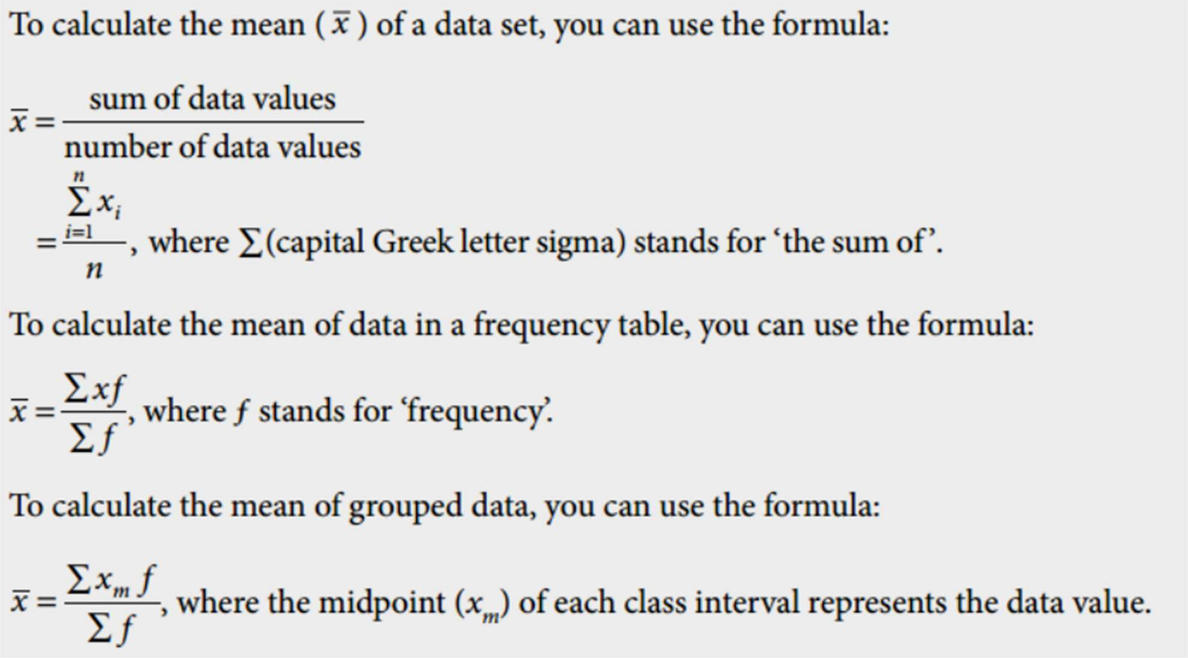
# 1 Measure of central tendency

## 1.1 the mean



To calculate the mean ($\bar{x}$) of a data set, you can use the formula:

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

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$, where $\sum$(capital Greek letter sigma) stands for 'the sum of'.

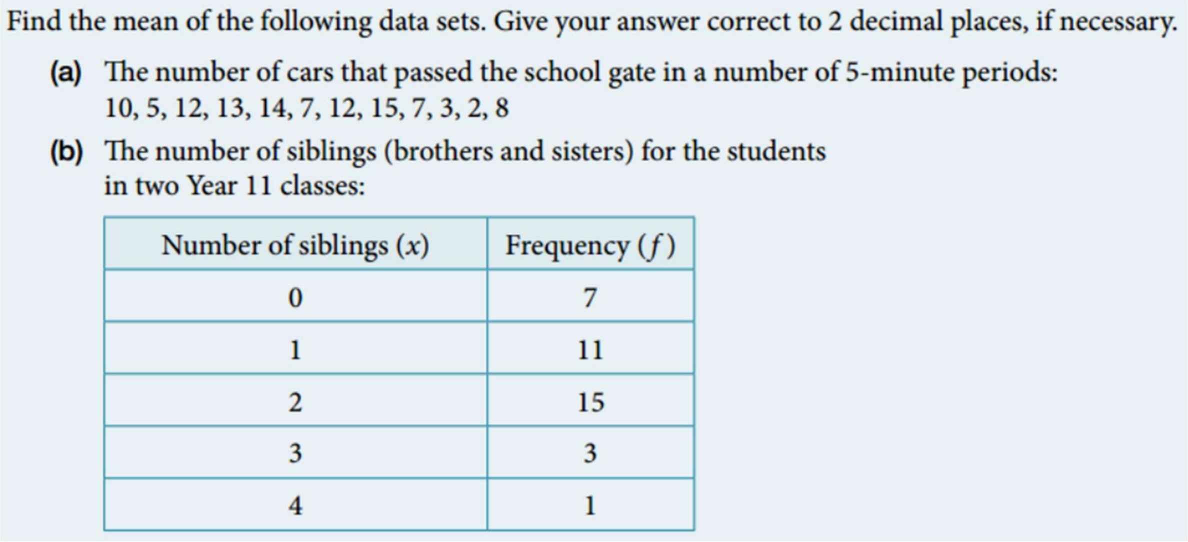
To calculate the mean of data in a frequency table, you can use the formula:

$$\bar{x} = \frac{\sum xf}{\sum f}$$, where $f$ stands for 'frequency'.

To calculate the mean of grouped data, you can use the formula:

$$\bar{x} = \frac{\sum x_m f}{\sum f}$$, where the midpoint ($x_m$) of each class interval represents the data value.

## Example of calculation of the mean for a discrete set of data

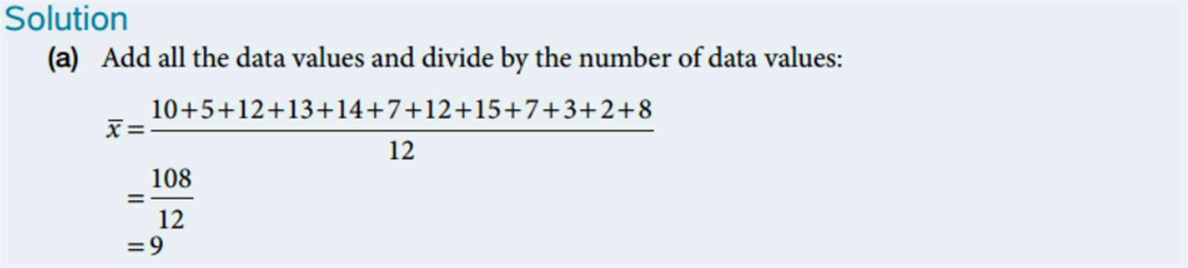


Find the mean of the following data sets. Give your answer correct to 2 decimal places, if necessary.

**(a)** The number of cars that passed the school gate in a number of 5-minute periods:
10, 5, 12, 13, 14, 7, 12, 15, 7, 3, 2, 8

**(b)** The number of siblings (brothers and sisters) for the students in two Year 11 classes:

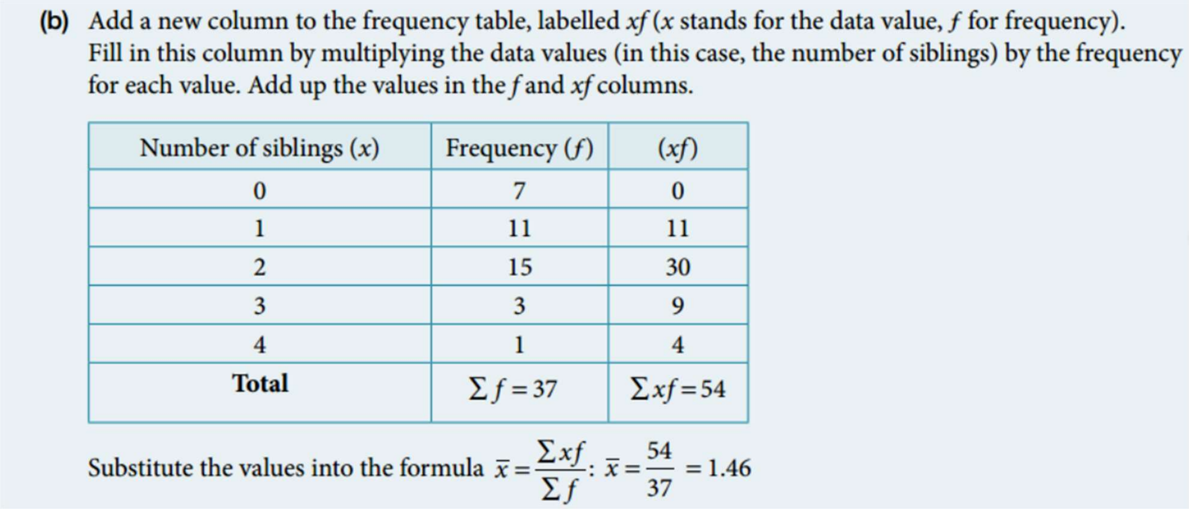
| Number of siblings ($x$) | Frequency ($f$) |
|---|---|
| 0 | 7 |
| 1 | 11 |
| 2 | 15 |
| 3 | 3 |
| 4 | 1 |



### Solution

**(a)** Add all the data values and divide by the number of data values:

$$\begin{aligned}\bar{x} &= \frac{10+5+12+13+14+7+12+15+7+3+2+8}{12}\\ &= \frac{108}{12}\\ &= 9\end{aligned}$$

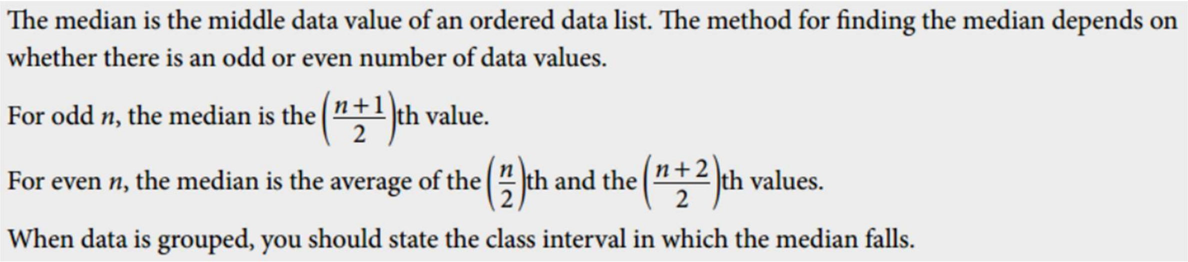


**(b)** Add a new column to the frequency table, labelled $xf$ ($x$ stands for the data value, $f$ for frequency). Fill in this column by multiplying the data values (in this case, the number of siblings) by the frequency for each value. Add up the values in the $f$ and $xf$ columns.

| Number of siblings ($x$) | Frequency ($f$) | ($xf$) |
|---|---|---|
| 0 | 7 | 0 |
| 1 | 11 | 11 |
| 2 | 15 | 30 |
| 3 | 3 | 9 |
| 4 | 1 | 4 |
| **Total** | $\sum f = 37$ | $\sum xf = 54$ |

Substitute the values into the formula $\bar{x} = \dfrac{\sum xf}{\sum f}$: $\bar{x} = \dfrac{54}{37} = 1.46$

## 1.2 the median



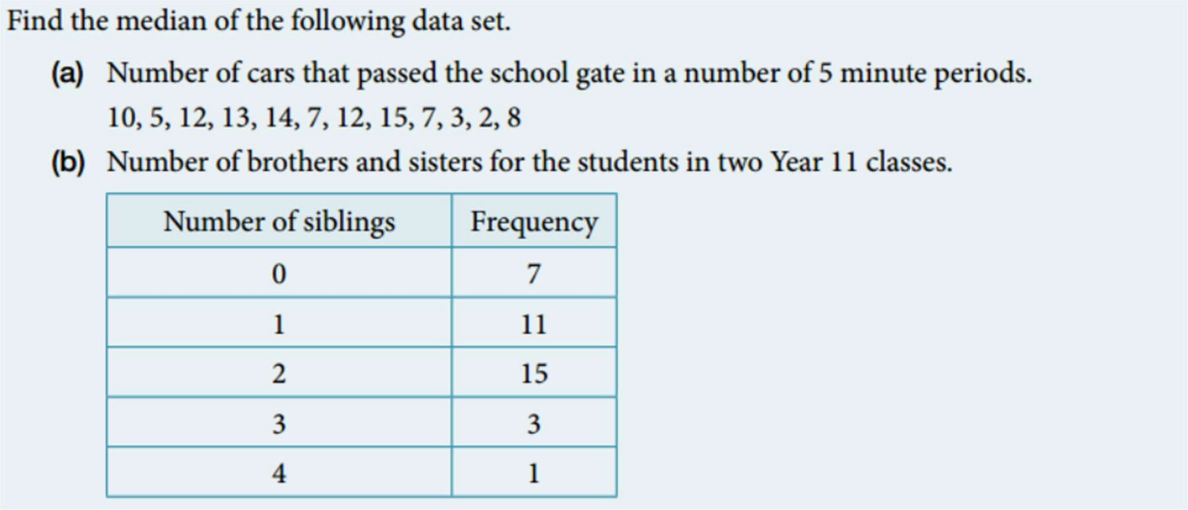
The median is the middle data value of an ordered data list. The method for finding the median depends on whether there is an odd or even number of data values.

For odd $n$, the median is the $\left(\frac{n+1}{2}\right)$th value.

For even $n$, the median is the average of the $\left(\frac{n}{2}\right)$th and the $\left(\frac{n+2}{2}\right)$th values.

When data is grouped, you should state the class interval in which the median falls.

## Example of calculation of the median



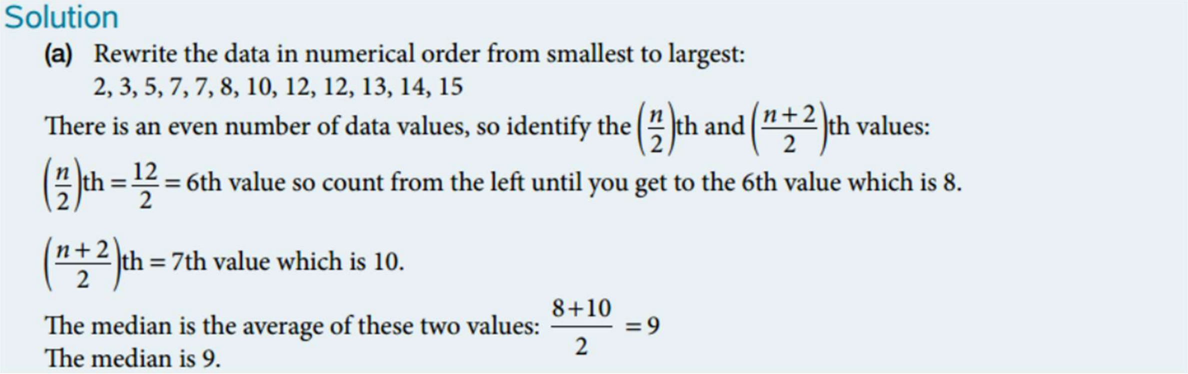
Find the median of the following data set.

**(a)** Number of cars that passed the school gate in a number of 5 minute periods.

10, 5, 12, 13, 14, 7, 12, 15, 7, 3, 2, 8

**(b)** Number of brothers and sisters for the students in two Year 11 classes.

| Number of siblings | Frequency |
|---|---|
| 0 | 7 |
| 1 | 11 |
| 2 | 15 |
| 3 | 3 |
| 4 | 1 |



### Solution

**(a)** Rewrite the data in numerical order from smallest to largest:

2, 3, 5, 7, 7, 8, 10, 12, 12, 13, 14, 15

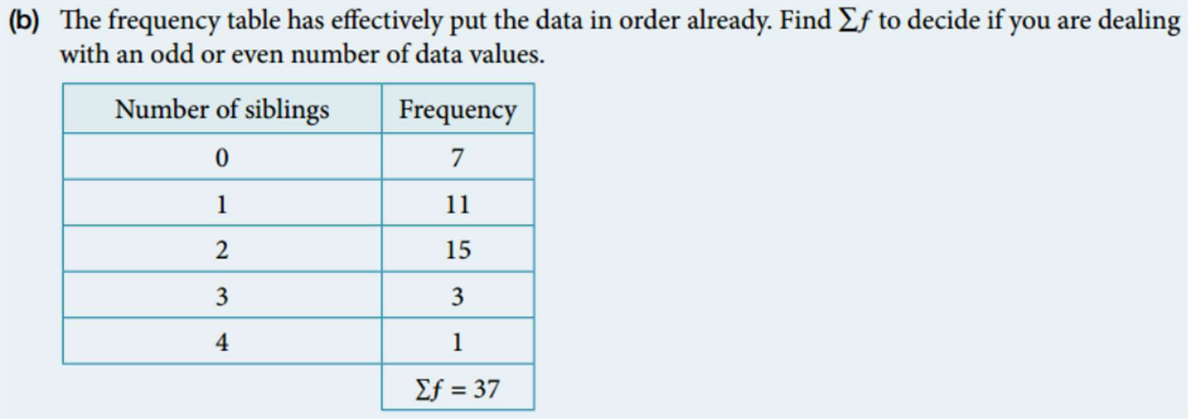
There is an even number of data values, so identify the $\left(\frac{n}{2}\right)$th and $\left(\frac{n+2}{2}\right)$th values:

$\left(\frac{n}{2}\right)$th $= \frac{12}{2} =$ 6th value so count from the left until you get to the 6th value which is 8.

$\left(\frac{n+2}{2}\right)$th = 7th value which is 10.

The median is the average of these two values: $\frac{8+10}{2} = 9$

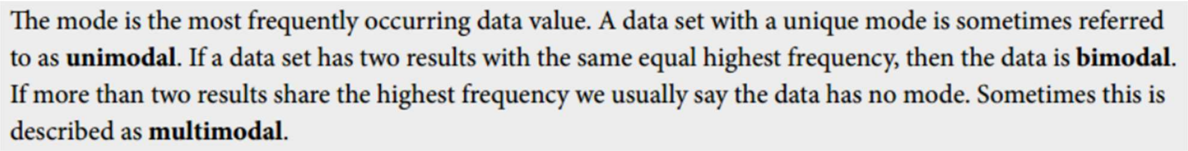
The median is 9.



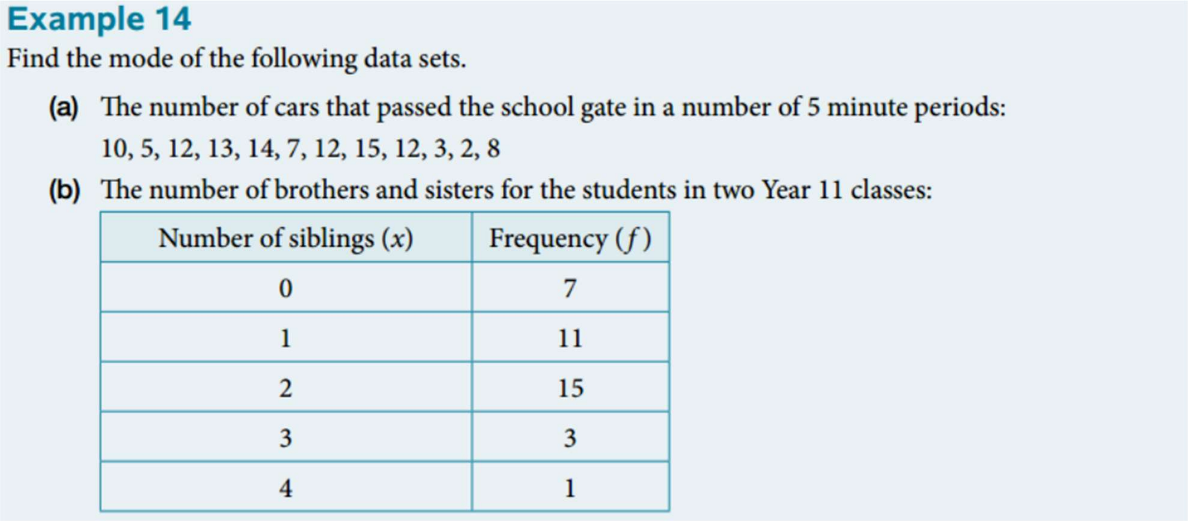
**(b)** The frequency table has effectively put the data in order already. Find $\sum f$ to decide if you are dealing with an odd or even number of data values.

| Number of siblings | Frequency |
|---|---|
| 0 | 7 |
| 1 | 11 |
| 2 | 15 |
| 3 | 3 |
| 4 | 1 |
| | $\sum f = 37$ |

## 1.3 the mode



The mode is the most frequently occurring data value. A data set with a unique mode is sometimes referred to as **unimodal**. If a data set has two results with the same equal highest frequency, then the data is **bimodal**. If more than two results share the highest frequency we usually say the data has no mode. Sometimes this is described as **multimodal**.



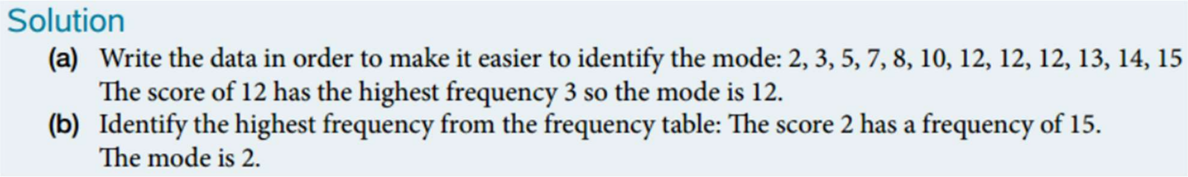
### Example 14

Find the mode of the following data sets.

**(a)** The number of cars that passed the school gate in a number of 5 minute periods:
10, 5, 12, 13, 14, 7, 12, 15, 12, 3, 2, 8

**(b)** The number of brothers and sisters for the students in two Year 11 classes:

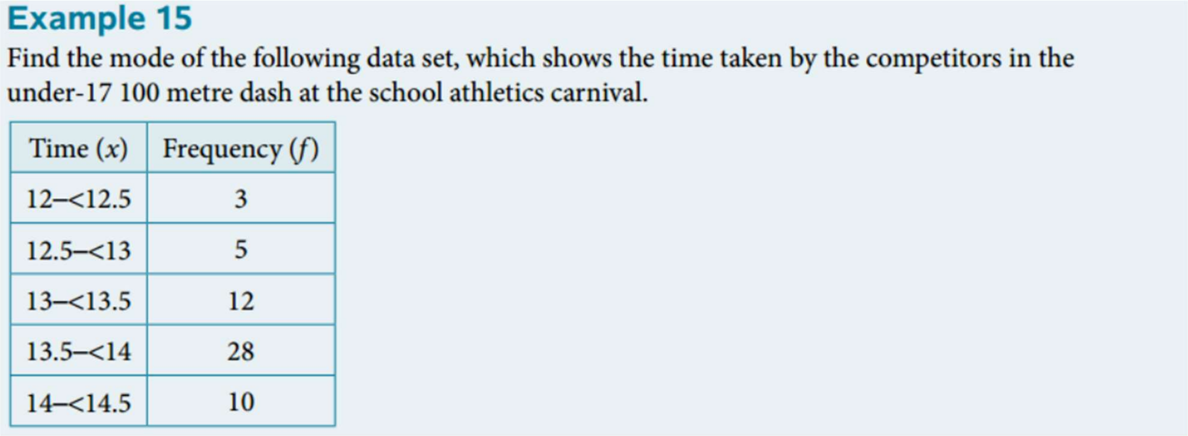
| Number of siblings ($x$) | Frequency ($f$) |
|---|---|
| 0 | 7 |
| 1 | 11 |
| 2 | 15 |
| 3 | 3 |
| 4 | 1 |



### Solution

**(a)** Write the data in order to make it easier to identify the mode: 2, 3, 5, 7, 8, 10, 12, 12, 12, 13, 14, 15
The score of 12 has the highest frequency 3 so the mode is 12.

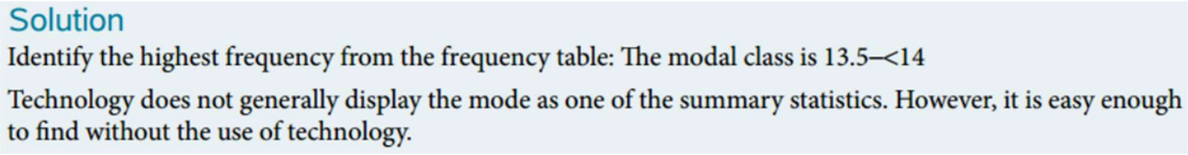
**(b)** Identify the highest frequency from the frequency table: The score 2 has a frequency of 15.
The mode is 2.



### Example 15

Find the mode of the following data set, which shows the time taken by the competitors in the under-17 100 metre dash at the school athletics carnival.

| Time ($x$) | Frequency ($f$) |
|---|---|
| 12–<12.5 | 3 |
| 12.5–<13 | 5 |
| 13–<13.5 | 12 |
| 13.5–<14 | 28 |
| 14–<14.5 | 10 |



### Solution

Identify the highest frequency from the frequency table: The modal class is 13.5–<14

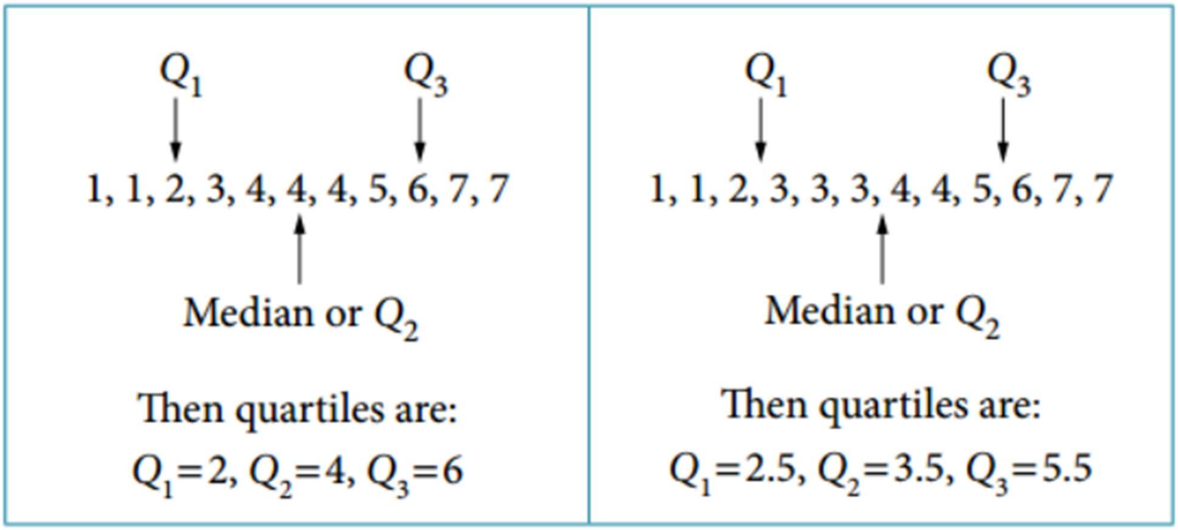
Technology does not generally display the mode as one of the summary statistics. However, it is easy enough to find without the use of technology.

### 1.4 Quartiles

The median divides the data set into two equal 'halves'. The data set could be divided into many equal groups called quantiles. In this subject you will only consider dividing the data into four groups called quartiles.

- The median is referred to as the second quartile, $Q_2$.
- The median of the lower half of the data is called the first quartile $Q_1$ or lower quartile $Q_L$.
- The median of the upper half of the data is called the third quartile $Q_3$ or upper quartile $Q_U$.

Remember to order the data from smallest to largest before you find these values.



Then quartiles are:
$Q_1=2$, $Q_2=4$, $Q_3=6$

Then quartiles are:
$Q_1=2.5$, $Q_2=3.5$, $Q_3=5.5$

## 2 Measure of spread

### 2.1 the range

A measure of spread gives an indication of how close or spread out the values are in a data set. The measure of spread you would have used most frequently is the **range**.

Range = maximum value – minimum value

The range is influenced by extreme values. For example, the following table shows the number of pets owned by each of the students in a class.

| Number of pets ($x$) | Frequency ($f$) |
|---|---|
| 0 | 5 |
| 1 | 9 |
| 2 | 4 |
| 3 | 2 |
| 4 | 2 |
| 5 | 1 |
| 6 | 1 |

Range = 6 – 0 = 6 pets

## 2.2 inter-quartile range (abbreviation IQR)

The interquartile range is a measure of spread based on the quartile values. It is often referred to by the abbreviation IQR.

Interquartile range = third quartile – first quartile

$\text{IQR} = Q_3 - Q_1$

The IQR represents the middle 50% of the ordered data set. It is a more reliable measure of spread than the range because it does not use extreme values at the start or end of the ordered data list.

### Example 16

Find the interquartile range for the data set: 1, 14, 20, 1, 105, 30, 27, 16, 24, 24, 29.

### Solution

Order the data from smallest to largest: 1, 1, 14, 16, 20, 24, 24, 27, 29, 30, 105

Identify $Q_1$, $Q_2$ (median) and $Q_3$ for the data set:

$$1, 1, \overset{\overset{Q_1}{\downarrow}}{14}, 16, 20, \underset{\underset{Q_2\text{(median)}}{\uparrow}}{24}, 24, 27, \overset{\overset{Q_3}{\downarrow}}{29}, 30, 105$$

$$\begin{aligned} \text{IQR} = Q_3 - Q_1 &= 29 - 14 \\ &= 15 \end{aligned}$$

## 2.3 five-number summary

The five-number summary for a data set consists of:

- the minimum value
- the lower quartile
- the median
- the upper quartile
- the maximum value

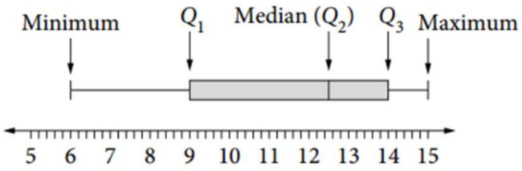
For the data in Example **16** this is:

Minimum = 1, $Q_1 = 14$, Median $Q_2 = 24$, $Q_3 = 29$, Maximum = 105

## 2.4 box plots

The five-number summary can be used to draw a statistical graph called a **box plot** or box-and-whisker plot.

A box plot needs to have a scale line associated with it. The 'box' is the central box structure and the 'whiskers' are the lines going out from the box to the extreme values.



### Example 18

Find the interquartile range and draw a box plot for the data in the table.

| Score | Frequency |
|---|---|
| 3 | 2 |
| 4 | 8 |
| 5 | 9 |
| 6 | 6 |
| 7 | 4 |
| 8 | 1 |

### Solution

Find the sum of the frequencies, $n$: $2 + 8 + 9 + 6 + 4 + 1 = 30$

Identify the position of the median, and its value: The median is between the 15th and 16th values.

The 15th and 16th values are both 5, so median = 5.

Identify the position of $Q_1$, and its value: There are 15 values in the lower half of the data; the middle value here is the 8th.

$Q_1 = 4$

Identify the position of $Q_3$, and its value: There are 15 values in the upper half of the data; the middle value here is the 8th from the end:

$Q_3 = 6$

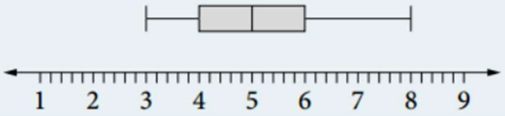
Calculate the interquartile range: $\text{IQR} = Q_3 - Q_1$

$= 6 - 4 = 2$

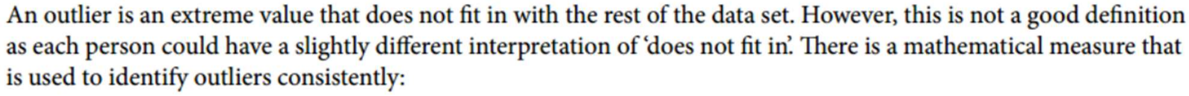
Write the five-number summary as a basis for drawing the box plot:

Minimum = 3<br>
$Q_1 = 4$<br>
Median $Q_2 = 5$<br>
$Q_3 = 6$<br>
Maximum = 8

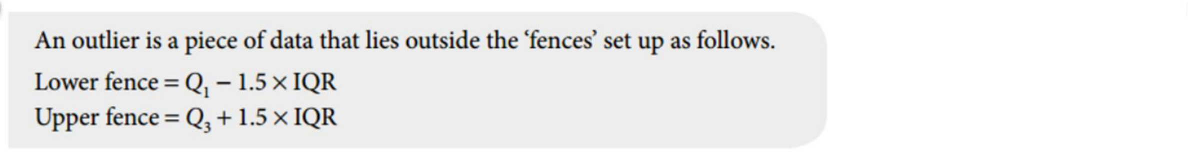
Draw the box plot. (Don't forget to include a scale line.)



## 2.5 outliers



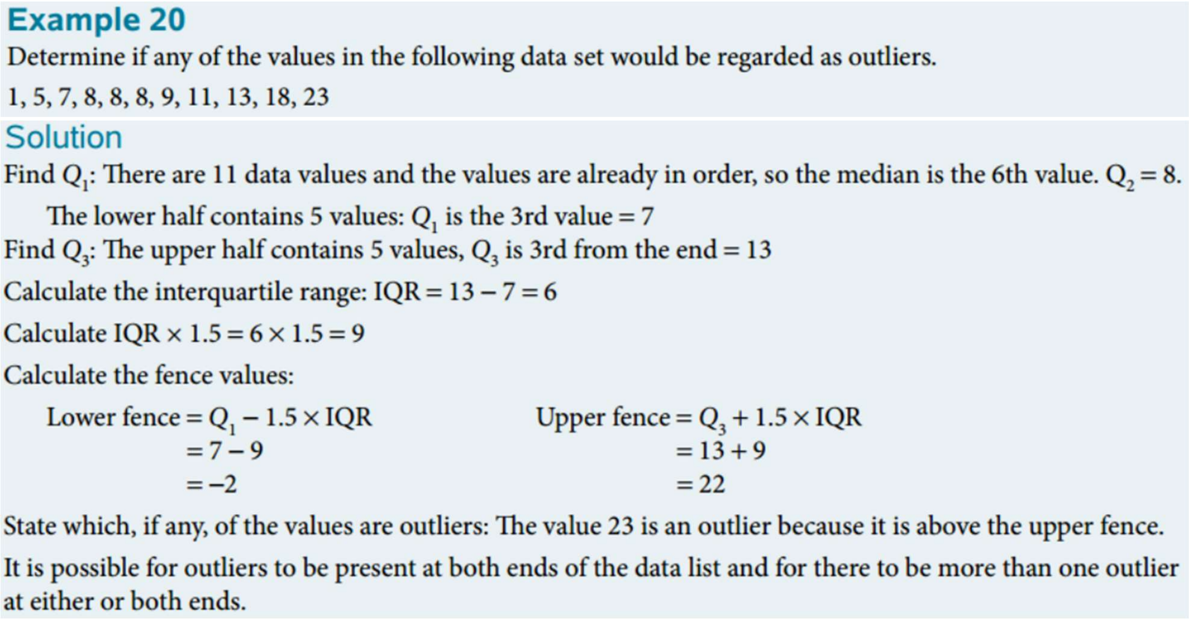
An outlier is an extreme value that does not fit in with the rest of the data set. However, this is not a good definition as each person could have a slightly different interpretation of 'does not fit in'. There is a mathematical measure that is used to identify outliers consistently:



An outlier is a piece of data that lies outside the 'fences' set up as follows.

Lower fence = $Q_1 - 1.5 \times \text{IQR}$

Upper fence = $Q_3 + 1.5 \times \text{IQR}$



### Example 20

Determine if any of the values in the following data set would be regarded as outliers.

1, 5, 7, 8, 8, 8, 9, 11, 13, 18, 23

### Solution

Find $Q_1$: There are 11 data values and the values are already in order, so the median is the 6th value. $Q_2 = 8$.

The lower half contains 5 values: $Q_1$ is the 3rd value = 7

Find $Q_3$: The upper half contains 5 values, $Q_3$ is 3rd from the end = 13

Calculate the interquartile range: IQR = 13 − 7 = 6

Calculate IQR × 1.5 = 6 × 1.5 = 9

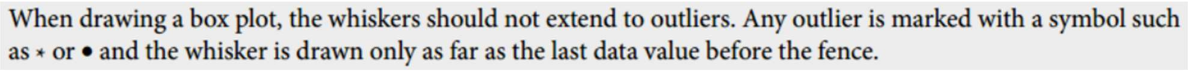
Calculate the fence values:

$$\begin{aligned} \text{Lower fence} &= Q_1 - 1.5 \times \text{IQR} \\ &= 7 - 9 \\ &= -2 \end{aligned} \qquad \begin{aligned} \text{Upper fence} &= Q_3 + 1.5 \times \text{IQR} \\ &= 13 + 9 \\ &= 22 \end{aligned}$$

State which, if any, of the values are outliers: The value 23 is an outlier because it is above the upper fence.

It is possible for outliers to be present at both ends of the data list and for there to be more than one outlier at either or both ends.

Important note:



When drawing a box plot, the whiskers should not extend to outliers. Any outlier is marked with a symbol such as ⋆ or • and the whisker is drawn only as far as the last data value before the fence.