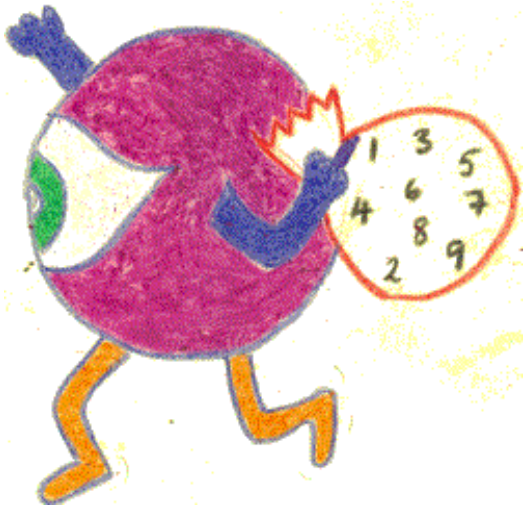
Relevant & Engaging Statistics & Data Handling

All the Statistical Facts, Formulae and information you need to know



Centre for Statistical Education

Chapter 9

Second Edition: published in the UK in 2009 by

The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
Plymouth
PL4 8AA
UK

We would like to thank the many contributors to this book. These include: Doreen Connor, Neville Davies, John Marriott, Alan Catley (Chapter 4) and Mark Crowley (Chapter 6) and those who reviewed and helped in it's production including Alison Davies, Claire Webster, Peter Holmes & Elizabeth Gibson.

This booklet is aimed at all secondary level teachers: there are hints and tips that we hope will be useful to support the teaching and learning of statistics and data handling. We hope that you find the material useful. Please email or write to us with suggestions for improvements. We will try to respond to all communications.

Doreen Connor
The Royal Statistical Society Centre for Statistical Education
The University of Plymouth
2009

email info@censusatschool.org.uk

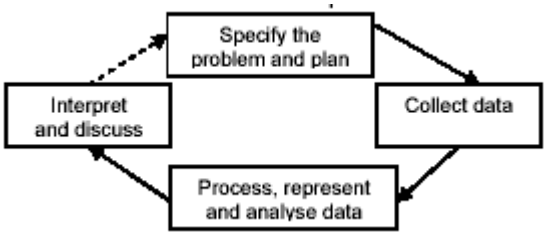
www.censusatschool.org.uk

www.rsscse.org.uk

Chapter 9

All the Statistical Facts, Formulae and information you need to know

Data are numbers in context and the goal of statistics is to get information from those data, usually through *problem solving*. The dotted line means that, following discussion, the problem may need to be re-formulated and at least one more iteration completed.



Types of Data

Discrete Data	Data that can only take certain values. e.g. having 0.3 of an egg or 0.26 sisters is not sensible.
Continuous Data	Data that is measured on a continuous scale and can be represented on a number line. e.g. time, length. This *should be called* measurement data.
Quantitative Data	Numerical data, literally data that have a quantity. This includes both discrete and measurement data.
Qualitative Data	Non numerical data. Literally data that has quality e.g. colours, types of trees etc.
Categorical Data	The data are in categories. This is often qualitative data as the categories are not numerical.
Bivariate Data	Data that has two variables recorded for each data point. E.g. height and weight. Often presented on a scatter diagram where each point is represented by a coordinate.

Outliers: These are unusual data points. They can be caused by mistakes or can be caused by natural variation and be simply an unusual data point. It is always wise to identify and investigate outliers, commenting on them and deciding if you want to keep them in the dataset or not.

Measures of Central Tendency (averages)

Average

A single value that is a representative of all the data. There are a number of different types of average and each is more useful in some situations than in others.

Mean
(NC level 5 for discrete data up to NC level 7 for grouped data)

Full title the arithmetic mean as you do arithmetic in order to work it out. Calculated by adding up all the data values and dividing by the number of values you have.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

For data from frequency tables $\bar{x} = \frac{\sum x_i f_i}{n}$ where $\sum f_i = n$ i.e. each value is multiplied by its frequency and added up before dividing by n the number of values you have.

If you have grouped data the mid point value of each group is used to calculate the mean.

Median
(NC level 5)

The value that is in the middle of the data when the data is arranged in order. If you have an odd number of values find the arithmetic mean of the two values either side of the median.

Mode
(NC level 4)

The data value you have the most of – the most frequently occurring value.

If you have grouped data the most frequently occurring group is used.

Geometric Mean

Only used in higher level work

Geometric mean = $\sqrt[n]{x_1 x_2 x_3 ..... x_n}$

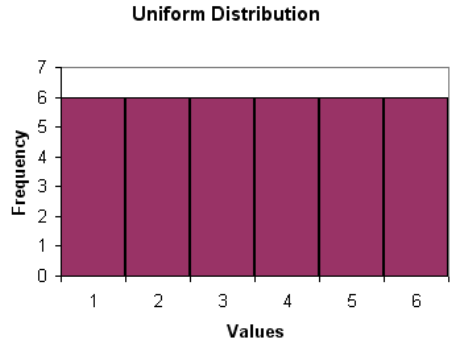
Harmonic Mean

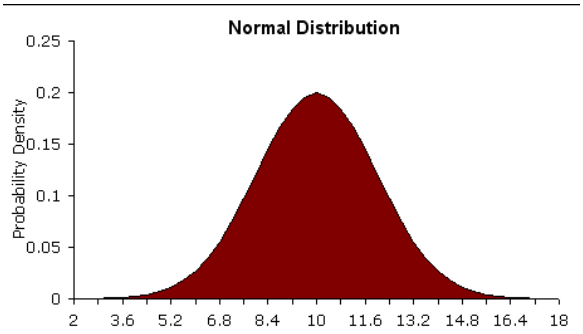
$$\frac{n1}{harmonic\ mean} = \frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}$$

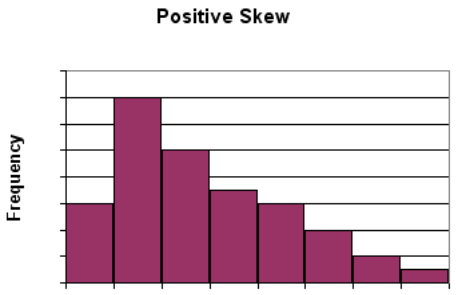
(In certain situations, the harmonic mean provides the truest average. For instance, if for half the ***distance*** of a trip you travel at 40 kilometres per hour and for the other half of the ***distance*** you travel at 60 kilometres per hour, then your average speed for the trip is given by the harmonic mean of 40 and 60, which is 48; that is, the total amount of time for the trip is the same as if you travelled the entire trip at 48 kilometres per hour. If you had travelled for half the ***time*** at one speed and the other half at another, the arithmetic mean, in this case 50 kilometres per hour, would provide the correct average.)

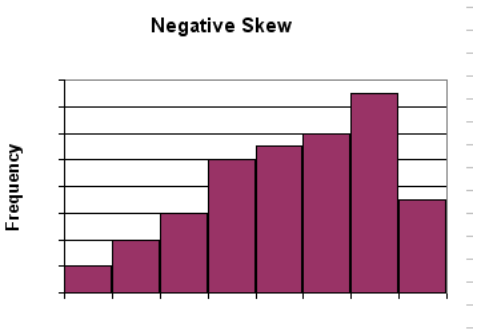
Distributions

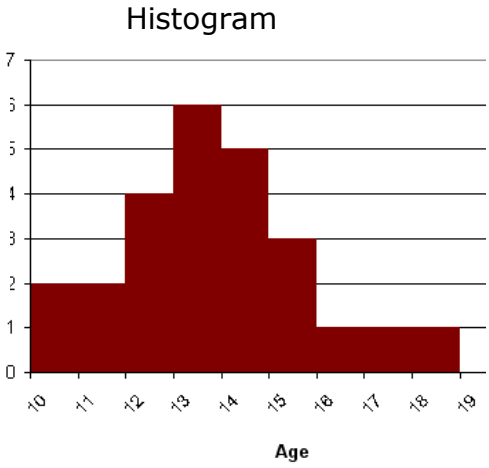
This is how the data is distributed around. There are several very important shapes to look out for.









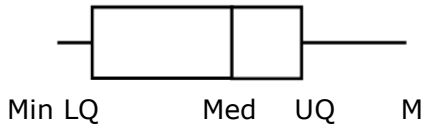


Here the **area** of each bar is proportional to the frequency. Frequency density can be calculated by dividing the frequency for a class by it's width

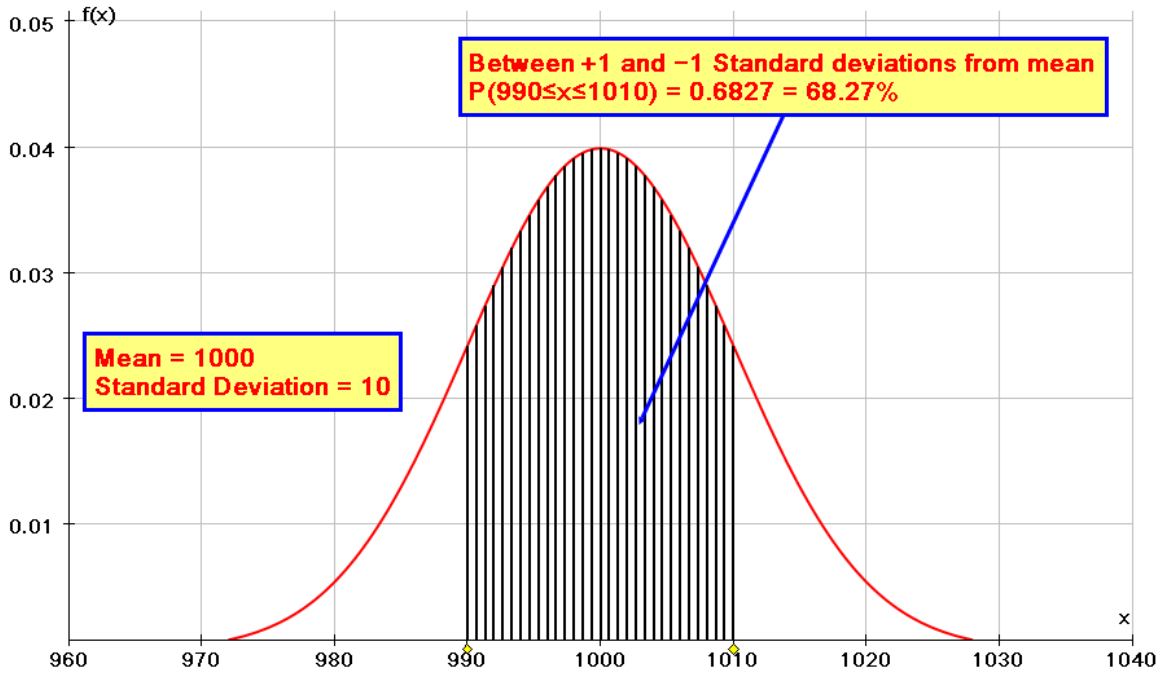
(NC level 5 to exp perform and continues through GCSE and A level Statistics modules.)

Measures of Dispersion (spread)

Range (NC level 4)	The difference between the largest and smallest data value
Inter-quartile Range (NC level 8)	The range of the middle half of your data. The top and bottom quarters are ignored. To find the inter-quartile range you take the lower quartile away from the upper quartile. (see box plot below)
Variance (A level S1)	This is the standard deviation squared and represents the *average* of the squared differences between data points and the mean $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation (A level S1)	This measures the spread of data about the mean, measured in the same units as the data. If many data points are close to the mean, then the standard deviation is small; if many data points are far from the mean, then the standard deviation is large. If all the data values are equal, then the standard deviation is zero. It is worked out by taking all of the differences between the individual data points and the mean, squaring them, adding them all up and dividing by 1 less than the number of values and then finally square rooting the answer. (the squaring and squarerooting are to ensure that positive and negative distances from the mean do not cancel each other out.) $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Box Plot (NC level 7)	If the sample data are ordered from smallest to largest then the: minimum (Min) is the smallest; lower quartile (LQ) is the ¼(n+1)-th value; median (Med) is the middle [or the ½(n+1)-th] value; upper quartile (UQ) is the ¾(n+1)-th value; maximum (Max) is the largest. These five values constitute a **five-number summary** of the data. They can be represented diagrammatically by a *box-and-whisker plot,* commonly called a *boxplot*.



The Normal Distribution



An example of the normal distribution, drawn by Autograph

For an interactive demonstration of the normal distribution go to the Autograph tutorial available from the Autograph website at
http://www.autograph-math.com/inaction/

(A level S1 although understanding implied at NC level 8)

Confidence Intervals

What can you say about the population when all you've got is a sample from it? The value of a statistic (e.g. a mean) worked out from a sample is obviously only one estimate of the value of the true mean of the population. If you drew a different sample, you'd get a different value.

The only way you can really get the true value is to measure everyone in the population. But it is possible to use your sample to calculate a range, or interval within which the population value is likely to fall. "Likely" is usually taken to be "95% of the time," and the range is called the **95% confidence interval**. The values at each end of the interval are called the **confidence limits**.

The **confidence interval is the likely range of the true value**. Note that there is *only one* true value, and that the confidence interval defines the range where it's most likely to be. The confidence interval is NOT the variability of the true value or of any other value between subjects. It is nothing like a standard deviation.

(A level S1)

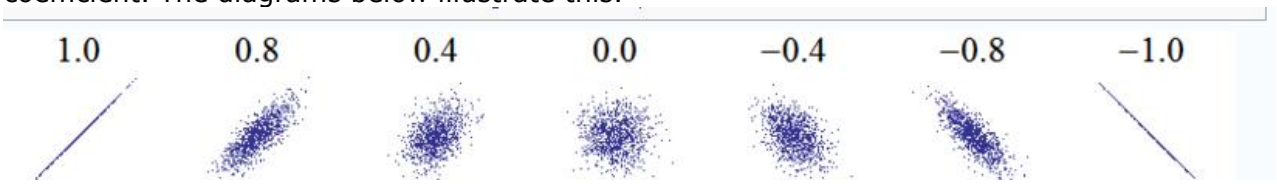
Population and Types of Sample

Population	The complete set of items under consideration.
Census	Data is taken from every member of the population under consideration
Survey	Data is taken from a sample of a population in order to investigate something
Sample	A set of items taken from a population
Simple Random Sample	A sample chosen so that each member of the population has an equal chance of being chosen. Each item is chosen entirely by chance.
Stratified sample	The population is divided into groups, or strata, and a sample, usually a random sample, is taken from each of these groups in the same proportion as the size of the strata.
Quota sample	The population is divided into groups and the number of items to take from each group is decided in advance. The sample must then fit the quota.
Cluster sample	The population is represented by a series of groups, or clusters that are thought to behave in roughly the same way as the population as a whole
Systematic sample	The sample is taken by taking items at regular intervals. e.g every 10^{th} item
Opportune sample	A convenient sample (often contains bias)

(NC Exp Performance GCSE H4.2d)

Correlation & Regression

Correlation shows the strength and direction of a linear association between two variables. If there is a causal relationship between the two variables, regression is fitting or modelling a line which shows this relationship.
Correlation can be shown in a pictorial way or by calculating a correlation coefficient. The diagrams below illustrate this.



Perfect positive correlation	Strong Positive correlation	Weak Positive correlation	No correlation	Weak negative correlation	Strong negative correlation	Perfect negative correlation

There are several different ways the coefficient can be calculated. The most common two are;

Pearson's Product moment correlation coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$$

so for every point the x and y distances from the mean are calculated and summed. This total is then divided by 1 less than the number of data points and the product of the standard deviations of x and y.

Spearman's rank correlation coefficient

Here rather than take the actual data points they are ranked. it is derived from the same idea of the Pearson's coefficient.

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

where d_i = rank(x_i)-rank(y_i) and n is the number of data items.

Regression

To fit a line to model the relationships statisticians tend to do this using a method of least squares saying that the least squares regression line of y on x is: $y - \bar{y} = b(x - \bar{x})$ where $b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

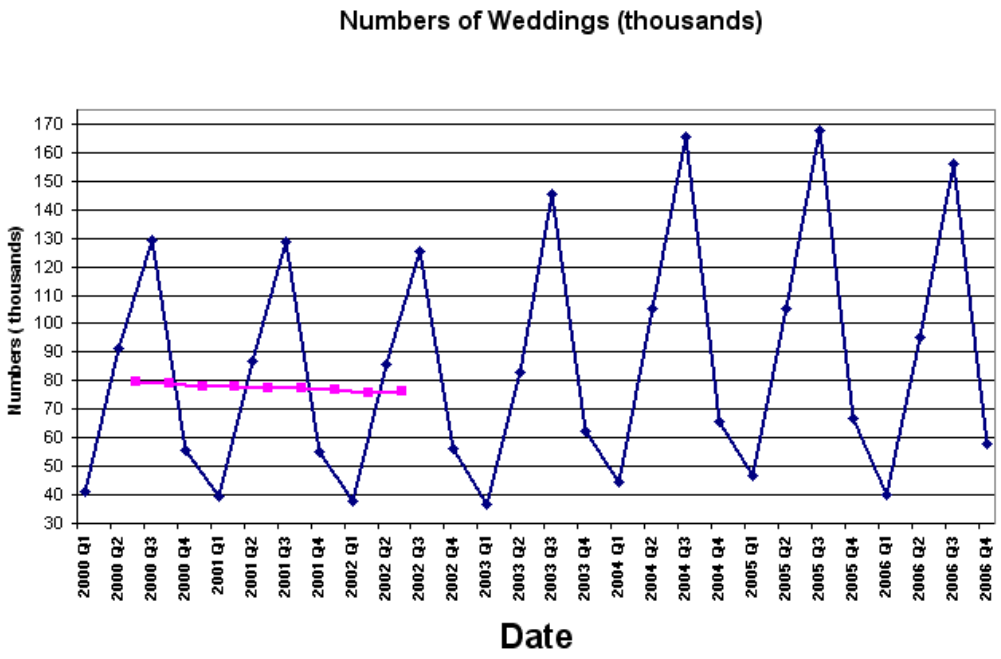
Mathematicians on the other hand will tend to adapt the already familiar
y = mx + c model.

Either way will result in the same equation

(Note: for bivariate data you are usually interested in whether one variable affects the other. The one affected is the DEPENDENT variable, the one causing the effect is the INDEPENDENT variable.)

Time Series

Much as the title implies time series data represents a series or set of observations or data points that are recorded over time. It may be the level of gas bills over a number of years or the maximum temperature over a year or the Retail price index etc. The data is analysed in order to look at trends and to make predictions about what might happen in the future.
Moving averages are calculated in order to smooth out fluctuations such as seasonal variation and to help with predictions. The moving averages themselves are also a time series.
In the example below the pink line shows the start of the 4 point moving average. 4 points are chosen in order to smooth out the very obvious seasonal trend that is occurring in this data. The number of points, 4, is chosen because the cycle observed in the raw data repeats itself after each four points. The points are plotted at the appropriate place so in this case the first moving average is plotted half way between point 2 and 3.



There are many good sources of real and relevant time series data available for you to use in the classroom available form the Internet. A couple of the best ones are climate and waste statistics available from the Defra website (for example currently http://defra.gov.uk/environment/statistics/globatmos/alltables.htm) and the Meteorological Office for weather data (for example you can find data going back many years for your local weather station from http://www.metoffice.gov.uk/climate/uk/stationcheck/index.html)

The National Lottery and the International Olympics records websites also contain a wealth of interesting time series data.

(GCSE Higher H 4.4f H 4.5b)

Hypothesis Tests

A hypothesis test involves testing a claim that has been made about a population by using a sample of that population.

For example:
'The life of XXX batteries is over 12 hours'
Or
'British Power claim that 90% of its customers enquiries are answered within 2 minutes'

The first step is to make a clear statement about the claim that we can test out. This is called the *null hypothesis* (H_0). Then we make the alternative claim and this is called the alternative hypothesis (H_1). The third step is then to test out our hypothesis and make a decision as to whether we can reject or accept our *null hypothesis*. The testing and decision making involves both using statistical distributions and particular statistics and also considered different types of error we may have made because we are using a sample of the population.

Let us investigate the probability **P** of getting a head from a coin. Then your null hypothesis, H_0, is what you are expecting to happen, i.e. you would test if **P**=0.5, and your alternative hypothesis, H_1, is something else that could happen, which you would like to test to see if it might occur. So, you could test if **P**≠0.5 (you may suspect that you have a biased coin and would like to test this theory out.)
If you then carry out your statistical tests, and it seems that there is evidence to suggest that your coin is indeed biased (e.g. if you threw the coin 100 times and only got 17 heads), then it looks like H_1 is true, and H_0 is not, and so you REJECT THE NULL HYPOTHESIS.

We do not intend to go into any further details here but would recommend the following websites where further information and more details can be found.

http://www.mathsrevision.net search for Hypothesis testing

http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html quite detailed

http://www.coventry.ac.uk/ec/research/discus/discus_8.html the DISCUSS workbook on Hypothesis tests

(A level S2, S3, S4)

Permutations and Combinations

This is how many ways items can be arranged. For example if you had 3 pieces of fruit for your snack, an apple (A), an orange (O) and a banana (B) there are actually 6 different orders in which you could eat them:

AOB ABO BAO BOA OBA OAB

But if you decide that you only wanted to eat any two out of the three pieces of fruit then you would have 3 choices:

AO AB OB

So the total number of ways of arranging ***n*** unlike objects is ***n!***

To select r objects out of a total of ***n***, the number of combinations is

$$^{n}C_{r} = \left(\frac{n}{r}\right) = \frac{n!}{r!(n-r)!}$$

If the order in which you select is important as well, then the total number of ways is the number of permutations, and this is:

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

So in the fruit example if I also have a pear (P), so that there are four pieces of fruit in total, I will have 4! or 24 different possible arrangements.

If I decide to eat any 2 pieces of fruit – i.e the order in which I eat these two pieces of fruit doesn't matter then I have 6 possible outcomes
AO AB AP OB OP BP

But if I am very fussy and think that the order is important and that eating an Apple followed by a banana is different from first eating the banana and then the apple I have 12 possible outcomes;

AB BA AO OA AP PA OB BO OP PO BP PB

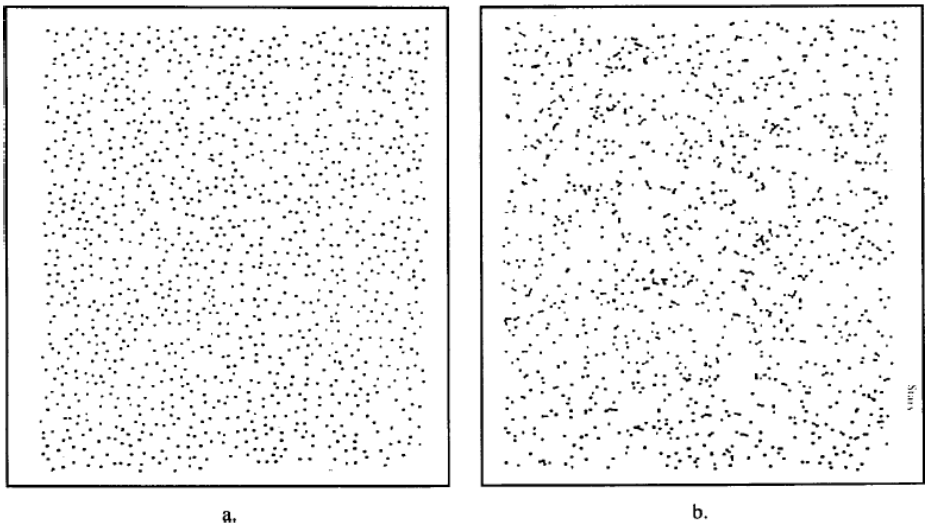
NOTE: The number of combinations are the numbers that appear in Pascal's triangle and these are also extremely important when using the Binomial distribution which is an extremely common distribution dealing with any number of independent events.

(A level S1)

Probabilities

Probability is the study of chance. It deals with likelihoods of events and variation and uncertainty. Chance and uncertainty pervade our real world but many learners leave school with very little grasp of this extremely important concept.
Within the mathematics classroom they will learn that the probability of a dice coming up with a 6 is 1/6, exactly the same as the probability of the dice coming up with a 2, but in the real world they know how hard it is to throw a 6 to start their favourite board game and so tend to leave the mathematical version in the classroom. They will learn that the likelihood of winning the jackpot in the lottery is 1 in 14 million but still spend their money on the lottery tickets, after all someone has to win!
It is also the case that many people badly underestimate the likelihood of surprising results and we, as mathematics teachers often do not help our learners to understand the true nature of chance and uncertainty and instead concentrate on simulations or practicals using coins and dice along with the theoretical probabilities that we can work out using arithmetic. Consequently, we as teachers must accept some of the blame for the total lack of understanding shown by the general public towards any chance event. There have been some very high profile incidents which highlight this issue and for many leave a sour taste in the mouth. To name but one such example when the prosecutors fallacy came into play to send innocent women to prison because a court believed the argument that a very small probability meant that the only explanation had to be that that the women had murdered their own children!
So beware of teaching probability without ensuring that you are teaching your learners an understanding of chance and uncertainty. Ensure they understand that random does **not** mean equally spread out. Ensure they understand that the mathematical theory of a situation is **not** what will happen in reality but that natural variation will occur. Use examples that demonstrate and show how mathematical models are developed but that they are just that, models, and need to be kept in their rightful place.

Look at the following two diagrams. Are these patterns random?



a. b.

Figure b is a random collection of points generated from two independent uniform distributions for their x and y co-ordinates respectively. Figure a is a plot of the same number of points generated through a fixed law, ordered by fields of inhibition around each dot (from a random start). Most students see a pattern in Figure b and randomness in Figure a while in reality the exact opposite is true.
(This example is taken from Chatterjee, **S.** (1996) Statistics and Intuition for the Classroom. ***Teaching Statistics,*** **18(2),** 34-38.)

An Example of Teaching Probability Theory

A useful and straightforward way in which to discuss probability with learners is to use survey data that they have an interest in. This can cover many probability concepts, some normally thought of as difficult for learners to understand, in an easily understandable and logical way. In what follows we provide an example that employs data from random samples obtained from phase 4 of CensusAtSchool.

The table below gives the responses from 350 learners to the question **"What type of book is your favourite?"** and the responses have been organised into a 2-way table according to the gender of the respondent.

	Favourite book choice				
Gender	Adventure	Crime	Fantasy/Horror	Romance	Total
F	57	10	55	73	195
M	50	27	68	10	155
Total	107	37	123	83	350

An initial question to the class could be

"What percentage of this group is female?"

The response should then be easily seen to be $100 \times \frac{195}{350}\% = 55.7\%$.
A simple discussion of how a learner from this group could be selected at random could then be followed by the question
"What is the chance that a randomly selected learner from this group is female?" and the answer to this is clearly also 55.7%.

A follow up question could be

"What is the chance that a randomly selected learner from this group chose 'Romance' as their favourite type of book?"

the answer to which is 23.7%.

A natural question arising from the above is

"What is the chance that a randomly selected learner from this group *both* chose Romance as their favourite book *and* is female?"

here the use of counts from inside the table are used and the answer is $100 \times \frac{73}{350}\% = 20.9\%$.

Because nearly half of the chosen group are male, and some of the male learners also chose Romance, the 'addition law' question

"What is the chance that a randomly selected learner from this group *either* chose Romance as their favourite book *or* is female?"

is of interest. Now, we can see that the argument used to find this answer can proceed as follows: if we simply take the total of the Romance column and the Female row to calculate $100\times\left(\frac{83+195}{350}\right)\%$ we have inadvertently ***double counted*** the group who *both* chose Romance as their favourite book *and* are female. There were 73 of those. So the correct calculation must be $100\times\left(\frac{83+195-73}{350}\right)\%=100\times\left(\frac{205}{350}\right)\%=58.6\%$

At any convenient stage during the above the idea of using numbers between 0 and 1 as measures of chance (probability) can be introduced and the use of the phrases "Proportion of the chosen group" and "probability that…" can be introduced.

One topic that is particularly difficult to explain to learners is conditional probability. By employing the approach above we can see how this concept can be dealt with quite easily using the ideas already developed.

We ask the question

"What is the chance that if a learner is randomly selected from those who chose Romance the chosen learner will be female?"

A natural intermediate stage is to ask

"What proportion of the Romance column is made up of Female learners?"

the answer to this then leads to the required probability $\frac{73}{83}=0.880$, which can also be presented as representing an 88.0% chance that a randomly selected 'Romance learner' would be female. And this of course means that there is a 12% chance that the randomly selected 'Romance learner' would be male!

Similarly, and providing an interesting discussion point, the probability that a randomly selected female learner would be one who chose Romance as her favourite is $\frac{73}{195}=0.374$, that is , there is a 37.4% chance of such an event occurring.

(Probability occurs from year 5 in the Primary Framework, throughout Key Stage 3 and 4, from level 5 to exceptional performance and has elements within all A level statistics modules.)