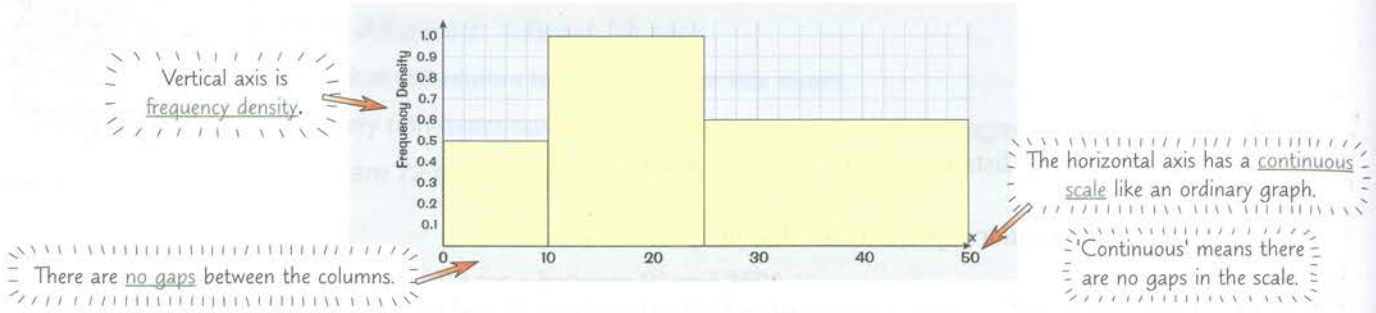
Histograms

Histograms are glorified bar charts. The main difference is that you plot the frequency density rather than the frequency. Frequency density is easy to find — you just divide the frequency by the width of the corresponding class.

Using frequency density means it's a column's area (and not its height) that represents the frequency.



To Draw a **Histogram** it's best to Draw a **Table** First

Getting histograms right depends on finding the right upper and lower boundaries for each class.

EXAMPLE Draw a histogram to represent the data below showing the masses of parcels (given to the nearest 100 g).

Mass of parcel (to nearest 100 g)	100 - 200	300 - 400	500 - 700	800 - 1100
Number of parcels	100	250	600	50

First draw a table showing the upper and lower class boundaries, plus the frequency density:

Smallest mass of parcel that will go in that class.

Biggest mass that will go in that class.

= ucb − lcb

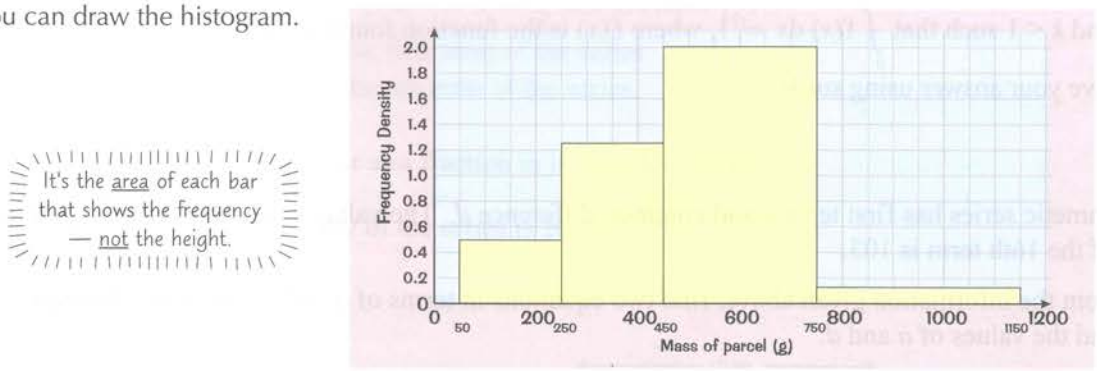
Mass of parcel	Lower class boundary (lcb)	Upper class boundary (ucb)	Class width	Frequency	Frequency density = frequency ÷ class width
100 - 200	50	250	200	100	0.5
300 - 400	250	450	200	250	1.25
500 - 700	450	750	300	600	2
800 - 1100	750	1150	400	50	0.125

= 250 ÷ 200

Look — no gaps between a ucb and the next lcb.

= 1150 − 750

Now you can draw the histogram.



Note: A class with a lower class boundary of 50 g and upper class boundary of 250 g can be written in different ways.

So you might see: "100 – 200 to nearest 100 g"
"50 ≤ mass < 250"
"50–", followed by "250–" for the next class and so on.

They all mean the same — just make sure you know how to spot the lower and upper class boundaries.

Stem and Leaf Diagrams

Stem and Leaf Diagrams look nothing like stems or leaves

Stem and leaf diagrams are an easy way to represent your data.
They come in two flavours — plain and back-to-back.

EXAMPLE The lengths in metres of cars in a car park were measured to the nearest 10 cm.
Draw a stem and leaf diagram to show the following data: 2.9, 3.5, 4.0, 2.8, 4.1, 3.7, 3.1, 3.6, 3.8, 3.7

It's best to do a rough version first, and then put the 'leaves' in order afterwards.

My 'stems' are the numbers before the decimal point, and my 'leaves' are the numbers after.

It's a good idea to cross out the numbers (in pencil) as you add them to your diagram.

2	9, 8
3	5, 7, 1, 6, 8, 7
4	0, 1

Digits after the decimal point — this row represents 4.0 m and 4.1 m.

Put the digits after the decimal point in order

2	8, 9
3	1, 5, 6, 7, 7, 8
4	0, 1

Key 2|9 means 2.9 m

Always give a key.

EXAMPLE The heights of boys and girls in a year 11 class are given to the nearest cm in the back-to-back stem and leaf diagram below. Write out the data in full.

First boy, 8|16|, has height 168 cm. The boys are read backwards.

Boys		Girls
	15	9
8	16	1, 5, 9
9, 8, 1	17	0, 2, 3, 5
5, 2	18	0
1	19	

First girl, |15|9, has height 159 cm.

Key 8|16|5 means
Boys 168 cm and girls 165 cm

Boys: 168, 171, 178, 179, 182, 185, 191

Girls: 159, 161, 165, 169, 170, 172, 173, 175, 180

EXAMPLE Construct a back-to-back stem and leaf diagram to represent the following data:
Boys' test marks: 34, 27, 15, 39, 20, 26, 32, 37, 19, 22
Girls' test marks: 21, 38, 37, 12, 27, 28, 39, 29, 25, 24, 31, 36

Boys		Girls
9, 5	1	2
2, 6, 0, 7	2	1, 7, 8, 9, 5, 4
7, 2, 9, 4	3	8, 7, 9, 1, 6

The 'stems' represent 10, 20 and 30.

Put the digits in the leaves in order

Boys		Girls
9, 5	1	2
7, 6, 2, 0	2	1, 4, 5, 7, 8, 9
9, 7, 4, 2	3	1, 6, 7, 8, 9

Key 0|2|1 means
Boys 20 and girls 21

First things first: remember — there are lies, damned lies and statistics...

Histograms shouldn't cause too many problems — this is quite a friendly topic really. The main things to remember are to work out the lower and upper boundaries of each class properly, and then make sure you use frequency density (rather than just the frequency). Stem and leaf diagrams — hah, they're easy, I do them in my sleep. Make sure you can too.

Location: Mean, Median and Mode

The mean, median and mode are measures of location or central tendency (basically... where the centre of the data lies).

The **Definitions** are really GCSE stuff

You more than likely already know them. But if you don't, learn them now — you'll be needing them loads.

Mean = $\bar{x} = \frac{\Sigma x}{n}$ or $\frac{\Sigma fx}{\Sigma f}$

The Σ (sigma) things just mean you add stuff up — so Σx means you add up all the values of x.

where each x is a data value, f is the frequency of each x-value (the number of times it occurs), and n is the total number of data values.

Median = middle data value when all the data values are placed in order of size.

Mode = most frequently occurring data value.

There are two ways to find the median (but they amount to the same thing):

Either: find the $\left(\frac{n+1}{2}\right)$th value in the ordered list.

If ½(n+1) isn't a whole number, take the average of the terms either side.

Or: (i) if $\frac{n}{2}$ is a whole number (i.e. n is even), then the median is the average of this term and the one above.

(ii) if $\frac{n}{2}$ is not a whole number (i.e. n is odd), just round the number up to find the position of the median.

EXAMPLE Find the mean, median and mode of the following list of data: 2, 3, 6, 2, 5, 9, 3, 8, 7, 2

Put in order first: 2, 2, 2, 3, 3, 5, 6, 7, 8, 9

Mode = 2

Mean = $\frac{2+2+2+3+3+5+6+7+8+9}{10} = 4.7$

Median = average of 5th and 6th values = **4**

Use a **Table** when there are a lot of **Numbers**

EXAMPLE The number of letters received one day in 100 houses was recorded. Find the mean, median and mode of the number of letters.

Number of letters	Number of houses
0	11
1	25
2	27
3	21
4	9
5	7

The number of letters received by each house is a **discrete** quantity (e.g. 3 letters). There isn't a **continuous** set of possible values between getting 3 and 4 letters (e.g. 3.45 letters).

The first thing to do is make a table like this one:

Number of letters x	Number of houses f		fx
0	11	(11)	0
1	25	(36)	25
2	27	(63)	54
3	21		63
4	9		36
5	7		35
totals	100		213

Multiply x by f to get this column.

Put the cumulative frequency (running total) in brackets — it's handy when you're finding the median. (But you can stop when you get past halfway.)

$\Sigma f = 100$

$\Sigma fx = 213$

① The mean is easy — just divide the total of the fx-column (sum of all the data values) by the total of the f-column (= n, the total number of data values).

Mean = $\frac{213}{100}$ = **2.13 letters**

② To find the position of the median, add 1 to the total frequency (= $\Sigma f = n$) and then divide by 2.
Here the median is in position: (100 + 1) ÷ 2 = 50.5.
So the median is halfway between the 50th and 51st data values.
Using your running total of f, you can see that the data values in positions 37 to 63 are all 2s. This means the data values at positions 50 and 51 are both 2 — so **Median = 2 letters**

③ The highest frequency is for 2 letters — so **Mode = 2 letters**

Location: Mean, Median and Mode

If the data's **Grouped** you'll have to **Estimate** the **Mean**

If the data's grouped, you can only estimate the mean and median, and identify a modal class.

There are no precise readings here — each reading's been put into one of these groups.

EXAMPLE The height of a number of trees was recorded. The data collected is shown in this table:

Height of tree to nearest m	0 - 5	6 - 10	11 - 15	16 - 20
Number of trees	26	17	11	6

Find an estimate of the mean height of the trees, and state the modal class.

To estimate the mean, you assume that every reading in a class takes the mid-class value (which you find by adding the lower class boundary to the upper class boundary and dividing by 2). It's best to make another table...

Height of tree to nearest m	Mid-class value x	Number of trees f	fx
0 - 5	2.75	26 (26)	71.5
6 - 10	8	17 (43)	136
11 - 15	13	11	143
16 - 20	18	6	108
	Totals	60 (= Σf)	458.5(= Σfx)

Lower class boundary (lcb) = 0.
Upper class boundary (ucb) = 5.5.
So the mid-class value = (0 + 5.5) ÷ 2 = 2.75.

Estimated mean = $\frac{458.5}{60}$ = **7.64 m**

The **modal class** is the class with the **highest frequency density**. In this example the modal class is **0 - 5 m**.

Linear Interpolation Means Assuming Values are **Evenly Spread**

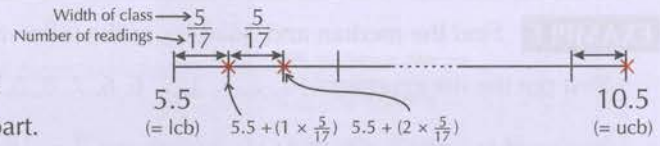
When you have grouped data, you can only estimate the median. To do this, you use (linear) interpolation.

The median position in the above example is (60 + 1) ÷ 2 = 30.5, so the median is the 30.5th reading (halfway between the 30th and 31st). Your 'running total' tells you the median must be in the '6 - 10' class.

Now you have to assume that all the readings in this class are evenly spread.

There are 26 trees before class 6 - 10, so the 30.5th tree is the 4.5th value of this class.

Divide the class into 17 equally wide parts (as there are 17 readings) and assume there's a reading at the end of each part.



Then you want the '4.5th reading' (which is '4.5 × width of 1 part' along).

So the **estimated median** = lower class boundary + (4.5 × width of each 'part') = $5.5 + \left[4.5 \times \frac{5}{17}\right]$ = **6.8 m** (to 1 d.p.)

The Mean, Median and Mode are useful for **Different Kinds** of Data

These three different averages are useful for different kinds of data.

Mean:
- The mean's a good average because you use all your data in working it out.
- But it can be heavily affected by extreme values / outliers.
- And it can only be used with quantitative data (i.e. numbers).

See page 109 for more about outliers.

Median: The median is not affected by extreme values, so this is a good average to use when you have outliers.

Mode:
- The mode can be used even with non-numerical data.
- But some data sets can have more than one mode (and if every value in a data set occurs just once, then the mode isn't very helpful at all).

I can't deny it — these pages really are 'about average'...

If you have large amounts of grouped data ($n > 100$, say), it's usually okay to use the value in position $\frac{n}{2}$ (rather than $\frac{n+1}{2}$) as the median. With grouped data, you can only estimate the median anyway, and if you have a lot of data, that extra 'half a place' doesn't really make much difference. But if in any doubt, use the value in position $\frac{n+1}{2}$ — that'll always be okay.

Variation: Interquartile Range

'Variation' means how spread out your data is. There are different ways to measure it.

The **Range** is a Measure of **Variation**

The range is about the simplest measure of variation you could imagine.

Range = highest value – lowest value

But the range is heavily affected by extreme values, so it isn't really the most useful way to measure variation.

Quartiles divide the data into **Four**

You've seen how the median divides a data set into two halves. Well, the quartiles divide the data into four parts — with 25% of the data less than the lower quartile, and 75% of the data less than the upper quartile.

There are various ways you can find the quartiles, and they sometimes give different results. But if you use the method below, you'll be fine.

1. To find the lower quartile (Q_1), first work out $\frac{n}{4}$.
   (i) if $\frac{n}{4}$ is a whole number, then the lower quartile is the average of this term and the one above.
   (ii) if $\frac{n}{4}$ is not a whole number, just round the number up to find the position of the lower quartile.

2. To find the upper quartile (Q_3), first work out $\frac{3n}{4}$.
   (i) if $\frac{3n}{4}$ is a whole number, then the upper quartile is the average of this term and the one above.
   (ii) if $\frac{3n}{4}$ is not a whole number, just round the number up to find the position of the upper quartile.

EXAMPLE Find the median and quartiles of the following data: 2, 5, 3, 11, 6, 8, 3, 8, 1, 6, 2, 23, 9, 11, 18, 19, 22, 7.

First put the list in order: 1, 2, 2, 3, 3, 5, 6, 6, 7, 8, 8, 9, 11, 11, 18, 19, 22, 23

You need to find Q_1, Q_2 and Q_3, so work out $\frac{n}{4} = \frac{18}{4}$, $\frac{n}{2} = \frac{18}{2}$, and $\frac{3n}{4} = \frac{54}{4}$.

The median is also known as Q_2.

1) $\frac{n}{4}$ is not a whole number (= 4.5), so round up and take the 5th term: $Q_1 = 3$

2) $\frac{n}{2}$ is a whole number (= 9), so find the average of the 9th and 10th terms: $Q_2 = \frac{7+8}{2} = 7.5$

3) $\frac{3n}{4}$ is not a whole number (= 13.5), so round up and take the 14th term: $Q_3 = 11$

If your data is grouped, you might need to use interpolation to find the quartiles. See page 105 for more info.

The **Interquartile Range** is Another Measure of **Variation**

Interquartile range (IQR) = upper quartile (Q_3) – lower quartile (Q_1)

The IQR shows the range of the 'middle 50%' of the data.

EXAMPLE Find the interquartile range of the data in the previous example.

$Q_1 = 3$ and $Q_3 = 11$, so the interquartile range $= Q_3 - Q_1 = 11 - 3 = 8$

Sing-a-long-a-stats — "Home, home on the interquartile range..."

Right then... the range and the interquartile range are both measures of how spread out your data is. The range is pretty crude, though — one freakily high or low value in your dataset and it can become completely misleading. The interquartile range is much better, and is easy to work out (easyish, anyway). Make sure all this is clear in your head before moving on.

Cumulative Frequency Graphs

Cumulative frequency means 'running total'. Cumulative frequency diagrams make medians and quartiles easy to find...

Use *Cumulative Frequency Graphs* to estimate the *Median* and *Quartiles*

EXAMPLE The ages of 200 students are shown. Draw a cumulative frequency graph and use it to estimate the median age, the interquartile range of ages, and how many students have already had their 18th birthday.

Age in completed years	11 - 12	13 - 14	15 - 16	17 - 18
Number of students	50	65	58	27

1 First draw a table showing the upper class boundaries and the cumulative frequency (CF):

Age in completed years	Upper class boundary (ucb)	Number of students, f	Cumulative frequency (CF)
Under 11	11	0	0
11-12	13	50	50
13-14	15	65	115
15-16	17	58	173
17-18	19	27	200

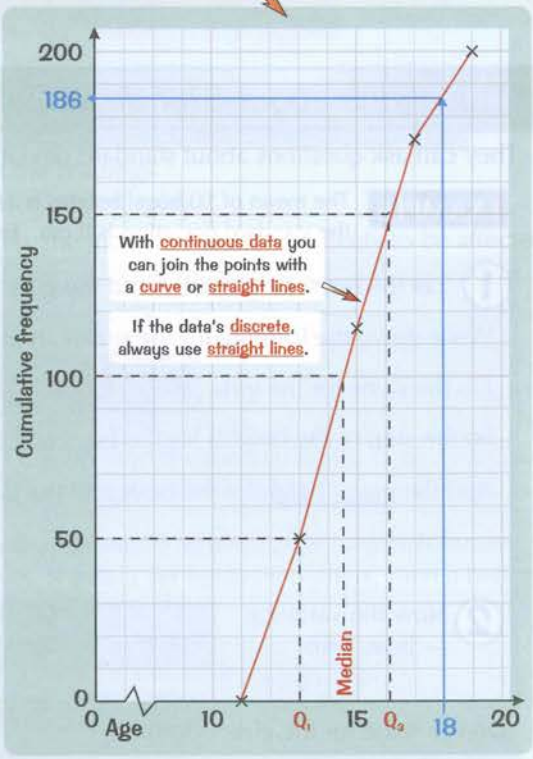
The first reading in a cumulative frequency table must be zero — so add this extra row to show the number of students with age less than 11 is 0.

CF is the number of students with age less than the ucb — it's basically a running total.

The last number in the CF column should always be the total number of readings.

People say they're '18' right up until their 19th birthday — so the ucb of class 17-18 is 19.

Next draw the axes — cumulative frequency always goes on the vertical axis. Here, age goes on the other axis. Then plot the upper class boundaries against the cumulative frequencies, and join the points.



Always plot the upper class boundary of each class.

2 To estimate the median from a graph, go to the median position on the vertical scale and read off the value from the horizontal axis.

Median position = $\frac{1}{2} \times 200 = 100$ so Median = 14.5 years

Because there are so many data values, and because you can only estimate the median (since your data values are in groups), you can say that the median is in position $\frac{n}{2}$ instead of $\frac{n+1}{2}$. And you can use a similar approximation for the position of the quartiles.

Then you can estimate the quartiles in the same way. Find their positions first:

Q_1 position = $\frac{1}{4} \times 200 = 50$,
and so the lower quartile, Q_1 = 13 years

Q_3 position = $\frac{3}{4} \times 200 = 150$,
and so the upper quartile, Q_3 = 16.2 years

The interquartile range (IQR) = $Q_3 - Q_1$. It measures variation. The smaller it is the less variation the data has.

IQR = $Q_3 - Q_1 = 16.2 - 13 = 3.2$ years

3 To estimate how many students have not yet had their 18th birthday, go up from 18 on the horizontal axis, and read off the number of students 'younger' than 18 (= 186).

Then the number of students who are 'older' than 18 is just 200 – 186 = 14 (approximately).

I don't like those frequency tables — I've always wanted to live in a classless society...

Cumulative frequency sounds a bit scarier than running total — but if you remember they're the same thing, then that'll help. And remember to plot the points at the upper class boundary — this makes sense if you remember that a cumulative frequency graph shows how many data-values are less than the figure on the x-axis. The rest is more or less easyish.

Variation: Standard Deviation

Standard deviation and variance both measure how spread out the data is from the mean — the bigger they are, the more spread out your readings are.

The **Formulas** look pretty **Tricky**

The formula is easier to use in these forms.

$$\text{Variance} = \frac{\sum(x-\bar{x})^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2 \quad \text{or} \quad \text{Variance} = \frac{\sum fx^2}{\sum f} - \bar{x}^2$$

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

The x-values are the data, $\bar{x}$ is the mean, f is the frequency of each x, and n (or $\sum f$) is the number of data values.

EXAMPLE Find the mean and standard deviation of the following numbers: 2, 3, 4, 4, 6, 11, 12

1) Find the total of the numbers first: $\sum x = 2 + 3 + 4 + 4 + 6 + 11 + 12 = 42$

2) Then the mean is easy: $\text{Mean} = \bar{x} = \frac{\sum x}{n} = \frac{42}{7} = 6$

3) Next find the sum of the squares: $\sum x^2 = 4 + 9 + 16 + 16 + 36 + 121 + 144 = 346$

4) Use this to find the variance: $\text{Variance} = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{346}{7} - 6^2 = \frac{346 - 252}{7} = \frac{94}{7}$

5) And take the square root to find the standard deviation: $\text{Standard deviation} = \sqrt{\frac{94}{7}} = 3.66$ to 3 sig. fig.

Questions about **Standard Deviation** can look a bit **Weird**

They can ask questions about standard deviation in different ways. But you just need to use the same old formulas.

EXAMPLE The mean of 10 boys' heights is 180 cm, and the standard deviation is 10 cm. The mean for 9 girls is 165 cm, and the standard deviation is 8 cm. Find the mean and standard deviation of the whole group of 19 girls and boys.

(1) Let the boys' heights be x and the girls' heights be y.

Write down the formula for the mean and put the numbers in for the boys: $\bar{x} = \frac{\sum x}{n} \Rightarrow 180 = \frac{\sum x}{10} \Rightarrow \sum x = 1800$

Do the same for the girls: $165 = \frac{\sum y}{9} \Rightarrow \sum y = 1485$

So the sum of the heights for the boys and the girls = $\sum x + \sum y = 1800 + 1485 = 3285$

And the mean height of the boys and the girls is: $\frac{3285}{19} = 172.9$ cm

Round the fraction to 1 d.p. to give your answer. But if you need to use the mean in more calculations, use the fraction (or your calculator's memory) so you don't lose accuracy.

(2) Now the variance — boys first: $\text{Variance} = \frac{\sum x^2}{n} - \bar{x}^2 \Rightarrow 10^2 = \frac{\sum x^2}{10} - 180^2 \Rightarrow \sum x^2 = 10 \times (100 + 32\,400) = 325\,000$

Do the same for the girls: $\text{Variance} = \frac{\sum y^2}{n} - \bar{y}^2 \Rightarrow 8^2 = \frac{\sum y^2}{9} - 165^2 \Rightarrow \sum y^2 = 9 \times (64 + 27\,225) = 245\,601$

Okay, so the sum of the squares of the heights of the boys and the girls is: $\sum x^2 + \sum y^2 = 325\,000 + 245\,601 = 570\,601$

So for all the heights, the variance is: $\text{Variance} = \frac{570\,601}{19} - \left(\frac{3285}{19}\right)^2 = 139.0\ \text{cm}^2$

Don't use the rounded mean (172.9) — you'll lose accuracy.

And finally the standard deviation of the boys and the girls is: $\text{standard deviation} = \sqrt{139.0} = 11.8$ cm

Phew.

People who enjoy this stuff are standard deviants...

The formula for the variance looks pretty scary, what with the x^2's and $\bar{x}$'s floating about. But it comes down to 'the mean of the squares minus the square of the mean'. That's how I remember it anyway — and my memory's rubbish.

Variation and Outliers

Use **Mid-Class Values** if your data's in a **Grouped Table**

With grouped data, assume every reading takes the mid-class value. Then use the frequencies to find $\sum fx$ and $\sum fx^2$.

EXAMPLE The heights of sunflowers in a garden were measured and recorded in the table below. Estimate the mean height and the standard deviation.

Height of sunflower, h (cm)	$150 \le h < 170$	$170 \le h < 190$	$190 \le h < 210$	$210 \le h < 230$
Number of sunflowers	5	10	12	3

Draw up another table, and include columns for the mid-class values x, as well as fx and fx^2:

Height of sunflower (cm)	Mid-class value, x	x^2	f	fx	fx^2
$150 \le h < 170$	160	25600	5	800	128000
$170 \le h < 190$	180	32400	10	1800	324000
$190 \le h < 210$	200	40000	12	2400	480000
$210 \le h < 230$	220	48400	3	660	145200
		Totals	30 ($= \Sigma f$)	5660 ($= \Sigma fx$)	1077200 ($= \Sigma fx^2$)

fx^2 means $f \times (x^2)$ — not $(fx)^2$.

Now you've got the totals in the table, you can calculate the mean and standard deviation:

$$\text{Mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{5660}{30} = 189 \text{ cm to 3 sig. fig.}$$

$$\text{Variance} = \frac{\sum fx^2}{\sum f} - \bar{x}^2 = \frac{1\,077\,200}{30} - \left(\frac{5660}{30}\right)^2 = 312 \text{ to 3 sig. fig.}$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = 17.7 \text{ cm to 3 sig. fig.}$$

Outliers can **Mess Up** some measures of **Central Tendency** and **Variation**

1) An outlier is a freak piece of data that lies a long way from the rest of the readings. If your data includes outliers, that might affect how you choose to describe or display it.

2) Some measures are more affected by outliers than others. For example, the mean is much more likely to be affected by outliers than the median, so for data with outliers, the median is usually a better measure of central tendency.

EXAMPLE Look at the following data set: 1, 2, 2, 3, 3, 5, 5, 5, 6, 7, 7, 7, 9, 9, 10, 10, 11, 12, 13, 85

The data set has mean 10.6 and median 7.
But that value of 85 is an outlier — it's much bigger than the other values.
If we ignore the outlier, the mean of the other values is 6.68, and the median is still 7 — the outlier has a big effect on the mean, but doesn't change the median at all.

3) Outliers can make the variance (and standard deviation) much larger than it would otherwise be — which means these freak pieces of data are having more influence than they deserve. If a data set contains outliers, then a better measure of variation is the interquartile range.

Let's look at the variation of the data set above. The variance with the outlier is 302.94, and without the outlier it's 12.22 — that's a pretty massive difference. On the other hand, the IQR with the outlier included is 6 — without it, the IQR is 7.

4) You might also need to think about outliers when you're deciding which sort of graph to use to display your data.

The box-and-whisker plot (see p111) for the data set above would have a really long 'whisker' on the right hand side. You couldn't tell from that whether the values above Q_3 are evenly spread, or if there's an outlier. But you could draw a histogram with a really short, wide bar on the right — this wouldn't tell you for sure that there was just one outlier, but it would make it clearer there aren't many values near the top of the range.

'Outlier' is the name I give to something that my theory can't explain...

Measures of location (or central tendency) and variation should capture the essential characteristics of a data set in just one or two numbers. So don't choose a measure that's heavily affected by freaky, far-flung outliers — it won't be much good.

Coding

Coding can make the Numbers much Easier

Coding means doing something to every reading (like adding or multiplying by a number) to make life easier.

Finding the mean of 1001, 1002 and 1006 looks hard(ish). But take 1000 off each number and finding the mean of what's left (1, 2 and 6) is much easier — it's 3. So the mean of the original numbers must be 1003. That's coding.

You usually change your original variable, x, to an easier one to work with, y (so here, if $x = 1001$, then $y = 1$).

Write down a formula connecting the two variables: e.g. $y = \frac{x - b}{a}$.

Then $\bar{y} = \frac{\bar{x} - b}{a}$ where $\bar{x}$ and $\bar{y}$ are the means of variables x and y.

Also **standard deviation of y's** $= \frac{\text{standard deviation of } x\text{'s}}{a}$

You can add/subtract a number, and multiply/divide by one as well — it all depends on what will make life easiest.

Note that if you don't multiply or divide your readings by anything (i.e. if $a = 1$), then the standard deviation isn't changed.

EXAMPLE Find the mean and standard deviation of: 1 000 020, 1 000 040, 1 000 010 and 1 000 050.

The obvious thing to do is subtract a million from every reading to leave 20, 40, 10 and 50.
Then make life even simpler by dividing by 10 — giving 2, 4, 1 and 5.

(1) So use the coding: $y = \frac{x - 1000\,000}{10}$. Then $\bar{y} = \frac{\bar{x} - 1000\,000}{10}$ and s.d. of $y = \frac{\text{s.d. of } x}{10}$.

s.d. = standard deviation

(2) Find the mean and standard deviation of the y values: $\bar{y} = \frac{2 + 4 + 1 + 5}{4} = 3$

$$\text{s.d. of } y = \sqrt{\frac{2^2 + 4^2 + 1^2 + 5^2}{4} - 3^2} = \sqrt{\frac{46}{4} - 9} = \sqrt{2.5} = 1.58 \text{ to 3 sig. fig.}$$

(3) Then use the formulas to find the mean and standard deviation of the original values:

$\bar{x} = 10\bar{y} + 1000\,000 = (10 \times 3) + 1000\,000 = 1000\,030$

s.d. of $x = 10 \times$ s.d. of $y = 10 \times 1.58 = 15.8$

And so variance of the x's = (s.d. of x's)^2 = (10 × s.d. of y's)^2
= 10^2 × (s.d. of y's)^2 = 10^2 × variance of the y's.

You can use coding with Summarised Data

This kind of question looks tricky at first — but use the same old formulas and it's a piece of cake.

EXAMPLE A set of 10 numbers (x-values) can be summarised as shown: $\sum(x - 10) = 15$ and $\sum(x - 10)^2 = 100$
Find the mean and standard deviation of the numbers.

(1) Okay, the obvious first thing to try is: $y = x - 10$ ⟹ That means: $\sum y = 15$ and $\sum y^2 = 100$

(2) Work out $\bar{y}$ and the standard deviation of the y's using the normal formulas: $\bar{y} = \frac{\sum y}{n} = \frac{15}{10} = 1.5$

$$\text{Variance of } y = \frac{\sum y^2}{n} - \bar{y}^2 = \frac{100}{10} - 1.5^2 = 10 - 2.25 = 7.75$$
so standard deviation of $y = \sqrt{7.75} = 2.78$ to 3 sig. fig.

(3) Then finding the mean and standard deviation of the x-values is easy: $\bar{x} = \bar{y} + 10 = 1.5 + 10 = 11.5$

The s.d. of x is the same as the s.d. of y since you've only subtracted 10 from every number. ⟹ s.d. of x = s.d. of y = 2.78 to 3 sig. fig.

I thought coding would be a little more... well, James Bond...

Coding data isn't hard — the only tricky thing can be to work out how best to code it, although there will usually be some pretty hefty clues in the question if you care to look. But remember that adding/subtracting a number from every reading won't change the variation (the variance or standard deviation), but multiplying/dividing readings by something will.

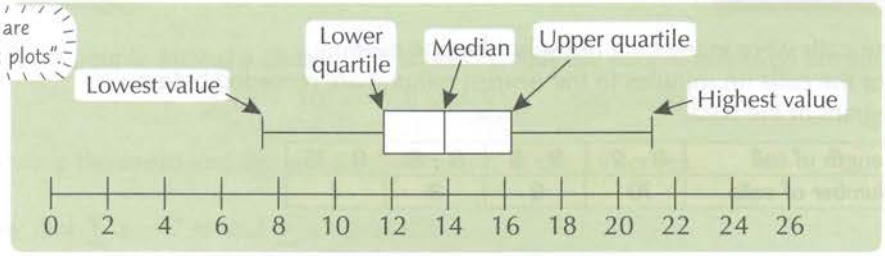
Comparing Distributions

To compare data sets, you need to know how to use all the formulas and what the results tell you.

Box-and-Whisker Plots are a Visual Summary of a Distribution

Box-and-whisker plots show the median and quartiles in an easy-to-look-at kind of way. They look like this:

Box-and-whisker plots are sometimes just called "box plots".



ALWAYS DRAW A SCALE

Use Location and Variation to Compare Distributions

EXAMPLE

This table summarises the marks obtained in Maths 'calculator' and 'non-calculator' papers. Compare the location and variation of the distributions.

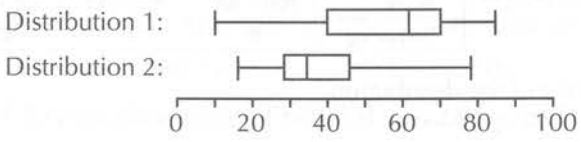
Calculator Paper		Non-calculator paper
28	Minimum	12
78	Maximum	82
40	Lower quartile, Q_1	35
58	Median, Q_2	42
70	Upper quartile, Q_3	56
55	Mean	46.1
21.2	Standard deviation	17.8

Location: The mean, the median and the quartiles are all higher for the calculator paper. This means that scores were generally higher on the calculator paper.

Although the maximum mark on the non-calculator paper was higher than the maximum mark on the calculator paper, this doesn't say anything about the results generally.

Variation: The interquartile range (IQR) for the calculator paper is $Q_3 - Q_1 = 70 - 40 = 30$.
The interquartile range (IQR) for the non-calculator paper is $Q_3 - Q_1 = 56 - 35 = 21$.
The range for the calculator paper is $78 - 28 = 50$.
The range for the non-calculator paper is $82 - 12 = 70$.
The IQR and the standard deviation are both bigger for the calculator paper, so it looks like the scores on the calculator paper are more spread out than for the non-calculator paper. Although the range is bigger for the non-calculator paper, this might not be a reliable guide to the variation since the data may well contain outliers.

EXAMPLE Compare the distributions represented by the box-and-whisker plots below.



Location: The median and the quartiles are higher for Distribution 1, showing that these data values are generally higher than for Distribution 2.

Variation: The interquartile range (IQR) and the range for Distribution 1 are bigger, showing that the values are more varied for Distribution 1 than for Distribution 2.

That's the end of the Data section — hurrah for that...

On exam day, you could be asked to compare two distributions. Just work out any measures of location and variation you can. Then say which distribution has a higher value for each measure, and what it means — e.g. a higher variance means the values are more spread out, while a higher mean means the scores are generally higher. And so on.

Random Events and Venn Diagrams

Random events happen by chance. Probability is a measure of how likely they are. It can be a chancy business.

A Random Event has Various Outcomes

1) In a trial (or experiment) the things that can happen are called outcomes (so if I time how long it takes to eat my dinner, 63 seconds is a possible outcome).
2) Events are 'groups' of one or more outcomes (so an event might be 'it takes me less than a minute to eat my dinner every day one week').
3) When all outcomes are equally likely, you can work out the probability of an event by counting the outcomes:

$$\text{P(event)} = \frac{\text{Number of outcomes where event happens}}{\text{Total number of possible outcomes}}$$

EXAMPLE Suppose I've got a bag with 15 balls in — 5 red, 6 blue and 4 green.

If I take a ball out without looking, then any ball is equally likely — there are 15 possible outcomes. Of these 15 outcomes, 5 are red, 6 are blue and 4 are green. And so...

$\text{P(red ball)} = \frac{5}{15} = \frac{1}{3}$ $\quad \text{P(blue ball)} = \frac{6}{15} = \frac{2}{5}$ $\quad \text{P(red or green ball)} = \frac{9}{15} = \frac{3}{5}$

If I then do 90 trials (i.e. I pick a ball out 90 times, replacing the ball each time), then I would expect to pick:

a red ball $\frac{1}{3} \times 90 = 30$ times $\quad$ a blue ball $\frac{2}{5} \times 90 = 36$ times $\quad$ either a red or a green ball $\frac{3}{5} \times 90 = 54$ times

You can also use relative frequencies to assign probabilities — you use the results of trials you've already carried out.

$$\text{P(event)} = \frac{\text{Number of trials where event happened}}{\text{Total number of trials carried out}}$$

The Sample Space is the Set of All Possible Outcomes

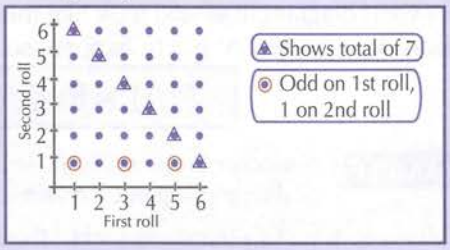
Drawing the sample space (called S) helps you count the outcomes you're interested in.

EXAMPLE The classic probability machine is a dice. If you roll it twice, you can record all the possible outcomes in a 6 × 6 table (a possible diagram of the sample space).

There are 36 outcomes in total. You can find probabilities by counting the ones you're interested in (and using the above formula). For example:

(i) The probability of an odd number and then a '1'. There are 3 outcomes that make up this event, so the probability is: $\frac{3}{36} = \frac{1}{12}$

(ii) The probability of the total being 7. There are 6 outcomes that correspond to this event, giving a probability of: $\frac{6}{36} = \frac{1}{6}$



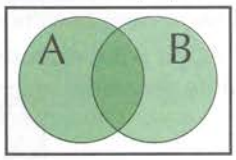
Venn Diagrams show which Outcomes correspond to which Events

Say you've got 2 events, A and B — a Venn diagram can show which outcomes satisfy event A, which satisfy B, which satisfy both, and which satisfy neither.

(i) All outcomes satisfying event A go in one part of the diagram, and all outcomes satisfying event B go in another bit.

(ii) If they satisfy 'both A and B', they go in the dark green middle bit, written $A \cap B$ (and called the intersection of A and B).

(iii) The whole of the green area is written $A \cup B$ — it means 'either A or B' (and is called the union of A and B).

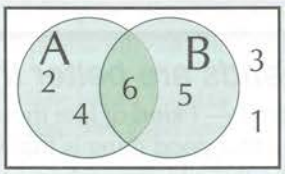


Again, you can work out probabilities of events by counting outcomes and using the formula above. You can also get a nice formula linking $P(A \cap B)$ and $P(A \cup B)$.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If you just add up the outcomes in A and B, you end up counting $A \cap B$ twice — that's why you have to subtract it.

EXAMPLE If you roll a dice, event A could be 'I get an even number', and B 'I get a number bigger than 4'. The Venn diagram would be:



$P(A) = \frac{3}{6} = \frac{1}{2}$ $\quad P(B) = \frac{2}{6} = \frac{1}{3}$ $\quad P(A \cap B) = \frac{1}{6}$ $\quad P(A \cup B) = \frac{4}{6} = \frac{2}{3}$

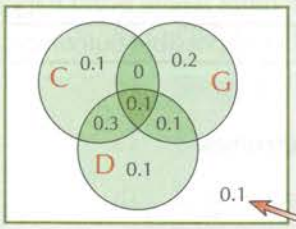
Here, I've just counted outcomes — but I could have used the formula.

Random Events and Venn Diagrams

You can also use Venn diagrams to show probabilities...

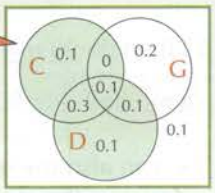
EXAMPLE A survey was carried out to find what pets people like.

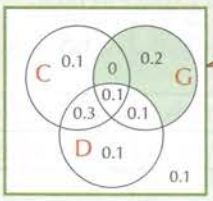
The probability they like dogs is 0.6. The probability they like cats is 0.5. The probability they like gerbils is 0.4. The probability they like dogs and cats is 0.4. The probability they like cats and gerbils is 0.1, and the probability they like gerbils and dogs is 0.2. Finally, the probability they like all three kinds of animal is 0.1.
You can draw all this in a Venn diagram. (Here I've used C for 'likes cats', D for 'likes dogs' and G for 'likes gerbils'.)



1) Stick in the middle one first — 'likes all 3 animals' (i.e. C ∩ D ∩ G).
2) Then do the 'likes 2 animals' probabilities by taking 0.1 from each of the given 'likes 2 animals' probabilities. (If they like 3 animals, they'll also be in the 'likes 2 animals' bits.)
3) Then do the 'likes 1 kind of animal' probabilities, by making sure the total probability in each circle adds up to the probability in the question.
4) Finally, subtract all the probabilities so far from 1 to find 'likes none of these animals'.

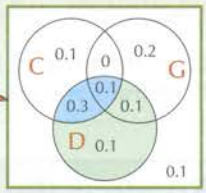
① From the Venn diagram, the probability that someone likes either dogs or cats is 0.7.





② The probability that someone likes gerbils but not dogs is 0.2.

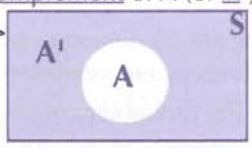
③ You can work out the probability that a dog-lover also like cats by ignoring everything outside the 'dogs' circle.



$$P(\text{dog-lover also like cats}) = \frac{0.3 + 0.1}{0.3 + 0.1 + 0.1 + 0.1} = \frac{2}{3}$$

The Complement of 'Event A' is 'Not Event A'

An event A will either happen or not happen. The event 'A doesn't happen' is called the complement of A (or A').
On a Venn diagram, it would look like this (because A ∪ A' = S, the sample space):



At least one of A and A' has to happen, so...

$$P(A) + P(A') = 1 \quad \text{or} \quad P(A') = 1 - P(A)$$

EXAMPLE A teacher keeps socks loose in a box. One morning, he picks out a sock. He calculates that the probability of then picking out a matching sock is 0.56. What is the probability of him not picking a matching sock?

Call event A 'picks a matching sock'. Then A' is 'doesn't pick a matching sock'. Now A and A' are complementary events (and P(A) = 0.56), so P(A) + P(A') = 1, and therefore P(A') = 1 − 0.56 = 0.44

Mutually Exclusive Events Have No Overlap

If two events can't both happen at the same time (i.e. P(A ∩ B) = 0) they're called mutually exclusive (or just 'exclusive').
If A and B are exclusive, then the probability of A or B is: P(A ∪ B) = P(A) + P(B). ← Use the formula from page 115, but put P(A ∩ B) = 0.
More generally,

For *n* exclusive events (i.e. only one of them can happen at a time):
$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

EXAMPLE Find the probability that a card pulled at random from a standard pack of cards (no jokers) is either a picture card (a Jack, Queen or King) or the 7, 8 or 9 of clubs.

Call event A — 'I get a picture card', and event B — 'I get the 7, 8 or 9 of clubs'.
Events A and B are mutually exclusive — they can't both happen. Also, $P(A) = \frac{12}{52} = \frac{3}{13}$ and $P(B) = \frac{3}{52}$.
So the probability of either A or B is: $P(A \cup B) = P(A) + P(B) = \frac{12}{52} + \frac{3}{52} = \frac{15}{52}$

Two heads are better than one — though only half as likely using two coins...

I must admit — I kind of like these pages. This stuff isn't too hard, and it's really useful for answering loads of questions. And one other good thing is that Venn diagrams look, well, nice somehow. But more importantly, when you're filling one in, the thing to remember is that you usually need to 'start from the inside and work out'.

Tree Diagrams

Tree diagrams — they blossom from a tiny question-acorn into a beautiful tree of possibility. Inspiring and useful.

Tree Diagrams Show Probabilities for **Two or More** Events

Each 'chunk' of a tree diagram is a trial, and each branch of that chunk is a possible outcome. Multiplying probabilities along the branches gives you the probability of a series of outcomes.

EXAMPLE If Susan plays tennis one day, the probability that she'll play the next day is 0.2. If she doesn't play tennis, the probability that she'll play the next day is 0.6. She plays tennis on Monday. What is the probability she plays tennis:

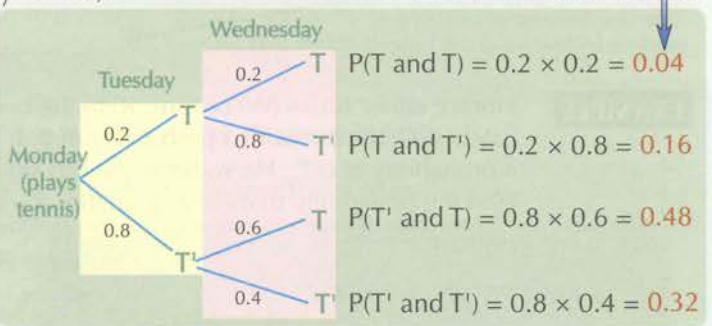
(i) on both the Tuesday and Wednesday of that week?
(ii) on the Wednesday of the same week?

Let T mean 'plays tennis' (and then T' means 'doesn't play tennis').

(i) Then the probability that she plays on Tuesday and Wednesday is P(T and T) = 0.2 × 0.2 = 0.04 (multiply probabilities since you need a series of outcomes — T and then T).

(ii) Now you're interested in either P(T and T) or P(T' and T). To find the probability of one event or another happening, you have to add probabilities: P(plays on Wednesday) = 0.04 + 0.48 = 0.52.

Notice that these add up to 1.



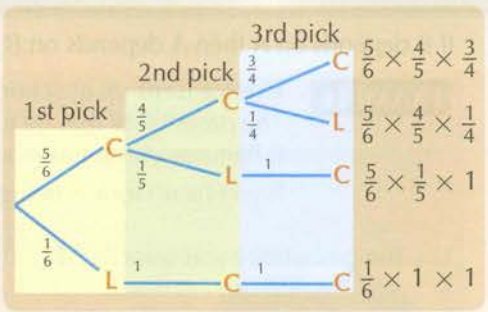
Sometimes a Branch is **Missing**

EXAMPLE A box of biscuits contains 5 chocolate biscuits and 1 lemon biscuit. George takes out 3 biscuits at random, one at a time, and eats them.

a) Find the probability that he eats 3 chocolate biscuits.

b) Find the probability that the last biscuit is chocolate.

Let C mean 'picks a chocolate biscuit' and L mean 'picks the lemon biscuit'.



After the lemon biscuit there are only chocolate biscuits left, so the tree diagram doesn't 'branch' after an 'L'.

a) Three chocolate biscuits is shown by only one 'path' along the branches.

$$P(C \text{ and } C \text{ and } C) = \frac{5}{6} \times \frac{4}{5} \times \frac{3}{4} = \frac{60}{120} = \frac{1}{2}$$

b) The third biscuit being chocolate is shown by 3 'paths' along the branches — so you can add up the probabilities:

$$P(\text{third biscuit is chocolate}) = \left(\frac{5}{6} \times \frac{4}{5} \times \frac{3}{4}\right) + \left(\frac{5}{6} \times \frac{1}{5} \times 1\right) + \left(\frac{1}{6} \times 1 \times 1\right) = \frac{1}{2} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}$$

There's a quicker way to do this, since there's only one outcome where the chocolate isn't picked last:

$$P(\text{third biscuit is not chocolate}) = \frac{5}{6} \times \frac{4}{5} \times \frac{1}{4} = \frac{1}{6}, \text{ so } P(\text{third biscuit is chocolate}) = 1 - \frac{1}{6} = \frac{5}{6}$$

Working out the probability of the complement of the event you're interested in is sometimes easier.

Sampling **with replacement** *— the probabilities stay the same*

In the above example, each time George takes a biscuit he eats it before taking the next one (i.e. he doesn't replace it) — this is sampling without replacement. Suppose instead that each time he takes a biscuit he puts it back in the box before taking the next one — this is sampling with replacement. All this means is that the probability of choosing a particular item remains the same for each pick.

So part a) above becomes:

$$P(C \text{ and } C \text{ and } C) = \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = \frac{125}{216} > \frac{1}{2}$$

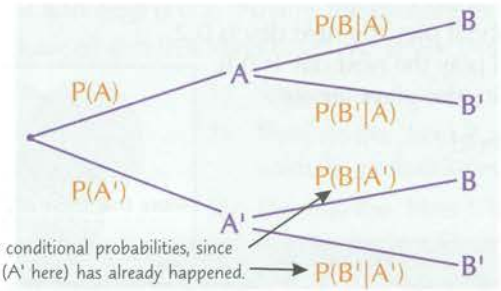
So the probability that George picks 3 chocolate biscuits is slightly greater when sampling is done with replacement. This makes sense because now there are, on average, more chocolate biscuits available for his 2nd and 3rd picks, so he is more likely to choose one.

Conditional Probability

After the first set of branches, tree diagrams actually show conditional probabilities. Read on...

P(B|A) means Probability of B, given that A has Already Happened

Conditional probability means the probability of something, given that something else has already happened. For example, P(B|A) means the probability of B, given that A has already happened. Back to tree diagrams...



If you multiply probabilities along the branches, you get:

i.e. P(A and B) $\Rightarrow$ $P(A \cap B) = P(A) \times P(B \mid A)$

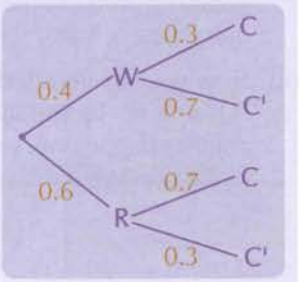
You can rewrite this as:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

EXAMPLE Horace either walks (W) or runs (R) to the bus stop. If he walks he catches (C) the bus with a probability of 0.3. If he runs he catches it with a probability of 0.7. He walks to the bus stop with a probability of 0.4. Find the probability that Horace catches the bus.

$$\begin{aligned} P(C) &= P(C \cap W) + P(C \cap R) \\ &= P(W)\,P(C \mid W) + P(R)\,P(C \mid R) \\ &= (0.4 \times 0.3) + (0.6 \times 0.7) = 0.12 + 0.42 = 0.54 \end{aligned}$$

This is easier to follow if you match each part of this working to the probabilities in the tree diagram.



If B is Conditional on A then A is Conditional on B

If B depends on A then A depends on B — and it doesn't matter which event happens first.

EXAMPLE Horace turns up at school either late (L) or on time (L'). He is then either shouted at (S) or not (S'). The probability that he turns up late is 0.4. If he turns up late the probability that he is shouted at is 0.7. If he turns up on time the probability that he is shouted at is 0.2.

If you hear Horace being shouted at, what is the probability that he turned up late?

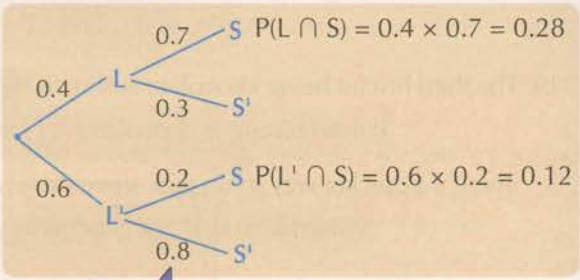
1) The probability you want is P(L | S).

Get this the right way round — he's already being shouted at.

2) Use the conditional probability formula: $P(L \mid S) = \frac{P(L \cap S)}{P(S)}$

3) The best way to find P(L ∩ S) and P(S) is with a tree diagram.

Be careful with questions like this — the information in the question tells you what you need to know to draw the tree diagram with L (or L') considered first. But you need P(L | S) — where S is considered first. So don't just rush in.



$$\begin{aligned} P(L \cap S) &= 0.4 \times 0.7 = 0.28 \\ P(S) &= P(L \cap S) + P(L' \cap S) = 0.28 + 0.12 = 0.40 \end{aligned}$$

4) Put these in your conditional probability formula to get:

$$P(L \mid S) = \frac{0.28}{0.4} = 0.7$$

These are all conditional probabilities — e.g. P(S' | L') = 0.8.

There's a Formula for Working this Out — but it's Easier to Use the Tree Diagram

This formula will be on the formula sheet in your exam.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

Here, this gives:

$$P(L \mid S) = \frac{P(L \cap S)}{P(S)} = \frac{P(S \mid L)P(L)}{P(S \mid L)P(L) + P(S \mid L')P(L')} = \frac{0.7 \times 0.4}{(0.7 \times 0.4) + (0.2 \times 0.6)} = \frac{0.28}{0.4} = 0.7$$

This is basically the same working as with the tree diagram above.

Independent Events

Independent Events Have **No Effect** on Each Other

If the probability of B happening doesn't depend on whether or not A has happened, then A and B are independent.

1) If A and B are independent, $P(A \mid B) = P(A)$.
2) If you put this in the conditional probability formula, you get: $P(A \mid B) = P(A) = \frac{P(A \cap B)}{P(B)}$

Or, to put that another way:

For independent events: $P(A \cap B) = P(A)P(B)$

EXAMPLE V and W are independent events, where $P(V) = 0.2$ and $P(W) = 0.6$.
a) Find $P(V \cap W)$. b) Find $P(V \cup W)$.

a) Just put the numbers into the formula for independent events: $P(V \cap W) = P(V)P(W) = 0.2 \times 0.6 = 0.12$

b) Using the formula on page 115: $P(V \cup W) = P(V) + P(W) - P(V \cap W) = 0.2 + 0.6 - 0.12 = 0.68$

Sometimes you'll be asked if two events are independent or not. Here's how you work it out...

EXAMPLE You are exposed to two infectious diseases — one after the other. The probability you catch the first (A) is 0.25, the probability you catch the second (B) is 0.5, and the probability you catch both of them is 0.2. Are catching the two diseases independent events?

You need to compare $P(A \mid B)$ and $P(A)$ — if they're different, the events aren't independent.

$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.5} = 0.4$ $\quad P(A) = 0.25$ $\quad P(A \mid B)$ and $P(A)$ are different, so they're not independent.

Take Your Time with **Tough** *Probability Questions*

EXAMPLE A and B are two events, with $P(A) = 0.4$, $P(B \mid A) = 0.25$, and $P(A' \cap B) = 0.2$.
a) Find: (i) $P(A \cap B)$, (ii) $P(A')$, (iii) $P(B' \mid A)$, (iv) $P(B \mid A')$, (v) $P(B)$, (vi) $P(A \mid B)$.
b) Say whether or not A and B are independent.

a) i) $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = 0.25$, so $P(A \cap B) = 0.25 \times P(A) = 0.25 \times 0.4 = 0.1$

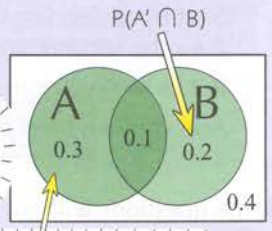
ii) $P(A') = 1 - P(A) = 1 - 0.4 = 0.6$

iii) $P(B' \mid A) = 1 - P(B \mid A) = 1 - 0.25 = 0.75$

iv) $P(B \mid A') = \frac{P(B \cap A')}{P(A')} = \frac{0.2}{0.6} = \frac{1}{3}$

v) $P(B) = P(B \mid A)P(A) + P(B \mid A')P(A') = (0.25 \times 0.4) + \left(\frac{1}{3} \times 0.6\right) = 0.3$

vi) $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.3} = \frac{1}{3}$



A Venn diagram sometimes makes it easier to see what's going on. Fill it in as you find anything out.

Since $P(B' \mid A) + P(B \mid A) = 1$.

$P(B \cap A') = P(A' \cap B)$

From (i) $P(A \cap B) = 0.1$, so you know this must be 0.3.

Or use the Venn diagram.

Or you could say that $P(A \cap B) = 0.1$, while $P(A)P(B) = 0.4 \times 0.3 = 0.12$ — they're different, which shows A and B are not independent.

b) If $P(B \mid A) = P(B)$, then A and B are independent.
But $P(B \mid A) = 0.25$, while $P(B) = 0.3$, so A and B are not independent.

Statisticians say: P(Having cake ∩ Eating it) = 0...

Probability questions can be tough. For tricky questions like the last one, try drawing a Venn diagram or a tree diagram, even if the question doesn't tell you to — they're really useful for getting your head round things and understanding what on earth is going on. And don't forget the tests for independent events — you're likely to get asked a question on those.

Arrangements and Selections

This page is a bit of a gentle introduction to this topic — it's basically about counting things.

n Different Objects can be Arranged in *n*! Different Ways...

There are $n!$ ("n factorial") ways of arranging n different objects, where $n! = n \times (n-1) \times (n-2) \times ... \times 3 \times 2 \times 1$.

EXAMPLE In how many ways can 4 different ornaments be arranged on a shelf?

You have 4 choices for the first ornament, 3 choices for the second ornament, 2 choices for the third ornament, and 1 choice for the last ornament.

So there are $4! = 4 \times 3 \times 2 \times 1 = 24$ arrangements.

...but Divide by *r*! if *r* of These Objects are the Same

If r of your n objects are identical, then the total number of possible arrangements is $n! \div r!$.

EXAMPLE In how many different ways can 5 objects be arranged in a line if 2 of those objects are identical?

Imagine those 2 identical objects were different.

Then there would be $5! = 120$ possible arrangements.

But because those 2 objects are actually identical, you can always swap them round without making a different arrangement. ($2! = 2 \times 1 = 2$.)

So there are really only $120 \div 2 = 60$ different ways to arrange the objects.

Arrangement Questions are about Counting Choices

Some arrangement questions are a bit more complicated.
For example, you might have more than one group of identical objects.

EXAMPLE In how many ways can the letters of the word STEEPLES be arranged?

Start by pretending the various S's and E's are different. Then there would be 8 choices for the first letter, 7 for the second, and so on. This would give $8! = 40\,320$ arrangements.

But in fact, you can swap those two S's around without getting a different arrangement, so divide by 2!.

And you can swap the three E's about too, so divide by 3!. This means there are $\frac{8!}{2! \times 3!} = 3360$ arrangements.

Some questions will put restrictions on the order you can arrange the objects in.

EXAMPLE Alice, Bernie, Camilla and Dave are going to sit on a 4-person bench, but Dave doesn't want to sit next to Bernie. How many acceptable arrangements are there?

First, you need to choose someone for the first seat — but there are two cases:
Either: (i) Dave or Bernie is in this first position, Or: (ii) Alice or Camilla is in this first position.

(i) Choose Dave or Bernie for position 1 (= 2 choices). Then there are also 2 choices for who sits in the next seat (i.e. Alice or Camilla). Now there are no more restrictions — you have 2 choices for Position 3 and 1 choice for Position 4 — giving $2 \times 2 \times 2 \times 1 = 8$ arrangements.

(ii) Choose Alice or Camilla for position 1 (= 2 choices). The next person must then be Dave or Bernie (= 2 choices). The next person must be either Alice or Camilla (whoever isn't in position 1) — so you have only 1 choice. And you also have only 1 choice for Position 4. This means there are $2 \times 2 \times 1 \times 1 = 4$ arrangements.

(Otherwise they'll be together in the last two places.)

This gives a grand total of $8 + 4 =$ 12 possible arrangements altogether.

You can use your fingers and toes for counting up to 5! ÷ 3!...

Hopefully that didn't seem too bad. But statistics (like maths generally) is one of those subjects where everything builds on what you've just learnt. So you need to commit all this to memory, and (preferably) understand why it's true too.

Arrangements and Selections

Choices. That's what this page is all about. Bear that in mind when I tell you that you must read all this very carefully.

In a **Permutation**, the **Order Matters**

First of all, you need to know that a permutation is an arrangement of things where the order matters. So AB and BA are different permutations of the letters A and B, for example.

EXAMPLE How many 3-digit permutations using the numbers 0-9 are there, if each digit can only appear once?

You have 10 choices for the first digit, 9 choices for the second digit, and 8 choices for the third digit.

So there are $10 \times 9 \times 8 = 720$ different permutations.

This is just $\frac{10 \times 9 \times 8 \times 7 \times ... \times 1}{7 \times 6 \times ... \times 1} = \frac{10!}{7!} = \frac{10!}{(10-3)!}$.

Always count the 'choices' you have at each point.

Permutations

When choosing r objects from n, the number of possible permutations is: ${}^nP_r = \frac{n!}{(n-r)!}$

Most calculators have a button for finding ${}^nP_r$.

In a **Combination**, the **Order Doesn't Matter**

In a combination, the order of things isn't important. So AB and BA are actually the same combination of A and B.

EXAMPLE How many 3-digit combinations using the numbers 0-9 are there, if each digit can only appear once?

From the example above, there are 720 permutations with 3 digits. But think of the permutation 123 — this is the same combination as 321. In fact, it's the same combination as any rearrangement of the digits 1, 2 and 3 (and there are $3! = 3 \times 2 \times 1 = 6$ permutations (or arrangements) of the digits 1, 2 and 3).

So the number of combinations of 3 digits must be $720 \div 6 = 120$. This is $\frac{10!}{7!3!} = \frac{10!}{(10-3)! \times 3!}$.

Combinations

When choosing r objects from n, the number of possible combinations is: ${}^nC_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}$

Your calculator will have a button for finding ${}^nC_r$ too.

Use **Binomial Coefficients** if There are **Only Two Types** of Object

The values of ${}^nC_r$ for any particular values of r and n are called binomial coefficients. They're useful for arrangement questions with two different types of object.

EXAMPLE a) In how many different ways can n objects of two types be arranged if r are of the first type?

b) How many ways are there to select 11 players from a squad of 16?

a) If the objects were all different, there would be $n!$ ways to arrange them. But r of the objects are of the same type and could be swapped around, so divide by $r!$. Since there are only two types, the other $(n-r)$ could also be swapped around — so divide by $(n-r)!$. This means there are $\frac{n!}{r!(n-r)!}$ arrangements.

b) This is basically a 'combinations' problem. Imagine the 16 players are lined up — then you could 'pick' or 'not pick' players by giving each of them a sign marked with a tick or a cross.

So just find the number of ways to arrange 11 ticks and 5 crosses — this is $\binom{16}{11} = \frac{16!}{11!5!} = 4368$.

In the Exam, you'll have no choice but to answer a question on this topic...

You must learn all this business about permutations and combinations. Once you've done that, you can calculate how many tickets you'd need to buy to be certain of winning the lottery jackpot. Quite a few, I'm guessing.

Arrangements and Selections

Now you know all about arrangements, permutations and combinations, here's the fun bit.
And by 'the fun bit', I mean 'the bit where you use them in probability questions'.

Probability Questions Introduce Restrictions

Probability questions involving arrangements, combinations and permutations are pretty straightforward. They involve random events, so you can use the formula from p115 — you just need to decide the best way to calculate the number of possible outcomes and the number that match your event.

EXAMPLE Pete is on an all-fruit diet, and is going to eat an apple, an orange, a banana, a peach and a pear for his lunch. He eats them one at a time, and selects which one to eat next at random. What is the probability that Pete eats the apple first and the banana last?

The total number of possible arrangements of the 5 pieces of fruit is $5! = 120$.

Now we need to find how many of those have the apple first and the banana last.
As the first and last positions are fixed, that means we're looking for the number of ways of arranging the orange, peach and pear in the 2nd, 3rd and 4th places.

So the number of arrangements with the apple first and the banana last is $3! = 6$.

So P(apple first and banana last) $= \frac{6}{120} = \frac{1}{20}$

EXAMPLE Emma has 10 cards marked with the digits 0-9.
She picks 4 cards at random without replacing them.
What is the probability that she only picks odd numbers?

To find the probability, we need to divide the number of possible selections that are all odd by the total number of possible selections.

The order the cards are picked doesn't matter, so this is a combination question.

The total number of possible combinations of 4 cards from 10 is ${}^{10}C_4 = \frac{10!}{4!\,6!} = 210$.

If all the numbers are odd, then there are only 5 cards she could pick.

The number of possible combinations of 4 cards from 5 is ${}^{5}C_4 = \frac{5!}{1!\,4!} = 5$.

So P(picking only odd numbers) $= \frac{5}{210} = \frac{1}{42}$.

EXAMPLE Four letters are picked at random without replacement from the letters A-H and displayed in the order they are chosen. What is the probability that the letters ABCD are displayed in alphabetical order?

The order in which the letters are picked matters here, so we need to find the number of permutations of 4 letters from 8. This is:

$\frac{8!}{4!} = 1680$

There's only 1 permutation with ABCD in alphabetical order,

so P(ABCD displayed in order) $= \frac{1}{1680}$

You'd get the same answer if you did this with a tree diagram:
P(A with 1st pick) $= \frac{1}{8}$, P(B with 2nd pick) $= \frac{1}{7}$, and so on,
so P(ABCD displayed in order) $= \frac{1}{8} \times \frac{1}{7} \times \frac{1}{6} \times \frac{1}{5} = \frac{1}{1680}$

Mind your Ps and Cs...

You might see Hollywood actors or famous rock stars mixing up combinations and permutations and think it's a cool thing to do. But it's not — nobody will be impressed if you do it. Seriously, I know I only said it a page ago, but it's important to learn the difference between them — there's a really good chance they'll come up on your S1 exam.

Probability Distributions

This stuff isn't hard — but it can seem a bit weird at times.

Getting your head round this **Basic Stuff** will help a bit

This first bit isn't particularly interesting. But understanding the difference between X and x (bear with me) might make the later stuff a bit less confusing. Might.

1) X (upper case) is just the name of a random variable. So X could be 'score on a dice' — it's just a name.
2) A random variable doesn't have a fixed value. Like with a dice score — the value on any 'roll' is all down to chance.
3) x (lower case) is a particular value that X can take. So for one roll of a dice, x could be 1, 2, 3, 4, 5 or 6.
4) Discrete random variables only have a certain number of possible values. Often these values are whole numbers, but they don't have to be. Usually there are only a few possible values (e.g. the possible scores with one roll of a dice).
5) A probability distribution is a table showing the possible values of x, plus the probability for each one.
6) A probability function is a formula that generates the probabilities for different values of x.

All the Probabilities **Add up to 1**

For a discrete random variable X:

$$\sum_{\text{all } x} P(X = x) = 1$$

This says that if you add up the probabilities of all the possible values of X, you get 1.

EXAMPLE The random variable X has probability function $P(X = x) = kx$ for $x = 1, 2, 3$. Find the value of k.

So X has three possible values ($x = 1$, 2 and 3), and the probability of each is kx (where you need to find k).

It's easier to understand with a table:

x	1	2	3
$P(X = x)$	$k \times 1 = k$	$k \times 2 = 2k$	$k \times 3 = 3k$

Now just use the formula: $\sum_{\text{all } x} P(X = x) = 1$ Here, this means: $k + 2k + 3k = 6k = 1$

i.e. $k = \frac{1}{6}$

Piece of cake.

EXAMPLE The discrete random variable X has the probability distribution shown below.

x	0	1	2	3	4
$P(X = x)$	0.1	0.2	0.3	0.2	a

Find: (i) the value of a, (ii) $P(X > 2)$ (iii) $P(2 \leq X < 4)$.

(i) Use the formula $\sum_{\text{all } x} P(X = x) = 1$ again.

From the table: $0.1 + 0.2 + 0.3 + 0.2 + a = 1$

$0.8 + a = 1$

$a = 0.2$

(ii) This is asking you to find the probability that 'X is greater than 2'.
So you need to add up the probabilities for $x = 3$ and $x = 4$.

$P(X > 2) = P(X = 3) + P(X = 4) = 0.2 + 0.2 = 0.4$

(iii) This is asking for the probability that 'X is greater than or equal to 2, but less than 4'.
Easy — just add up the probabilities again.

$P(2 \leq X < 4) = P(X = 2) + P(X = 3) = 0.3 + 0.2 = 0.5$

Careful with the inequality signs — you need to include $x = 2$ but not $x = 4$.

Probability Distributions

EXAMPLE An unbiased six-sided dice has faces marked 1, 1, 1, 2, 2, 3.
The dice is rolled twice. Let X be the random variable "sum of the two scores on the dice".
Show that $P(X = 4) = \frac{5}{18}$. Find the probability distribution of X.

① Make a table showing the 36 possible outcomes.
You can see from the table that 10 of these have the outcome $X = 4$

... so $P(X = 4) = \frac{10}{36} = \frac{5}{18}$

Score on roll 1

Score on roll 2 +	1	1	1	2	2	3
1	2	2	2	3	3	4
1	2	2	2	3	3	4
1	2	2	2	3	3	4
2	3	3	3	4	4	5
2	3	3	3	4	4	5
3	4	4	4	5	5	6

Don't forget to change the fractions into their simplest form.

② Use the table to work out the probabilities for the other outcomes and then fill in a table summarising the probability distribution. So...

... $\frac{9}{36}$ of the outcomes are a score of 2
... $\frac{12}{36}$ of the outcomes are a score of 3
... $\frac{4}{36}$ of the outcomes are a score of 5
... $\frac{1}{36}$ of the outcomes are a score of 6

x	2	3	4	5	6
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{5}{18}$	$\frac{1}{9}$	$\frac{1}{36}$

Do Complicated questions ***Bit by bit***

EXAMPLE A game involves rolling two fair dice. If the sum of the scores is greater than 10 then the player wins 50p.
If the sum is between 8 and 10 (inclusive) then they win 20p. Otherwise they get nothing.
If X is the random variable "amount player wins", find the probability distribution of X.

There are 3 possible values for X (0, 20 and 50) and you need the probability of each.
To work these out, you need the probability of getting various totals on the dice.

① You need to know $P(8 \leq \text{score} \leq 10)$ — the probability that the score is between 8 and 10 inclusive (i.e. including 8 and 10) and $P(11 \leq \text{score} \leq 12)$ — the probability that the score is greater than 10.

This means working out: $P(\text{score} = 8)$, $P(\text{score} = 9)$, $P(\text{score} = 10)$, $P(\text{score} = 11)$ and $P(\text{score} = 12)$. Use a table...

② Score on dice 1

Score on dice 2 +	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	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

There are 36 possible outcomes...
...5 of these have a total of 8 — so the probability of scoring 8 is $\frac{5}{36}$
...4 have a total of 9 — so the probability of scoring 9 is $\frac{4}{36}$
...the probability of scoring 10 is $\frac{3}{36}$
...the probability of scoring 11 is $\frac{2}{36}$
...the probability of scoring 12 is $\frac{1}{36}$

③ To find the probabilities you need, you just add the right bits together:

$P(X = 20\text{p}) = P(8 \leq \text{score} \leq 10) = \frac{5}{36} + \frac{4}{36} + \frac{3}{36} = \frac{12}{36} = \frac{1}{3}$

$P(X = 50\text{p}) = P(11 \leq \text{score} \leq 12) = \frac{2}{36} + \frac{1}{36} = \frac{3}{36} = \frac{1}{12}$

To find $P(X = 0)$ just take the total of the two probabilities above from 1 (since $X = 0$ is the only other possibility).

$P(X = 0) = 1 - \left[\frac{12}{36} + \frac{3}{36}\right] = 1 - \frac{15}{36} = \frac{21}{36} = \frac{7}{12}$

④ Now just stick all this info in a table (and check that the probabilities all add up to 1):

x	0	20	50
$P(X = x)$	$\frac{7}{12}$	$\frac{1}{3}$	$\frac{1}{12}$

Useful quotes: All you need in life is ignorance and confidence, then success is sure...*
I said earlier that the 'counting the outcomes' approach was useful — well there you go. And if you remember how to do that, then you can work out a probability distribution. And if you can work out one of those, then you can often begin to unravel even fairly daunting-looking questions. But most of all, REMEMBER THAT ALL THE PROBABILITIES ADD UP TO 1.

* Mark Twain

Expected Values, Mean and Variance

This is all about the mean and variance of random variables — not a load of data. It's a tricky concept, but bear with it.

Discrete Random Variables have an 'Expected Value' or 'Mean'

You can work out the expected value (or 'mean') $E(X)$ for a discrete random variable X. $E(X)$ is a kind of 'theoretical mean' — it's what you'd expect the mean of X to be if you took loads of readings. In practice, the mean of your results is unlikely to match the theoretical mean exactly, but it should be pretty near.

Remember, 'discrete' just means it can only take a certain number of values.

If the possible values of X are $x_1, x_2, x_3, \ldots$ then the expected value of X is:

$$\text{Mean} = \text{Expected Value, } E(X) = \sum x_i P(X = x_i) = \sum x_i p_i$$

$p_i = P(X = x_i)$

EXAMPLE The probability distribution of X, the number of daughters in a family of 3 children, is shown in the table. Find the expected number of daughters.

x_i	0	1	2	3
p_i	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$\text{Mean} = \sum x_i p_i = \left[0 \times \tfrac{1}{8}\right] + \left[1 \times \tfrac{3}{8}\right] + \left[2 \times \tfrac{3}{8}\right] + \left[3 \times \tfrac{1}{8}\right] = 0 + \tfrac{3}{8} + \tfrac{6}{8} + \tfrac{3}{8} = \tfrac{12}{8} = 1.5$$

So the expected number of daughters is 1.5 — which sounds a bit weird.
But all it means is that if you check a large number of 3-child families, the mean will be close to 1.5.

The Variance measures how Spread Out the distribution is

You can also find the variance of a random variable. It's the 'expected variance' of a large number of readings.

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \sum x_i^2 p_i - \left[\sum x_i p_i\right]^2$$

This formula needs $E(X^2) = \sum x_i^2 p_i$ — take each possible value of x, square it, multiply it by its probability and then add up all the results.

EXAMPLE Work out the variance for the '3 daughters' example above:

First work out $E(X^2)$:

$$E(X^2) = \sum x_i^2 p_i = \left[0^2 \times \tfrac{1}{8}\right] + \left[1^2 \times \tfrac{3}{8}\right] + \left[2^2 \times \tfrac{3}{8}\right] + \left[3^2 \times \tfrac{1}{8}\right] = 0 + \tfrac{3}{8} + \tfrac{12}{8} + \tfrac{9}{8} = \tfrac{24}{8} = 3$$

The standard deviation (s.d.) of a random variable is the square root of its variance: s.d. $= \sqrt{\text{Var}(X)}$

Now you take away the mean squared: $\text{Var}(X) = E(X^2) - [E(X)]^2 = 3 - 1.5^2 = 3 - 2.25 = 0.75$

EXAMPLE X has the probability function $P(X = x) = k(x + 1)$ for $x = 0, 1, 2, 3, 4$.
Find the mean and variance of X.

1. First you need to find k — work out all the probabilities and make sure they add up to 1.

$P(X = 0) = k \times (0 + 1) = k$. Similarly, $P(X = 1) = 2k$, $P(X = 2) = 3k$, $P(X = 3) = 4k$, $P(X = 4) = 5k$.

So $k + 2k + 3k + 4k + 5k = 1$, i.e. $15k = 1$, and so $k = \frac{1}{15}$

Now you can work out p_1, p_2, p_3... where $p_1 = P(X = 1)$ etc.

2. Now use the formulas — find the mean $E(X)$ first:

$$E(X) = \sum x_i p_i = \left[0 \times \tfrac{1}{15}\right] + \left[1 \times \tfrac{2}{15}\right] + \left[2 \times \tfrac{3}{15}\right] + \left[3 \times \tfrac{4}{15}\right] + \left[4 \times \tfrac{5}{15}\right] = \tfrac{40}{15} = \tfrac{8}{3}$$

For the variance you need $E(X^2)$:

$$E(X^2) = \sum x_i^2 p_i = \left[0^2 \times \tfrac{1}{15}\right] + \left[1^2 \times \tfrac{2}{15}\right] + \left[2^2 \times \tfrac{3}{15}\right] + \left[3^2 \times \tfrac{4}{15}\right] + \left[4^2 \times \tfrac{5}{15}\right] = \tfrac{130}{15} = \tfrac{26}{3}$$

And finally: $\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{26}{3} - \left[\frac{8}{3}\right]^2 = \frac{14}{9}$

The Geometric Distribution

When you think of the geometric distribution, think "fail, fail,..., fail, succeed!"
So it's a bit like England at the Football World Cup, only the geometric distribution ends in a success.

The Geometric Distribution Models "The Number of Trials Before a Success"

EXAMPLE I'm rolling a fair dice. Find the probability that I first roll a six:
a) on the 4th throw,
b) on the nth throw.

a) If the first 'six' occurs on the 4th throw, then the first 3 throws must all have landed on 'not a six'.

So $P(\text{first 'six' on 4th throw}) = \left(\frac{5}{6}\right)^3 \times \frac{1}{6} = \frac{125}{1296}$ ⟸ This is the probability of needing 4 trials for the first success.

b) If the first 'six' occurs on the nth throw, then the first $(n-1)$ throws must all have landed on 'not a six'.

So $P(\text{first 'six' on } n\text{th throw}) = \left(\frac{5}{6}\right)^{n-1} \times \frac{1}{6}$ ⟸ This is the probability of needing n trials for the first success.

Learn the Conditions for a Geometric Probability Distribution

Geometric Distribution Geo(p)

A random variable X follows a Geometric Distribution as long as these conditions are satisfied:

1) There is a sequence of independent trials with only two possible outcomes ('success' and 'failure').
2) There is a constant probability (p) of success at each trial.
3) X is the number of trials before the first success occurs (including the 'successful' trial).

In this case $\mathbf{P(X = x) = p(1-p)^{x-1}}$ for $x = 1, 2, 3...$

You can write $\mathbf{X \sim Geo(p)}$. The expected value is given by $\mathbf{E(X) = \frac{1}{p}}$.

Random variables following a geometric distribution have an infinite number of possible values.

Make Sure You Understand these Geometric Distribution Examples

EXAMPLE X is a discrete random variable, and $X \sim$ Geo(0.4).
a) Find: (i) $P(X = 7)$, (ii) $P(X > 3)$, (iii) $E(X)$.
b) Show that $P(X \text{ is a multiple of } 2) = \frac{q}{1+q}$, where $q = 1 - p$.

a) (i) $P(X = 7) = 0.4 \times 0.6^6 = 0.019$ (to 3 d.p.)

(ii) $P(X > 3) = 1 - P(X \leq 3) = 1 - (P(X = 1) + P(X = 2) + P(X = 3)) = 1 - (0.4 + 0.4 \times 0.6 + 0.4 \times 0.6^2) = 0.216$

(iii) $E(X) = \frac{1}{p} = \frac{1}{0.4} = 2.5$

b) $$\begin{aligned} P(X \text{ is a multiple of } 2) &= P(X = 2) + P(X = 4) + P(X = 6) + \ldots \\ &= (1-p)p + (1-p)^3p + (1-p)^5p + \ldots \\ &= (1-p)p\{1 + (1-p)^2 + (1-p)^4 + \ldots\} \\ &= (1-p)p \times \frac{1}{1-(1-p)^2} \\ &= (1-p)p \times \frac{1}{p(2-p)} = \frac{1-p}{2-p} = \frac{q}{1+q} \end{aligned}$$

Factorise by taking $(1-p)p$ outside the big brackets.

The thing in the big brackets is just the sum to infinity of a geometric series (p65). So you can write the sum using the formula: $S_\infty = \frac{a}{1-r}$. Here the ratio (r) is $(1-p)^2$.

Geometry means "measuring the earth" — so why is this a "geometric distribution"...

There's nothing too horrendous about the geometric distribution really. If you didn't follow the sum to infinity stuff, head back to the Core 2 bit of this book and check it out — it's a cracking read. Make sure you learn the conditions for the geometric distribution — they won't always tell you if you're dealing with a geometric distribution in the exam.

The Binomial Probability Function

This page involves counting the number of different arrangements of things — and it uses binomial coefficients to do it. If you need a reminder about either of those topics, look back to pages 120-121.

Use *Binomial Coefficients* to Count Arrangements of 'Successes' and 'Failures'

A while ago, you learnt that if p = P(something happens), then $1 - p$ = P(that thing doesn't happen). You'll need that fact now.

EXAMPLE I toss a fair coin 5 times. Find the probability of:
a) 0 heads, b) 1 head, c) 2 heads.

First, note that each coin toss is independent of the others.
That means you can multiply individual probabilities together.

On n tosses, with r heads and $(n - r)$ tails, there are ${}^nC_r$ or $\binom{n}{r}$ ways to arrange them (see p.121). This is the binomial coefficient.

a) $\text{P(0 heads)} = \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} = 0.5^5 = 0.03125$

P(tails) = P(heads) = 0.5.

b) $\text{P(1 head)} = \text{P(heads)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)}$
$+ \text{P(tails)} \times \text{P(heads)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)}$
$+ \text{P(tails)} \times \text{P(tails)} \times \text{P(heads)} \times \text{P(tails)} \times \text{P(tails)}$
$+ \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(heads)} \times \text{P(tails)}$
$+ \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(tails)} \times \text{P(heads)}$

These are the $\binom{5}{1} = 5$ ways to arrange 1 head and 4 tails.

So $\text{P(1 head)} = 0.5 \times (0.5)^4 \times \binom{5}{1} = 0.03125 \times \frac{5!}{1!4!} = 0.15625$

= P(heads) × [P(tails)]^4 × ways to arrange 1 head and 4 tails.

c) $\text{P(2 heads)} = [\text{P(heads)}]^2 \times [\text{P(tails)}]^3 \times \text{ways to arrange 2 heads and 3 tails} = (0.5)^2 \times (0.5)^3 \times \binom{5}{2} = 0.3125$

The *Binomial Probability Function* gives P(r successes out of n trials)

The previous example really just shows why this thing in a box must be true.

Binomial Probability Function

$$\text{P}(r \text{ successes in } n \text{ trials}) = \binom{n}{r} \times [\text{P(success)}]^r \times [\text{P(failure)}]^{n-r}$$

This is the probability function for a binomial distribution — see next page for more info.

EXAMPLE I roll a fair dice 5 times. Find the probability of rolling: a) 2 sixes, b) 3 sixes, c) 4 numbers less than 3.

Again, note that each roll of a dice is independent of the other rolls.

a) For this part, call "roll a 6" a success, and "roll anything other than a 6" a failure.

Then $\text{P(roll 2 sixes)} = \binom{5}{2} \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^3 = \frac{5!}{2!3!} \times \frac{1}{36} \times \frac{125}{216} = 0.161$ (to 3 d.p.).

b) Again, call "roll a 6" a success, and "roll anything other than a 6" a failure.

Then $\text{P(roll 3 sixes)} = \binom{5}{3} \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2 = \frac{5!}{3!2!} \times \frac{1}{216} \times \frac{25}{36} = 0.032$ (to 3 d.p.).

Notice how $\binom{5}{2} = \binom{5}{3}$. In fact, $\binom{n}{r} = \binom{n}{n-r}$.

c) This time, success means "roll a 1 or a 2", while failure is now "roll a 3, 4, 5 or 6".

Then $\text{P(roll 4 numbers less than 3)} = \binom{5}{4} \times \left(\frac{1}{3}\right)^4 \times \left(\frac{2}{3}\right) = \frac{5!}{4!1!} \times \frac{1}{81} \times \frac{2}{3} = 0.041$ (to 3 d.p.).

Let this formula for success go to your head — and then keep it there...

This page is all about finding the probabilities of different numbers of successes in n trials. Now then... if you carry out n trials, there are $n + 1$ possibilities for the number of successes (0, 1, 2, ..., n). This 'family' of possible results along with their probabilities is sounding suspiciously like a probability distribution. Oh rats... I've given away what's on the next page.

The Binomial Distribution

I know you're having so much fun learning all about random variables... well, there's more to come. This page is about discrete random variables following a binomial distribution (whose probability function you saw on p129).

There are **5 Conditions** for a **Binomial Distribution**

Binomial Distribution: B(n, p)

A random variable X follows a Binomial Distribution as long as these 5 conditions are satisfied:

1) There is a fixed number (n) of trials.
2) Each trial involves either "success" or "failure".
3) All the trials are independent.
4) The probability of "success" (p) is the same in each trial.
5) The variable is the total number of successes in the n trials.

Binomial random variables are discrete, since they only take values 0, 1, 2... n.

n and p are the two parameters of the binomial distribution. (Or n is sometimes called the 'index'.)

In this case, $P(X = x) = \binom{n}{x} \times p^x \times (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$, and you can write $X \sim B(n, p)$.

EXAMPLE: Which of the random variables described below would follow a binomial distribution? For those that do, state the distribution's parameters.

a) **The number of faulty items (T) produced in a factory per day, if the probability of each item being faulty is 0.01 and there are 10 000 items produced every day.**
Binomial — there's a fixed number (10 000) of trials with two possible results ('faulty' or 'not faulty'), a constant probability of 'success', and T is the total number of 'faulty' items.
So (as long as faulty items occur independently) $T \sim B(10\,000, 0.01)$.

b) **The number of red cards (R) drawn from a standard 52-card deck in 10 picks, not replacing the cards each time.**
Not binomial, since the probability of 'success' changes each time (as I'm not replacing the cards).

c) **The number of red cards (R) drawn from a standard 52-card deck in 10 picks, replacing the cards each time.**
Binomial — there's a fixed number (10) of independent trials with two possible results ('red' or 'black/not red'), a constant probability of success (I'm replacing the cards), and R is the number of red cards drawn. $R \sim B(10, 0.5)$.

d) **The number of times (T) I have to toss a coin before I get heads.**
Not binomial, since the number of trials isn't fixed.

e) **The number of left-handed people (L) in a sample of 500 randomly chosen people, if the fraction of left-handed people in the population as a whole is 0.13.**
Binomial — there's a fixed number (500) of independent trials with two possible results ('left-handed' or 'not left-handed'), a constant probability of success (0.13), and L is the number of left-handers. $L \sim B(500, 0.13)$.

EXAMPLE: When I toss a grape in the air and try to catch it in my mouth, my probability of success is always 0.8. The number of grapes I catch in 10 throws is described by the discrete random variable X.

a) How is X distributed? Name the type of distribution, and give the values of any parameters.
b) Find the probability of me catching at least 9 grapes.

a) There's a fixed number (10) of independent trials with two possible results ('catch' and 'not catch'), a constant probability of success (0.8), and X is the total number of catches.
Therefore X follows a binomial distribution, $X \sim B(10, 0.8)$.

b) P(at least 9 catches)

$$= P(9 \text{ catches}) + P(10 \text{ catches})$$
$$= \left\{\binom{10}{9} \times 0.8^9 \times 0.2^1\right\} + \left\{\binom{10}{10} \times 0.8^{10} \times 0.2^0\right\}$$
$$= 0.268435... + 0.107374... = 0.376 \text{ (to 3 d.p.)}.$$

Binomial distributions come with 5 strings attached...

There's a big, boring box at the top of the page with a list of 5 conditions in — and you do need to know it, unfortunately. There's only one way to learn it — keep trying to write down the 5 conditions until you can do it in your sleep.

Using Binomial Tables

Your life is just about to be made a whole lot easier. So smile sweetly and admit that statistics isn't all bad.

Look up Probabilities in **Binomial Tables**

EXAMPLE I have an unfair coin. When I toss this coin, the probability of getting heads is 0.35. Find the probability that it will land on heads fewer than 3 times when I toss it 12 times in total.

If the random variable X represents the number of heads I get in 12 tosses, then $X \sim B(12, 0.35)$.
You need to find $P(X \leq 2)$.

① You could work this out 'manually'...

$$P(0 \text{ heads}) + P(1 \text{ head}) + P(2 \text{ heads}) = \left\{\binom{12}{0} \times 0.35^0 \times 0.65^{12}\right\} + \left\{\binom{12}{1} \times 0.35^1 \times 0.65^{11}\right\} + \left\{\binom{12}{2} \times 0.35^2 \times 0.65^{10}\right\}$$
$$= 0.0057 + 0.0368 + 0.1088 = 0.1513$$

② But it's much quicker to use tables of the binomial cumulative distribution function (c.d.f.). The c.d.f. of a distribution is a function that gives the probability that X will be less than or equal to a particular value. So the tables show $P(X \leq x)$, for $X \sim B(n, p)$.

- First find the table for the correct values of n and p. Then the table gives you a value for $P(X \leq x)$.
- Here: $n = 12$ and $p = 0.35$.

n = 12																								
p	0.05	0.1	0.15	1/6	0.2	0.25	0.3	1/3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	2/3	0.7	0.75	0.8	5/6	0.85	0.9	0.95
x = 0	0.5404	0.2824	0.1422	0.1122	0.0687	0.0317	0.0138	0.0077	0.0057	0.0022	0.0008	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.8816	0.6590	0.4435	0.3813	0.2749	0.1584	0.0850	0.0540	0.0424	0.0196	0.0083	0.0032	0.0011	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9804	0.8891	0.7358	0.6774	0.5583	0.3907	0.2528	0.1811	0.1513	0.0834	0.0421	0.0193	0.0079	0.0028	0.0008	0.0005	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.9978	0.9744	0.9078	0.8748	0.7946	0.6488	0.4925	0.3931	0.3467	0.2253	0.1345	0.0730	0.0356	0.0153	0.0056	0.0039	0.0017	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
4	0.9998	0.9957	0.9761	0.9636	0.9274	0.8424	0.7237	0.6315	0.5833	0.4382	0.3044	0.1938	0.1117	0.0573	0.0255	0.0188	0.0095	0.0028	0.0006	0.0002	0.0001	0.0000	0.0000
5	1.0000	0.9995	0.9954	0.9921	0.9806	0.9456	0.8822	0.8223	0.7873	0.6652	0.5269	0.3872	0.2607	0.1582	0.0846	0.0664	0.0386	0.0143	0.0039	0.0013	0.0007	0.0001	0.0000
6	1.0000	0.9999	0.9993	0.9987	0.9961	0.9857	0.9614	0.933	0.9154	0.8418	0.7393	0.6128	0.4731	0.3348	0.2127	0.1777	0.1178	0.0544	0.0194	0.0079	0.0046	0.0005	0.0000
7	1.0000	1.0000	0.9999	0.9998	0.9994	0.9972	0.9905	0.98	0.9745	0.9427	0.8883	0.8062	0.6956	0.5618	0.4167	0.3685	0.2763	0.1576	0.0726	0.0364	0.0239	0.0043	0.0002
8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9983	0.9	0.9944	0.9847	0.9644	0.9270	0.8655	0.7747	0.6533	0.6069	0.5075	0.3512	0.2054	0.1252	0.0922	0.0256	0.0022
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9	0.9992	0.9972	0.9921	0.9807	0.9579	0.9166	0.8487	0.8189	0.7472	0.6093	0.4417	0.3226	0.2642	0.1109	0.0196
10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968	0.9917	0.9804	0.9576	0.9460	0.9150	0.8416	0.7251	0.6187	0.5565	0.3410	0.1184
11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9992	0.9978	0.9943	0.9923	0.9862	0.9683	0.9313	0.8878	0.8578	0.7176	0.4596
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

- You need $P(X \leq 2)$.
- The table tells you this is 0.1513.

See p150 for more binomial tables.

Practise using those Binomial Tables

Binomial tables can be a bit awkward. Make sure you know how to find out what you want to know.

EXAMPLE I have a different unfair coin. When I toss this coin, the probability of getting tails is 0.6. The random variable X represents the number of tails in 12 tosses, so $X \sim B(12, 0.6)$.

If I toss this coin 12 times, use the table above to find the probability that:

a) it will land on tails at least 9 times,
b) it will land on heads exactly 9 times,
c) it will land on heads at least 6 times,
d) it will land on tails more than 3 but fewer than 6 times.

a) $P(X \geq 9) = 1 - P(X < 9) = 1 - P(X \leq 8) = 1 - 0.7747 = 0.2253$

1) P(event happens) = 1 – P(event doesn't happen),
2) P(X < 9) = P(X ≤ 8), as X takes whole number values.

b) This means exactly 3 tails.
$P(X = 3) = P(X \leq 3) - P(X \leq 2) = 0.0153 - 0.0028 = 0.0125$

Use P(A or B) = P(A) + P(B) with the mutually exclusive events "X ≤ 2 " and "X = 3" to get P(X ≤ 3) = P(X ≤ 2) + P(X = 3).
Or you can think of it as "subtracting P(X ≤ 2) from P(X ≤ 3) leaves just P(X = 3)".

c) At least 6 heads means 6 or fewer tails. $P(X \leq 6) = 0.3348$

d) $P(3 < X < 6) = P(X \leq 5) - P(X \leq 3) = 0.1582 - 0.0153 = 0.1429$

Statistical tables are the original labour-saving device...

...as long as you know what you're doing. Careful, though — it's easy to trip yourself up. Basically, as long as you can find the right value of n and p in a table, you can use those tables to work out anything you might need. So hurrah for tables.

Mean and Variance of B(n, p)

You know from page 127 what the mean (or expected value) and variance of a random variable are.
And you also know what the binomial distribution is. Put those things together, and you get this page.

For a Binomial Distribution: Mean = np

This formula will be in your formula booklet, but it's worth committing to memory anyway.

Mean of a Binomial Distribution

If $X \sim B(n, p)$, then:

Mean (or Expected Value) = $\mu = E(X) = np$

Greek letters (e.g. μ) often show something based purely on theory rather than experimental results.

Remember... the expected value is the value you'd expect the random variable to take on average if you took loads and loads of readings. It's a "theoretical mean" — the mean of experimental results is unlikely to match it exactly.

EXAMPLE If $X \sim B(20, 0.2)$, what is $E(X)$?

Just use the formula: $E(X) = np = 20 \times 0.2 = 4$

EXAMPLE What's the expected number of sixes when I roll a fair dice 30 times? Interpret your answer.

If the random variable X represents the number of sixes in 30 rolls, then $X \sim B(30, \frac{1}{6})$.

So the expected value of X is $E(X) = 30 \times \frac{1}{6} = 5$

If I were to repeatedly throw the dice 30 times, and find the average number of sixes in each set of 30 throws, then I would expect it to end up pretty close to 5.
And the more sets of 30 throws I did, the closer to 5 I'd expect the average to be.

Notice that the probability of getting exactly 5 sixes on my next set of 30 throws = $\binom{30}{5} \times \left(\frac{1}{6}\right)^5 \times \left(\frac{5}{6}\right)^{25} = 0.192$
So I'm much more likely not to get exactly 5 sixes (= 1 − 0.192 = 0.808).
This is why it only makes sense to talk about the mean as a "long-term average", and not as "what I expect to happen next".

For a Binomial Distribution: Variance = npq

Variance of a Binomial Distribution

If $X \sim B(n, p)$, then:

Variance = $Var(X) = \sigma^2 = np(1 - p) = npq$

Standard Deviation = $\sigma = \sqrt{np(1-p)} = \sqrt{npq}$

For a binomial distribution, P(success) is usually called p, and P(failure) is sometimes called q (= 1 − p).

EXAMPLE If $X \sim B(20, 0.2)$, what is $Var(X)$?

Just use the formula: $Var(X) = np(1 - p) = 20 \times 0.2 \times 0.8 = 3.2$

EXAMPLE If $X \sim B(25, 0.2)$, find: a) $P(X \leq \mu)$, b) $P(X \leq \mu - \sigma)$, c) $P(X \leq \mu - 2\sigma)$

$E(X) = \mu = 25 \times 0.2 = 5$, and $Var(X) = \sigma^2 = 25 \times 0.2 \times (1 - 0.2) = 4$, which gives $\sigma = 2$.

So, using tables (for $n = 25$ and $p = 0.2$): a) $P(X \leq \mu) = P(X \leq 5) = 0.6167$

b) $P(X \leq \mu - \sigma) = P(X \leq 3) = 0.2340$

c) $P(X \leq \mu - 2\sigma) = P(X \leq 1) = 0.0274$

See page 150.

For B(n, p) — the variance is always less than the mean...

Nothing too fancy there really. A couple of easy-to-remember formulas, and some stuff about how to interpret these figures which you've seen before anyway. So learn the formulas, put the kettle on, and have a cup of tea while the going's good.

Binomial Distribution Problems

That's everything you need to know about binomial distributions (for now).
So it's time to put it all together and have a look at the kind of thing you might get asked in the exam.

EXAMPLE 1: Selling Double Glazing

A double-glazing salesman is handing out leaflets in a busy shopping centre. He knows that the probability of each passing person taking a leaflet is always 0.3. During a randomly chosen one-minute interval, 30 people passed him.

a) Suggest a suitable model to describe the number of people (X) who take a leaflet.
b) What is the probability that more than 10 people take a leaflet?
c) How many people would the salesman expect to take a leaflet?
d) Find the variance and standard deviation of X.

a) During this one-minute interval, there's a fixed number (30) of independent trials with two possible results ("take a leaflet" and "do not take a leaflet"), a constant probability of success (0.3), and X is the total number of people taking leaflets. So $X \sim B(30, 0.3)$.

Use binomial tables for this — see p150.

b) $P(X > 10) = 1 - P(X \leq 10) = 1 - 0.7304 = 0.2696$

c) The number of people the salesman could expect to take a leaflet is $E(X) = np = 30 \times 0.3 = 9$

d) Variance $= np(1 - p) = 30 \times 0.3 \times (1 - 0.3) = 6.3$ Standard deviation $= \sqrt{6.3} = 2.51$ (to 2 d.p.)

EXAMPLE 2: Multiple-Choice Guessing

A student has to take a 25-question multiple-choice exam, where each question has five possible answers, of which only one is correct. He believes he can pass the exam by guessing answers at random.

a) How many questions could the student be expected to guess correctly?
b) If the pass mark is 10, what is the probability that the student will pass the exam?
c) The examiner decides to set the pass mark so that it is at least 3 standard deviations above the expected number of correct guesses. What should the minimum pass mark be?

Let X be the number of correct guesses over the 25 questions. Then $X \sim B(25, 0.2)$.

Define your random variable first, and say how it will be distributed.

a) $E(X) = np = 25 \times 0.2 = 5$

b) $P(X \geq 10) = 1 - P(X < 10) = 1 - P(X \leq 9) = 1 - 0.9827 = 0.0173$

c) $\text{Var}(X) = np(1 - p) = 25 \times 0.2 \times 0.8 = 4$ — so the standard deviation $= \sqrt{4} = 2$.
So the pass mark needs to be at least $5 + (3 \times 2) = 11$.

EXAMPLE 3: An Unfair Coin

I am spinning a coin that I know is three times as likely to land on heads as it is on tails.

a) What is the probability that it lands on tails for the first time on the third spin?
b) What is the probability that in 10 spins, it lands on heads at least 7 times?

Careful... this doesn't need you to use one of the binomial formulas (it's a geometric distribution).

You know that P(heads) = 3 × P(tails), and that P(heads) + P(tails) = 1.
This means that P(heads) = 0.75 and P(tails) = 0.25.

a) P(lands on tails for the first time on the third spin) $= 0.75 \times 0.75 \times 0.25 = 0.141$ (to 3 d.p.).

b) If X represents the number of heads in 10 spins, then $X \sim B(10, 0.75)$.

$P(X \geq 7) = 1 - P(X < 7) = 1 - P(X \leq 6) = 1 - 0.2241 = 0.7759$

Proof that you shouldn't send a monkey to take your multi-choice exams...

You can see now how useful a working knowledge of statistics is. Ever since you first started using CGP books, I've been banging on about how hard it is to pass an exam without revising. Well, now you can prove I was correct using a bit of knowledge and binomial tables. Yup... statistics can help out with some of those tricky situations you face in life.

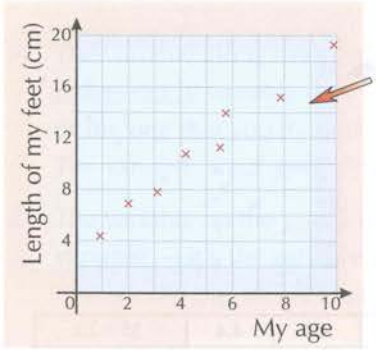
Correlation

Correlation is all about how closely two quantities are linked. And it can involve a fairly hefty formula.

Draw a Scatter Diagram to see Patterns in Data

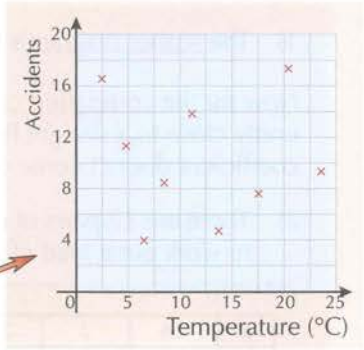
Sometimes variables are measured in pairs — maybe because you want to find out how closely they're linked.
These pairs of variables might be things like: — 'my age' and 'length of my feet', or
— 'temperature' and 'number of accidents on a stretch of road'.

You can plot readings from a pair of variables on a scatter diagram — this'll tell you something about the data.



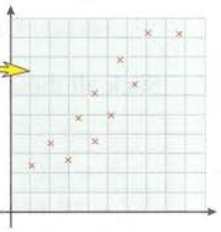
The variables 'my age' and 'length of my feet' seem linked — all the points lie close to a line. As I got older, my feet got bigger and bigger (though I stopped measuring when I was 10).

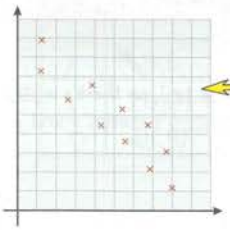
It's a lot harder to see any connection between the variables 'temperature' and 'number of accidents' — the data seems scattered pretty much everywhere.



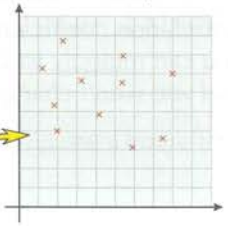
Correlation is a measure of How Closely variables are Linked

1) Sometimes, as one variable gets bigger, the other one also gets bigger — then the scatter diagram might look like the one on the right. Here, a line of best fit would have a positive gradient. The two variables are positively correlated (or there's a positive correlation between them).





2) But if one variable gets smaller as the other one gets bigger, then the scatter diagram might look like this one — and the line of best fit would have a negative gradient. The two variables are negatively correlated (or there's a negative correlation between them).



3) And if the two variables aren't linked at all, you'd expect a random scattering of points — it's hard to say where the line of best fit would be. The variables aren't correlated (or there's no correlation).

The Product-Moment Correlation Coefficient (r) measures Correlation

1) The Product-Moment Correlation Coefficient (PMCC, or r, for short) measures how close to a straight line the points on a scatter graph lie.

2) The PMCC is always between +1 and –1.
If all your points lie exactly on a straight line with a positive gradient (perfect positive correlation), $r = +1$.
If all your points lie exactly on a straight line with a negative gradient (perfect negative correlation), $r = -1$.
(In reality, you'd never expect to get a PMCC of +1 or –1 — your scatter graph points might lie pretty close to a straight line, but it's unlikely they'd all be on it.)

3) If $r = 0$ (or more likely, pretty close to 0), that would mean the variables aren't correlated.

4) The formula for the PMCC is a real stinker. But some calculators can work it out if you type in the pairs of readings, which makes life easier. Otherwise, just take it nice and slow.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum[x-\bar{x}][y-\bar{y}]}{\sqrt{(\sum[x-\bar{x}]^2)(\sum[y-\bar{y}]^2)}} = \frac{\sum xy - \frac{[\sum x][\sum y]}{n}}{\sqrt{\left(\sum x^2 - \frac{[\sum x]^2}{n}\right)\left(\sum y^2 - \frac{[\sum y]^2}{n}\right)}}$$

This is the easiest one to use, but it's still a bit hefty. Fortunately, it'll be on your formula sheet.

See pages 139-140 for more about S_{xy}, S_{xx} and S_{yy}.

Correlation

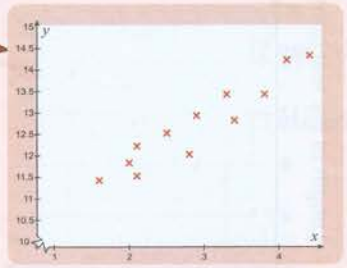
Don't rush questions on correlation. In fact, take your time and draw yourself a nice table.

EXAMPLE Illustrate the following data with a scatter diagram, and find the product-moment correlation coefficient (r) between the variables x and y.

If $p = 4x - 3$ and $q = 9y + 17$, what is the PMCC between p and q?

x	1.6	2.0	2.1	2.1	2.5	2.8	2.9	3.3	3.4	3.8	4.1	4.4
y	11.4	11.8	11.5	12.2	12.5	12.0	12.9	13.4	12.8	13.4	14.2	14.3

1) The scatter diagram's the easy bit — just plot the points.



Now for the correlation coefficient. From the scatter diagram, the points lie pretty close to a straight line with a positive gradient — so if the correlation coefficient doesn't come out pretty close to +1, we'd need to worry...

2) There are 12 pairs of readings, so $n = 12$. That bit's easy — now you have to work out a load of sums. It's best to add a few extra rows to your table...

x	1.6	2	2.1	2.1	2.5	2.8	2.9	3.3	3.4	3.8	4.1	4.4	$35 = \Sigma x$
y	11.4	11.8	11.5	12.2	12.5	12	12.9	13.4	12.8	13.4	14.2	14.3	$152.4 = \Sigma y$
x^2	2.56	4	4.41	4.41	6.25	7.84	8.41	10.89	11.56	14.44	16.81	19.36	$110.94 = \Sigma x^2$
y^2	129.96	139.24	132.25	148.84	156.25	144	166.41	179.56	163.84	179.56	201.64	204.49	$1946.04 = \Sigma y^2$
xy	18.24	23.6	24.15	25.62	31.25	33.6	37.41	44.22	43.52	50.92	58.22	62.92	$453.67 = \Sigma xy$

Stick all these in the formula to get:

$$r = \frac{\left[453.67 - \frac{35 \times 152.4}{12}\right]}{\sqrt{\left[110.94 - \frac{35^2}{12}\right] \times \left[1946.04 - \frac{152.4^2}{12}\right]}} = \frac{9.17}{\sqrt{8.857 \times 10.56}} = 0.948 \text{ (to 3 s.f.)}$$

This is pretty close to 1, so there's a strong positive correlation between x and y.

3) Correlation coefficients aren't affected by linear transformations — you can multiply variables by a number, and add a number to them, and you won't change the PMCC between them.

So if p and q are given by $p = 4x - 3$ and $q = 9y + 17$, then the PMCC between p and q is also 0.948.

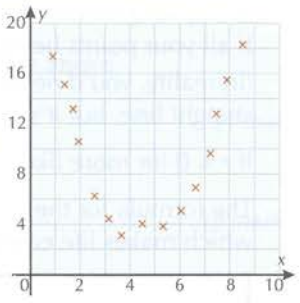
Don't make **Sweeping Statements** using Statistics

1) A high correlation coefficient doesn't necessarily mean that one factor causes the other.

EXAMPLE The number of televisions sold in Japan and the number of cars sold in America may well be correlated, but that doesn't mean that high TV sales in Japan cause high car sales in the US.

2) The PMCC is only a measure of a linear relationship between two variables (i.e. how close they'd be to a straight line if you plotted a scatter diagram).

EXAMPLE In the diagram on the right, the PMCC would be pretty low, but the two variables definitely look linked. It looks like the points lie on a parabola (the shape of an x^2 curve) — not a straight line.



What's a statistician's favourite soap — Correlation Street... (Boom boom)

It's worth remembering that the PMCC assumes that both variables are normally distributed — chances are you won't get asked a question about that, but there's always the possibility that you might, so learn it.

Linear Regression

Linear regression is just fancy stats-speak for 'finding lines of best fit'. Not so scary now, eh...

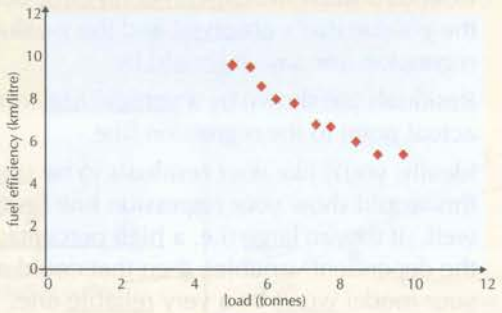
Decide which is the **Independent Variable** and which is the **Dependent**

EXAMPLE The data below shows the load on a lorry, x (in tonnes), and the fuel efficiency, y (in km per litre).

x	5.1	5.6	5.9	6.3	6.8	7.4	7.8	8.5	9.1	9.8
y	9.6	9.5	8.6	8.0	7.8	6.8	6.7	6	5.4	5.4

1) The variable along the x-axis is the explanatory or independent variable — it's the variable you can control, or the one that you think is affecting the other. The variable 'load' goes along the x-axis here.

2) The variable up the y-axis is the response or dependent variable — it's the variable you think is being affected. In this example, this is the fuel efficiency.



The **Regression Line** (Line of Best Fit) is in the form **$y = a + bx$**

To find the line of best fit for the above data you need to work out some sums.
Then it's quite easy to work out the equation of the line. If your line of best fit is $y = a + bx$, this is what you do...

① First work out these four sums — a table is probably the best way: $\sum x$, $\sum y$, $\sum x^2$, $\sum xy$.

x	5.1	5.6	5.9	6.3	6.8	7.4	7.8	8.5	9.1	9.8	$72.3 = \Sigma x$
y	9.6	9.5	8.6	8	7.8	6.8	6.7	6	5.4	5.4	$73.8 = \Sigma y$
x^2	26.01	31.36	34.81	39.69	46.24	54.76	60.84	72.25	82.81	96.04	$544.81 = \Sigma x^2$
xy	48.96	53.2	50.74	50.4	53.04	50.32	52.26	51	49.14	52.92	$511.98 = \Sigma xy$

② Then work out S_{xy}, given by: $S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$

and S_{xx}, given by: $S_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$

These are the same as the terms used to work out the PMCC (see p.137).

③ The gradient (b) of your regression line is given by: $b = \frac{S_{xy}}{S_{xx}}$

④ And the intercept (a) is given by: $a = \bar{y} - b\bar{x}$.

⑤ Then the regression line is just: $y = a + bx$.

Loads of calculators will work out regression lines for you — but you still need to know this method, since they might give you just the sums from Step 1.

EXAMPLE Find the equation of the regression line of y on x for the data above.

The 'regression line of y on x' means that x is the independent variable, and y is the dependent variable.

1) Work out the sums: $\sum x = 72.3$, $\sum y = 73.8$, $\sum x^2 = 544.81$, $\sum xy = 511.98$.

2) Then work out S_{xy} and S_{xx}: $S_{xy} = 511.98 - \frac{72.3 \times 73.8}{10} = -21.594$, $S_{xx} = 544.81 - \frac{72.3^2}{10} = 22.081$

3) So the gradient of the regression line is: $b = \frac{-21.594}{22.081} = -0.978$ (to 3 sig. fig.)

Remember: $\bar{x} = \frac{\sum x}{n}$

4) And the intercept is: $a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{73.8}{10} - (-0.978) \times \frac{72.3}{10} = 14.451 = 14.5$ (to 3 sig. fig.)

5) This all means that your regression line is: $y = 14.5 - 0.978x$

The regression line always goes through the point $(\bar{x}, \bar{y})$.

This tells you: (i) for every extra tonne carried, you'd expect the lorry's fuel efficiency to fall by 0.978 km per litre, and (ii) with no load ($x = 0$), you'd expect the lorry to do 14.5 km per litre of fuel.

Assuming the trend continues down to $x = 0$.

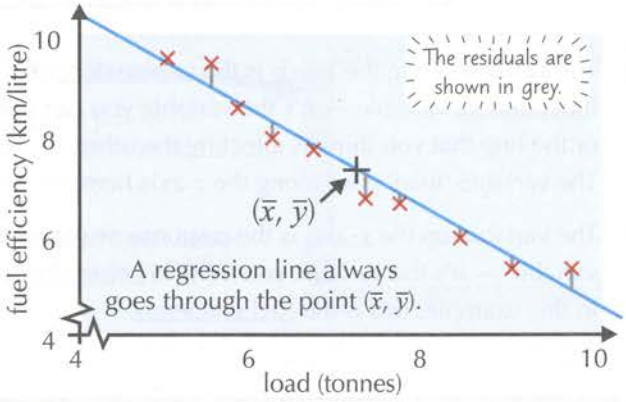
Linear Regression

Residuals — the difference between Practice and Theory

A residual is the difference between an observed y-value and the y-value predicted by the regression line.

Residual = Observed y-value – Estimated y-value

1) Residuals show the experimental error between the y-value that's observed and the y-value your regression line says it should be.
2) Residuals are shown by a vertical line from the actual point to the regression line.
3) Ideally, you'd like your residuals to be small — this would show your regression line fits the data well. If they're large (i.e. a high percentage of the dependent variable), then that could mean your model won't be a very reliable one.



EXAMPLE For the fuel efficiency example on the last page, calculate the residuals for: (i) $x = 5.6$, (ii) $x = 7.4$.

(i) When $x = 5.6$, the residual = $9.5 - (-0.978 \times 5.6 + 14.451) = 0.526$ (to 3 sig. fig.)

(ii) When $x = 7.4$, the residual = $6.8 - (-0.978 \times 7.4 + 14.451) = -0.414$ (to 3 sig. fig.)

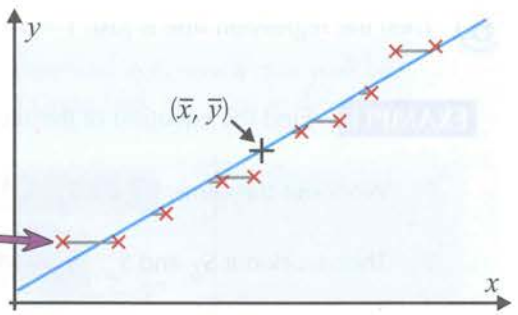
A positive residual means the regression line is too low for that value of x.
A negative residual means the regression line is too high.

This kind of regression is called Least Squares Regression, because you're finding the equation of the line which minimises the sum of the squares of the residuals (i.e. $\sum e_i^2$ is as small as possible, where the e_i are the residuals).

You can also Find the Regression Line of x on y

The formulas on the previous page give you the 'regression line of y on x', which you use when x is the independent variable and y the dependent variable. But if y is actually the independent variable, then you need the 'regression line of x on y'.

1) In that case, your regression line will be $x = c + dy$, where: $d = \frac{S_{xy}}{S_{yy}}$ and $c = \bar{x} - d\bar{y}$.
2) And your residuals will look like this:



You can't just rearrange the regression line of y on x to get the regression line of x on y. You must work it out from scratch.

I predicted I'd win a million on the lottery — but the residual turned out to be large...
Residuals are errors in the dependent variable — not the independent variable. The regression equations on the previous page will be in your formula booklet, so you don't need to learn them, but practise using and interpreting them.

More About Regression and Correlation

Use Regression Lines With Care

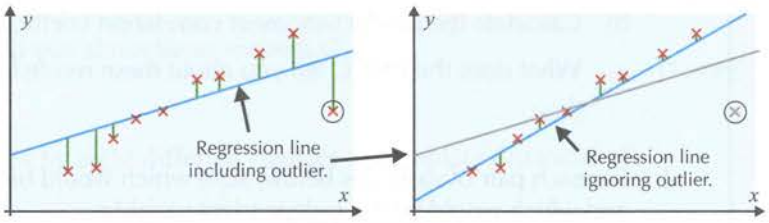
You can use your regression line to predict values of the dependent variable.
But it's best not to do this for values of the independent variable outside the range of your original table of values.

EXAMPLE Use your regression equation from p139 to estimate the value of y when: (i) $x = 7.6$, (ii) $x = 12.6$

(i) When $x = 7.6$, $y = -0.978 \times 7.6 + 14.5 = 7.1$ (to 2 sig. fig.). This should be a pretty reliable guess, since $x = 7.6$ falls in the range of x we already have readings for — this is called interpolation.

(ii) When $x = 12.6$, $y = -0.978 \times 12.6 + 14.5 = 2.2$ (to 2 sig. fig.). This may well be unreliable since $x = 12.6$ is bigger than the biggest x-value we already have — this is called extrapolation.

Outliers can also be a problem — they can have a big effect on the regression line's equation, and drag it far away from the rest of the data values.
Here, the circled data value is an outlier.



Spearman's Rank Correlation Coefficient (SRCC or r_s) works with Ranks

You can use the SRCC (or r_s for short) when your data is a set of ranks. (Ranks are the positions of the values when you put them in order — e.g. from biggest to smallest, or from best to worst, etc.)

EXAMPLE At a dog show, two judges put 8 labradors (A-H) in the following orders, from best to worst. Calculate the SRCC between the sets of ranks.

Position	1st	2nd	3rd	4th	5th	6th	7th	8th
Judge 1:	B	C	E	A	D	F	G	H
Judge 2:	C	B	E	D	F	A	G	H

First, make a table of the ranks of the 8 labradors — i.e. for each dog, write down where it came in the show.

Dog	A	B	C	D	E	F	G	H
Rank from Judge 1:	4	1	2	5	3	6	7	8
Rank from Judge 2:	6	2	1	4	3	5	7	8

Now for each dog, work out the difference (d) between the ranks from the two judges — you can ignore minus signs.

Dog	A	B	C	D	E	F	G	H
d	2	1	1	1	0	1	0	0

Take a deep breath, and add another row to your table — this time for d^2:

Dog	A	B	C	D	E	F	G	H	Total = Σd^2
d^2	4	1	1	1	0	1	0	0	8

Then the SRCC is:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

This formula is given on the formula sheet.

You can ignore minus signs when you work out d, since only d^2 is used to work out the SRCC.

So here, $r_s = 1 - \frac{6 \times 8}{8 \times (8^2-1)} = 1 - \frac{48}{504} = 0.905$ (to 3 sig. fig.).

Interpret r_s in the same way as you'd interpret the PMCC (see p137).

— this is close to +1, so the judges ranked the dogs in a pretty similar way.

99% of all statisticians make sweeping statements...

Be careful with that extrapolation business — it's like me saying that because I grew at an average rate of 10 cm a year for the first few years of my life, by the time I'm 50 I should be 5 metres tall. (There's still time, but I can't see that happening.)