AVERAGES AND RANGE

Averages are measures that give us information about data. Along with the range they allow us to make comparisons between data.

Check first that you:
- understand the different types of data:
  discrete data is data that can only take certain values
  continuous data is data that can take any value
- can order numbers
- understand frequency tables including those for grouped data.

Averages and discrete data There are 3 types of averages that we use:

The median
- This average is the middle value when the data is arranged in order of size.
- If there are two middle values then we find the mean of the two values by adding them and dividing the total by 2.

The mode
- This average is the most common value.
- It is possible have more than one mode and it is also possible to have no mode when all values appear the same number of times.

The mean
- This average is found by dividing the total of all the values by the number of values.

The range is a measure that tells us about the spread of the data. The range = highest – lowest value

E.g. Find the median, mode, mean and range of the following test scores: 12, 15, 12, 13, 10, 15, 19, 8

We place the numbers in ascending order: 8, 10, 12, 12, 13, 15, 15, 19

The median 8, 10, 12, (12, 13) 15, 15, 19 $\frac{12+13}{2} = 12{\cdot}5$ **The range** $19 - 8 = 11$

The mode 8, 10, 12, 12, 13, 15, 15, 19 12 and 15 **The mean** $\frac{8+10+12+12+13+15+15+19}{8} = \frac{104}{8} = 13$

Remember the range tells us how close together the data values are. The smaller the range the more consistent the data is.

Frequency tables Rather than listing the individual data values, it can be easier to display it in a frequency table. E.g. The table shows the number of children in 75 households on Heol Hir. Find the median, mean, mode and range.

Number of children	Frequency
0	16
1	25
2	20
3	9
4	2
5	3

Take care! When finding the mode or range, don't confuse the frequency values with the actual data values.

Median

Adding the frequencies tells us how many households there are.

$16 + 25 + 20 + 9 + 2 + 3 = 75$

The median is the middle value and its position is given by: $\frac{75+1}{2} = 38$.

Number of children (x)	Frequency (f)	Cumulative frequency (cf)	
0	16	16	1st - 16th
1	25	16 + 25 = 41	17th - 41st
2	20	41 + 20 = 61	42nd - 61st
3	9	61 + 9 = 70	62nd - 70th
4	2	70 + 2 = 72	71st - 72nd
5	3	72 + 3 = 75	73rd - 75th

38th

The data is already ordered therefore we add the frequencies (cumulative frequency) to find where the 38th household lies. From the table we see that the median number of children for a household in Heol Hir is 1.

Mean

To find the mean we need the total of all the children living in Heol Hir and then we divide this number by the total number of households.
If we multiply the number of children (x) with the frequency (f) and then add these values we get the total number of children.

Number of children (x)	Frequency (f)	No of children × frequency (fx)
0	16	$0 \times 16 = 0$
1	25	$1 \times 25 = 25$
2	20	$2 \times 20 = 40$
3	9	$3 \times 9 = 27$
4	2	$4 \times 2 = 8$
5	3	$5 \times 3 = 15$
Total	$\sum f = 75$	$\sum fx = 115$

$\text{Mean} = \frac{\text{total number of children}}{\text{total number of households}}$ or $\frac{\sum fx}{\sum f} = \frac{115}{75} = 1{\cdot}5$ (1 d.p.)

Σ means the 'sum of'

Mode

The mode is number of children with the highest frequency. The modal number of children per household is 1 with a frequency of 25.

Range

The highest number of children is 5 and lowest number of children is 0. Therefore the range = 5 – 0 = 5

Continuous grouped data Continuous data i.e. data that can be measured such as height or weight, will also be displayed in a frequency table but the data will be grouped in equal class intervals. E.g. The table shows the time taken for pupils in Year 11 to travel to school. Find an estimate for the mean length of travel time for pupils in Year 11.

Time, t (minutes)	Frequency
$0 < t \leq 10$	24
$10 < t \leq 20$	16
$20 < t \leq 30$	35
$30 < t \leq 40$	11

When the data is grouped, the individual values are not known. Therefore, we find the midpoint of the class interval and use this as an estimate of every value recorded in that group.

Time (minutes)	Midpoint x	Frequency f	Midpoint × frequency (fx)
$0 < t \leq 10$	5	24	$5 \times 24 = 120$
$10 < t \leq 20$	15	16	$15 \times 16 = 240$
$20 < t \leq 30$	25	35	$25 \times 35 = 875$
$30 < t \leq 40$	35	11	$35 \times 11 = 385$
Total		$\sum f = 86$	$\sum fx = 1620$

$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{1620}{86}$

= 19 minutes

(to the nearest minute)

Use the same method as in the example above to find the median or modal class but *remember* to write the class interval as your answer e.g for length of travel time for Year 11 pupils the modal class is $20 < t \leq 30$.

Remember with grouped data we can only find estimates as we don't know the individual data values.

STATISTICS

Constructing and interpreting bar charts, vertical line diagrams, line graphs and pictograms

A graph or a chart can highlight features or trends that a table does not show so well.

To display qualitative or categorical data, you can use a bar chart, pie chart or pictogram.

To display discrete data, you can use a vertical line graph.

Constructing a bar chart

This table shows the number of hours Ben spent on his jobs around the house in one week.

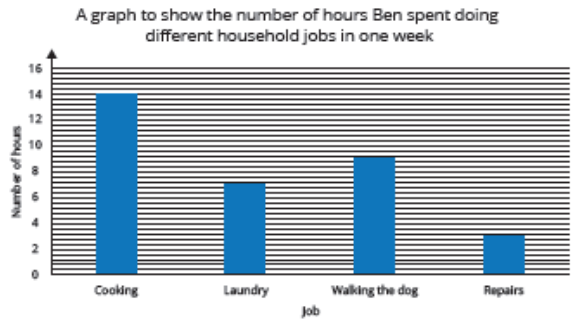
Job	Cooking	Laundry	Walking the dog	Repairs
Hours worked	14	7	9	3

The hours worked is the frequency, and this should go on the y-axis.

The bar widths should be the same.

The scale on the y-axis should be uniform.

Both axes should be labelled, and a title should be given.



Remember:

The type of data you have will determine the type of graph you need to draw. Make sure you understand the difference between qualitative and quantitative data, and for qualitative data, the difference between continuous and discrete data. Have a look at the WJEC Knowledge Organiser on Sorting Data if you need a reminder.

Constructing a line graph

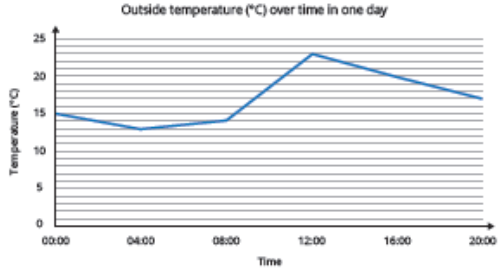
When the data is continuous, a line graph would be appropriate in order to display the data.

For example, you can use a line graph to show the temperature over a period of time. On these graphs, the time should always go on the x-axis.

Example

The table shows the outside temperature, recorded every four hours for one summer's day in Wales.

Time	00:00	04:00	08:00	12:00	16:00	20:00
Temperature (°C)	15	13	14	23	20	17



Constructing a vertical line diagram

This type of diagram is similar to a bar chart, but instead of bars, lines are used. The chart will indicate that the data is discrete (as the lines don't touch each other).

Example

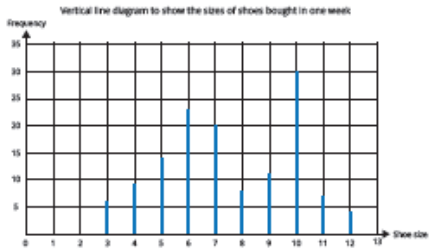
The following table shows the sizes of shoes that different people bought at a shop in one particular week.

Size	3	4	5	6	7	8	9	10	11	12
Frequency	6	9	14	23	20	8	11	30	7	4

The frequency should go on the y-axis.

The scale on the y-axis should be uniform.

Both axes should be labelled, and a title should be given.



Check that you can:

- interpret scales on axes
- understand the difference between qualitative and quantitative data, and continuous and discrete data.

Constructing pictograms

When you want to display your data as a pictogram, you must first pick a symbol that is easy to draw. It should also be symmetrical so that it is clear how many objects it represents if you draw a fraction of it. Write a key to let your reader know how many objects your symbol represents.

Here are some good symbols to use:

Here are some symbols which aren't so good as they are not symmetrical:

Example

The following table shows the type of snack chosen by some pupils at break time.

Type of snack	Apple	Orange	Toast	Yoghurt
Frequency	12	15	7	10

Here is a pictogram to display these results.

Key: represents two pupils.

Snack	
Apples	
Orange	
Toast	
Yoghurt	

Interpreting graphs:

It is important to look carefully at the graph and check that you fully understand the information being displayed. You should:

- read the title
- inspect the axes
- check the scales
- look for a key
- take accurate readings.

CUMULATIVE FREQUENCY DIAGRAMS

When working with continuous grouped data we estimate the different averages as the individual data values are not known, only the class interval they fall into. A cumulative frequency diagram is a good way of estimating one of those averages, the median, and also the interquartile range.

Check first that you:

- understand the difference between the averages as a measure of central tendency and the range as a measure of the spread of the data
- understand how data is grouped into class intervals
- can find averages of grouped data
- can read scales on a graph
- can plot points on a graph
- can find $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$ of a number.

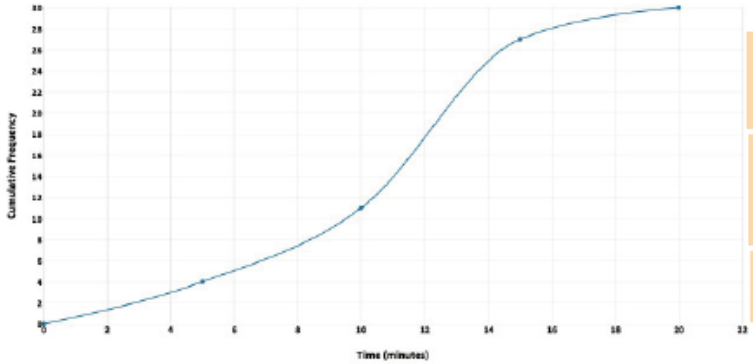
Drawing a cumulative frequency diagram

E.g. The frequency table shows the length of time in minutes it took pupils to complete a puzzle. Draw a cumulative frequency diagram for this data.

Time (minutes)	$0 < t \leq 5$	$5 < t \leq 10$	$10 < t \leq 15$	$15 < t \leq 20$
Frequency	4	7	16	3

We draw a cumulative frequency diagram by plotting the upper boundary of each class against the cumulative frequency therefore we use the information in the table to create a cumulative frequency table.

Time (minutes)	≤ 0	≤ 5	≤ 10	≤ 15	≤ 20
Cumulative frequency	**0**	**0 + 4 = 4**	4 + 7 = **11**	11 + 16 = **27**	27 + 3 = **30**



In this example time should be on the horizontal axis and the cumulative frequency should be on the vertical axis.

Plot each upper group boundary against the cumulative frequency ensuring that the cumulative frequency of the last point is equal to the total frequency.

Connect the points using a ruler or by free hand to get a smooth curve.

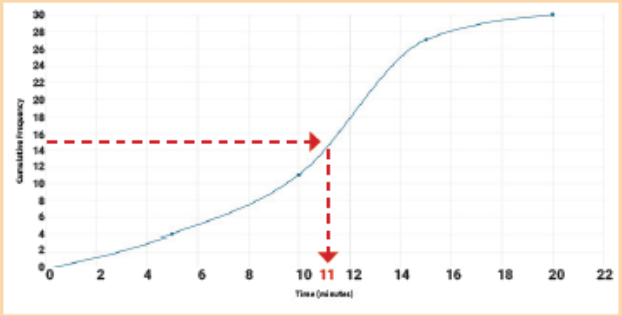
Using a cumulative frequency diagram

We use a cumulative frequency diagram to find the following:

Median

This is the middle value of the data. It is located at ½ of the total frequency.

To find the median, draw a straight line from this ½ way value on the vertical axis (cumulative frequency) across to the curve. Where it meets the curve, draw a line down to the horizontal axis and read off the value. In the diagram below, the total frequency is 30 so we draw a line from 15 on the vertical axis to the curve.



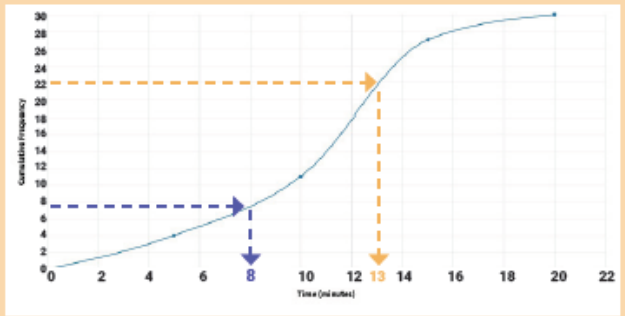
We can see that the median length of time it took to solve the puzzle was **11** minutes to the nearest minute.

Interquartile range

This is a measure of the spread of the middle 50% of the data.

Interquartile range = upper quartile − lower quartile.

The lower quartile is found at a ¼ of the total frequency and the upper quartile is found at ¾ of the total frequency. In the diagram below the total frequency is 30, so we draw a line from 7·5 (LQ) and 22·5 (UQ) on the vertical axis, to the curve and then draw a line down to the horizontal axis to read off the two values.



We can see that the LQ is **8** and the UQ is **13** therefore the interquartile range is 13 − 8 = 5 minutes.

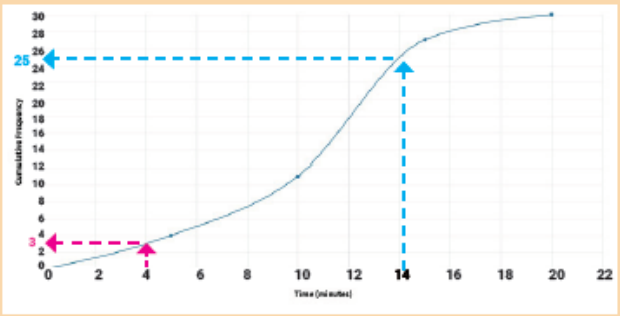
Less than and more than values

In this example, if we wanted to find the number of pupils that took

a) less than 4 minutes,

b) more than 14 minutes to complete the puzzle,

we would draw lines from the horizontal axis (time) to the curve and read off the values on the vertical axis to get the cumulative frequency. The diagram always shows values for 'less than', therefore to obtain a 'more than' value, the 'less than' value needs to be subtracted from the total frequency.



a) 3 students

b) 30 − 25 = 5 students

You may be asked to use cumulative frequency diagrams to compare two sets of data, ***remember*** to compare the median values for each set and also the interquartile range.
Remember a smaller interquartile range shows more consistent data and a larger interquartile range less consistent data.

STATISTICS

Grouped frequency distributions — This is where data is arranged in groups within a range of values.

Check that you can:

- recognise the inequality symbols $\leqslant$, $\geqslant$, $<$ and $>$
- interpret simple inequalities such as $x > 7$ or $x \leqslant 4$
- group discrete and continuous data into a grouped frequency table
- display grouped data in a frequency diagram.

Displaying grouped data: frequency polygons

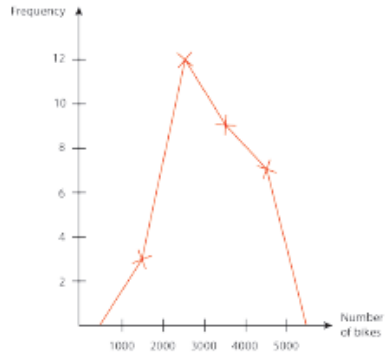
A visual way of displaying the data in grouped frequency tables is by constructing a **frequency polygon**. Frequency polygons are best used to display more than one set of frequencies on the same graph.

The information in the following example is taken from question 7 in the Mathematics – Numeracy Unit 2, Intermediate Tier, Autumn 2018 paper.

Here are the recorded number of bikes made each day by *Tube Cycles*:

Number of bikes, b	Frequency
$1000 \leqslant b < 2000$	3
$2000 \leqslant b < 3000$	12
$3000 \leqslant b < 4000$	9
$4000 \leqslant b < 5000$	7

Here is a frequency polygon of this information:



The frequency polygon is constructed by joining the mid-points of each class interval.

The mid-point is the value halfway between the limits of the class interval.

For example, for the class interval $1000 \leqslant b < 2000$, the midpoint is 1500.

Note that:

- the frequency always goes on the vertical axis
- each axis is labelled
- there is a linear, uniform scale on each axis
- you should never label the horizontal axis with the groups
- the scales do not need to start at 0 (although they do in this example).

Estimating the mean from a grouped frequency table

To estimate the mean from a grouped frequency table, we must follow these steps:

1. Find the mid-point of each class interval.
2. Multiply the mid-point for each group with the frequency for that group.
3. Find the total of the *frequency × mid-point* values.
4. Find the total of the frequencies.
5. Divide the total of the *frequency × mid-point* by the total of the frequencies.

The information in the following worked example has been taken from question 9 in the Mathematics Unit 3, Higher Tier, Winter 2015 examination paper.

The amount of money that some customers spend in a supermarket on a Saturday afternoon is shown in this table.

Find the estimate of the mean amount of money that the customers spent in a supermarket on a Saturday afternoon.

Amount, s (£)	Frequency
$0 < s \leqslant 20$	5
$20 < s \leqslant 40$	19
$40 < s \leqslant 60$	34
$60 < s \leqslant 80$	12
$80 < s \leqslant 100$	12
$100 < s \leqslant 120$	10
$120 < s \leqslant 140$	8

After following the first four steps to estimate the mean, the table will look like this:

Amount, s (£)	Frequency	Midpoint	*Frequency × midpoint*
$0 < s \leqslant 20$	5	10	5 × 10 = 50
$20 < s \leqslant 40$	19	30	19 × 30 = 570
$40 < s \leqslant 60$	34	50	34 × 50 = 1700
$60 < s \leqslant 80$	12	70	12 × 70 = 840
$80 < s \leqslant 100$	12	90	12 × 90 = 1080
$100 < s \leqslant 120$	10	110	10 × 110 = 1100
$120 < s \leqslant 140$	8	130	8 × 130 = 1040
	100		6380

Then for the final step, divide the total of the *frequency × midpoint* by the total of the frequencies:

Estimate of the mean = £6380 ÷ 100 = £63.80

Remember to look at the table for the correct units. Also, check that your answer is sensible!

Your answer should be between £0 and £140.

Remember:

A grouped frequency table is a table that contains data that has been organised into different groups or class intervals.

Finding the class interval which contains the median from a grouped frequency table

If the data is in a grouped frequency table, we do not know the exact value of each item of data, just which group it belongs to. Therefore, we cannot find the exact value for the median, but we can find the group that contains the median.

The median is the middle number or value in a set of data which has been placed in ascending (or descending) order. The middle value is the $\frac{n+1}{2}$th value, where n is the total frequency. We always use the median as the $\frac{n+1}{2}$th value for small sets of data. For example, with 11 data values arranged in order, the median is the $\frac{11+1}{2}$ = 6th data value. For 30 data values, it would be the 15½th data value (midway between the 15th and 16th).

For sets of data with many data values, it is usually acceptable to use the $\frac{n}{2}$th value instead of the $\frac{n+1}{2}$th value as they should be very close.

Example

The amount of money that some customers spend in a supermarket on a Saturday afternoon is shown in this table.

Find the group that contains the median amount of money.

Amount, s (£)	Frequency
$0 < s \leqslant 20$	5
$20 < s \leqslant 40$	19
$40 < s \leqslant 60$	34
$60 < s \leqslant 80$	12
$80 < s \leqslant 100$	12
$100 < s \leqslant 120$	10
$120 < s \leqslant 140$	8

Answer

The total frequency is 100. We need to find the $\frac{100}{2}$th = 50th value in the table. To do this, we will work our way down the frequency column, adding up the frequencies as we go until we hit 50.

Amount, s (£)	Frequency	
$0 < s \leqslant 20$	5	5
$20 < s \leqslant 40$	19	5 + 19 = 24
$40 < s \leqslant 60$	34	34 + 24 = 58
$60 < s \leqslant 80$	12	
$80 < s \leqslant 100$	12	
$100 < s \leqslant 120$	10	
$120 < s \leqslant 140$	8	

The 50th value is in the $40 < s \leqslant 60$ group.

The group that contains the median amount of money = £40 < s ⩽ £60.

Note: if we found the $\frac{100+1}{2}$th value, then the median would be the 50·5th value, which lies between the 50th and 51st value. Both of these values lie in the £40 < $s \leqslant$ £60 group.

STATISTICS

Grouped frequency distributions — This is where data is arranged in groups within a range of values.

Check that you can:

- recognise the inequality symbols $\leqslant$, $\geqslant$, $<$ and $>$
- interpret simple inequalities such as $x > 7$ or $x \leqslant 4$.

It is often easier to interpret data if it is put into a frequency table.

Where we have a large amount of data and the number of different data values is too many to list individually, we can construct a **grouped frequency table**.

If the data is continuous, then the class intervals or groups will contain inequalities such as $<$ and $\leqslant$.

Make sure you understand the following definitions.

Frequency table

A frequency table contains a set of observations showing how often something occurs.

Continuous data

If a set of data is continuous, any number within a specific range can appear in that set.

Discrete data

If a set of data is discrete, only specific numbers can appear in that set.

Inequality

An inequality contains two expressions separated by one of the following symbols:

$\leqslant$ Less than or equal to	$<$ Less than
$\geqslant$ Greater than or equal to	$>$ Greater than

Grouped frequency table using discrete data

Below is an example of a grouped frequency table using **discrete data**. Here are the number of customers that visited a market stall over 30 consecutive days:

16	9	15	14	8	17
9	19	5	19	12	5
10	13	13	20	11	9
12	15	6	4	13	17
17	15	3	15	11	18

Here is the completed grouped frequency table for this data:

Number of customers	Frequency
0 – 4	2
5 – 8	4
9 – 12	8
13 – 16	9
17 – 20	7

The modal group is 13 – 16.

Grouped frequency table using continuous data

Below is an example of a grouped frequency table using **continuous data**.

The monthly average temperatures for a city are recorded each month for a year.

5·3 °C	5·1 °C	6·5 °C	9·0 °C
10·8 °C	14·7 °C	16·4 °C	16·1 °C
15·0 °C	11·5 °C	8·1 °C	5·9 °C

We can put these temperatures into a grouped frequency table.

Temperature is an example of continuous data.

Below is an example of a grouped frequency table with six class intervals or groups.

The temperature is represented by x.

It is very important that you look at the inequalities carefully.

Temperature, x °C	Frequency
$5 < x \leqslant 7$	
$7 < x \leqslant 9$	
$9 < x \leqslant 11$	
$11 < x \leqslant 13$	
$13 < x \leqslant 15$	
$15 < x \leqslant 17$	

This interval or group will contain all the temperatures that are greater than **11 °C, but less than or equal to 13 °C. Note that 11 °C itself will be in the $9 < x \leqslant 11$ class interval.**

Look at how the following table is different.

Temperature, x °C	Frequency
$5 < x \leqslant 7$	
$7 < x \leqslant 9$	
$9 < x \leqslant 11$	
$11 < x \leqslant 13$	
$13 < x \leqslant 15$	
$15 < x \leqslant 17$	

This interval or group will contain all the temperatures that are greater than or equal to **11 °C, but less than 13 °C. Note that 13 °C itself will be in the $13 \leqslant x < 15$ class interval.**

Where the width of groups is the same, the modal group is the one which has the highest frequency.

Frequency diagram

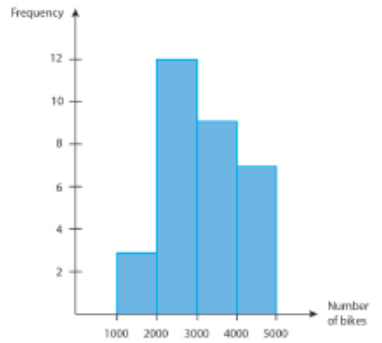
A visual way of displaying the data in grouped frequency tables is by constructing a **frequency diagram**.

The information in the following example is taken from question 7 in the Mathematics – Numeracy Unit 2, Intermediate Tier, Autumn 2018 paper.

Here are the recorded number of bikes made each day by *Tube Cycles*:

Number of bikes, b	Frequency
$1000 \leqslant b < 2000$	3
$2000 \leqslant b < 3000$	12
$3000 \leqslant b < 4000$	9
$4000 \leqslant b < 5000$	7

Here is a frequency diagram of this information:



The bar is drawn at the group limits, e.g. 1000 and 2000 for the first bar.

The height is the frequency, e.g. 3 for the first bar.

There are no gaps between the bars in the frequency diagram as the data is continuous.

Note:

- The frequency always goes on the vertical axis.
- Each axis is labelled.
- There is a linear, uniform scale on each axis. You should never label the horizontal axis with the groups.
- The scales do not need to start at 0 (although they do in this example).

Remember:

If a set of data is continuous, any number within a specific range can appear in that set. When you draw a frequency diagram for continuous data, there are **no gaps** between the bars. If a set of data is discrete, only specific numbers can appear in that set.

HIGHER PROBABILITY

Understanding conditional probability, and the difference between dependent and independent events.
How to use tree diagrams for dependent events and to calculate probability of dependent events.

Check that you can:

- *multiply two or more fractions together*
- *calculate the probability of compound events.*

INDEPENDENT EVENTS

This is the term used to describe events that have no connection to each other, but it is possible to calculate their probability.
Two events are independent when the outcome of the first event does not affect the outcome of the second event.
The multiplication rule for independent events states that:

$$P(A \cap B) = P(A) \times P(B).$$

For example, the probability of throwing a six on a dice and throwing a head on a fair coin is:

$$P(\text{six and head}) = P(\text{six}) \times P(\text{head})$$
$$\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}.$$

The probability of throwing a six on a dice twice is:

$$P(\text{two sixes}) = P(\text{six}) \times P(\text{six})$$
$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

DEPENDENT EVENTS

This term is used in the context of probability. It describes an event that is affected by a previous event.
Consider the following situation. There are five counters in a bag, two are red and three are blue. Every time a counter is removed from the bag **without replacement**, the probability of choosing a blue counter or a red counter changes.
For example, two counters are removed from the bag without replacement. The probability of choosing a red counter the first time and a red counter the second time is:

$$\frac{2}{5} \times \frac{1}{4} = \frac{2}{20} = \frac{1}{10}$$

The probability of choosing a red counter the first time and a blue counter the second time is:

$$\frac{2}{5} \times \frac{3}{4} = \frac{6}{20} = \frac{3}{10}$$

When the probability of an event is **dependent on the outcome of another event**, it is described as **conditional probability**.

EXAMPLE

A bag contains three green counters and seven purple counters.

Three counters are taken from the bag and placed on a table.
What is the probability that the second counter placed on the table is purple if:

a) The first counter put on the table was green?
Selecting a purple counter after one green counter has already been selected means that there are still seven purple counters remaining, but only nine counters in total.

$$\text{Answer} = \frac{7}{9}$$

b) The first counter put on the table was purple?
Selecting a purple counter after one purple counter has already been selected means that there are still six purple counters remaining, but only nine counters in total.

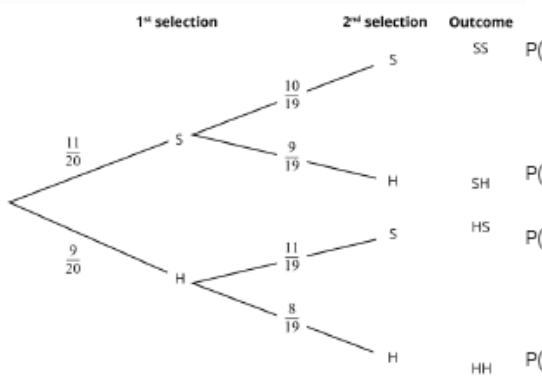
$$\text{Answer} = \frac{6}{9} = \frac{2}{3}$$

Sampling or selecting items **without replacement** is an example of where you would have conditional probability. Consider the counters taken from the bag in the example. When the first counter is taken, depending on its colour, it will affect the probability of selecting a purple or green counter the second time and so on. So, when considering conditional probability, you need to remember that the probability of the second event will alter depending on the outcome of the first event (or any further events depending on previous events).

USING TREE DIAGRAMS FOR DEPENDENT EVENTS

EXAMPLE

A box of chocolates contains 11 soft centres and 9 hard centres. One chocolate is chosen at random and eaten. A second chocolate is then chosen, again at random, and eaten.
Below is a tree diagram showing all the possible outcomes. The probabilities of each selection have been added to each branch.



$$P(SS) = \frac{11}{20} \times \frac{10}{19} = \frac{110}{380} = \frac{11}{38}$$

$$P(SH) = \frac{11}{20} \times \frac{9}{19} = \frac{99}{380}$$

$$P(HS) = \frac{9}{20} \times \frac{11}{19} = \frac{99}{380}$$

$$P(HH) = \frac{9}{20} \times \frac{8}{19} = \frac{72}{380} = \frac{18}{95}$$

Take note of the second set of branches. As the first chocolate selected has been eaten, it isn't possible for this chocolate to be selected a second time. This is why the totals are out of 19. The probabilities of all the possible outcomes have been calculated using the information on the tree diagram and are displayed alongside the end of each branch.

Remember to check that all possible outcomes add to one.

$$\frac{11}{38} + \frac{99}{380} + \frac{99}{380} + \frac{18}{95} = 1$$

Note, it is often easier to check that the sum of the possible outcomes is one if the probabilities are not simplified.

$$\frac{110}{380} + \frac{99}{380} + \frac{99}{380} + \frac{72}{380} = \frac{380}{380} = 1$$

REMEMBER! The probabilities at the end of the branches should all sum to 1.

STATISTICS

Histograms — A histogram is used instead of a frequency diagram when the class intervals of a grouped frequency table are of unequal width.

Check that you can:

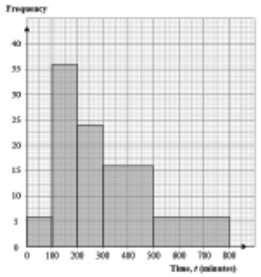
- calculate areas of squares and rectangles
- find the missing length or width of a square or rectangle, given the area and either a length or width
- group data into a grouped frequency table
- transfer data into a graph format, selecting appropriate scales for the axes.

Histograms

The number of minutes a group of people spent watching a certain television channel on a particular day is shown on the frequency table below.

Time, t (minutes)	Frequency
$0 \leqslant t < 100$	6
$100 \leqslant t < 200$	36
$200 \leqslant t < 300$	24
$300 \leqslant t < 500$	16
$500 \leqslant t < 800$	6

An **incorrect** frequency diagram which attempts to represent this information is shown below:



Notice that the class intervals are not equal.

The first three class intervals have a class width of 100 minutes.

$300 \leqslant t < 500$ has a class width of 200 minutes.

$500 \leqslant t < 800$ has a class width of 300 minutes.

Because the class intervals are unequal, the frequency diagram seen above does not represent the data fairly. Look at the size of the bars for $0 \leqslant t < 100$ and $500 \leqslant t < 800$, where both have a frequency of six. This implies that there are six people in each of the groups $500 \leqslant t < 600$, $600 \leqslant t < 700$ and $700 \leqslant t < 800$ which is clearly incorrect.

Instead of a frequency diagram, a **histogram** is used when the class intervals are of unequal width.

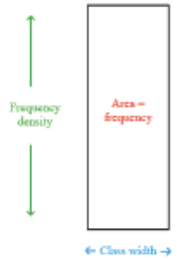
In a histogram, the **area of the bars represents the frequency**.

By multiplying the width and the height of the bar together, we get the frequency of that class interval.

The width of the bar is represented by the class, or interval width.

The height of the bar is called **frequency density**.

To find the frequency density of a class interval, we divide the area (frequency) by the class width.



Finding the frequency densities

The class width of the first class interval = 100 – 0 = 100

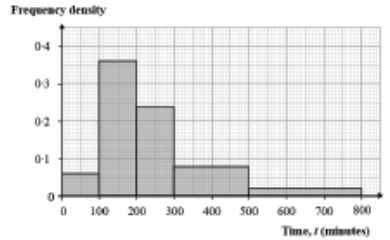
The frequency density of the first class interval = 6 ÷ 100

= 0·06

Time, t (minutes)	Frequency	Class width	*Frequency density = frequency ÷ class width*
$0 < t \leqslant 100$	6	100	0·06
$100 < t \leqslant 200$	36	100	0·36
$200 < t \leqslant 300$	24	100	0·24
$300 < t \leqslant 500$	16	200	0·08
$500 < t \leqslant 800$	6	300	0·02

We can use this information to draw a histogram. The frequency density will always be on the vertical axis.

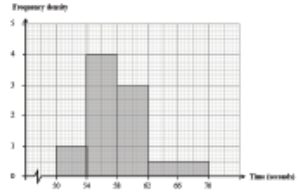
Here you can see the completed histogram.



Example

The time taken to run a distance of 400 m was recorded for each member of a running club.

The histogram below shows the results for the members who are under 30 years of age.



To calculate how many members of the running club are under 30 years of age, we need to look at the areas of the bars.

The area of each bar gives the frequency. To find the total frequency, find the area of each bar and add them together.

Remember:
Frequency = class width × frequency density

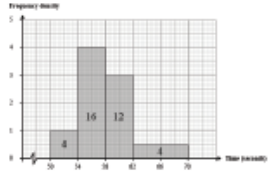
$4 \times 1 = 4$

$4 \times 4 = 16$

$4 \times 3 = 12$

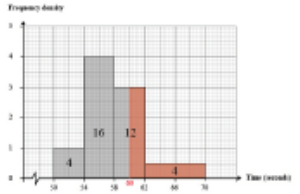
$8 \times 0{\cdot}5 = 4$

$4 + 16 + 12 + 4 =$ 36 members



These frequencies can be written in the bars for use later.

To calculate (to the nearest whole number) an estimate of the percentage of members that took more than one minute to run 400 m, we need to find the area of the bars shaded red in the diagram below.



As the frequencies have already been written on the bars, the answer is $12 + 2 + 4 = 10$.

This can also be calculated using $2 \times 3 = 6$ and $8 \times 0{\cdot}5 = 4$.

The number of members that took more than one minute to run 400 m

$= 6 + 4$

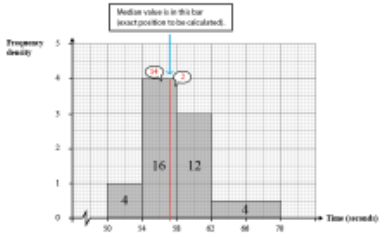
$= 10$.

The percentage of members that took more than one minute to run 400 m

$= \frac{10}{36} \times 100\% = 27{\cdot}777777\ldots\%$

$= 28\%$ to the nearest whole number.

To calculate an estimate of the median time taken by the under-30s to run 400 m, we consider the following. As the median value is the middle value, we need to find the value which divides the total area in half. We therefore need the value with areas of $36 \div 2 = 18$ both below and above it.



By looking at the areas of the bars, since $4 + 16 = 20$, the median lies in the bar between 54 and 58 seconds.

For the second bar we need a frequency of 14, so we need to find w.

Where:
$4 \times w = 14$ then $w = 14 \div 4 = 3{\cdot}5$.

So, our estimate of the median is
$54 + 3{\cdot}5 = 57{\cdot}5$ seconds.

Remember:

In a histogram, the area of the bars represents the frequency.
Frequency density = frequency ÷ class width.

PROBABILITY

Probability tells us how likely an event will happen.

We can use a scale from 0 to 1 to describe the probability of an event.

We can estimate the probability that an even can occur.

Check that you can:

- *convert fractions to decimals and percentages*
- *add and subtract fractions*
- *find a fraction of a number.*

FINDING PROBABILITY

There are several terms we use to describe the probability of an event happening.

We can show probability on a probability scale which is always between 0 (impossible) and 1 (certain).

E.g. A fair coin has an *even chance* of landing on Heads.

Probabilities can be written as fractions, decimals or percentages.

If an event has a probability of 0, then the event is impossible.

If the event has a probability of 1 (or 100%), it is certain of happening.

The probability of an event happening is:

$$\frac{\text{the number of times the outcome can happen}}{\text{the total number of possible outcomes}}$$

The probability of choosing a red ball is:

$$\frac{\text{the number of red balls in the bag}}{\text{the total number of balls in the bag}} = \frac{6}{10}.$$

We can write this as: $P(\text{red}) = \frac{6}{10}$.

We could also write it as a decimal: $P(\text{red}) = 0{\cdot}6$, or as a percentage: $P(\text{red}) = 60\%$.

What is the probability of choosing a purple ball from the bag?

TOTAL PROBABILITY IS ALWAYS 1

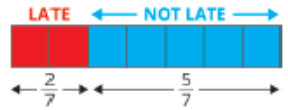
If you throw a fair coin, the probability it lands on heads is $\frac{1}{2}$ and the probability it lands on tails is also $\frac{1}{2}$.

Throwing a coin only has two outcomes and the sum of the probabilities of **all possible outcomes is 1**.

So $P(\text{head}) + P(\text{tail}) = \frac{1}{2} + \frac{1}{2} = 1$.

Example 1: The probability that Mrs Khan is late for work is $\frac{2}{7}$.

There are two possible outcomes for Mrs Khan: being late and not being late.



The sum of the probabilities must add up to 1, so the probability of not being late is:

$$1 - \frac{2}{7} = \frac{5}{7}.$$

Example 2: The probability that a football team wins its next game is 0·7.

There are two possible outcomes for the football team: winning and not winning. Note that 'not winning' could mean losing or drawing the game.

The sum of the probabilities must add up to 1. So, the probability of not winning is:

$$1 - 0{\cdot}7 = 0{\cdot}3.$$

Example 3: The probability that it will rain in a town tomorrow is 89%.

There are two possible outcomes for the weather in the town tomorrow: raining and not raining.

The sum of the probabilities must add up to 1, so the probability of it not raining is:

$$100\% - 89\% = 11\%.$$

FINDING THE EXPECTED NUMBER OF TIMES AN EVENT WILL OCCUR

We know the probability of rolling a 5 with a fair, six-sided dice is $\frac{1}{6}$.

We can use this to calculate how many times we would expect to see a 5 if we rolled the dice 300 times.

Expected number of 5s rolled = probability of rolling a 5 × number of times the dice is rolled.

Expected number of 5s rolled = $\frac{1}{6} \times 300 = 50$.

REMEMBER! If you are adding and subtracting fractions, you need to make the denominators equal.

STATISTICS

Questionnaires and surveys — A questionnaire is a set of questions to ask an individual. A survey is the process of collecting, analysing and interpreting data from many individuals.

Check that you can:

- work with numerical data
- group data and present it in a table
- think critically about information you are given and express your ideas clearly in written form.

Make sure you understand the meanings of these key words.

Hypothesis

A hypothesis is a statement proposed on the basis of little or no evidence. This is often the starting point for an enquiry or investigation. A hypothesis can be tested to find out whether it is true or not. For example, a hypothesis that large dogs are better at catching a ball can be tested using hundreds of different sized dogs.

Biased

A question is biased if it tries to suggest the answer or does not allow a full range of answers. A questionnaire is biased if it contains biased questions or the sample of people chosen does not represent the whole population under study.

Questionnaires and surveys

If you need to collect information to prove a **hypothesis** or to collect people's opinions, one way this can be done is to create a **questionnaire** and undertake a **survey**.

Questionnaires need to be designed carefully to ensure that they are:

- easy to understand
- not **biased**
- not personal.

Surveys need to be planned carefully to ensure that a representative sample of the population under study is included.

The population under study is the whole group of people whose opinions are required.

Examples:

For a survey of favourite school subject, the population under study would be all pupils in school.

For a survey of favourite food of toddlers, the population under study would be parents or carers of toddlers.

Questionnaires should not be biased or have leading questions.

For example, you want school pupils' opinions on which subject taught in school is the most important.

You ask some pupils:

'I think Mathematics is the most important subject taught in school. Which of the subjects taught in school do you think is the most important?'

This question is biased towards Mathematics. It has an opinion, therefore making it a 'leading question'. Some pupils may just agree with the person asking the question. Therefore, the results of the survey may be unreliable.

A better question would be:

'Which of the subjects taught in school do you think is the most important?'

You should consider:

- who you are asking
- when you are asking them (i.e. the time of day)
- where you are asking them (i.e. the location).

If we consider the question *'Which of the subjects taught in school do you think is the most important?'*, you want the opinions of school pupils on the topic. Therefore, you would not go and stand on a high street during a school day and ask the first 100 people that walked past! You would get a better idea if you went to a school and asked school pupils.

If you wish to get the opinions that represent a wide range of different people of all ages, then you need to consider carrying out the survey in a place and at a time that ensures that the people likely to be there are representative of the whole population under study.

Are the questions appropriate?

Some people may be offended if you asked them their age, where they live or any other personal question for that matter. You must think of whether the question is relevant to what you are trying to find out.

Is it easy to collect the answers?

You may wish to include categories with your questions. These categories make it easier to collect data from the questionnaire.

If there aren't any categories, you may receive a number of different responses, and these may be difficult to analyse.

Remember:

You should make sure your questionnaire is not biased and that it does not ask leading questions. Having a choice of categories will help you to collect and analyse the responses to your questions.

Choice of categories

Having a choice of categories will make it easier to collect and analyse the responses.

How many times a week do you go to the supermarket?
0
1 – 3
4 – 7
More than 7

Your categories must not overlap.

Here is an incorrect example:

How many times a week do you go to the supermarket?
0
1 – 3
3 – 7
More than 7

Which box would you tick if you visited the supermarket **three** times a week?

Your categories must not have gaps and you should have considered all possible responses.

Here is an incorrect example:

How many times a week do you go to the supermarket?
0
1 – 2
4 – 7

Which box would you tick if you visited the supermarket **three** times a week?

Which box would you tick if you visited the supermarket **eight** times a week?

You need to be specific about the timescale.

Here is an incorrect example:

How many times do you go to the supermarket?
0
1 – 3
4 – 7
More than 7

What is the timescale in this question? Some people may answer based on how often they visit a supermarket in a day, a week, a month or some may even consider how often in a year!

RELATIVE FREQUENCY

How to find the relative frequency or estimation of probability of an event as the proportion of times it has occurred.

Comparing an estimated probability from experimental results with a theoretical probability.

Check that you can:

- *determine the probability of an event*
- *convert between fractions, percentages and decimals.*

Probability is the term used when discussing how likely it is that an event will happen.

Probabilities can be written as fractions, decimals or percentages.

EXAMPLE

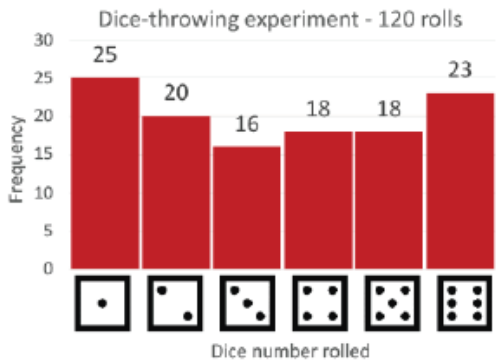
If a fair, six-sided dice is thrown, the theoretical probability of throwing a four is:

$$= \frac{1}{6}$$

$$= 0{\cdot}1666$$

$$= 16{\cdot}666...\%$$

However, if a fair, six-sided dice is thrown 120 times, you may not get exactly twenty 4s ($\frac{1}{6} \times 120$), although it should be close to this amount. If you do an experiment of throwing a dice 120 times, you might get a result that looks like the following:



The number of times a 4 appears is the **relative frequency** or **experimental probability** of throwing a 4.

The number 4 appeared 18 times, so the relative frequency or the experimental probability of throwing a 4 is:

$$= \frac{18}{120}$$

$$= 0{\cdot}15$$

$$= 15\%$$

The more times we throw the dice, the more reliable the relative frequency is as an estimate of the probability.

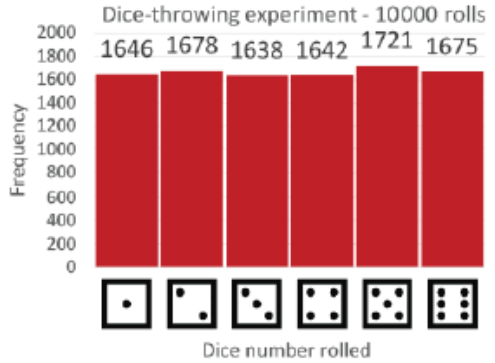
Here are the results after throwing the dice 10000 times. The relative probability of throwing a 4 is:

$$= \frac{1642}{10000}$$

$$= 0{\cdot}1642$$

$$= 16{\cdot}42\%$$

We can see from the above that as the number of trials (throws) increases, the relative frequency of throwing a 4 gets closer to the theoretical probability of throwing a 4, which is $\frac{1}{6}(\frac{1}{6} = 0{\cdot}1666 = 16{\cdot}66...\%)$.



Relative frequency (also referred to as the experimental probability of an event) is used to estimate the probability of an event, if conducting an experiment, test or survey, or when we do not know the theoretical probability.

We can estimate the probability of an event by using the following formula:

$$\text{Relative frequency of an event} = \frac{\text{the number of times an event happens}}{\text{the total number of trials}}$$

Relative frequency can be written as fractions, decimals or percentages.
It is important to remember that the most accurate estimate for the probability, or the relative frequency, of an event is found by using the greatest number of trials.

REMEMBER!

The most accurate estimate for the probability, or the relative frequency, of an event is found by using the greatest number of trials.

RELATIVE FREQUENCY

How to find the relative frequency or estimation of probability of an event as the proportion of times it has occurred.

Comparing an estimated probability from experimental results with a theoretical probability.

Plotting and interpreting a graphical representation of relative frequency against the number of trials.

Check that you can:

- *determine the probability of an event*
- *convert between fractions, percentages and decimals.*

Probability is the term used when discussing how likely it is that an event will happen.

Probabilities can be written as fractions, decimals or percentages.

EXAMPLE
If a fair, six-sided dice is thrown, the theoretical probability of throwing a four is:

$$= \frac{1}{6}$$

$$= 0{\cdot}1666$$

$$= 16{\cdot}666...\%$$

However, if a fair, six-sided dice is thrown 120 times, you may not get exactly twenty 4s ($\frac{1}{6} \times 120$), although it should be close to this amount. If you do an experiment of throwing a dice 120 times, you might get a result that looks like the following:

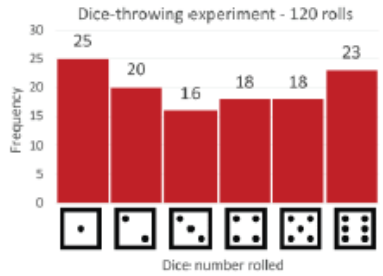
The number of times a 4 appears is the **relative frequency** or **experimental probability** of throwing a 4.

The number 4 appeared 18 times, so the relative frequency or the experimental probability of throwing a 4 is:

$$= \frac{18}{120}$$

$$= 0{\cdot}15$$

$$= 15\%$$



The more times we throw the dice, the more reliable the relative frequency is as an estimate of the probability.

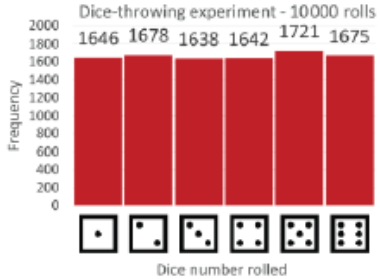
Here are the results after throwing the dice 10000 times.

The relative probability of throwing a 4 is:

$$= \frac{1642}{10000}$$

$$= 0{\cdot}1642$$

$$= 16{\cdot}42\%$$



We can see from the above that as the number of trials (throws) increases, the relative frequency of throwing a 4 gets closer to the theoretical probability of throwing a 4, which is $\frac{1}{6}$ ($\frac{1}{6} = 0{\cdot}1666 = 16{\cdot}66...\%$).

Relative frequency (also referred to as the experimental probability of an event) is used to estimate the probability of an event, if conducting an experiment, test or survey, or when we do not know the theoretical probability.

We can estimate the probability of an event by using the following formula:

$$\text{Relative frequency of an event} = \frac{\text{the number of times an event happens}}{\text{the total number of trials}}$$

Relative frequency can be written as fractions, decimals or percentages.

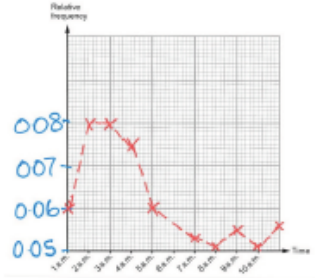
It is important to remember that the most accurate estimate for the probability, or the relative frequency, of an event is found by using the greatest number of trials.

GRAPHICAL REPRESENTATION OF RELATIVE FREQUENCY AGAINST THE NUMBER OF TRIALS

(Example adapted from (Legacy) Unit 1, Foundation tier, Winter 2017, Question 10, Worked example). A machine is used to pack tins of coffee beans. To check the machine, 100 tins of coffee beans are selected on the hour for 10 consecutive hours. There should be 800 coffee beans in each tin. Each hour, the number of tins containing exactly 800 coffee beans is recorded.

Time	1 a.m.	2 a.m.	3 a.m.	4 a.m.	5 a.m.	6 a.m.	7 a.m.	8 a.m.	9 a.m.	10 a.m.
Number of the 100 tins with exactly 800 coffee beans	6	10	8	6	0	2	4	8	2	10

If we record and plot the relative frequencies for the information shown in the previous table, the results would look like this:



We can use the plot to best estimate the probability that a tin selected at random will contain exactly 800 coffee beans. The last plot on the graph uses all the data from the experiment. It therefore gives the most accurate estimate of the probability: **0·056**

REMEMBER! The most accurate estimate for the probability, or the relative frequency, of an event is found by using the greatest number of trials.

PROBABILITY: SAMPLE SPACE

How to list all possible outcomes of compound events in a sample space diagram.

How to calculate probabilities and solve questions using a sample space diagram.

Check that you can:

- *determine the probability of an event.*

COMPOUND EVENTS

Flipping a coin AND spinning a fair 4-sided spinner (with sections labelled 1, 2, 3 and 4) is an example of a compound event. It is a combination of more than one event.

- The outcome when flipping a coin is head or tail.
- The outcome when spinning the 4-sided spinner is 1 or 2 or 3 or 4.
- One possible outcome of this compound event is a head and a 1.

When we list all possible outcomes, we need to be systematic and change one item at a time.

Here is one way of listing all the outcomes:
(starting with head and then changing the outcome of the spinner each time).

Head, 1
Head, 2
Head, 3
Head, 4
Tail, 1
Tail, 2
Tail, 3
Tail, 4

Is there a way of checking you have listed all the possible outcomes?

2 outcomes × 4 outcomes = 8 outcomes

If a player wins a prize if the coin lands on tails and the spinner shows the number 4, what is the probability of winning a prize by playing the game once?

Number of outcomes showing a tail and a 4 = 1 outcome (Tail, 4).

Total number of possible outcomes = 8

The probability of winning a prize = $\frac{1}{8}$

REMEMBER! A sample space diagram helps you to list all the outcomes of a compound event in a systematic fashion.

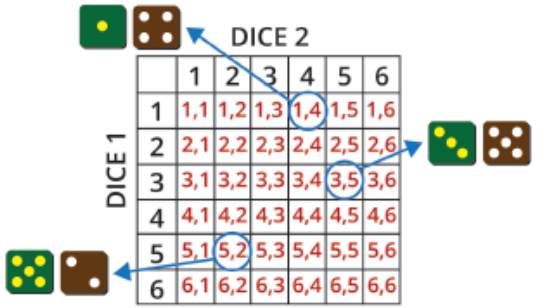
SAMPLE SPACE DIAGRAMS

When listing the outcomes of a compound event, sometimes it is better to use a sample space diagram. This will help make sure that you include all the possible outcomes.

Example 1
Two fair dice are rolled, and the scores are noted.
Answer
Find how many outcomes each event has.

- **The first dice has 6 outcomes.**
- **The second dice has 6 outcomes.**
- **This means there are 6 × 6 = 36 outcomes in total.**

Draw a table 6 × 6 and label 'Dice 1' and 'Dice 2'.
Complete the table with all the possible outcomes. This is called the **sample space**.



DICE 2

DICE 1	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

Example 2
Two fair dice are rolled, and the scores are noted.
What is the probability that both scores are even numbers?
Answer
We can use the sample space to help answer this question.
Circle all the outcomes where both of the scores are even numbers.

DICE 2

DICE 1	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

Both scores are even = 9

Total number of possible outcomes = 36

The probability that both scores are even numbers:

$= \frac{9}{36} \quad \left(= \frac{1}{4}\right)$

SAMPLING

When working with data we have to ensure that we collect relevant information. We may need to collect data from a large population but it may not be possible to gather this information from everyone. This could be due to time, money or other constraints. Therefore, it is useful to use a sample of the population.

Check first that you:

- understand the difference between qualitative data (words) and quantitative data (numbers)
- can identify what information is needed to test a hypothesis
- can design and criticise questions for a questionnaire
- understand fairness and bias.

Sampling techniques In order to ensure that we get a fair and non-biased sample, we use one of the following sampling techniques:

Simple random sampling

- This is when each member of the population has the same chance of being selected for the sample.
- The sample may be chosen by drawing names from a hat.
- It could also be chosen by giving all individuals in the population a number and then using a calculator, computer or random number tables to generate numbers for individuals to be chosen.

Systematic random sampling

- This is very similar method to random sampling but the population would first be ordered according to specific criteria such as listing names of people in the population in alphabetical order.
- The sample would be drawn by selecting every n^{th} person. For example every 10^{th} person in the list.

Stratified random sampling

- This method of sampling is used when each member of the population can be distributed into a certain group such as gender or age.
- A random sample is taken from every group but the number of members selected from each of the groups should be in proportion to the size of the group within the whole population.
- The number selected from each group is given by:

$$\text{No. selected from group} = \frac{\text{size of group}}{\text{size of population}} \times \text{size of sample}$$

Remember that the population is the whole of the data but a sample is just a part of this data.

Remember to ensure that the sample size is large! This is important to ensure that it is representative of the whole population we are gathering information on.

Simple random sampling We will look at two methods of randomly selecting numbers to choose a sample of 40 when all the members of a population of 200 have been previously numbered 1 – 200.

Random numbers generated by a calculator

To generate random numbers on the calculator we use the RAN button.

In this particular case we want to generate numbers up to 200 therefore we type:

200 Shift RAN

A number is generated and the corresponding numbered member of the population is selected for the sample. If a decimal is generated it will need to be rounded.

We continue with this method until we have a sample of 40.

Random numbers table

16405	82950	30197	64500	91130	58106	26415	72358
80937	54607	31138	68716	83922	27129	28755	41225
16405	82950	30197	64500	81200	98166	24435	62318
10538	64917	51678	28736	43424	57328	88137	14201
82551	81560	34246	15420	23452	78109	54010	

To select our sample using a random numbers table like the one shown above, we read the digits in groups of three to generate 3-digit numbers (because the population is a 3-digit number). Looking at the first line **16405 82950 30197 64500 91130 58106** ...

The numbers generated are 164, 058, 295, 030, 197, 645, 009, 113, 058, 106 ...

We ignore 295 and 645 as they are larger than 200 (the size of the population). 58 is repeated so we only include this the once.

The sample would therefore include the members numbered 164, 58, 30, 197, 9, 113 and 106.

We continue reading the digits in groups of three until we have our sample of 40.

Stratified random sampling E.g. The table shows the number of students of each nationality studying at The Welsh International College.

Nationality	American	British	Chinese	French	Spanish
Number of students	29	48	31	18	26

The Dean of the college would like to carry out a survey on student wellbeing. A stratified random sample of 30 students is to be used. How many students of each nationality will be included in the sample?

First we find the size of the population, i.e., the number of students at the college:
29 + 48 + 31 + 18 + 26 = 152

We then use: $\text{No. selected from group} = \frac{\text{size of group}}{\text{size of population}} \times \text{size of sample}$
to find the number of students of each nationality that needs to be in the sample.

$\text{No. of American students} = \frac{29}{152} \times 30$
= 6 (to nearest person)

$\text{No. of French students} = \frac{18}{152} \times 30$
= 4 (to nearest person)

$\text{No. of British students} = \frac{48}{152} \times 30$
= 9 (to nearest person)

$\text{No. of Spanish students} = \frac{26}{152} \times 30$
= 5 (to nearest person)

$\text{No. of Chinese students} = \frac{31}{152} \times 30$
= 6 (to nearest person)

Check that the number for each group add to give the sample size needed, in this case 30. It's possible to have 1 more or 1 less then needed. This will be due to rounding the answer to the nearest person, therefore you may have to adjust an answer so that it fits with the sample size.

Sampling

Understanding the reasons we use sampling, and two different sampling techniques – simple random sampling, and systematic sampling.

Check that you can:

- *understand the difference between qualitative data (words) and quantitative data (numbers)*
- *identify what information is needed to test a hypothesis*
- *design and criticise questions for a questionnaire*
- *understand fairness and bias.*

Why do we need to sample?

When investigating a particular question or **hypothesis**, we have to ensure that we collect relevant information or data.

We may need to collect data from a large **population under study**. The population under study is the whole group of people whose opinions are required. It may not be possible to gather the information from the whole population under study. This could be due to time, money or other constraints.

For example, if a pupil wanted to ask the opinion of other pupils in a particular class on a subject, then this is possible if the pupil had enough time and there weren't too many pupils in the class. However, it would be very difficult to collect the opinions of all the 1100 pupils in a school.

When it is not possible to collect data from a population under study, it is useful to use a **sample** of this population.

There are a number of ways to choose the sample. The sample should be:

- fair and **unbiased**
- large enough in size to be representative of the whole population under study.

However, it is important to remember that different samples will give you different results. If your sample is large enough, and you have used a good, unbiased sampling method, the sample should reasonably represent the whole population under study.

Sampling techniques

In order to ensure that we get a fair and non-biased sample, we use one of the following sampling techniques:

Simple random sampling

- This is when each member of the population has the same chance of being selected for the sample.
- The sample may be chosen by drawing names from a hat.
- It could also be chosen by giving all individuals in the population a number and then using a calculator, computer or random number tables to generate numbers for individuals to be chosen.

Systematic sampling

- This is a very similar method to random sampling, but the population would first be ordered according to specific criteria such as listing names of people in the population in alphabetical order.
- The sample would be drawn by selecting every n^{th} person. For example, every 10^{th} person in the list.

Example

Lisa wants a random sample of the 600 people who work in her office building.
Give some examples of **simple random sampling** methods she could use.

Answer

- Assign a number to all of the office workers and use a calculator to get 50 random numbers.
- Put everyone's name in a hat and pick 50.

A simple random sample is when each member of the population under study has the same chance or probability of being selected for the sample.

The following options are not random, as not all the office workers have the same chance of being chosen:

- Choose the first 50 people who arrive at the office.
- Choose 50 people whose surname begins with J or T.
- List all the office workers in alphabetical order and choose every 5^{th} name on the list.

Example

A company manufactures bottles in a factory.
In order to ensure high quality, the manager wants to take a systematic sample of seven bottles from the first 100 produced.

Explain how the manager could use systematic sampling to obtain this sample.

Answer

Number the bottles 1 to 100.
The sampling interval $= 100 \div 7 = 14 \cdot 285$. We can round this number down to get 14.
Pick one number at random from the first 14.
If the random starting number is 11, you then pick every 14^{th} number from there on until you have an additional six numbers.
So, the sample in ascending order would be 11, 25, 39, 53, 67, 81, 95.
The corresponding numbered bottle is selected for the sample.

REMEMBER!
The population is the whole of the data and a sample is just a part of this data. You should ensure that the sample size is large. This is important to ensure that it is representative of the whole population we are gathering information on.

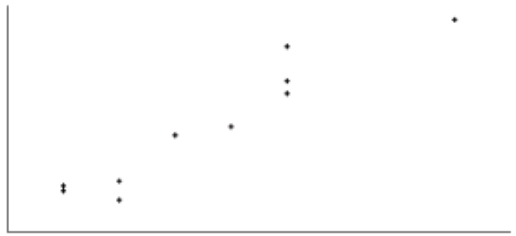
SCATTER DIAGRAMS

Scatter diagrams are used to see if a relationship exists between two sets of data or variables.

Check first that you:

- understand coordinates and can plot points on a graph
- can understand and read scales on a graph
- know how to find the mean by dividing the total of the all values by the number of values.

Correlation This describes the type of relationship between two sets of data.

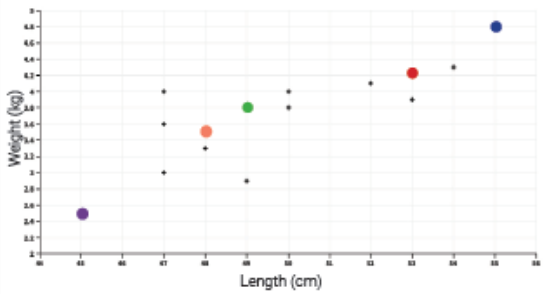
Positive Correlation	Negative correlation	No Correlation
		
As the values of one set of data increase, the values of the other set of data also increase. E.g. the time exercising and number of calories burned.	As the values of one set of data increase, the values of the other set of data decrease. E.g. the value of a car and its mileage.	There is no relationship between the values of the two sets of data. E.g. the height and salary of employees of a company.

Can you *give* your own examples? Can you *explain* why a scatter diagram does or doesn't show correlation?

Drawing a scatter diagram The table shows the length and corresponding weight of baby boys born at Cwmbrân Hospital.

Length (cm)	45	55	49	48	53	49	48	50	53	47	47	50	47	52	54
Weight (kg)	2·5	4·8	3·8	3·5	4·2	2·9	3·3	4	3·9	4	3·6	3·8	3	4·1	4·3

a) Draw a scatter diagram to display this data.



Take care when reading the scale on each of the axes. They may not be the same!

We plot the length (horizontal axis) against the weight (vertical axis). Here are the first five points highlighted on the diagram:

Length (cm)	45	55	49	48	53
Weight (kg)	2·5	4·8	3·8	3·5	4·2

We don't connect the points on a scatter diagram.

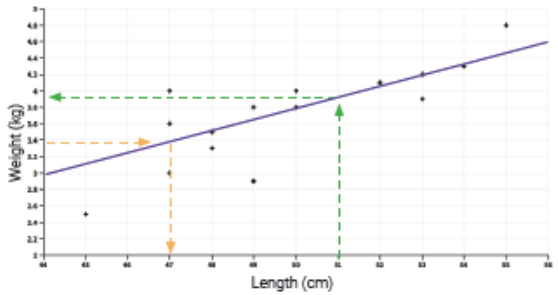
b) Describe the relationship between the weight and length of baby boys shown by the scatter diagram.

The diagram shows positive correlation between the length and weight of baby boys. As the length of the baby increases the weight also increases.

Line of best fit

When there is positive or negative correlation, we can draw a line of best fit by eye on the scatter diagram. This line allows us to estimate a value of one variable if we know a value of the other variable.

c) Draw a line of best fit by eye on the scatter diagram for the length and weight of baby boys born at Cwmbrân Hospital.



Using a ruler, draw a straight line that follows the trend of the data. You should try and get as many points as possible on the line. You should also try and get an equal number of points lying above the line and lying below the line.

d) Use the line of best fit to estimate the weight of a baby that is 51 cm in length.

Draw a straight line from 51 cm on the horizontal axis (length) to the line of best fit. Where they meet draw a line across to the vertical axis to read off an estimate for the weight. 3·9 kg

e) Use the line of best fit to estimate the length of a baby that weighs 3·4 kg.

Draw a straight line from 3·4 kg on the vertical axis (weight) to the line of best fit. Where they meet draw a line down to the horizontal axis to read off an estimate for the length. 47 cm

Take care: the line of best fit won't necessarily need to go through the point of intersection of the axes.

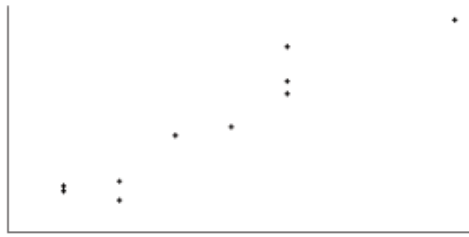
SCATTER DIAGRAMS

Scatter diagrams are used to see if a relationship exists between two sets of data or variables.

Check first that you:

- understand coordinates and can plot points on a graph
- can understand and read scales on a graph
- know how to find the mean by dividing the total of the all values by the number of values.

Correlation This describes the type of relationship between two sets of data.

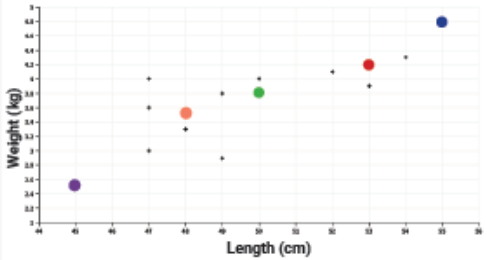
Positive correlation	Negative correlation	No correlation
		
As the values of one set of data increase, the values of the other set of data also increase. E.g. the time exercising and number of calories burned.	As the values of one set of data increase, the values of the other set of data decrease. E.g. the value of a car and its mileage.	There is no relationship between values of the two sets of data. E.g. the height and salary of employees of a company.

Can you ***give*** your own examples? Can you ***explain*** why a scatter diagram does or doesn't show correlation?

Drawing a scatter diagram The table shows the length and corresponding weight of baby boys born at Cwmbrân Hospital.

Length (cm)	45	55	49	48	53	49	48	50	53	47	47	50	47	52	54
Weight (kg)	2·5	4·8	3·8	3·5	4·2	2·9	3·3	4	3·9	4	3·6	3·8	3	4·1	4·3

a) Draw a scatter diagram to display this data.



Take care when reading the scale on each of the axes. They may not be the same!

We plot the length (horizontal axis) against the weight (vertical axis). Here are the first five points highlighted on the diagram.

Length (cm)	45	55	49	48	53
Weight (kg)	2·5	4·8	3·8	3·5	4·2

We don't connect the points on a scatter diagram.

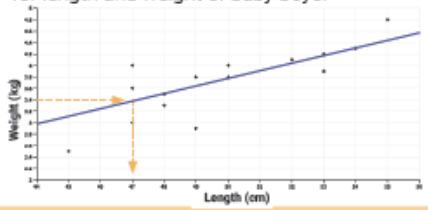
b) Describe the relationship between the weight and length of baby boys shown by the scatter diagram.

The diagram shows positive correlation between the length and weight of baby boys. As the length of the baby increases the weight also increases.

Line of best fit When there is positive or negative correlation we can draw a line of best fit by eye on the scatter diagram. This line allows us to estimate a value of one variable if we know a value of the other variable.

Line of best fit by eye

c) Draw a line of best fit by eye for the scatter diagram for length and weight of baby boys.



Draw a straight line that follows the trend of the data. You should try and get as many as possible of the points on the line and an equal number of points above and below the line.

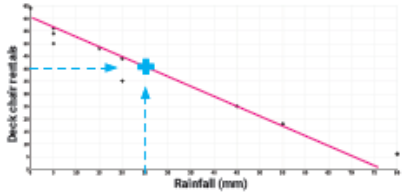
d) Use the line of best fit to estimate the length of a baby that weighs 3·4kg.

Draw a straight line from 3·4kg on the vertical axis (weight) to the line of best fit. Where they meet draw a line down to the horizontal axis to read off an estimate for the length. 47cm

Line of best fit using the mean

The scatter diagram below shows recorded rainfall (to the nearest 5mm) at Barry Island and the number of deck chair rentals over a period of 10 days.

Draw a line of best fit knowing the mean rainfall was 25mm and the mean number of deck chair rentals was 40.



Plot the meant rainfall value, 25mm on the horizontal axis against the mea number of deck chat rentals, 40 on the vertical axis. Draw a straight line that goes through this mean point that follows the trend of the data. You should also try and get as many as possible of the points on the line and equal number of points above and below the line.

Take care: the line of best fit won't necessarily need to go through the point of intersection of the axes.
Remember if the mean isn't given, we need to find it by adding the values for one variable on the horizontal axis and dividing by the total number of values. Then we repeat for the other variable on the vertical axis.

STATISTICS

Sorting data — Classifying different types of data and how to sort and present data in tables and using tally charts.

Check that you can:

- put numbers in order from smallest to largest
- understand what it means to 'classify' something.

Classifying data

Data can be divided into two main types: **qualitative data** and **quantitative data**.

Qualitative data deals with data that can only be written in words. Examples include hair colour, favourite football team and type of weather.

Quantitative data deals with numbers and things you can measure.

Quantitative data can then be divided into two types: **discrete data** and **continuous data**.

Discrete data can only have certain values. This type of data is usually found by counting. The number of pupils in a class would be discrete because you can only have whole numbers. Shoe sizes would also be discrete, you can have half sizes, but nothing in between them. For example, you can buy a shoe size 7 or 7·5, but not size 7·341.

Continuous data can have any value within a certain range. This type of data is usually found by measuring. The length of a pencil would be an example of continuous data. It is important to realise that your ruler is only marked to the nearest millimetre, but there are measurements between these values.

Types of data: an example

Dyfan has received his school report from his teacher – it contains three pieces of information. For each one, state whether the data is **qualitative** or **quantitative**. If the data is quantitative, also state whether it is **discrete** or **continuous**.

a) The number of homework assignments Dyfan has been given this term.

Answer

The number of homework assignments Dyfan has been given this term can be counted, therefore this data is **quantitative**. It can also only be a whole number. Therefore, this is **discrete data**.

b) How long Dyfan took to talk through his class presentation.

Answer

The length of Dyfan's class presentation is given as a time which is **quantitative**; the result could be any value. Therefore, it is **continuous data**.

c) A paragraph written by Dyfan's teacher explaining how well he is doing in school.

Answer

This paragraph describes something and gives the teacher's opinion of Dyfan's progress in school. This data is not a number and cannot be counted. Therefore, this is **qualitative data**.

Sorting data: grouping data

Example

Here are the marks obtained by 100 students in their mathematics test.

72 61 63 66 62 68 69 64 65 67 69 56 60 66 62 57 72 67 65 70
64 66 71 73 67 65 64 63 61 58 64 62 69 66 65 63 63 59 61 64
65 57 66 71 68 70 67 66 60 62 65 58 63 68 64 61 62 65 66 59
62 65 65 60 64 61 64 69 62 64 62 63 68 67 65 62 65 68 61 63
62 72 62 66 66 65 63 67 66 63 63 66 65 63 62 62 66 64 62 62

When the data is presented in this form, it can be hard to make sense of it.

When data is collected, it is usually for a purpose. Say you were looking for the most popular score or the range of scores, the answer would be much easier to spot if the results were displayed as a table. This is known as tabulating data.

Remember, frequency means how many times something happens.

Score	Frequency	Score	Frequency
56	1	65	13
57	2	66	12
58	2	67	6
59	2	68	5
60	3	69	4
61	6	70	2
62	15	71	2
63	11	72	3
64	10	73	1

The most popular score (the modal score) is 62. This is the one with the highest frequency, 15.

The range of scores is the highest score minus the lowest score, which is:

73 – 56 = 17

It could be even longer with some data tables, which is not very useful.

To avoid having such long tables, it is better to group the data.

Remember:

When you hear the term **quantitative data**, think about the word **quantity** (the number of things).

The **number** of trees in a forest is an example of **discrete data**, as it will always be given as a whole number.

Distance, for example, is **continuous.** Even though your GPS may only display whole numbers or decimal numbers to one decimal place, it is possible to have **any value**.

Qualitative data describes something in words, but this data is not a number and cannot be counted.

Grouping discrete data

Test scores like those you just saw are an example of discrete data, as each score is a whole number.

It is always best to keep the group widths the same.

Score	Frequency
56–58	5
59–61	11
62–64	36
65–67	31
68–70	11
71–73	6

The only disadvantage to grouping data is that you lose accuracy slightly.

When data is continuous, we need to use a different notation for the boundaries of our groups.

Grouping continuous data

Example

The time taken by pupils in a sports lesson to complete 50 star jumps is an example of continuous data.

Again, it is always best to keep the group widths the same, but this time, we need to consider that time can take any value.

Time taken, t (seconds)	Frequency
$30 \leqslant t < 35$	2
$35 \leqslant t < 40$	6
$40 \leqslant t < 45$	7
$45 \leqslant t < 50$	8
$50 \leqslant t < 55$	10
$55 \leqslant t < 60$	5

Look at the group below taken from the table.

$30 \leqslant t < 35$	2

This information tells us that two pupils took from 30 seconds up to, but not including, 35 seconds to complete the exercise. A pupil who took 35 seconds would be included in the second group.

The only disadvantage to grouping data is that you lose accuracy slightly. Take for instance the group below.

$50 \leqslant t < 55$	10

You can see that 10 pupils took between 50 and 55 seconds to complete the exercise, but you do not get to know their exact times.

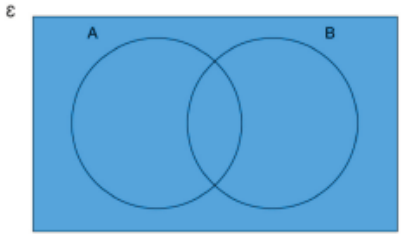
VENN DIAGRAMS: SET NOTATION

Set notation is an easier and quicker way to list the values or number of values in each subset of a set of data, which could also be represented using a Venn diagram.

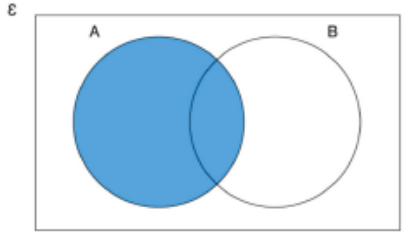
Check that you can:

- *use the correct terminology associated with Venn diagrams*
- *sort sets of data and do so using Venn diagrams.*

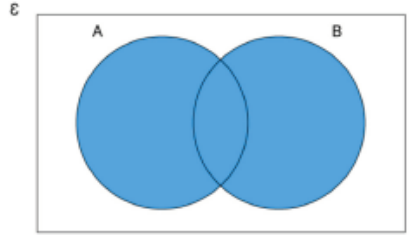
You will have already come across the **universal set**, which is denoted by the symbol 'ε' (epsilon). The universal set contains all of the objects.



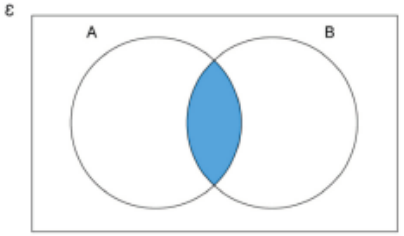
Here is Set A shown in a Venn diagram. Notice how some values are also in Set B, where the circles overlap.



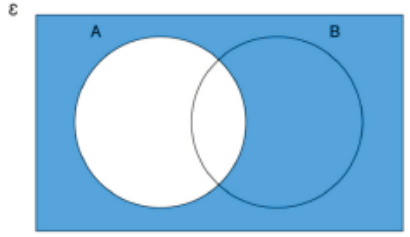
The **union** of Set A and Set B are the objects that are in Set A or Set B (or both). This can be written as A ∪ B.



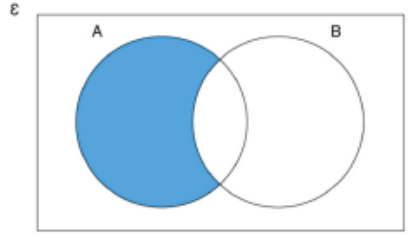
The **intersection** of a Venn diagram is the region where the values in the sets overlap. This can be written as A ∩ B.



The values not in Set A are represented by A' (A dash). This is called the **complement of Set A**.



The values that are in Set A, but not in Set B, are represented by A ∩ B'.



REMEMBER!

The universal set contains all of the objects within a specified data set. Any values/objects that are part of the universal set, but do not fit into any of the subsets denoted by the circles in the Venn diagram, should be placed outside the circles but within the rectangle. Don't forget the letter epsilon, ε, next to the rectangle to indicate that it is the universal set!

VENN DIAGRAMS

Venn diagrams can be used to sort sets of data.

Check that you can:

- *recognise factors, multiples, prime numbers and square numbers*
- *recognise numerical and described characteristics of data values.*

Terminology of Venn diagrams

The first step in drawing a Venn diagram is to draw a rectangle. This rectangle will contain all the data values that are to be considered for the Venn diagram. The name given to these data values is the universal set, and it is denoted by the Greek letter epsilon, which is written as the symbol ε. This symbol, ε, is usually placed at the side of the rectangle.

Inside this rectangle, we can then draw circles to represent sets of values, which contain values from the universal set. Although these values form sets, they are sometimes referred to as subsets of the universal set. As some values may belong to more than one of these sets, the circles will normally intersect.

A set is the name given to a collection of data values, which could be numerical values, described values or objects.

We use curly brackets { } to list the values within a set.

Example

Here, the universal set = all the people shown.

ε = {Mali, Erhan, Anita, Dewi}

The circles inside the rectangle are used to group these four people from the universal set.

The people listed in the different regions within these circles are called subsets.

From the example above, the subsets are:

people with black hair = {Erhan, Mali}
people who wear glasses = {Mali, Dewi}.

Anita does not have black hair and she does not wear glasses; therefore, her name is placed outside the circles but still within the rectangle.

The **intersection** of two or more sets within a Venn diagram is the region where the sets overlap.

This is represented by the red region in the diagram above. Here, we have the name of the person who has black hair AND also wears glasses.

Intersection of both sets = red region = {Mali}

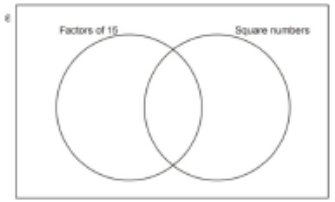
The **union** of a Venn diagram contains the objects that are in either set. Here, we have the names of the people who have black hair OR wear glasses OR both.

Union of both sets = blue + red + green regions = {Mali, Erhan, Dewi}

Sorting sets of data using Venn diagrams

Example from Mathematics Unit 1, Foundation Tier, Autumn 2018, Question 12

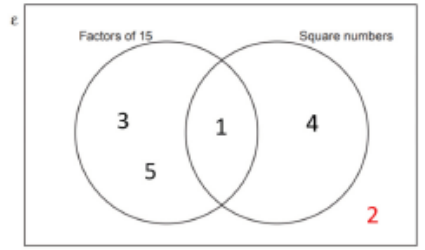
12. Place the numbers 1, 2, 3, 4, 5 in the Venn diagram below. [2]



Answer
Factors of 15 = {1, 3, 5}
Square numbers = {1, 4}

The value, or values, that are common to both sets is given by factors of 15 AND square numbers = {1}.

2 is not a factor of 15 or a square number, so it should be placed outside of the circles but still within the rectangle.



REMEMBER!

The universal set is denoted by the symbol ε. When sorting data, you should only consider the data within the universal set you are given. Even if certain values don't fit into a circle, they should still be placed within the rectangle to show they are part of the universal set.

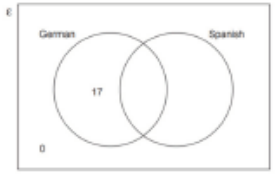
Using Venn diagrams to solve problems

Example from Mathematics Unit 2, Intermediate Tier, Autumn 2020 Question 8

8. Each of 30 students studies German, Spanish or both languages.

A student is chosen at random.
The probability that the student studies both German and Spanish is $\frac{1}{3}$.

Complete the Venn diagram. [2]



Answer
There are 30 students. This is our universal set.
At the end, we must check that all regions in the Venn diagram add up to 30.

$\frac{1}{3} \times 30 = 10$

This means that 10 students study both German and Spanish. Therefore, 10 is placed in the middle. The only region that is left to complete is the number of students that study only Spanish.

All the regions in the Venn diagram must add up to 30.

Therefore,

$? = 30 - 17 - 10 - 0 = 13$