | 30–31 | 32–33 | 34–36 | 37–39 |
|---|---|---|---|---|
| Frequency | 2 | 25 | 30 | 13 |

**a** Find an estimate for the upper quartile.

**b** Find an estimate for the 10th percentile.

**a** Upper quartile: $\frac{3 \times 70}{4} = 52.5$th value

Using interpolation:


$\frac{Q_3 - 33.5}{36.5 - 33.5} = \frac{52.5 - 27}{57 - 27}$

$\frac{Q_3 - 33.5}{3} = \frac{25.5}{30}$

$Q_3 = 36.05$

**b** The 10th percentile is the 7th data value.

$\frac{P_{10} - 31.5}{33.5 - 31.5} = \frac{7 - 2}{27 - 2}$

$\frac{P_{10} - 31.5}{2} = \frac{5}{25}$

$P_{10} = 31.9$

The endpoints on the line represent the class boundaries.

The values on the bottom are the cumulative frequencies for the previous classes and this class.

**Problem-solving**

Use proportion to estimate $Q_3$. The 52.5th value lies $\frac{52.5 - 27}{57 - 27}$ of the way into the class, so $Q_3$ lies $\frac{Q_3 - 33.5}{36.5 - 33.5}$ of the way between the class boundaries. Equate these two fractions to form an equation and solve to find $Q_3$.

**Notation** You can write the 10th percentile as $P_{10}$.

## Exercise 2C

**1** From the large data set, the daily mean pressure (hPa) during the last 16 days of July 2015 in Perth is recorded. The data is given below:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1024 | 1022 | 1021 | 1013 | 1009 | 1018 | 1017 | 1024 |
| 1027 | 1029 | 1031 | 1025 | 1017 | 1019 | 1017 | 1014 |

**a** Find the median pressure for that period.

**b** Find the lower and upper quartiles.

**2** Rachel records the number of CDs in the collections of students in her year. The results are in the table below.

| Number of CDs | 35 | 36 | 37 | 38 | 39 |
|---|---|---|---|---|---|
| Frequency | 3 | 17 | 29 | 34 | 12 |

Find $Q_1$, $Q_2$ and $Q_3$.

**Hint** This an ungrouped frequency table so you do not need to use interpolation. Use the rules for finding the median and quartiles of **discrete** data.

(E) **3** A hotel is worried about the reliability of its lift. It keeps a weekly record of the number of times it breaks down over a period of 26 weeks. The data collected is summarised in the table opposite.

| Number of breakdowns | Frequency |
|---|---|
| 0–1 | 18 |
| 2–3 | 7 |
| 4–5 | 1 |

Use interpolation to estimate the median number of breakdowns.

**(2 marks)**

**4** The weights of 31 Jersey cows were recorded to the nearest kilogram. The weights are shown in the table.

| Weight of cattle (kg) | 300–349 | 350–399 | 400–449 | 450–499 | 500–549 |
|---|---|---|---|---|---|
| Frequency | 3 | 6 | 10 | 7 | 5 |

**a** Find an estimate for the median weight.

**b** Find the lower quartile, $Q_1$.

**c** Find the upper quartile, $Q_3$.

**d** Interpret the meaning of the value you have found for the upper quartile in part **c**.

(E) **5** A roadside assistance firm kept a record over a week of the amount of time, in minutes, people were kept waiting for assistance. The times are shown below.

| Time waiting, $t$ (minutes) | $20 \leqslant t < 30$ | $30 \leqslant t < 40$ | $40 \leqslant t < 50$ | $50 \leqslant t < 60$ | $60 \leqslant t < 70$ |
|---|---|---|---|---|---|
| Frequency | 6 | 10 | 18 | 13 | 2 |

**a** Find an estimate for the mean wait time. **(1 mark)**

**b** Calculate the 65th percentile. **(2 marks)**

The firm writes the following statement for an advertisement:

> Only 10% of our customers have to wait longer than 56 minutes.

**c** By calculating a suitable percentile, comment on the validity of this claim. **(3 marks)**

(E) 6 The table shows the recorded wingspans, in metres, of 100 endangered Californian condors.

| Wingspan, $w$ (m) | $1.0 \leqslant w < 1.5$ | $1.5 \leqslant w < 2.0$ | $2.0 \leqslant w < 2.5$ | $2.5 \leqslant w < 3.0$ | $3.0 \leqslant w$ |
|---|---|---|---|---|---|
| Frequency | 4 | 20 | 37 | 28 | 11 |

a Estimate the 80th percentile and interpret the value. **(3 marks)**

b State why it is not possible to estimate the 90th percentile. **(1 mark)**

## 2.3 Measures of spread

A measure of spread is a measure of how spread out the data is. Here are two simple measures of spread.

**Notation** Measures of spread are sometimes called **measures of dispersion** or **measures of variation**.

- **The range is the difference between the largest and smallest values in the data set.**
- **The interquartile range (IQR) is the difference between the upper quartile and the lower quartile, $Q_3 - Q_1$.**

The range takes into account all of the data but can be affected by extreme values. The interquartile range is not affected by extreme values but only considers the spread of the middle 50% of the data.

- **The interpercentile range is the difference between the values for two given percentiles.**

The 10th to 90th interpercentile range is often used since it is not affected by extreme values but still considers 80% of the data in its calculation.

### Example 6

The table shows the masses, in tonnes, of 120 African bush elephants.

| Mass, $m$ (t) | $4.0 \leqslant m < 4.5$ | $4.5 \leqslant m < 5.0$ | $5.0 \leqslant m < 5.5$ | $5.5 \leqslant m < 6.0$ | $6.0 \leqslant m < 6.5$ |
|---|---|---|---|---|---|
| Frequency | 13 | 23 | 31 | 34 | 19 |

Find estimates for:

a the range b the interquartile range c the 10th to 90th interpercentile range.

a Range is 6.5 − 4.0 = 2.5 tonnes.

The largest possible value is 6.5 and the smallest possible value is 4.0.

b $Q_1$ = 30th data value: 4.87 tonnes.
$Q_3$ = 90th data value: 5.84 tonnes.
The interquartile range is therefore 5.84 − 4.87 = 0.97 tonnes.

Use interpolation: $\frac{Q_1 - 4.5}{5.0 - 4.5} = \frac{30 - 13}{23}$

Use interpolation: $\frac{Q_3 - 5.5}{6.0 - 5.5} = \frac{90 - 67}{34}$

c 10th percentile = 12th data value: 4.46 tonnes.
90th percentile = 108th data value: 6.18 tonnes.
The 10th to 90th interpercentile range is therefore 6.18 − 4.46 = 1.72 tonnes.

Use interpolation to find the 10th and 90th percentiles, then work out the difference between them.

## Exercise 2D

(P) **1** The lengths of a number of slow worms were measured, to the nearest mm.
The results are shown in the table.

| Lengths of slow worms (mm) | Frequency |
|---|---|
| 125–139 | 4 |
| 140–154 | 4 |
| 155–169 | 2 |
| 170–184 | 7 |
| 185–199 | 20 |
| 200–214 | 24 |
| 215–229 | 10 |

**a** Work out how many slow worms were measured.

**b** Estimate the interquartile range for the lengths of the slow worms.

**c** Calculate an estimate for the mean length of slow worms.

**d** Estimate the number of slow worms whose length is more than one interquartile range above the mean.

**Problem-solving**

For part **d**, work out $\bar{x}$ + IQR, and determine which class interval it falls in. Then use proportion to work out how many slow worms from that class interval you need to include in your estimate.

(E) **2** The table shows the monthly income for workers in a factory.

| Monthly income, $x$ (£) | $900 \leqslant x < 1000$ | $1000 \leqslant x < 1100$ | $1100 \leqslant x < 1200$ | $1200 \leqslant x < 1300$ |
|---|---|---|---|---|
| Frequency | 3 | 24 | 28 | 15 |

**a** Calculate the 34% to 66% interpercentile range. **(3 marks)**

**b** Estimate the number of data values that fall within this range. **(2 marks)**

(E) **3** A train travelled from Lancaster to Preston. The times, to the nearest minute, it took for the journey were recorded over a certain period. The times are shown in the table.

| Time for journey (minutes) | 15–16 | 17–18 | 19–20 | 21–22 |
|---|---|---|---|---|
| Frequency | 5 | 10 | 35 | 10 |

**a** Calculate the 5% to 95% interpercentile range. **(3 marks)**

**b** Estimate the number of data values that fall within this range. **(1 mark)**

(P) **4** From the large data set, the daily mean temperature (°C) for Leeming during the first 10 days of June 1987 is given below:

14.3 12.7 12.4 10.9 9.4 13.2 12.1 10.3 10.3 10.6

**a** Calculate the median and interquartile range. **(2 marks)**

The median daily mean temperature for Leeming during the first 10 days of May 1987 was 9.9 °C and the interquartile range was 3.9 °C.

**b** Compare the data for May with the data for June. **(2 marks)**

The 10% to 90% interpercentile range for the daily mean temperature for Leeming during July 1987 was 5.4 °C.

**c** Estimate the number of days in July 1987 on which the daily mean temperature fell within this range. **(1 mark)**

## 2.4 Variance and standard deviation

Another measure that can be used to work out the spread of a data set is the **variance**. This makes use of the fact that each data point deviates from the mean by the amount $x - \bar{x}$.

- **Variance** $= \dfrac{\Sigma(x-\bar{x})^2}{n} = \dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2 = \dfrac{S_{xx}}{n}$

  **where** $S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}$

**Notation** $S_{xx}$ is a **summary statistic**, which is used to make formulae easier to use and learn.

The second version of the formula, $\dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2$, is easier to work with when given raw data.

It can be thought of as 'the mean of the squares minus the square of the mean'.

The third version, $\dfrac{S_{xx}}{n}$, is easier to use if you can use your calculator to find $S_{xx}$ quickly.

The units of the variance are the units of the data squared. You can find a related measure of spread that has the same units as the data.

- **The standard deviation is the square root of the variance:**

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{S_{xx}}{n}}$$

**Notation** $\sigma$ is the symbol we use for the standard deviation of a data set. Hence $\sigma^2$ is used for the variance.

### Example 7

The marks gained in a test by seven randomly selected students are:

3 4 6 2 8 8 5

Find the variance and standard deviation of the marks of the seven students.

$\Sigma x = 3 + 4 + 6 + 2 + 8 + 8 + 5 = 36$

$\Sigma x^2 = 9 + 16 + 36 + 4 + 64 + 64 + 25 = 218$

variance, $\sigma^2 = \dfrac{218}{7} - \left(\dfrac{36}{7}\right)^2 = 4.69$

standard deviation, $\sigma = \sqrt{4.69} = 2.17$

Use the 'mean of the squares minus the square of the mean'.

$\sigma^2 = \dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2$

- **You can use these versions of the formulae for variance and standard deviation for grouped data that is presented in a frequency table:**
  - $\sigma^2 = \dfrac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \dfrac{\Sigma f x^2}{\Sigma f} - \left(\dfrac{\Sigma f x}{\Sigma f}\right)^2$
  - $\sigma = \sqrt{\dfrac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = \sqrt{\dfrac{\Sigma f x^2}{\Sigma f} - \left(\dfrac{\Sigma f x}{\Sigma f}\right)^2}$

  **where $f$ is the frequency for each group and $\Sigma f$ is the total frequency.**

## Example 8

Shamsa records the time spent out of school during the lunch hour to the nearest minute, $x$, of the female students in her year.
The results are shown in the table.

| Time spent out of school (min) | 35 | 36 | 37 | 38 |
|---|---|---|---|---|
| Frequency | 3 | 17 | 29 | 34 |

Calculate the standard deviation of the time spent out of school.

$\Sigma fx^2 = 3 \times 35^2 + 17 \times 36^2 + 29 \times 37^2 + 34 \times 38^2 = 114\,504$

$\Sigma fx = 3 \times 35 + 17 \times 36 + 29 \times 37 + 34 \times 38 = 3082$

$\Sigma f = 3 + 17 + 29 + 34 = 83$

$\sigma^2 = \frac{114\,504}{83} - \left(\frac{3082}{83}\right)^2 = 0.741\,47...$

$\sigma = \sqrt{0.741\,47...} = 0.861$ (3 s.f.)

The values of $\Sigma fx^2$, $\Sigma fx$ and $\Sigma f$ might be given with the question.

$\sigma^2$ is the variance, and $\sigma$ is the standard deviation.

Use $\sigma^2 = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2$

If the data is given in a grouped frequency table, you can calculate **estimates** for the variance and standard deviation of the data using the **midpoint** of each class interval.

## Example 9

Andy recorded the length, in minutes, of each telephone call he made for a month. The data is summarised in the table below.

| Length of telephone call ($l$ min) | $0 < l \leqslant 5$ | $5 < l \leqslant 10$ | $10 < l \leqslant 15$ | $15 < l \leqslant 20$ | $20 < l \leqslant 60$ | $60 < l \leqslant 70$ |
|---|---|---|---|---|---|---|
| Frequency | 4 | 15 | 5 | 2 | 0 | 1 |

Calculate an estimate of the standard deviation of the length of telephone calls.

| Length of telephone call ($l$ min) | Frequency | Midpoint $x$ | $fx$ | $fx^2$ |
|---|---|---|---|---|
| $0 < l \leqslant 5$ | 4 | 2.5 | $4 \times 2.5 = 10$ | $4 \times 6.25 = 25$ |
| $5 < l \leqslant 10$ | 15 | 7.5 | 112.5 | 843.75 |
| $10 < l \leqslant 15$ | 5 | 12.5 | 62.5 | 781.25 |
| $15 < l \leqslant 20$ | 2 | 17.5 | 35 | 612.5 |
| $20 < l \leqslant 60$ | 0 | 40 | 0 | 0 |
| $60 < l \leqslant 70$ | 1 | 65 | 65 | 4225 |
| total | 27 | | 285 | 6487.5 |

$\Sigma fx^2 = 6487.5$ $\Sigma fx = 285$ $\Sigma f = 27$

$\sigma^2 = \frac{6487.5}{27} - \left(\frac{285}{27}\right)^2 = 128.858\,02$

$\sigma = \sqrt{128.858\,02} = 11.35$

You can use a table like this to keep track of your working.

**Online**

Work this out in one go on your calculator. You might need to enter the values manually for the midpoint of each class interval.

## Exercise 2E

**1** Given that for a variable $x$: $\Sigma x = 24$ $\Sigma x^2 = 78$ $n = 8$

Find:

**a** the mean **b** the variance $\sigma^2$ **c** the standard deviation $\sigma$.

(E) **2** Ten collie dogs are weighed ($w$ kg). The summary data for the weights is:

$\Sigma w = 241$ $\Sigma w^2 = 5905$

Use this summary data to find the standard deviation of the collies' weights. **(2 marks)**

**3** Eight students' heights ($h$ cm) are measured. They are as follows:

165 170 190 180 175 185 176 184

**a** Work out the mean height of the students.

**b** Given $\Sigma h^2 = 254\,307$ work out the variance. Show all your working.

**c** Work out the standard deviation.

(P) **4** For a set of 10 numbers: $\Sigma x = 50$ $\Sigma x^2 = 310$

For a different set of 15 numbers: $\Sigma x = 86$ $\Sigma x^2 = 568$

Find the mean and the standard deviation of the combined set of 25 numbers.

(E) **5** Nahab asks the students in his year group how much pocket money they get per week. The results, rounded to the nearest pound, are shown in the table.

| Number of £s | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|
| Frequency | 14 | 8 | 28 | 15 | 20 |

**a** Use your calculator to work out the mean and standard deviation of the pocket money. Give units with your answer. **(3 marks)**

**b** How many students received an amount of pocket money more than one standard deviation above the mean? **(2 marks)**

(E) **6** In a student group, a record was kept of the number of days of absence each student had over one particular term. The results are shown in the table.

| Number of days absent | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Frequency | 12 | 20 | 10 | 7 | 5 |

Use your calculator to work out the standard deviation of the number of days absent. **(2 marks)**

(E/P) **7** A certain type of machine contained a part that tended to wear out after different amounts of time. The time it took for 50 of the parts to wear out was recorded. The results are shown in the table.

| Lifetime, $h$ (hours) | $5 < h \leqslant 10$ | $10 < h \leqslant 15$ | $15 < h \leqslant 20$ | $20 < h \leqslant 25$ | $25 < h \leqslant 30$ |
|---|---|---|---|---|---|
| Frequency | 5 | 14 | 23 | 6 | 2 |

The manufacturer makes the following claim:

90% of the parts tested lasted longer than one standard deviation below the mean.

Comment on the accuracy of the manufacturer's claim, giving relevant numerical evidence.

**Problem-solving**

You need to calculate estimates for the mean and the standard deviation, then estimate the number of parts that lasted longer than one standard deviation below the mean.

**(5 marks)**

(E) **8** The daily mean windspeed, $x$ (kn) for Leeming is recorded in June 2015. The summary data is:

$\Sigma x = 243 \quad \Sigma x^2 = 2317$

**a** Use your calculator to work out the mean and the standard deviation of the daily mean windspeed in June 2015. **(2 marks)**

The highest recorded windspeed was 17 kn and the lowest recorded windspeed was 4 kn.

**b** Estimate the number of days in which the windspeed was greater than one standard deviation above the mean. **(2 marks)**

**c** State one assumption you have made in producing this estimate. **(1 mark)**

## 2.5 Coding

Coding is a way of simplifying statistical calculations. Each data value is coded to make a new set of data values which are easier to work with.

In your exam, you will usually have to code values using a formula like this: $y = \frac{x - a}{b}$

where $a$ and $b$ are constants that you have to choose or are given with the question.

When data is coded, different statistics change in different ways.

- **If data is coded using the formula $y = \frac{x - a}{b}$**
  - **the mean of the coded data is given by $\bar{y} = \frac{\bar{x} - a}{b}$**
  - **the standard deviation of the coded data is given by $\sigma_y = \frac{\sigma_x}{b}$, where $\sigma_x$ is the standard deviation of the original data.**

**Hint** You usually need to find the mean and standard deviation of the **original data** given the statistics for the **coded data**. You can rearrange the formulae as:

- $\bar{x} = b\bar{y} + a$
- $\sigma_x = b\sigma_y$

### Example 10

A scientist measures the temperature, $x$ °C, at five different points in a nuclear reactor. Her results are given below:

332 °C  355 °C  306 °C  317 °C  340 °C

**a** Use the coding $y = \frac{x - 300}{10}$ to code this data.

**b** Calculate the mean and standard deviation of the coded data.

**c** Use your answer to part **b** to calculate the mean and standard deviation of the original data.

**a**

| Original data, $x$ | 332 | 355 | 306 | 317 | 340 |
|---|---|---|---|---|---|
| Coded data, $y$ | 3.2 | 5.5 | 0.6 | 1.7 | 4.0 |

When $x = 332$, $y = \frac{332 - 300}{10} = 3.2$.

**b** $\Sigma y = 15$, $\Sigma y^2 = 59.74$

$\bar{y} = \frac{15}{5} = 3$

$\sigma_y^2 = \frac{59.74}{5} - \left(\frac{15}{5}\right)^2 = 2.948$

$\sigma_y = \sqrt{2.948} = 1.72$ (3 s.f.)

Substitute into $\bar{y} = \frac{\bar{x} - a}{b}$ and solve to find $\bar{x}$.

You could also use $\bar{x} = b\bar{y} + a$ with $a = 300$, $b = 10$ and $\bar{y} = 3$.

**c** $3 = \frac{\bar{x} - 300}{10}$ so $\bar{x} = 30 + 300 = 330$ °C

$1.72 = \frac{\sigma_x}{10}$ so $\sigma_x = 17.2$ °C (3 s.f.)

Substitute into $\sigma_y = \frac{\sigma_x}{b}$ and solve to find $\sigma_x$.

You could also use $\sigma_x = b\sigma_y$ with $\sigma_y = 1.72$ and $b = 10$.

## Example 11

From the large data set, data on the maximum gust, $g$ knots, is recorded in Leuchars during May and June 2015.

The data was coded using $h = \frac{g - 5}{10}$ and the following statistics found:

$S_{hh} = 43.58$ $\quad \bar{h} = 2$ $\quad n = 61$

Calculate the mean and standard deviation of the maximum gust in knots.

$2 = \frac{\bar{g} - 5}{10}$

Use the formula for the mean of a coded variable: $\bar{h} = \frac{\bar{g} - a}{b}$ with $a = 5$ and $b = 10$.

$\bar{g} = 2 \times 10 + 5 = 25$ knots

$\sigma_h = \sqrt{\frac{43.58}{61}} = 0.845\ldots$

Calculate the standard deviation of the coded data using $\sigma_h = \sqrt{\frac{S_{hh}}{n}}$, then use the formula for the standard deviation of a coded variable: $\sigma_h = \frac{\sigma_g}{b}$ with $b = 10$.

$\sigma_h = \frac{\sigma_g}{10}$

$\sigma_g = \sigma_h \times 10 = 8.45$ knots (3 s.f.)

## Exercise 2F

**1** A set of data values, $x$, is shown below:

110 90 50 80 30 70 60

**a** Code the data using the coding $y = \frac{x}{10}$.

**b** Calculate the mean of the coded data values.

**c** Use your answer to part **b** to calculate the mean of the original data.

**2** A set of data values, $x$, is shown below:

52 73 31 73 38 80 17 24

**a** Code the data using the coding $y = \frac{x-3}{7}$.

**b** Calculate the mean of the coded data values.

**c** Use your answer to part **b** to calculate the mean of the original data.

(E) **3** The coded mean price of televisions in a shop was worked out. Using the coding $y = \frac{x-65}{200}$ the mean price was 1.5. Find the true mean price of the televisions. **(2 marks)**

**4** The coding $y = x - 40$ gives a standard deviation for $y$ of 2.34.
Write down the standard deviation of $x$.

**Watch out** Adding or subtracting constants does not affect how spread out the data is, so you can ignore the '−40' when finding the standard deviation for $x$.

(P) **5** The lifetime, $x$, in hours, of 70 light bulbs is shown in the table.

| **Lifetime, $x$ (hours)** | $20 < x \leqslant 22$ | $22 < x \leqslant 24$ | $24 < x \leqslant 26$ | $26 < x \leqslant 28$ | $28 < x \leqslant 30$ |
|---|---|---|---|---|---|
| **Frequency** | 3 | 12 | 40 | 10 | 5 |

The data is coded using $y = \frac{x-1}{20}$.

**a** Estimate the mean of the coded values $\bar{y}$.

**b** Hence find an estimate for the mean lifetime of the light bulbs, $\bar{x}$.

**c** Estimate the standard deviation of the lifetimes of the bulbs.

**Problem-solving**

Code the midpoints of each class interval. The midpoint of the $22 < x \leqslant 24$ class interval is 23, so the coded midpoint will be $\frac{23-1}{20} = 1.1$.

(E) **6** The weekly income, $i$, of 100 women workers was recorded.

The data was coded using $y = \frac{i-90}{100}$ and the following summations were obtained:

$\Sigma y = 131$, $\Sigma y^2 = 176.84$

Estimate the standard deviation of the actual women workers' weekly income. **(2 marks)**

(E) **7** A meteorologist collected data on the annual rainfall, $x$ mm, at six randomly selected places.

The data was coded using $s = 0.01x - 10$ and the following summations were obtained:

$\Sigma s = 16.1$, $\Sigma s^2 = 147.03$

Work out an estimate for the standard deviation of the actual annual rainfall. **(2 marks)**

(P) **8** A teacher standardises the test marks of his class by adding 12 to each one and then reducing the mark by 20%.

If the standardised marks are represented by $t$ and the original marks by $m$:

**a** write down a formula for the coding the teacher has used. **(1 mark)**

The following summary statistics are calculated for the standardised marks:

$n = 28$ $\bar{t} = 52.8$ $S_{tt} = 7.3$

**b** Calculate the mean and standard deviation of the original marks gained. **(3 marks)**

E/P **9** From the large data set, the daily mean pressure, $p$ hPa, in Hurn during June 2015 is recorded.

The data is coded using $c = \frac{p}{2} - 500$ and the following summary statistics are obtained:

$n = 30 \quad \bar{c} = 10.15 \quad S_{cc} = 296.4$

Find the mean and standard deviation of the daily mean pressure. **(4 marks)**

## Mixed exercise 2

**1** The mean science mark for one group of eight students is 65. The mean mark for a second group of 12 students is 72. Calculate the mean mark for the combined group of 20 students.

**2** The data shows the prices ($x$) of six shares on a particular day in the year 2007:

807 967 727 167 207 767

**a** Code the data using the coding $y = \frac{x - 7}{80}$.

**b** Calculate the mean of the coded data values.

**c** Use your answer to part **b** to calculate the mean of the original data.

**3** The coded mean of employees' annual earnings (£ $x$) for a store is 18. The coding used was $y = \frac{x - 720}{1000}$. Work out the uncoded mean earnings.

**4** Different teachers using different methods taught two groups of students. Both groups of students sat the same examination at the end of the course. The students' marks are shown in the grouped frequency table.

| **Exam mark** | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 | 70–79 | 80–89 |
|---|---|---|---|---|---|---|---|
| **Frequency group *A*** | 1 | 3 | 6 | 6 | 11 | 10 | 8 |
| **Frequency group *B*** | 1 | 2 | 4 | 13 | 15 | 6 | 3 |

**a** Work out an estimate of the mean mark for group $A$ and an estimate of the mean mark for group $B$.

**b** Write down whether or not the answer to **a** suggests that one method of teaching is better than the other. Give a reason for your answer.

**5** The lifetimes of 80 batteries, to the nearest hour, are shown in the table below.

| **Lifetime (hours)** | 6–10 | 11–15 | 16–20 | 21–25 | 26–30 |
|---|---|---|---|---|---|
| **Frequency** | 2 | 10 | 18 | 45 | 5 |

**a** Write down the modal class for the lifetime of the batteries.

**b** Use interpolation to find the median lifetime of the batteries.

The midpoint of each class is represented by $x$ and its corresponding frequency by $f$, giving $\Sigma fx = 1645$.

**c** Calculate an estimate of the mean lifetime of the batteries.

Another batch of 12 batteries is found to have an estimated mean lifetime of 22.3 hours.

**d** Estimate the mean lifetime for all 92 batteries.

**6** A frequency distribution is shown below.

| Class interval | 1–20 | 21–40 | 41–60 | 61–80 | 81–100 |
|---|---|---|---|---|---|
| Frequency | 5 | 10 | 15 | 12 | 8 |

Use interpolation to find an estimate for the interquartile range.

**7** A frequency distribution is shown below.

| Class interval | 1–10 | 11–20 | 21–30 | 31–40 | 41–50 |
|---|---|---|---|---|---|
| Frequency | 10 | 20 | 30 | 24 | 16 |

**a** Use interpolation to estimate the value of the 30th percentile.

**b** Use interpolation to estimate the value of the 70th percentile.

**c** Hence estimate the 30% to 70% interpercentile range.

(E) **8** The times it took a random sample of runners to complete a race are summarised in the table.

| Time taken ($t$ minutes) | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 |
|---|---|---|---|---|---|
| Frequency | 5 | 10 | 36 | 20 | 9 |

**a** Use interpolation to estimate the interquartile range. **(3 marks)**

The midpoint of each class was represented by $x$ and its corresponding frequency by $f$ giving:

$\Sigma fx = 3740 \qquad \Sigma fx^2 = 183\,040$

**b** Estimate the variance and standard deviation for this data. **(3 marks)**

**9** The heights of 50 clover flowers are summarised in the table.

| Heights in mm ($x$) | $90 \leqslant x < 95$ | $95 \leqslant x < 100$ | $100 \leqslant x < 105$ | $105 \leqslant x < 110$ | $110 \leqslant x < 115$ |
|---|---|---|---|---|---|
| Frequency | 5 | 10 | 26 | 8 | 1 |

**a** Find $Q_1$. **b** Find $Q_3$. **c** Find the interquartile range.

**d** Use $\Sigma fx = 5075$ and $\Sigma fx^2 = 516\,112.5$ to find the standard deviation.

(E/P) **10** The daily mean temperature is recorded in Camborne during September 2015.

| Temperature, $t$ (°C) | $11 < t \leqslant 13$ | $13 < t \leqslant 15$ | $15 < t \leqslant 17$ |
|---|---|---|---|
| Frequency | 12 | 14 | 4 |



**a** Use your calculator to find estimates for the mean and standard deviation of the temperatures. **(3 marks)**

**b** Use linear interpolation to find an estimate for the 10% to 90% interpercentile range. **(3 marks)**

**c** Estimate the number of days in September 2015 where the daily mean temperature in Camborne is more than one standard deviation greater than the mean. **(2 marks)**

(E) **11** The daily mean windspeed, $w$ knots was recorded at Heathrow during May 2015. The data were coded using $z = \frac{w - 3}{2}$.

Summary statistics were calculated for the coded data:

$n = 31 \qquad \Sigma z = 106 \qquad S_{zz} = 80.55$

**a** Find the mean and standard deviation of the coded data. **(2 marks)**

**b** Work out the mean and standard deviation of the daily mean windspeed at Heathrow during May 2015. **(2 marks)**

(E) **12** 20 endangered forest owlets were caught for ringing. Their wingspans ($x$ cm) were measured to the nearest centimetre.

The following summary statistics were worked out:

$\Sigma x = 316$ $\Sigma x^2 = 5078$

**a** Work out the mean and the standard deviation of the wingspans of the 20 birds. **(3 marks)**

One more bird was caught. It had a wingspan of 13 centimetres.

**b** Without doing any further calculation, say how you think this extra wingspan will affect the mean wingspan. **(1 mark)**

20 giant ibises were also caught for ringing. Their wingspans ($y$ cm) were also measured to the nearest centimetre and the data coded using $z = \frac{y - 5}{10}$.

The following summary statistics were obtained from the coded data:

$\Sigma z = 104$ $S_{zz} = 1.8$

**c** Work out the mean and standard deviation of the wingspans of the giant ibis. **(5 marks)**

**Challenge**

A biologist recorded the heights, $x$ cm, of 20 plant seedlings.
She calculated the mean and standard deviation of her results:

$\bar{x} = 3.1$ cm $\sigma = 1.4$ cm

The biologist subsequently discovered she had written down one value incorrectly. She replaced a value of 2.3 cm with a value of 3.2 cm.

Calculate the new mean and standard deviation of her data.

## Summary of key points

1 The **mode** or **modal class** is the value or class that occurs most often.

2 The **median** is the middle value when the data values are put in order.

3 The **mean** can be calculated using the formula $\bar{x} = \frac{\Sigma x}{n}$.

4 For data given in a frequency table, the mean can be calculated using the formula $\bar{x} = \frac{\Sigma xf}{\Sigma f}$.

5 To find the **lower quartile** for discrete data, divide $n$ by 4. If this is a whole number, the lower quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.

6 To find the **upper quartile** for discrete data, find $\frac{3}{4}$ of $n$. If this is a whole number, the upper quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.

7 The **range** is the difference between the largest and smallest values in the data set.

8 The **interquartile range** (IQR) is the difference between the upper quartile and the lower quartile, $Q_3 - Q_1$.

9 The **interpercentile range** is the difference between the values for two given percentiles.

10 **Variance** $= \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{S_{xx}}{n}$ where $S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

11 The **standard deviation** is the square root of the variance:

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{S_{xx}}{n}}$$

12 You can use these versions of the formulae for variance and standard deviation for grouped data that is presented in a frequency table:

$$\sigma^2 = \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2 \qquad \sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

where $f$ is the frequency for each group and $\Sigma f$ is the total frequency.

13 If data is coded using the formula $y = \frac{x-a}{b}$

- the mean of the coded data is given by $\bar{y} = \frac{\bar{x}-a}{b}$
- the standard deviation of the coded data is given by $\sigma_y = \frac{\sigma_x}{b}$ where $\sigma_x$ is the standard deviation of the original data.