

GCSE Data rules and tools

Table of Contents

GCSE Data rules and tools.....	1	Frequency polygon.....	21
Symbols and conventions.....	2	Comparing frequency distributions.....	22
Data Handling.....	3	Stem and leaf diagrams.....	23
Averages: Mean, median, mode and range.....	3	Median and range from stem and leaf diagram.....	24
Median for large data set.....	4	Scatter diagrams.....	25
Which average is the best to use?...4		Describing the pattern.....	26
Number puzzles involving the mean and median.....	5	Line of best fit.....	27
Data: Qualitative, quantitative, continuous and discrete.....	6	Making predictions.....	28
Data collection cycle.....	7	Two way tables.....	29
Collecting data.....	7	Sampling and bias.....	29
Questionnaire questions.....	8	Random samples.....	30
Data collection form.....	9	Bias.....	30
Primary and secondary data.....	9	Stratified samples.....	30
Pictogram.....	10	Number of measurements for reliability	31
Tally charts and frequency.....	11	Probability.....	32
Discrete frequency tally chart.....	11	Fair and biased coins, dice.....	32
Grouped frequency tally chart.....	12	Likelihood words.....	33
Frequency.....	12	Probability scale.....	33
Mean from frequency table.....	13	Spinners.....	34
Discrete frequency table.....	13	Making up events that fit on a scale	35
Grouped frequency table for continuous data.....	14	Probabilities adding to 1.....	35
Median from frequency table.....	15	Decimal probabilities.....	36
Discrete frequency table.....	15	Find the missing probability.....	36
Grouped frequency table.....	16	Probability of something not happening.....	36
Mode or modal group from a frequency table.....	16	Combined events.....	36
Bar charts.....	17	List all the possible outcomes of tossing two coins.....	37
Stacked bar chart.....	17	Sample space diagram: rolling two dice, add the scores.....	37
Clustered bar chart.....	18	Expected Frequency.....	38
Pie chart.....	18	Relative frequency.....	38
Pie chart number puzzles.....	19		
Frequency Bar charts.....	20		

Symbols and conventions

Symbol with example	Meaning	Notes and associated words
$3 + 4$	Adding, addition	Sum, total
$11 - 6$	Subtracting, subtraction	Difference
2×5	Multiply, multiplication	Product
$15 \div 5$	Divide, division	Quotient (rare!)
$15.239261 \approx 15$	Approximately, approximation	Estimate, round off, round up
$5 > 3$	Greater than	
$2 < 1000$	Less than	
$21 \geq 17$	Greater than or equal to	
$2 \leq 10$	Less than or equal to	
$2 \neq 7$	Not equal to	
\equiv	Equivalent, equivalence	Used mainly in algebra
2^{10}	Power, index, indices	The 10 is small and above the line of writing
15°C	Degrees Celsius	The little circle is different to a zero. If temperature then there has to be a letter after the little circle to say which temperature scale. $^{\circ}\text{C}$ is degrees Celsius, $^{\circ}\text{F}$ is degrees Fahrenheit, $^{\circ}\text{K}$ is degrees Kelvin (degrees above absolute zero, H level)
45°	Degrees	Without a letter after it, the little circle stands for the degrees you measure with a protractor, the size of an angle

Data Handling

Data handling is the branch of GCSE Maths that deals with how to make sense of information you collect.

Averages give you a 'typical value' for some quantity. For instance the average adult male height is 175cm and the average adult female height is 162cm in the UK.

Once you have an average, you need to know what the **variation about the average** is. If you design the seats on buses, you might want to know how many very tall people there are.

Most organisations collect statistics on their clients to ensure that they are complying with government policies and that they are meeting their targets. Knowing how to read the statistics published by hospitals, schools and care homes will help you to make choices about the services you may need.

Averages: Mean, median, mode and range

The average of a set of results provides you with a 'typical value' for the quantity that you are measuring.

There are different kinds of average to use for different types of data.

When you quote an average for some data that you have collected, you should also give some idea of how spread out that data is.

The range of the data is an easy way to measure the spread of data.

I have summarised the definitions of the three kinds of average in the table below.

Mean	Add up all the numbers to find the total. Divide the total by the number of results.	Five people have heights in cm as follows, 173, 165, 182, 159, 170 Their mean height is $\frac{173+165+182+159+170}{5}=169.8 \text{ cm}$ Leaving answer to 1d.p. Is ok
Median	Put the numbers in order of size. Pick the middle number.	Putting the heights in order of size 159, 165, 170, 173, 182 The median height is 170 Even number of data: suppose you had 6 heights 150, 159, 165, 170, 173, 182 Pick the middle two, 165, 170 and find their mean $\frac{165+170}{2}=167.5$ So 167.5 is median.
Mode	Find the most common value in the set of data	The number of letters received each day in a small office is 12, 14, 12, 12, 17 The modal number of letters is 12
Range	The difference between the largest and smallest value in the data.	Five people have heights in cm as follows, 173, 165, 182, 159, 170 The range in height is $182 - 159 = 13 \text{ cm}$

Median for large data set

Suppose you have to find the median of (say) 100 numbers. Or 1000.

Let the number of numbers be n

Find the value of $\frac{n+1}{2}$

That number is the position of the median in a list of your numbers that is in order of size.

Example: Here are the heights in cm of a mini bus full of people:

169 171 176 170 171 176 166 174 173 176 166 168 163 165 177.

Find the median height.

Stage 1: Put data in order of size

163, 165, 166, 166, 168, 169, 170, 171, 171, 173, 174, 176, 176, 176, 177

Stage 2: calculate the position of the median $n = 15$ so $\frac{n+1}{2} = \frac{15+1}{2} = 8$

Stage 3: The 8th number gives the median, which corresponds to 171cm

Example with even number of data: here are the heights of 14 people

163, 165, 166, 166, 168, 169, 170, 171, 171, 173, 174, 176, 176, 176

Find the median height

Stage 1: Put data in order of size (already done to save reading!)

Stage 2: Calculate position of the median $n = 14$ so $\frac{n+1}{2} = \frac{14+1}{2} = 7.5$

Stage 3: The 7.5th position tells us to take the mean of 7th and 8th item $\frac{170+171}{2} = 170.5$

So the median for this second set of heights is 170.5cm

Which average is the best to use?

Suppose you had 5 people working for a small company, and their weekly wages looked like this

£200, £200, £240, £480, £1 300

No prizes for guessing which person owns the company!

The mean weekly wage is: $\frac{200+200+240+480+1300}{5} = £ 484$.

This figure is close to one person's wage but is not really 'typical' of all the wages.

The median weekly wage is the wage of the $\frac{5+1}{2} = 3$ rd person in the list, which is £240.

That figure is more typical.

Moral: if there is a small number of extreme values in your data, the median is a better average to use than the mean.

Number puzzles involving the mean and median

The mean is the total of the data values divided by the number of data values.

As a formula: $Mean = \frac{Total}{Number}$

You can rearrange that formula (see Algebra below) to work *backwards* from the mean to find the total

Example: There are 7 people in a room and the average age is 24. The oldest person leaves the room and the average age falls to 16. How old was the person who left the room?

Always work back to the total

Stage 1: Find the total age of the people in the room before the person left

$$24 \times 7 = 168 \text{ years}$$

Stage 2: Find the total age of the people in the room after the person left

$$16 \times 6 = 96 \text{ years}$$

Stage 3: Subtract the two totals, the result must be the age of the person who left the room. $168 - 96 = 72$ years old.

Sometimes you have to make up numbers with averages of certain values

Example: Write down five numbers that have mean 12 and median 8.

Stage 1: Mean 12 suggests that the total is $12 \times 5 = 60$

Stage 2: Median 8 suggests that one number is 8, that two numbers are smaller than 8 and that 2 numbers are bigger than 8

Stage 3: $60 - 8 = 52$, so lets have 7 and 6 as the two smaller numbers. $52 - 13 = 39$ so lets have 19 and 20 as the two larger numbers

Stage 4: My data set looks like 6, 7, 8, 19, 20. Check: adds to 60, has median 8

A popular twist is to ask you to write down some numbers with a given median and range.

Example: Write down seven numbers with median 14 and range 10.

Stage 1: Smallest to largest has to be 10, and the middle number has to be 14

Stage 2: Pick 8 for the smallest and the largest has to be 18 to give range of 10

Stage 3: We need 2 numbers larger than 8 but smaller than 14, and then 2 numbers larger than 14 but smaller than 18

Stage 4: My data set looks like 8, 10, 12, 14, 15, 17, 18

Challenge: find a different set of numbers that work for each of the last two examples.

Data: Qualitative, quantitative, continuous and discrete

Data is information that you collect about people or things.

Suppose you asked people in the class about their eye colour, shoe size, and then measured their height in cm. Suppose there are 24 people in the class. You would have 24 sets of three data items (including yourself).

We call the three things you are collecting and measuring the **variables**. One variable is eye colour, another is shoe size (say English) and the third is height.

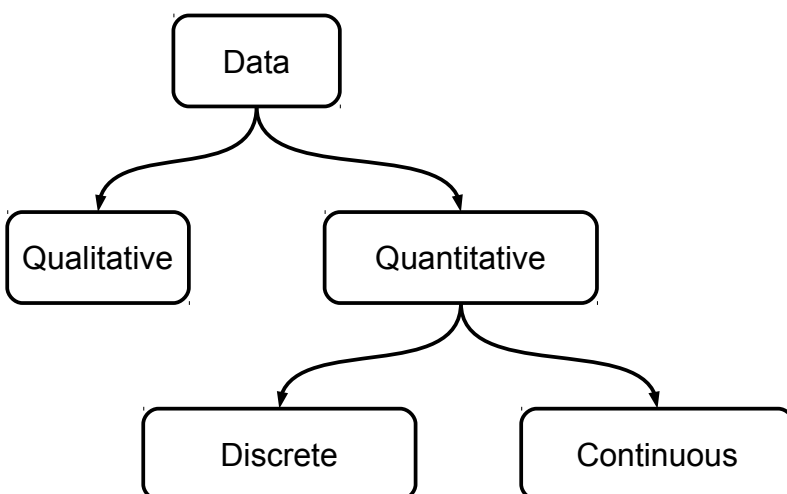
Eye colour is called a **qualitative** variable because you don't give it a number, eye colour has a name.

Shoe size and height are **quantitative** variables because they are measured using a number, and one person can be said to have a larger shoe size or height than another.

Shoe sizes are 'lumpy'. You can have a shoe size of 4 or 4½ or 5 or 5½ but there is nothing between (say) 5 and 5½. Quantitative variables that are 'lumpy' like this are called **discrete** variables. The word discrete is spelt like concrete (which is lumpy!)

Heights, or weights or anything that you measure can vary as you increase the accuracy of the measurement. If you use one of the height measuring devices they have at the doctors, you will find that you are taller in the morning than you are in the evening (your vertebrae compress lightly over the day). For any two people you measure the height of, you can usually find a third person with a height in between. Quantitative variables that are 'smooth' like this are called **continuous** variables.

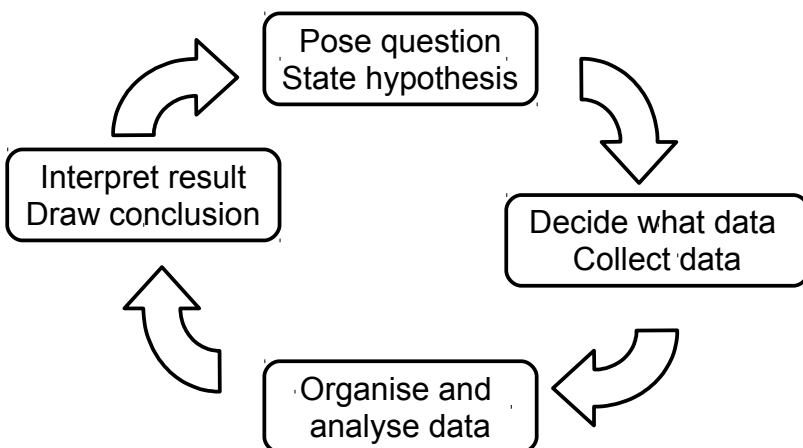
The chart below summarises the kinds of data you can collect



Challenge: write down five examples each of a qualitative variable, a quantitative discrete variable and a quantitative continuous variable.

Challenge: is money a discrete or continuous variable?

Data collection cycle



People don't just go round with clipboards writing down data. They start with a question they are trying to ask, then decide what data to collect, then organise the data somehow then finally say what it all means, hopefully answering the original question.

The data collection cycle (see diagram above) shows the main stages:

Stage 1: An example of a question might be

“does the 33 bus really run every 5 minutes during the day?”

The hypothesis that comes from the question might be

“there are 12 33 buses each hour from the Dale End bus stop”

Stage 2: We need to record which buses go from the Dale End bus stop during a typical week day. We need to record the time of departure of each bus.

Stage 3: We could use our records to produce a table of the number of buses each hour OR we could work out how many minutes there were between each 33 bus and take an average.

Stage 4: We would look at the table or look at the average time between the 33 buses and try to answer the question.

Collecting data

There are two main ways to collect data yourself, asking people what they think about something and observing what actually happens.

Asking people usually takes the form of a questionnaire.

When you observe what happens, a prepared data collection form is a good idea.

Questionnaire questions

People have short memories, and they don't like writing things down.

People don't read a lot of words or complex instructions

People will lie if you ask them about their wages or age or other personal things

The same words mean different things to different people

When you devise a questionnaire, you should

- Give people boxes to tick, and make sure you cover all possibilities and that the various options don't overlap. See the examples.
- Don't ask vague questions like 'how often do you take the bus', but ask 'how many times did you travel by bus in the last week' and give them boxes to tick.

Example: Age groups. Suppose you are asking people in College about the canteen.

You might want to know how old each person is so that you can see if there are differences between younger and older students.

You could include the question below

How old are you? (tick one)

Under 16 [] 16 to 18 [] 19 to 24 [] 24+ []

The **choices cover the whole range** and the **choices don't overlap**

Challenge: find a GCSE textbook and look up an exercise about finding the mistakes in questionnaire questions and about writing good questionnaire questions. Complete the exercise.

Make sure you answer each question. If you are asked 'what is wrong with the questionnaire question below' you need to describe the mistake, NOT provide a corrected question!

Data collection form

When you collect data by observation, you need a form to record what you see.

Example: suppose you wanted to know what kind of drinks the College canteen sold each hour on a given day.

The canteen sells water, orange juice, tea and coffee

The canteen is open from 11am to 3pm.

You have two dimensions of data (which hour and what drink) so you make a two dimensional table.

	Water	Orange juice	Tea	Coffee
11:00 to 11:59				
12:00 to 12:59				
13:00 to 13:59				
14:00 to 14:59				

You could position yourself near the till and put a tally mark in each type of drink during the appropriate hour.

I've added 5 water drinks sold between 11 and 12 to show how the data would be entered.

Primary and secondary data

Primary data is data you obtain for yourself either by questionnaire or by observation. If there is a team of you doing the data collection, your data is still primary, even though you personally may not have collected it.

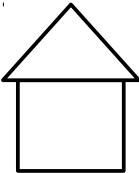
Secondary data is data from other sources such as government statistics, other published research, data collected by other teams earlier on or in different Colleges.

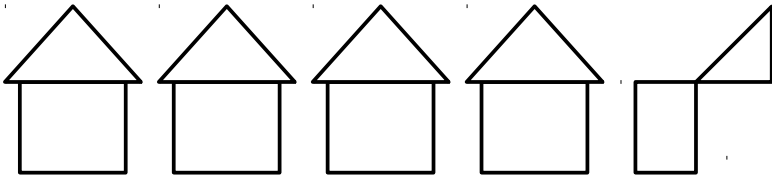
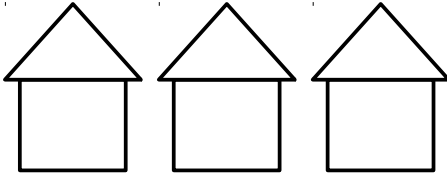
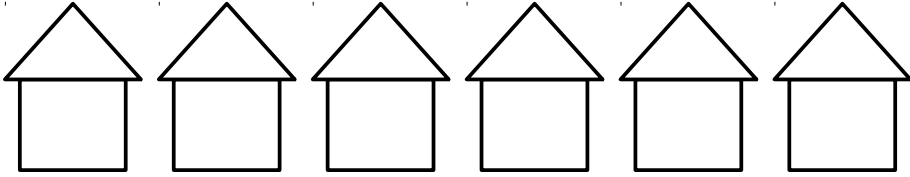
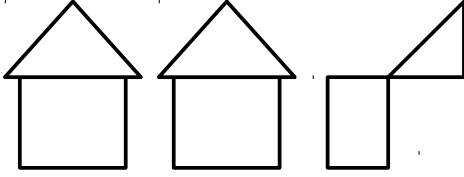
Pictogram

A pictogram is a kind of bar chart made from pictures.

Pictograms are used to communicate statistical information to children or adults of all educational levels. Pictograms are commonly used in health education and similar contexts.

Example: number of each type of house in an inner-city ward

Key  = 1000 houses

House type	Number
Flat	
Terraced house	
Semi-detached house	
Detached house	

The pictogram has a **key** to show how many houses each symbol represents

Half symbols can be used

The pictogram above shows 4 500 flats in the ward.

Challenge: how many more semi-detached houses than detached houses are there in the ward?

Challenge: look out for pictograms in leaflets and in the newspaper.

Tally charts and frequency

Tally marks form one of the oldest types of counting from before place notation and even Roman numerals. Tallies are scratched in bundles of five, so

|||| | |||| | |||| | |||

means $5 + 5 + 5 + 3 = 18$.

Discrete frequency tally chart

Suppose you rolled a dice 100 times and recorded the results...

2 5 5 5 5 1 6 4 4 1 6 1 2 6
 3 1 2 1 2 6 2 1 3 1 2 1 3 4
 1 4 4 1 6 1 6 3 1 6 4 4 4 4
 1 4 3 6 2 6 5 6 1 1 3 4 6 1
 1 1 6 5 5 4 6 1 6 6 5 5 5 4
 6 3 3 2 1 6 2 6 4 4 2 5 5 4
 2 4 2 2 6 1 4 6 1 2 5 4 2 6
 2 6

You could go through the results and count how many 1s you found, then go through them again and count how many 2s and so on.

Making a tally chart is much more systematic and quicker as you only have to go through the data once.

To tally the data, you make a row in the table for each possible dice score.

Then you work along the table systematically and put a tally on the correct row for each number, so I put a mark in the 2 row, then in the 5 row (4 of those) then in the 1 row...

I work along each row and then down the rows.

I cross off the number in the table after I tally it just in case I get distracted part way through the tallying.

My example tally chart below took about 3 minutes to draw up.

Note how I checked the total of the tallies was 100.

Score	Tally	F
1		22
2		16
3		8
4		19
5		13
6		22
		<hr/> 100 <hr/> 2

The column at the end of the tally chart records the **frequency**, the number of times that score came up.

Grouped frequency tally chart

Continuous variables like height are unlikely to have many repeated values even when recorded to the nearest cm.

Below are the heights in centimetres of 100 adults from the UK

153 173 161 160 173 168 148 159 146 149 187 172 167 170 175 170 155 168
 172 159 152 161 160 157 160 178 180 181 157 164 179 178 153 161 156 157
 176 175 160 174 152 168 181 164 185 179 170 159 155 154 179 170 162 155
 159 165 178 176 184 157 181 179 191 158 167 172 171 159 160 164 145 149
 148 158 171 163 192 168 168 168 169 158 147 161 178 192 155 166 161 162
 149 165 151 183 172 168 163 168 147 171

The smallest height is 145 and the tallest person is 192cm, the range is 47cm

We can best organise the data in groups of 10cm from 140 to 200cm using a grouped frequency tally chart like the one below

	Tallies	F
$140 \leq h < 150$		4
$150 \leq h < 160$		24
$160 \leq h < 170$		30
$170 \leq h < 180$		26
$180 \leq h < 190$		8
$190 \leq h < 200$		3
		<hr/> 100 <hr/>

The groups are labelled like $140 \leq h < 150$.

The symbols are read as “140 is less than or equal to the height, which is less than 150”.

That particular group only has the heights between 140 and 149 in it. In theory, someone with height 149.9999 cm would fit into the group as well, so when you draw a frequency bar chart from the frequency table, you would make the bar 10cm wide on the graph scale.

Frequency

The word 'frequency' is used in Maths to mean 'the number of data items that have this value'. On your radio, it is the number of cycles per second that the radio wave oscillates at – quite different. The phrase 'number of...' is sometimes used instead of frequency.

Mean from frequency table

You work out the mean of a list of data by adding up the data values and dividing the total by the number of data values.

You can calculate the mean from the frequency table tallied from a large set of data.

You have to multiply each different value by the frequency of that value, and then divide by the total number of results, which is the total of the frequencies.

Lets look at the two frequency tables we have already tallied up.

Discrete frequency table

The dice scores look like this in a frequency table

Dice score	Frequency
1	22
2	16
3	8
4	19
5	13
6	22

So we know that the score 1 occurred 22 times and contributes 22 to the total score

The score 2 occurred 16 times, so that row of the table contributes 32 to the total

The score 3 occurred 8 times, and that row of the table contributes 24 to the total

You can add columns to the frequency table to record the contribution of each row...

Dice score	Frequency	$X \times F$	Contribution to total score
1	22	1×22	22
2	16	2×16	32
3	8	3×8	24
4	19	4×19	76
5	13	5×13	65
6	22	6×22	132
Totals	100		351

So the mean is the total score divided by the number of dice rolls, or $351 \div 100 = 3.51$

Challenge: Students often get mixed up with calculating the mean or total from a frequency table so find a GCSE textbook and work through some questions in an exercise. Check your answers.

Grouped frequency table for continuous data

With grouped data, you have to assume a value of the variable for each group.

The result can only be an estimate of the mean, because you had to assume a value of height to stand for each of the groups.

Example: calculate an estimate of the mean of the heights of the 100 people summarised in this frequency table

Height (cm)	Frequency
$140 \leq h < 150$	9
$150 \leq h < 160$	24
$160 \leq h < 170$	30
$170 \leq h < 180$	26
$180 \leq h < 190$	8
$190 \leq h < 200$	3

For each of the height groups, we use the midpoint to stand for all the heights in that group.

The midpoint is half way through the group, just the mean of the lower and upper boundary for the height group.

The midpoint for the first group, $140 \leq h < 150$, is just $\frac{(140+150)}{2} = 145$ cm

Height (cm)	Midpoint	Frequency	$X \times F$	Contribution to total height
$140 \leq h < 150$	145	9	145×9	1305
$150 \leq h < 160$	155	24	155×24	3720
$160 \leq h < 170$	165	30	165×30	4950
$170 \leq h < 180$	175	26	175×26	4550
$180 \leq h < 190$	185	8	185×8	1480
$190 \leq h < 200$	195	3	195×3	585
Totals		100		16590

So the mean estimate is the total height divided by the number of people measured,
 $16590 \div 100 = 165.9$ cm

If you add up the original heights the total is 16574 cm and the real mean is 165.74 cm.

The estimate is pretty close!

Median from frequency table

The median average is the value that the middle data value has when you put the list of data values in order of size.

If there is an even number of data items, then you take the mean of the middle two data items to find the median.

The frequency table puts the data values in order of size for us, so finding the median value is just about 'counting through' the frequency values to find the first one that is larger than the half way point.

Discrete frequency table

Example: The scores from 100 dice rolls are tallied into a frequency table. Find the median score.

Dice score	Frequency
1	22
2	16
3	8
4	19
5	13
6	22

The median dice score will be in the $\frac{100+1}{2} = 50.5$ th position in a list of the scores in order of size (see 'median for large dataset' above).

So, we add up the frequencies until we get a value larger than 50;

$$22 + 16 = 38 + 8 = 46 + 19 = 65$$

So the median is in the row of the table with frequency 19, which is the score 4 row.

Both the 50th and 51st scores are 4, so the median score is 4.

Grouped frequency table

Example: The heights of 100 adults have been tallied into a frequency table. Find the median height from the table.

Height (cm)	Frequency
$140 \leq h < 150$	9
$150 \leq h < 160$	24
$160 \leq h < 170$	30
$170 \leq h < 180$	26
$180 \leq h < 190$	8
$190 \leq h < 200$	3

The median height will be in the $\frac{100+1}{2} = 50.5$ th position in a list of the heights in order of size (see 'median for large dataset' above).

So, we add up the frequencies until we get a value larger than 50;

$$9 + 24 = 33 + 30 = 66$$

So we know the media height is somewhere in the row that has frequency 30, or somewhere in the $160 \leq h < 170$ group. You could take the midpoint of that group as a

value for the median, $\frac{(160+170)}{2} = 165$ cm .

Mode or modal group from a frequency table

You find the row with the largest frequency and then write down the value of the data item or group for that row.

Example: the heights of 100 people have been recorded in the frequency table below. Write down the modal group.

Height (cm)	Frequency
$140 \leq h < 150$	9
$150 \leq h < 160$	24
$160 \leq h < 170$	30
$170 \leq h < 180$	26
$180 \leq h < 190$	8
$190 \leq h < 200$	3

The modal group is

$160 \leq h < 170$

Because 30 is the largest frequency

Don't write 30 as the answer!

Bar charts

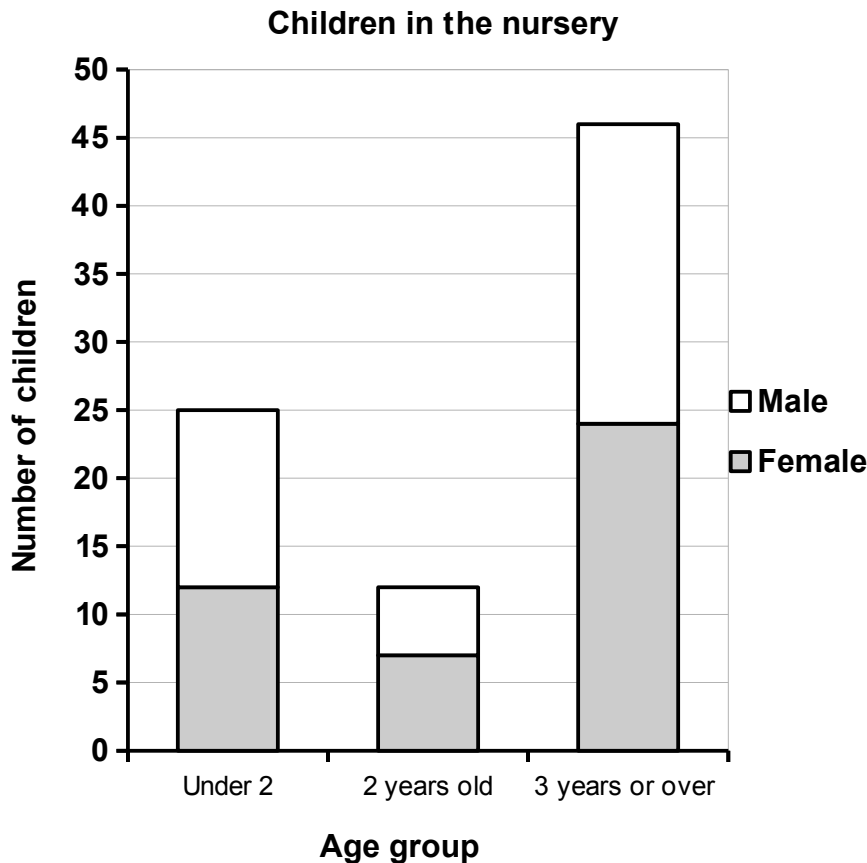
A bar chart is a good way of displaying data in tables.

Example: The table below shows the age and gender breakdown of children in a nursery. Draw a bar chart to illustrate the data.

	Female	Male
Under 2	12	13
2 years old	7	5
3 years or over	24	22

Below are two different ways of presenting this data as a bar chart

Stacked bar chart



The chart has a **title**, the **axes are labelled**, the **frequency axis starts from zero** so we don't overemphasise the differences between the age groups.

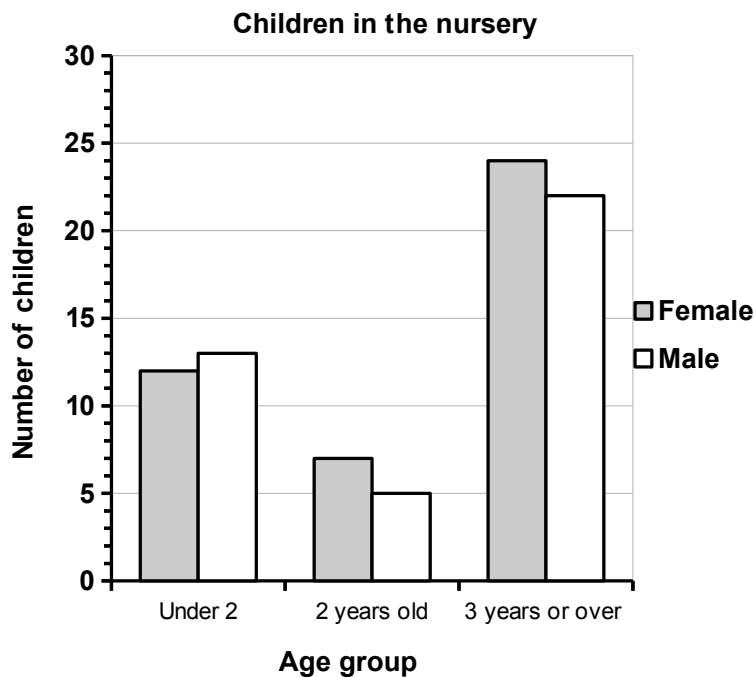
There is a **key** so you know that the grey part of each bar represents the number of girls and the white part of each bar represents the number of boys.

The whole length of the bar tells you the total number of children in the age group for that bar.

This chart emphasises the total numbers in each of the age groups.

Clustered bar chart

Here the male and female bars are side by side, so the emphasis is on comparing the number of children in each gender group rather than the total number in each age group.



Exam questions often ask you to read numbers from the charts and then use the numbers in another calculation.

Example: How many more female than male two year olds are there?

Example: Which age group has less female children than male children?

Pie chart

Example: draw a pie chart of this data about the types of cars in the College car park

Type of car	Number of cars
Smart cars	3
Two door sports	5
Four door saloon	10
Estate	6

Calculate the angle

Stage 1: Add up the numbers of cars $3 + 5 + 10 + 6 = 24$

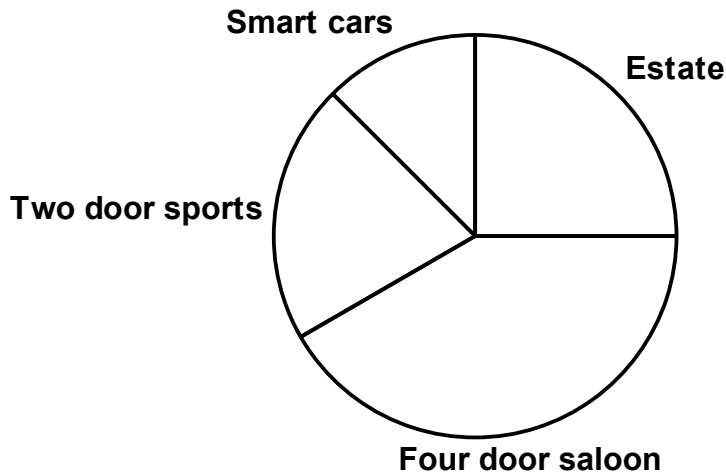
Stage 2: For each row in table, divide the number by the total and multiply by 360

Here are the first two calculations $\frac{3}{24} \times 360 = 45$ and $\frac{5}{24} \times 360 = 75$

Draw the pie chart

You need to do this in a lesson with support from a tutor
Below is what the pie chart for this data would look like

Pie chart of type of cars in car park



Do NOT label the sectors with the number of degrees!

You can label with the names of each row of the table

You can add the number of cars if you want to

The pie chart tells you something about the proportions of the various types of car

Pie chart number puzzles

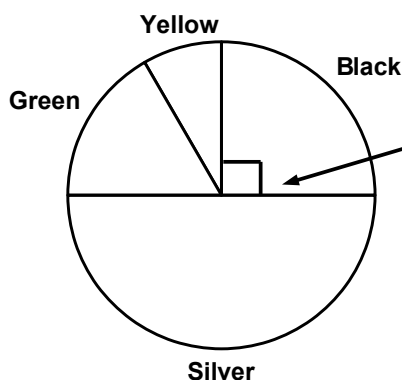
Sometimes a GCSE exam question will give you the pie chart, and ask you to work backwards to find other information.

Example: The pie chart below shows the colours of cars in the car park.

There were 20 black cars when the survey was completed.

Find the total number of cars included in the survey.

Pie chart of colour of cars in car park



Notice the right angle sign, so you know that one quarter of the pie is worth 20 cars.

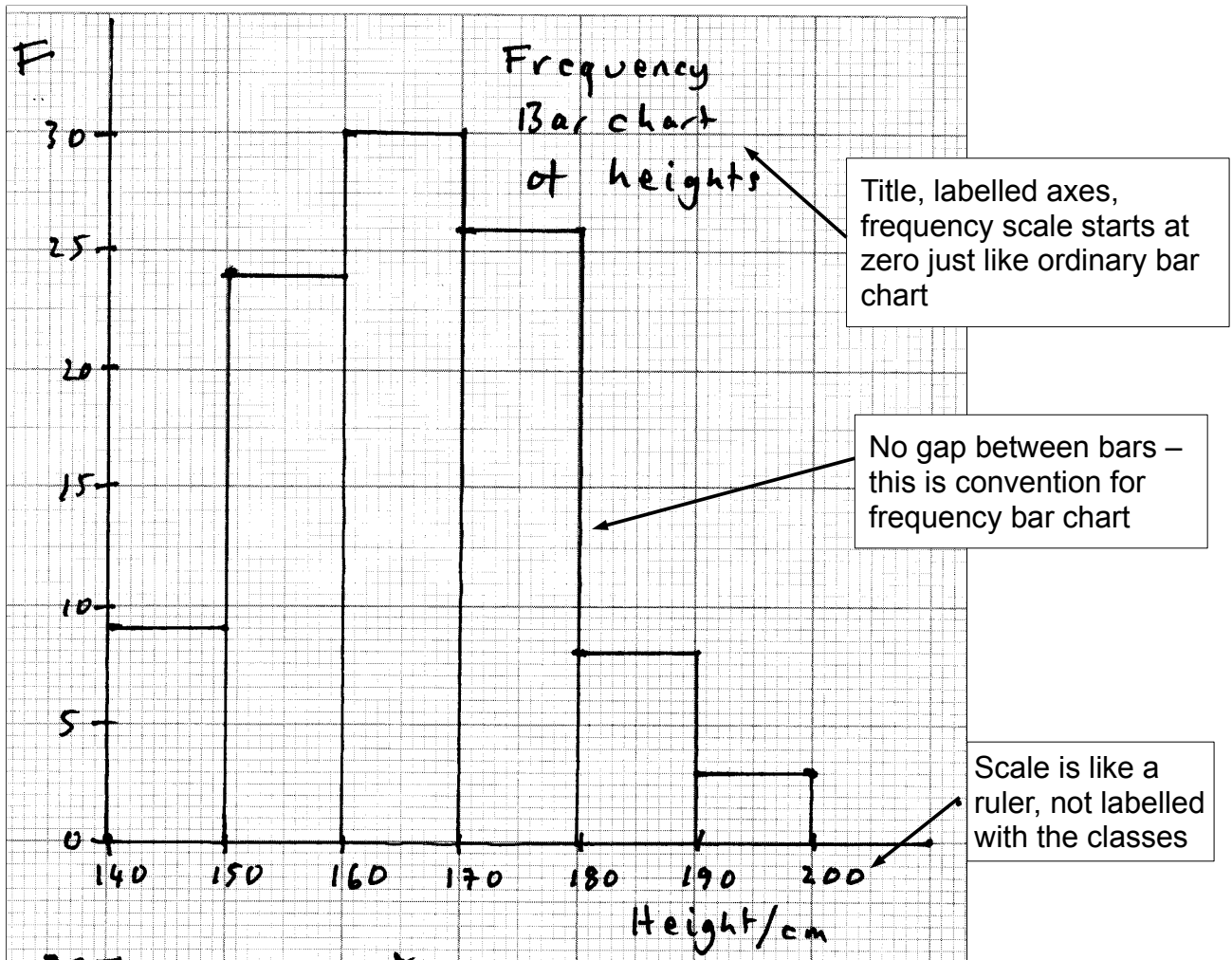
So $4 \times 20 = 80$ cars in the survey.

Frequency Bar charts

Example: The table below shows the heights of 100 people tallied into height groups. Draw a frequency bar chart of this data.

Height (cm)	Frequency
$140 \leq h < 150$	9
$150 \leq h < 160$	24
$160 \leq h < 170$	30
$170 \leq h < 180$	26
$180 \leq h < 190$	8
$190 \leq h < 200$	3

My plot. I'm using a thick pen here so the picture is clear. You could use a sharp HB pencil!



Challenge: print some graph paper and draw your own frequency bar chart from this data.

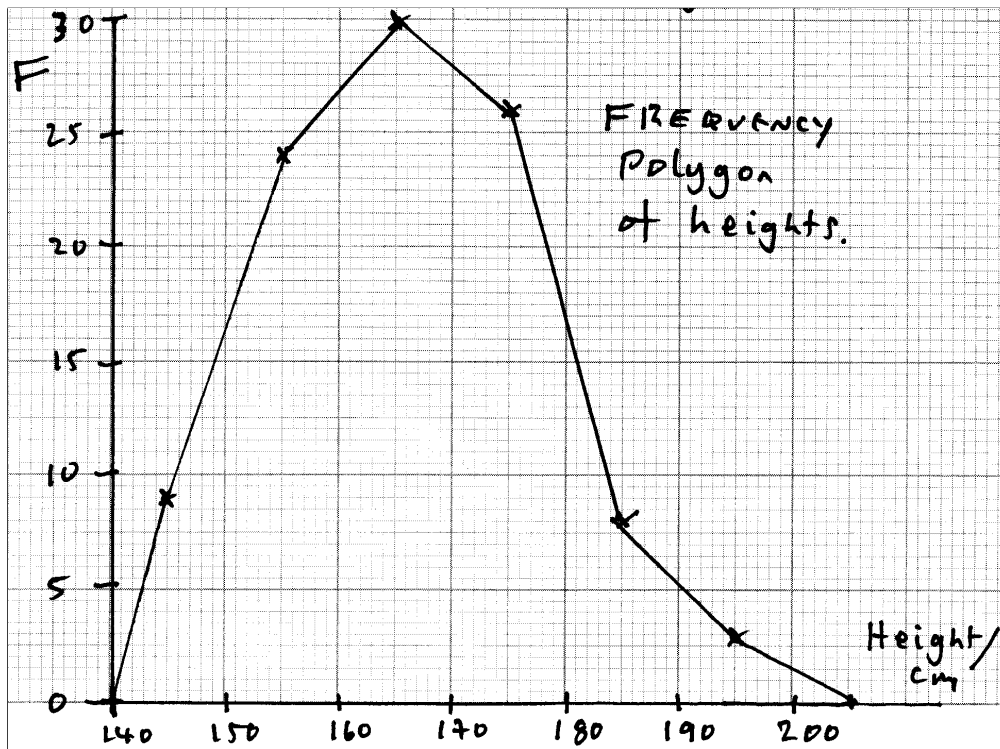
Note: A Histogram is a similar kind of chart. Some books may use that word.

Frequency polygon

Example: The table below shows the heights of 100 people tallied into height groups. Draw a frequency polygon for this data.

Height (cm)	Frequency	Midpoint
$140 \leq h < 150$	9	145
$150 \leq h < 160$	24	155
$160 \leq h < 170$	30	165
$170 \leq h < 180$	26	175
$180 \leq h < 190$	8	185
$190 \leq h < 200$	3	195

My Plot. I've used a thick pen here so the picture is clear. You could use a sharp pencil!



The frequency polygon is very similar to the frequency bar chart.

Instead of bars, you plot points at the midpoint of each group

You join the midpoints together with straight lines

You draw 'extra' lines to the horizontal axis for the first and the last point.

You can present two sets of data on the same axes using a frequency polygon.

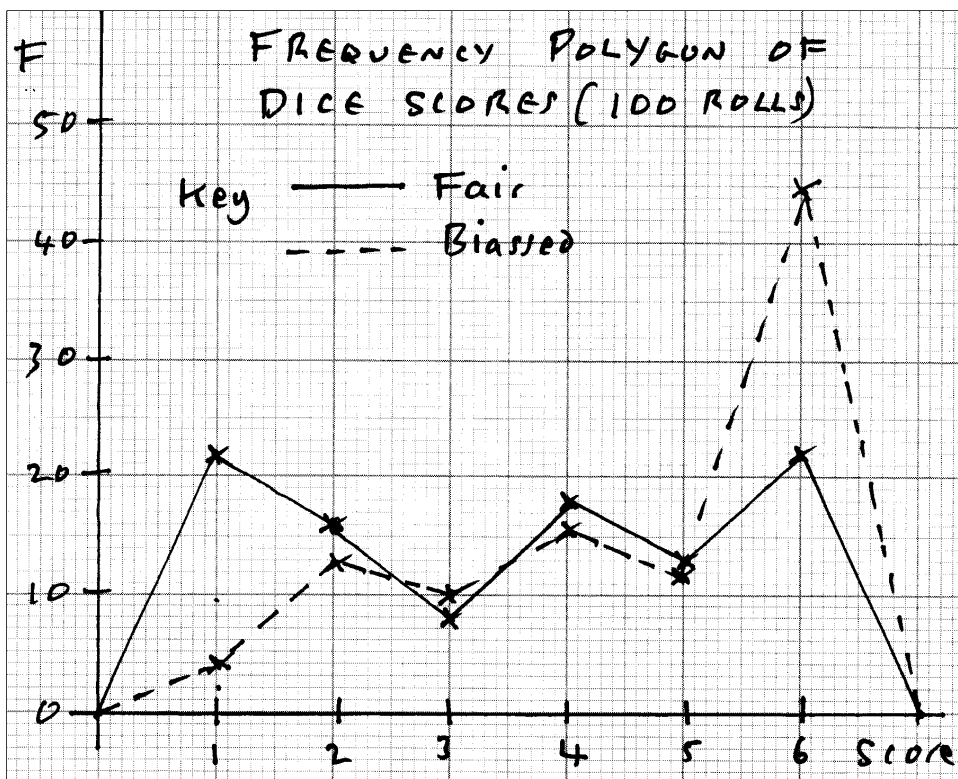
Challenge: print some graph paper and draw your own version of this chart.

Comparing frequency distributions

Example: A fair dice is rolled 100 times and the scores written down and tallied. A second dice that might be biased is also rolled 100 times and the scores tallied. Draw the frequency polygons representing both sets of scores on the same axes.

Dice score	Fair dice Frequency	Biassed dice Frequency
1	22	
2	16	
3	8	
4	19	
5	13	
6	22	

My plot below uses a thick pen for visibility. Your plots could use a sharp pencil.



Notice the key that allows people to see which polygon represents which data.

See how the allegedly biased dice has higher than expected number of 6 scores, and a lower than expected number of 1 scores.

On a dice, the dots on opposite faces always add to 7, so 1 is opposite to 6.

The probability section explains how you calculate the expected number of each score, and explains more about bias.

Stem and leaf diagrams

A stem and leaf diagram is like a frequency bar chart, but is constructed from the numbers themselves.

The stem and leaf diagram is one of a number of statistical diagrams built up using letters and symbols so they could be 'drawn' by early text based personal computers.

Other examples (not on the Foundation syllabus) include the dot plot and the box and whisker plot.

These new diagrams are still used despite our modern computers that can produce graphics easily.

These new kinds of diagram are useful for smaller data sets (30 numbers or less).

Example: The weights of 23 people were recorded to the nearest kilogramme. Draw an ordered stem and leaf diagram from this data.

Weights (Kg) 62, 70, 61, 58, 67, 66, 75, 53, 73, 61, 71, 84, 69, 68, 74, 66, 78, 82, 56, 87, 74, 78, 59

The lowest weight is 53 Kg and the highest weight is 87 Kg.

We use the tens figure as the 'stems' and the units figure as the 'leaves'

We make a line for each of the tens figures

Then draw a line after the 10s figure and copy the units figures of each of the numbers onto the appropriate row... a bit like a tally chart but using the units instead of tallies.

5 | 8 3 6 9

6 | 2 1 7 6 1 9 8 6

7 | 0 5 3 1 4 8 4 8

8 | 4 2 7

As you can see, the leaves are not sorted into order of size.

So you draw a second stem and leaf diagram and put the leaves (units figures) in order of size

5 | 3 6 8 9

6 | 1 1 2 6 6 7 8 9

7 | 0 1 3 4 4 5 8 8

8 | 2 4 7

The underlined 8 in the first row of the stem and leaf diagram is worth 58.

The underlined 4 on the 8 row is worth 84

You need to add a key to the stem and leaf diagram so someone could work out what the 10s were and what the units were.

Key: 8 | 4 represents 84Kg.

Turn the page anticlockwise. The stem and leaf diagram looks a bit like a bar chart.

Median and range from stem and leaf diagram

Example: Below is a stem and leaf diagram of the weights of 23 people.
Find the **median weight** and the **range of weights** from the diagram.

Stem and leaf diagram of weights in Kg of 23 people

5 | 3 6 8 9

6 | 1 1 2 6 6 7 8 9

7 | 0 1 3 4 4 5 8 8

8 | 2 4 7

Key: 8 | 4 represents 84Kg

Median: The stem and leaf diagram puts the weights in order of size for us

$n = 23$ so $\frac{n+1}{2} = \frac{24}{2} = 12^{\text{th}}$ number in the stem and leaf diagram is the median weight.

Counting through to the 12th leaf gives 9 on the 6 row.

The median is 69Kg

Range: Range = Largest – Smallest data value = 87 – 53 = 34Kg

(5 | 3 = 53Kg and 8 | 7 = 87Kg, we just pick the smallest leaf on the first row and the largest leaf on the last row of the stem and leaf diagram)

Other facts from stem and leaf diagram

Example: How many people weighed *more than* 75 kg?

Just count the number of leaves *after* the 5 on the 7 row

I counted 5 leaves, so 5 people weighed more than 75Kg

Challenge: Google 'stem and leaf diagram'. Find the BBC GCSE Maths Bitesize page about stem and leaf diagrams and work through the page.

Scatter diagrams

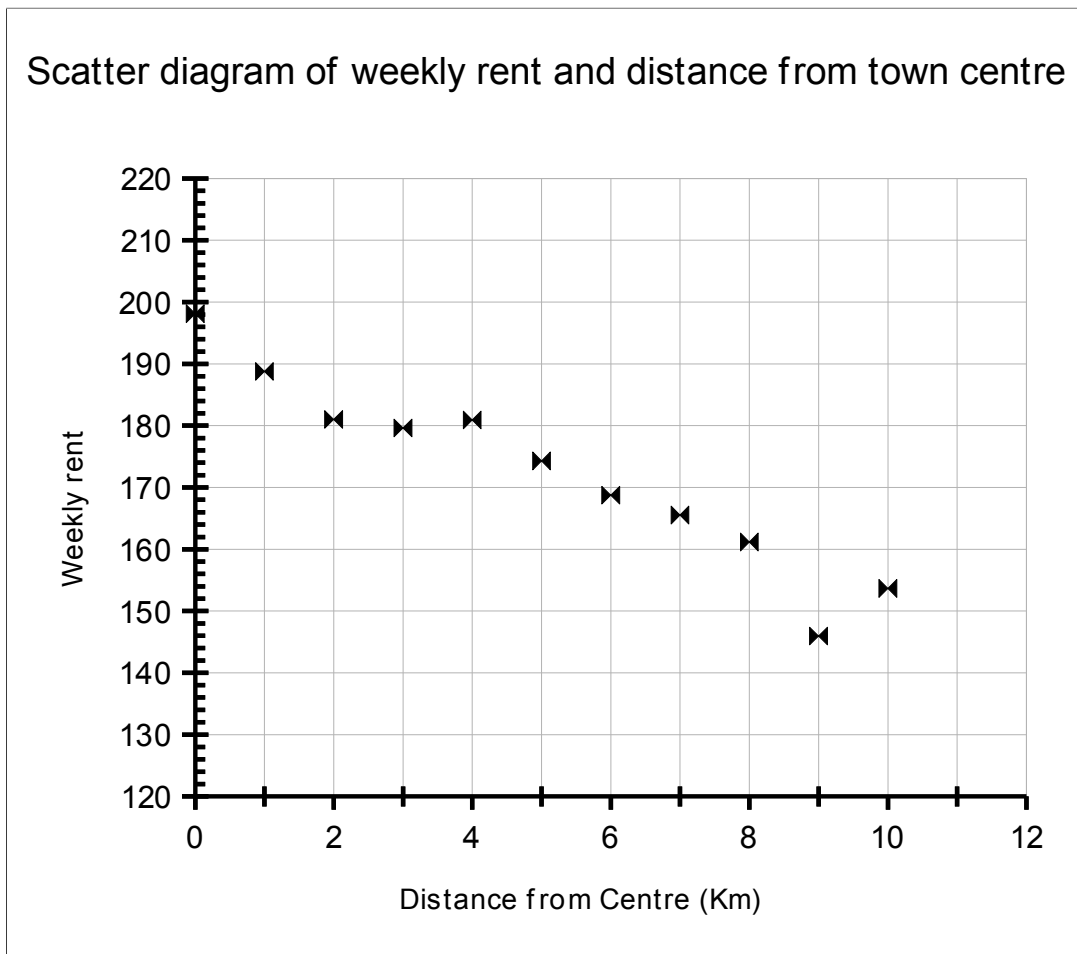
Scatter diagrams show you the relationship between two variables.

Example: Freda checks the rent per week for one bedroom apartments and notices a relationship between the rent and the distance in Km from the town centre.

Freda puts her findings into a table.

Distance Km	0	1	2	3	4	5	6	7	8	9	10
Rent	198	189	181	180	181	174	169	166	161	146	154

Draw a scatter diagram to display this data.



Notice how the vertical axis does not start from zero. Axes in a scatter diagram do not have to start from zero, the starting point is chosen to show the full range of the data.

Each distance and the corresponding rent is plotted as a point, so for instance the apartment that was 4km from the centre has a rent of £181

The points form a pattern. The pattern is not perfect, the rents don't fall on a straight line exactly, but often the pattern can tell you something about the relationship.

In this example, the rent drops the further away from the centre of town the apartment is. But there are anomalies (perhaps just larger rooms or better decoration).

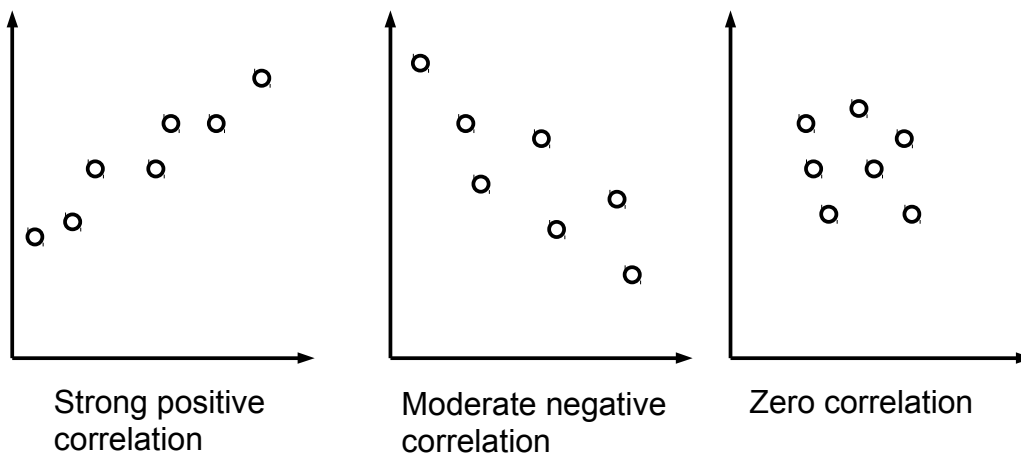
Describing the pattern

Correlation is the word used to describe the relationship between two variables.

The example above about the apartments was an example of **negative correlation** because the flats got cheaper as the distance from the town centre increased.

If you plot the height and weight of (say) sports students on a scatter diagram, there will be a **positive correlation** between height and weight, so taller students will tend to be heavier than shorter students.

There will be **random 'scatter'** though, the pattern might be quite hard to see.



The three sketches above show the possibilities for describing the correlations between two variables.

There is a 'strength' word: 'strong' or 'moderate' or 'weak', that describes how scattered the points are

Then there is a direction word: 'positive' for graphs that go uphill and 'negative' for graphs that go downhill, see gradient later.

If you drop rice grains on a piece of graph paper, you will usually end up with an example of zero correlation. There is no evidence for a relationship between the variables.

Sometimes, your rice grains will form a pattern that shows a strong correlation just by chance. That is why people often say '**correlation does not imply causation**'. They mean that just because an experiment shows a strong correlation, you can't say that one variable controls the other variable unless you find out what the mechanism is.

Challenge: find or think up examples of positive and negative correlation.

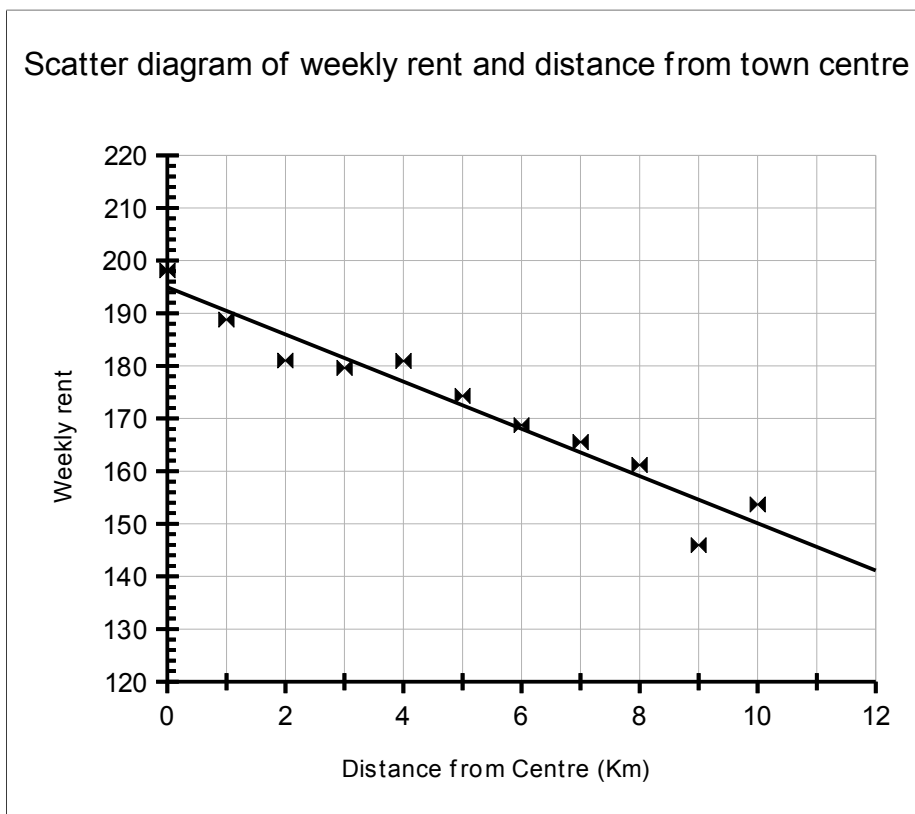
Line of best fit

If you have a strong correlation, then you can draw a line of best fit through your points.

Example: Recall Freda's data about apartment rents and distance from town centre.

Distance Km	0	1	2	3	4	5	6	7	8	9	10
Rent	198	189	181	180	181	174	169	166	161	146	154

Add a line of best fit to the scatter diagram of this data.



As you can see, the line of best fit is a single straight line that passes through the scattered data points in a way that 'averages out' the scatter.

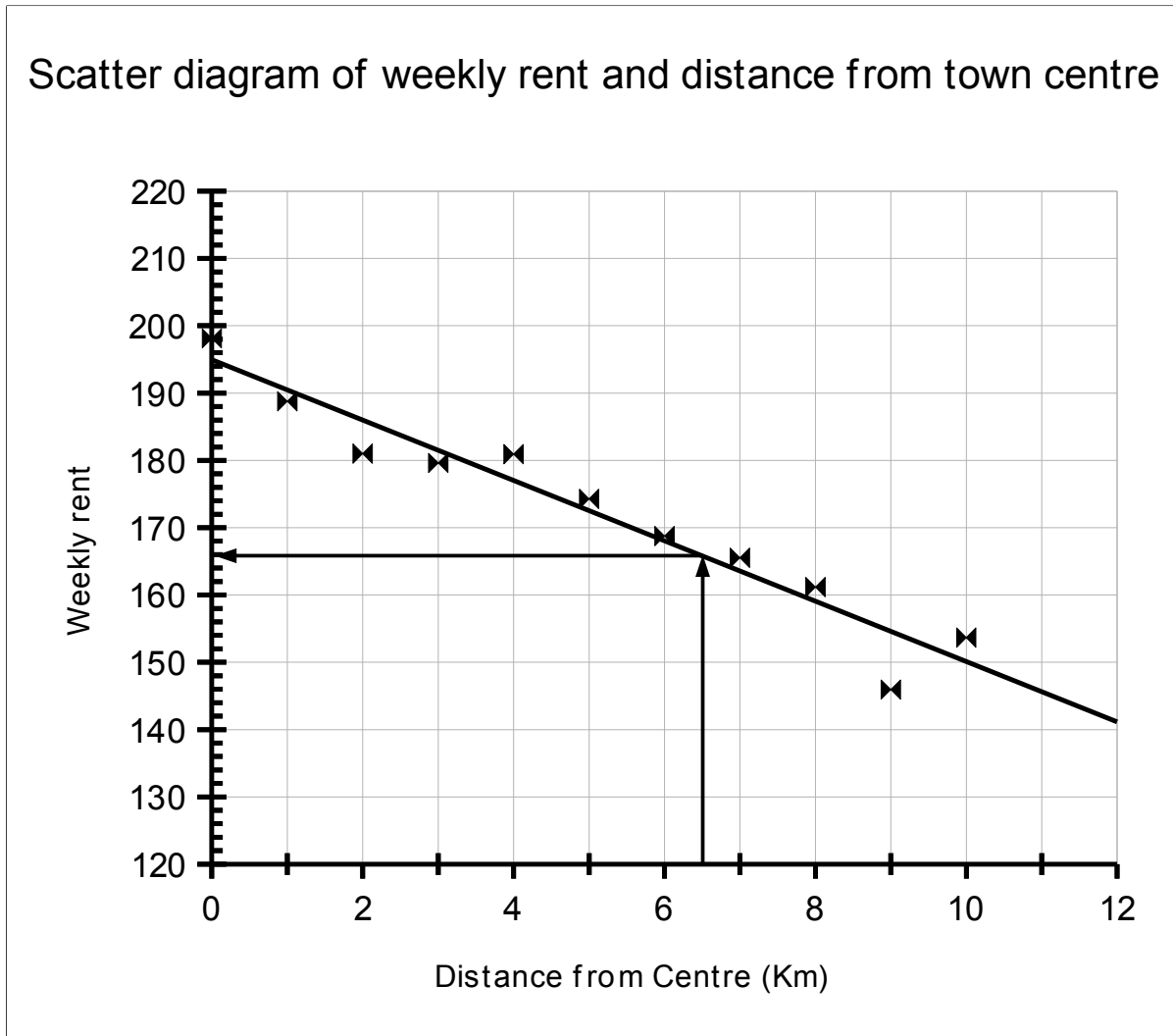
The 'rules' for drawing a straight line are

- The line should **follow the trend** of the data
- There are *roughly* **equal numbers of points above and below the line**
- There are **no special points**; the line does not have to pass through any particular point
- The line does not have to pass through the origin (in the case of positive correlation)
- You should only draw a line of best fit if the pattern of the points shows a moderate or strong correlation

Making predictions

Once you have drawn a line of best fit, you can use your line to estimate the value of one variable corresponding to a value of the second.

Example: Use the line of best fit to estimate the rent for an apartment that is 6.5 km from the centre of town.



Stage 1: Draw a line up from 6.5 km until the line meets the line of best fit

Stage 2: Then draw a line across to the rent axis

Stage 3: Read off the rent value: I estimate it to be £166

You can of course work the other way, finding the distance for a given rent.

Challenge: add a line of best fit to your scatter diagram.

Use the line of best fit to estimate the rent for an apartment that is 6.5 km from the town centre

Your value will be a bit different to mine because of how you draw the line of best fit.

This is expected, and you must show your lines on a scatter diagram in an exam.

Two way tables

Example: A College offers a choice of one of 4 subjects, English, Maths, Science and IT. Each subject has a first year course and a second year course.

The table below shows some of the information in a two way table, with subjects across the top and the years of study down the side.

Complete the table.

	English	Maths	Science	IT	Totals
Year 1	15		12	7	44
Year 2	15			8	46
Totals		25	20		90

Stage 1: check what you know, and fill in the easy gaps.

For instance, the first row has one number missing and a total, so you can easily work out the number of maths students was $44 - 15 - 12 - 7 = 10$ students

Stage 2: Now you know that there were 10 students in maths in the first year, and that there were 25 students in total, you know that there were 15 maths students in the second year.

Stage 3: There were $20 - 12 = 8$ science students in the second year.

Stage 4: The last two gaps are simple addition.

The complete table is shown below.

You should check the rows and columns add up to their respective totals.

	English	Maths	Science	IT	Totals
Year 1	15	10	12	7	44
Year 2	15	15	8	8	46
Totals	30	25	20	15	90

Sampling and bias

If you want to find out what people in the West Midlands think about (say) obesity you can't really ask all of them.

In statistics 'people in the West Midlands' would be called the **population**, and you would pick a small **sample** from the larger population.

The way you select people for your sample is called the sampling method and may have an effect on the reliability of your conclusions.

Random samples

If you wrote down the names of everyone in the West Midlands and put the slips of paper in a (very large) hat and shuffled them really well, and picked 500 names from the hat, you would have a random sample of the population.

You could then ask the people in the sample to fill in a questionnaire and have some confidence in the results.

Bias

In practice, marketing organisations use the **electoral roll** to select samples of adults in the West Midlands.

You can ask for your entry in the electoral roll not to be made available to outside organisations.

People who ask for their entries not to be made available are not included in surveys, so their opinions are not represented.

It may be there is something special about the people who decide to have their electoral roll entries not made available.

This means that the sample based on the electoral roll may not fully represent the population of adults in the West Midlands. The sample is called **biased**.

Example: Oscar thinks that there are too many heavy lorries on the road by his house.

He tallies the number of lorries and other vehicles on Sunday at 1pm for one hour.

Suggest two ways that Oscar could improve the reliability of his results.

Suggestion 1: Observe the vehicles on several days through out the week

Suggestion 2: Observe the vehicles at more than one time during the day, say 8am, 12 noon and 4pm.

Stratified samples

A stratified sample means that you divide the population into groups, then make sure that the same percentage of each group is included in the sample as is present in the population.

Using a stratified sample ensures that all the groups in the population are represented in the sample.

Example: You want to know the views of children, young people and adults about a proposed play centre in a small village.

The population of the village is broken down by age like this

Age group	Population
Children under 16	350
16 to 19 year olds	150
Adults over 19	900

Explain how you could pick a stratified sample of 50 people to take part in a survey about the play centre.

Stage 1: Work out the total number of people in the village, $350 + 150 + 900 = 1400$

Stage 2: Divide the number of people in each group by the total to find the fraction of the population in that group then multiply the decimal fraction by the size of the sample you need.

$$\frac{350}{1400} \times 50 = 12.5 \quad \text{So round up to 13 people in the under 16 age group}$$

$$\frac{150}{1400} \times 50 = 5.3 \quad \text{So round off to 5 people in the 16 to 19 age group}$$

$$\frac{900}{1400} \times 50 = 32.1 \quad \text{So round off to 32 people in the over 19 group}$$

Your sample will then represent the age profile of the village.

You should really still use a random selection method within each age group, but organisations often make do with a **convenient sample**.

Number of measurements for reliability

You need at least 100 repeated measurements to decide a statistical question.

Probability

If you toss a coin, most people would say that there is an 'even' chance of a head, or that there is a 1 in 2 chance of heads, or that there is a fifty-fifty chance of a head.

In Maths, you measure probabilities as a fraction between 0 and 1.

So you would say that there is a probability of $\frac{1}{2}$ of a head when you toss a fair coin.

Probability is the basis of insurance, and probability concepts are widely used in medical research and other branches of science.

Probability words: An **event** can have one or more **outcomes**.

Probability formula: The probability of an outcome is $\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}}$

Example: Write down the probability of rolling an even number on a dice

Stage 1: How many favourable outcomes are there?

Even numbers on a dice are 2, 4, 6 so three favourable outcomes. A dice has 6 faces.

Stage 2: Write as a fraction using the formula $\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}} = \frac{3}{6} = \frac{1}{2}$

I always write the 'raw' fraction and then show the simplified fraction.

Fair and biased coins, dice

A **fair coin** has an equal chance of showing a head or a tail when tossed.

A **biased coin** has more chance of showing one of the sides than the other.

A **fair dice** has an equal chance of showing each of the 6 sides when rolled.

A **biased dice** will show more of one face than expected and less of the opposite face than expected.

Opposite faces of a dice add up to 7, so if a lead plug is added just under the surface of the 1 face, that face will tend to be the bottom face when you roll the dice. You will see more 6s than expected with that dice.

If you want to check if a coin is biased, you will have to toss the coin at least 100 times.

See later under Expected Frequency.

Likelihood words

The likelihood words usually used at GCSE Foundation level include

Impossible, Unlikely, Evens, Likely, Certain

Example: How would you describe the likelihood of throwing a 5 on a dice?

Answer: unlikely. There is one way of scoring a 5 on a dice so that is one chance out of six possible chances.

People often say 'likely' for this example because they see the number 5.

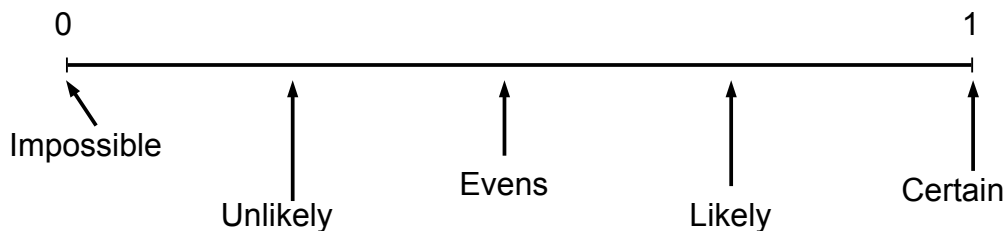
The number 5 is just a label for one face of the dice.

Probability scale

The probability scale is illustrated with a line starting at zero and finishing at 1.

Intermediate points will represent fractions

The likelihood words can be arranged along the scale as shown.



You might also be asked to draw arrows in pointing to specific fractions.

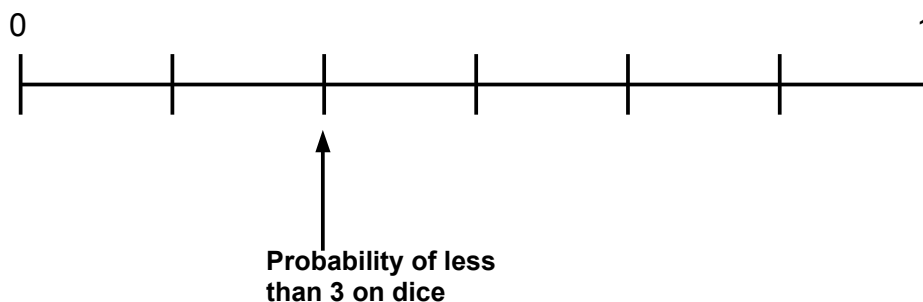
Example: Draw an arrow on the probability scale below showing the probability of throwing a number less than 3 on a dice.



Stage 1: notice how the probability scale has been broken up into sixths.

Stage 2: Less than 3 on a dice means 1 or 2. So two ways of getting the desired outcome out of 6.

Stage 3: As a fraction two out of 6 is $\frac{2}{6}$ so draw the arrow two sections away from the zero end...



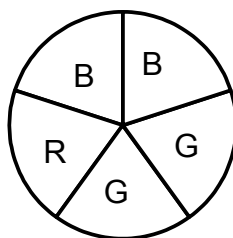
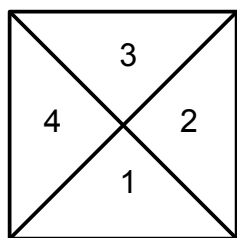
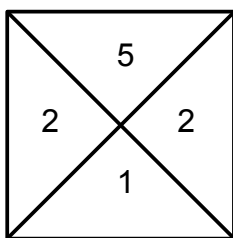
Spinners

A spinner is a piece of card with a cocktail stick pushed through it and sectors marked around the card so that when you spin the spinner each sector has an equal chance of landing.

You can make spinners with as many sectors as you want.

You can label the sectors in any way you want as well.

Look at these examples of spinners.



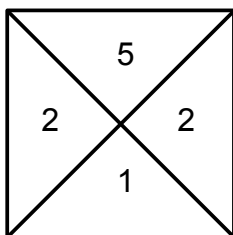
The 5 sector **circular spinner** is labelled with letters that could represent colours (I'm limited to black and white reproduction here).

The **middle spinner** is traditionally labelled with numbers 1 to 4. You can think of it as a four sided dice.

The **first spinner** has unusual numbers! The spinner is more likely to land on 2 because that number is used in two sectors.

The **first spinner is not biased** by the way, the spinner has an equal chance of landing on any one of the four sectors.

Example: Look at the spinner shown below. Write down the probability of the spinner landing on 2 when it is spun.



Stage 1: There are two sectors labelled with the number 2 so two out of four chances

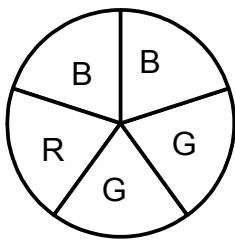
Stage 2: Write as a fraction using the probability formula

$$\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}} = \frac{2}{4} = \frac{1}{2}$$

Challenge: get a laminated spinner sheet from me and try making and using a spinner.

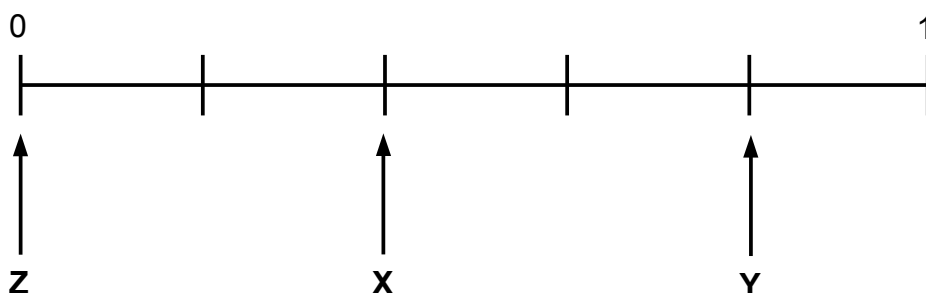
Making up events that fit on a scale

Example: You have a five sided spinner labelled with letters as shown...



Below is a probability scale with three marked probabilities, X, Y and Z

Write down an outcome from spinning the spinner that matches each of the probabilities.



Stage 1: X points to a probability of $\frac{2}{5}$, so either “spinner lands on B” or “spinner lands on G” would be fine

Stage 2: Y points to a probability of $\frac{4}{5}$, so we need a combination of outcomes that gives four sectors.

The only combination that gives four sectors is “spinner lands on G or B”

Stage 3: Z points to a probability of zero.

We need to make up an impossible outcome, something like “spinner lands on A”

Probabilities adding to 1

Example: what is the probability of a dice NOT landing on a 3?

Stage 1: Favourable outcomes include 1, 2, 4, 5, 6, so five favourable outcomes

Stage 2:
$$\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}} = \frac{5}{6}$$

Stage 3: Probability of rolling a 3 is
$$\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}} = \frac{1}{6}$$

Stage 4: Notice that
$$1 - \frac{1}{6} = \frac{6}{6} - \frac{1}{6} = \frac{5}{6}$$

Moral: the probability of something NOT happening is $1 - \text{Probability that it happens}$

Decimal probabilities

You can use decimal numbers to talk about probabilities.

$1 - 0.3 = 0.7$ (think 30p off a pound is 70p)

Find the missing probability

Example: The probabilities of a spinner landing on the letters A, B, C, D are given in the table below. Calculate the probability of the spinner landing on the letter D

Letter	A	B	C	D
Probability	0.1	0.3	x	2x

Stage 1: all the probabilities add to 1, so

So $0.1 + 0.3 + x + 2x = 0.4 + 3x = 1$

So $1 - 0.4 = 0.6 = 3x$

So $x = 0.2$

And probability of letter D = 0.4

(see Algebra for the symbols)

Probability of something not happening

Example: If the probability of the classroom projector working = 0.97, then the probability of it NOT working is $1 - 0.97 = 0.03$

(Think pounds and pence, what change do you get from £1 if it costs 97p?)

Now imagine that written on the till receipt)

In symbols $P(\text{something happens}) + P(\text{the thing not happening}) = 1$

Challenge: A bag contains three yellow balls, two red balls and five blue balls.

You pick one ball from the bag at random.

What is the probability of the ball you pick not being red?

Combined events

Suppose you spin two spinners, or toss two coins

You can calculate the probabilities of these combined events, for instance the probability of tossing two heads or the probability of getting one head and one tail.

One method is to systematically **list all the equally likely outcomes**

Another method is to use something called a **sample space diagram**

List all the possible outcomes of tossing two coins

Example: list all the possible outcomes of tossing two different coins.

Use your list to find the probability of getting at least one head.

Stage 1: One coin can land on H or on T, so there are $2 \times 2 = 4$ outcomes

Stage 2: Add the second coin.

For each outcome of this second coin, the first coin has outcomes HT, so the four outcomes of both coins are...

HH, HT, TH, TT

Stage 4: These four outcomes are all equally likely.

“At least one head” means all the outcomes with one head, HT, TH *and* the TT outcome.

The probability is $\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}} = \frac{3}{4}$

Biology link: You can model Mendelian inheritance in a monohybrid cross by tossing two coins.

Suppose Heads represents the dominant allele and Tails represents the recessive allele.

There are three phenotypes with an H and one with a T, so the ratio of offspring that show the trait to those that do not is 3:1

Sample space diagram: rolling two dice, add the scores

Example: In a game, you roll two dice and add their scores.

Draw a sample space diagram showing all the possible scores and calculate the probability of scoring more than 8.

Stage 1: Draw a two way table with 6 rows and 6 columns to represent the scores from each of the dice

Stage 2: Fill in the table with the sum of each pair of scores

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	<u>9</u>
4	5	6	7	8	<u>9</u>	<u>10</u>
5	6	7	8	<u>9</u>	<u>10</u>	<u>11</u>
6	7	8	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>

Stage 3: Count the entries in the table that are more than 8 (i.e. 9, 10, 11, 12).

I've underlined the entries, there are $4 + 3 + 2 + 1 = 10$ out of 36

$$\frac{\text{Number of favourable outcomes}}{\text{All equally likely outcomes}} = \frac{10}{36} = \frac{5}{18}$$

Expected Frequency

The expected frequency of an outcome is the number of times you would expect to see that outcome when you repeat the event.

Example: You toss a coin 100 times. How many heads would you expect to see?

Common sense answer: 50

Why the common sense answer works

Probability of getting a head is $\frac{1}{2}$ and $\frac{1}{2}$ of 100 is 50

The formula: Expected frequency of outcome = probability \times number of trials.

Example: The probability of the photocopier jamming is 0.03.

I use the photocopier once a day.

I work 240 days a year.

How many days would I expect the photocopier to jam on?

Answer: $0.03 \times 240 = 72$ days

Relative frequency

Reality might not match expectations (see Expected Frequency above).

You might *expect* to see 50 heads when you toss a fair coin 100 times, but I would be very surprised if you did see exactly 50 heads.

You might get 49 heads, or even as few as 40 or as many as 60 without any doubts about the coin.

Relative frequency is the fraction of times you actually observe the outcome.

Example: I toss a coin 10 times and I get 7 heads. Calculate the relative frequency of heads in this experiment.

Answer: Relative frequency = observed outcomes \div number of trials = $7 \div 10 = 0.7$

The relative frequency is a fraction.

It is a measure of probability based on observation rather than on prediction from the geometry of a spinner or dice

Some textbooks call relative frequency the 'experimental probability'. Other books do not distinguish probability and relative frequency at all.

As you increase the number of trials, you might expect the relative frequency to get closer to the probability.

Mega challenge: Find a GCSE textbook and turn to the probability chapter.

Try the questions at the back of the book and check your answers.