2. Statistics

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write"

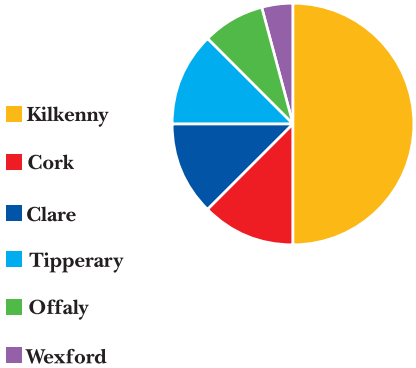
Statistician S.S. Wilks paraphrasing from H.G. Wells's 1903 book "Mankind in the Making"

Statistics is the collection, presentation and analysis of data.
The vast majority of decisions made today are taken after considering research. Governments, companies, institutions of all sorts need 'facts' before they make their decisions. The 'facts' are usually presented graphically by statisticians. Lists of numbers are difficult to read, so statisticians take the raw data, the numbers, and produce easy-to-read, bar charts, trend graphs, pie charts and a whole variety of different visual ways of presenting the data, to make it more useful.

Use of Statistics

You will already have used charts to show facts clearly.

Example 1: Draw a **Pie Chart** of the Men's All Ireland Hurling winners for the 24 years between 1991 and 2014 inclusive:

Data		Convert to Degrees	County Wins
Kilkenny	12	$\frac{12}{24}\times 360 = 180°$	
Cork	3	$\frac{3}{24}\times 360 = 45°$	
Clare	3	$\frac{3}{24}\times 360 = 45°$	
Tipperary	3	$\frac{3}{24}\times 360 = 45°$	
Offaly	2	$\frac{2}{24}\times 360 = 30°$	
Wexford	1	$\frac{1}{24}\times 360 = 15°$	

Men's Hurling All Ireland Winners 1991 – 2014

PROJECT 2.1 **Draw Pie Charts**

Look up the winners of the Football All Ireland Championships for both the men and the women for the 24 years between 1991 and 2014. Draw a Pie Chart for each to compare which counties are strongest in the men's Football Championship with those of the women's Football Championship for this period.

Statistics have obvious uses like this but advanced statistics can be used to do amazing things.

Language Translation

An interesting example of a possible use of statistics is a project Google is reported to be currently working on. They hope to be able to translate from one language to another in real time. They hope to use statistics derived from the billions of pages of text used on their sites every day to create this service. They do not intend to use rules or grammar to translate between the 47 languages. They are using the statistical probability of one word being beside another word to do this, e.g. you speak in English on the phone and the person at the other end hears you in Chinese, she speaks back in Chinese but you hear her in English.

Even if you think you have little interest in maths, everyone is affected by statistics during their life; e.g. if you get 40% in your Geography exam you might think that this is not a very good mark. However if the class average is only 34%, then it is actually quite a good mark.

Averages

The average of a set of numbers is the most useful way of **analysing** them. We will look at three ways of finding the average and investigate which is the most suitable in different situations:

Mean = $\frac{\text{sum of all numbers}}{\text{amount of numbers}}$ (the mean is what most people call the average)

Mode = the mode is the number or value which occurs most often

Median = when all the values are listed in order of size, the median is the middle one (or the average of the two middle ones)

Example 2: The marks for 28 pupils in a history exam are: 35, 38, 43, 46, 46, 49, 50, 55, 56, 56, 58, 60, 60, 61, 61, 62, 64, 65, 65, 67, 68, 70, 71, 72, 82, 82, 82, 84.

Find the mean, mode, median and comment on suitability of each average:

$$\text{Mean} = \frac{35+38+\ldots+84}{28} = \frac{1708}{28} = 61$$

$$\text{Median} = \frac{61+61}{2} = 61$$

$$\text{Mode} = 82$$

For a pupil wishing to compare her mark to the average student in the class, both the **mean** and the **median** give a good estimation in this example. The **mode** in this case is poor, because, by chance, there were three pupils scoring 82.

Example 3: An American professional basketball team publishes its players' wages. They have one exceptional star player. The wages per annum are: \$360,000, \$380,000, \$400,000, \$420,000, \$460,000, \$490,000, \$500,000, \$500,000, \$520,000, \$530,000, \$600,000, \$5,640,000. Find the mean, mode, median and comment on the suitability of each average

$$\text{Mean} = \frac{350,000 + 380,000 + \ldots + 5,640,000}{12} = \frac{10,800,000}{12} = \$900,000$$

$$\text{Median} = \frac{\$490,000 + \$500,000}{2} = \$495,000$$

Mode = \$500,000 (only figure which occurs twice)

In this example, for an ordinary player considering signing to play for the team, the median or the mode gives him a better idea of what he may earn. The mean is affected by the very high wages of one exceptional player. This is called outlier.

Normal Distribution

Example 4: 100 pupils were shown a 20 digit number for 30 seconds and then asked to recall the number. The number of digits they could remember and repeat before they made a mistake was recorded:

Digits remembered	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency	2	0	2	3	7	5	8	10	15	13	11	9	5	4	2	2	0	0	1	1	0

Plot the data on a bar chart, the frequency goes on the *y*-axis. Find the three averages, for this data, the mean, the mode and the median.

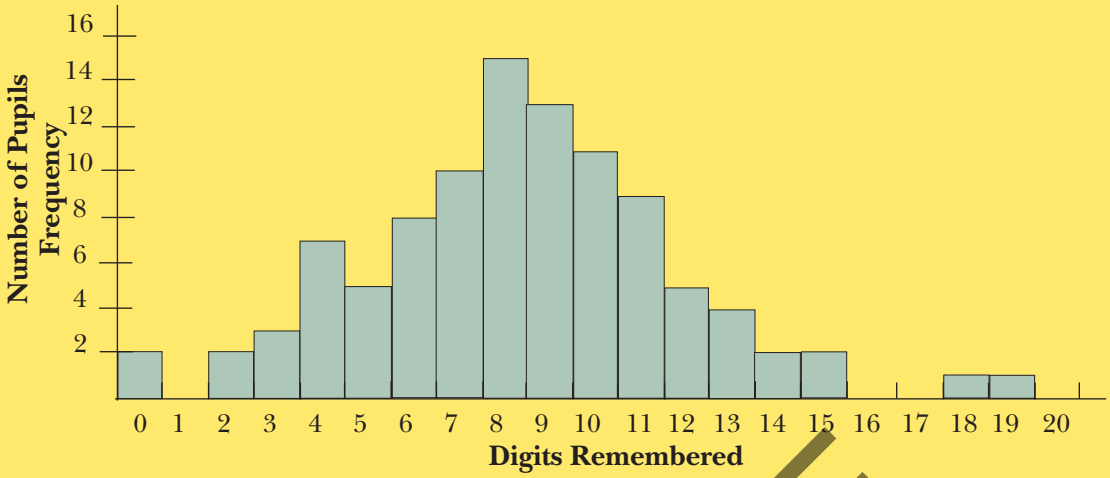
$$\text{Mean} = \frac{\text{total number of digits remembered}}{\text{total number of pupils}}$$

$$= \frac{2\times0 + 0\times1 + 2\times2 + 3\times3 + 7\times4 \ldots 1\times19 + 0\times20}{2+0+2+3\ldots+1+0} = 8{\cdot}37$$

Mode = the most common number of numbers remembered = 8

Median = line all the numbers remembered in a row and the median is the middle one, or the average of the two middle ones if it is an even number i.e. $= \frac{8+8}{2} = 8$

The data in this example should produce a bar chart which is close to the classical bell shape of a normal distribution. The data should be roughly symmetrical about the mean value.



Standard Deviation

The Standard Deviation of a set of numbers gives an indication of the spread of the data. The mean value gives an indication of the average. The Standard Deviation gives more information. It tells how far the figures are, on average, from the mean. It can be calculated from the formula:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$\sigma =$ standard deviation

$\sum =$ the sum of

$x =$ each value in the data set

$\bar{x} =$ the mean of all values in the data set

$n =$ the number of values in the data set

Example 5: Eight pupils from a Transition Year class recorded their pulse rate per minute when at rest. They then recorded the pulse rate for eight family members of different ages. The results were as follows:

Transition Year Pupils	72	70	74	80	74	75	76	79
Family Members	64	58	88	73	89	68	92	68

Find the **mean** and **standard deviation** of both sets of data and comment on the results.

$$\text{Mean (Pupils)} = \frac{72+70+74+80+74+75+76+79}{8} = \frac{600}{8} = \mathbf{75} \text{ (Mean)}$$

$$\text{Mean (Family members)} = \frac{64+58+88+73+89+68+92+68}{8} = \frac{600}{8} = \mathbf{75} \text{ (Mean)}$$

SET 1 (Pupils)			
x_1	$\bar{x}$	$(x_1 - \bar{x})$	$(x_1 - \bar{x})^2$
72	75	-3	9
70	75	-5	25
74	75	-1	1
80	75	5	25
74	75	-1	1
75	75	0	0
76	75	1	1
79	75	4	16
			78

SET 2 (Family Members)			
x_2	$\bar{x}$	$(x_2 - \bar{x})$	$(x_2 - \bar{x})^2$
64	75	-11	121
58	75	-17	289
88	75	13	169
73	75	-2	4
89	75	14	196
68	75	-7	49
92	75	17	289
68	75	-7	49
			1166

Standard Deviation $= \sqrt{\frac{78}{8}} = 3{\cdot}12$ Standard Deviation $= \sqrt{\frac{1166}{8}} = 12{\cdot}07$

Comment: By chance both groups had the same mean pulse rate. The Transition Year pupils all had a similar rate, causing the standard deviation to be low, 3·12. The Family Members, some possibly very young or very old had a much wider spread, resulting in a relatively high standard deviation of 12·07.

The following project is based on Example 4; it should take approximately two class periods to complete:

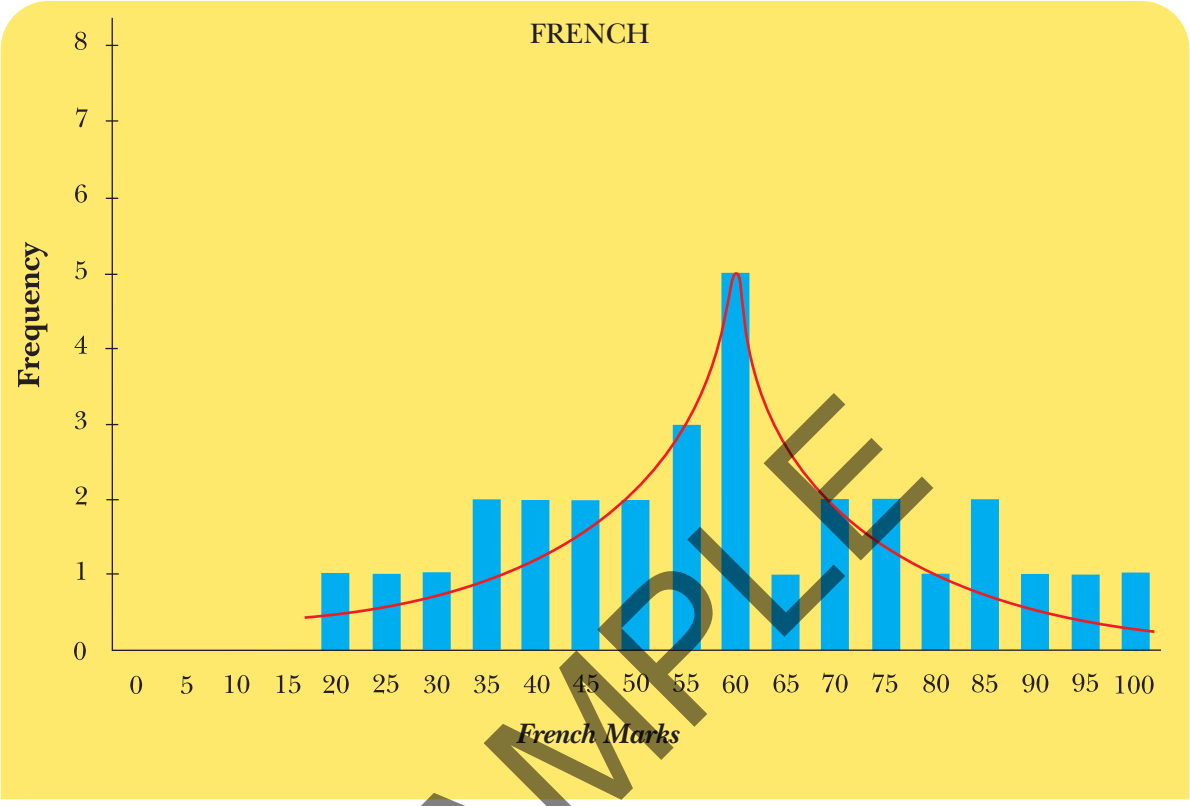
PROJECT 2.2 **Recall the Digits**

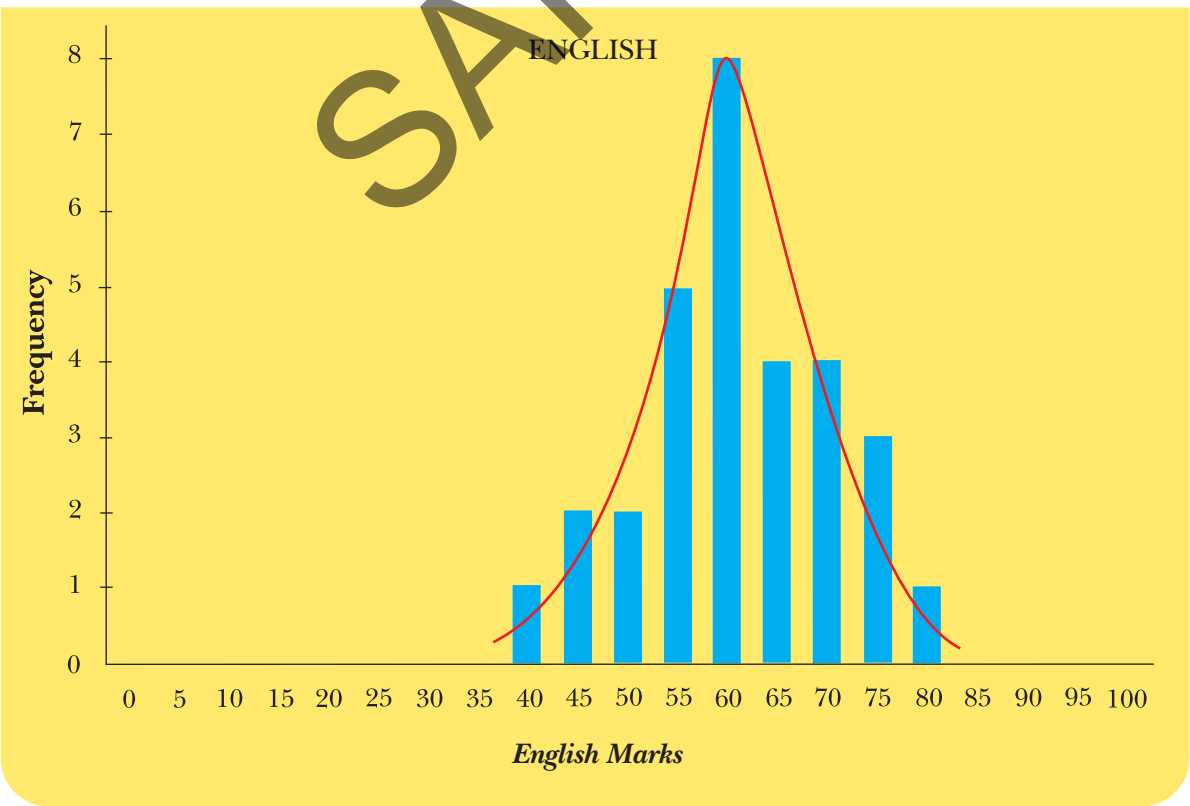
Do this as a class group. The teacher writes a 20 digit random number on the board. The class look at it for 30 seconds; the number is then covered. The pupils then write down (independently) as much of the random number as they can remember. Reveal the random number again and each pupil records how many of the digits they got correct before their first error. The exercise is repeated a number of times with a different 20 digit number each time. Record all the data on the board. Plot a Bar Chart of the data and find the average number of digits remembered. Discuss the shape of the Bar Chart.

Example 6: A class of 30 pupils get a short mid-term exam in English and one in French. The teacher grades the exam to the nearest 5%. The results are as follow:

French Marks										English Marks									
20	25	30	35	35	40	40	45	45	50	40	45	45	50	50	55	55	55	55	55
50	55	55	55	60	60	60	60	60	65	60	60	60	60	60	60	60	60	65	65
70	70	75	75	80	85	85	90	95	100	65	65	70	70	70	70	75	75	75	80

Compare the results in French and English by (i) finding the mean (an average) of each and (ii) by plotting a bar chart for each using the same scale.





Compare results:

(i) the average mark (the mean) in both classes is similar, one is 59 the other is 61. Therefore by looking **only** at the mean, you might conclude that classes have a similar ability in French and English.

(ii) Comparing the results using the Bar Charts, give us extra information. Yes, the charts confirm that the average mark is approximately 60% for both subjects. However, the charts clearly show that the **spread** of the marks is much greater in French than in English.

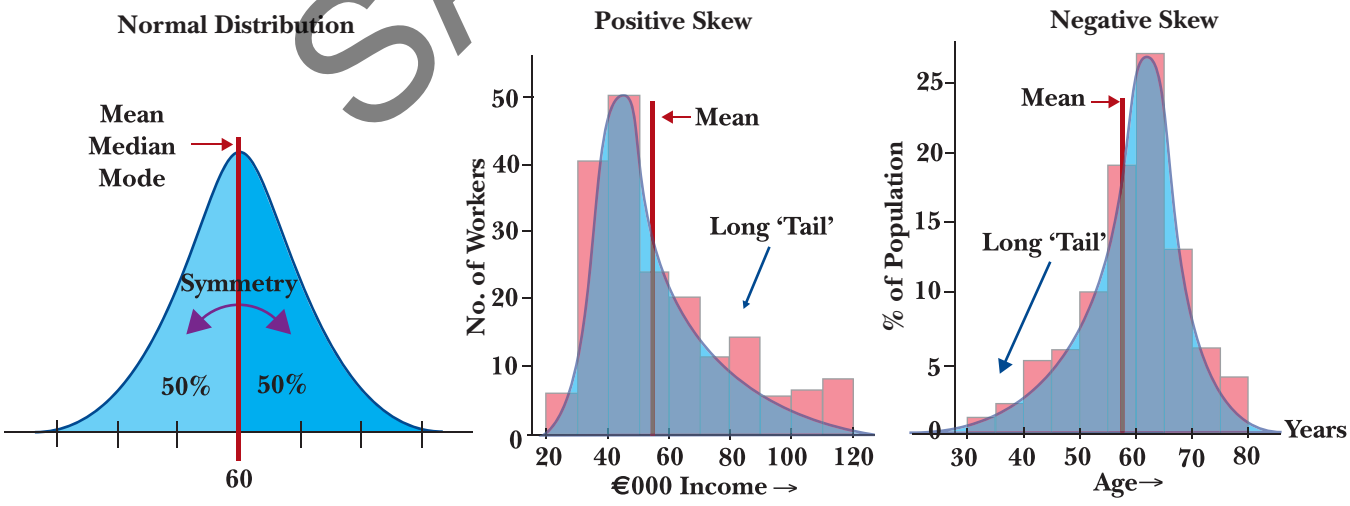
You can now conclude that the grades are significantly different for English and French even though the average is similar. Most pupils in the English class got 60% or within 10% of 60%. While most pupils in the French class got more than 10% above or below the 60% mark.

PROJECT 2.3 **Calculate Standard Deviation**

Using the data from Example 6, quantify the spread by calculating the Standard Deviation for the French and English marks

NOTE

Note on describing the shapes of charts:
A perfect normal distribution is symmetrical, the highest point will be the mean. Positive skew means that there is a long tail to the right. Negative skew means that there is a long tail to the left.

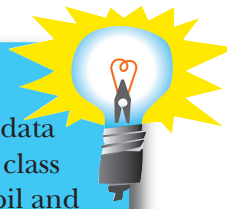


Example A: marks in the exam in Example 5 are normally distributed about the mean.

Example B: A plot of the income of the population of a small town will be skewed positively with a small number of people earning a lot.

Example C: the retirement age of populations in most developed countries will look like this, peaking at around 62 years.

NOTE



A full class period will be required to collect the data for the following project. Over the course of the class each pupil should be able to test every other pupil and get a full set of data for themselves. (50 pieces of data will give a good graph, so if you have 20 readings yourself, (Primary Data), take 30 readings from other students in the class, (Secondary Data).

PROJECT 2.4 **Reaction Time**

Use a 50 cm half metre stick or a 30 cm ruler.

Work in pairs to collect data for this project.

Ask another pupil to hold their hand out with their thumb and forefinger separated by 1 cm. Then place the 0 cm mark on the ruler between their thumb and forefinger, tell them you are going to drop the ruler sometime in the next 10 seconds and she has to catch it as quickly as possible. At a random time drop the ruler, record the distance it fell (to the nearest cm.) by reading the number of centimetres immediately above their fingers. Repeat this exercise for everyone in the class and record all the data on a bar chart, with "distance the ruler fell" recorded on the x-axis, and frequency on the y-axis. Use some secondary data from other pupils to ensure you have enough data.

What shape is the graph? Normal, Positive Skew or Negative Skew?
Can you explain the shape?

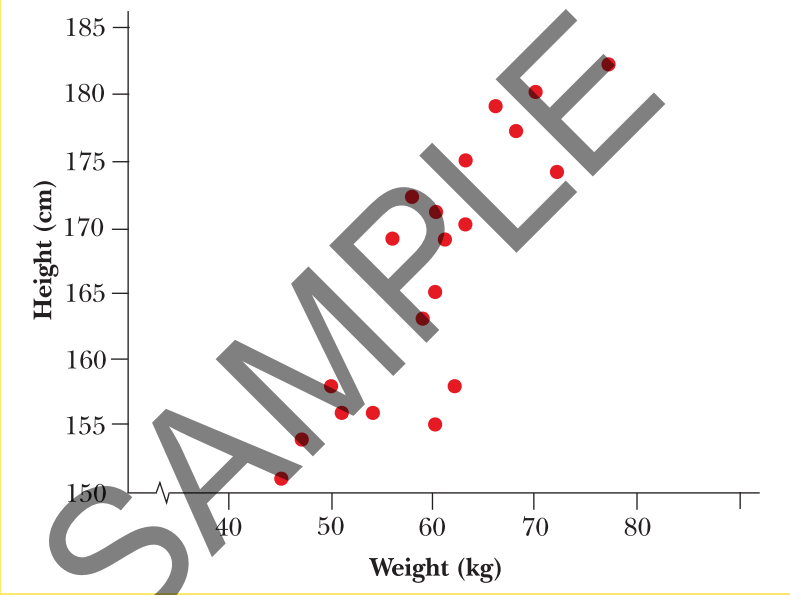
Scatter Graphs (Correlation)

Scatter Graphs show the relationship between two sets of data.

Example 7: 20 Transition Year boys were weighed and their heights measured. Plot the results on a Scatter Graph putting weight on the *x*-axis. Comment on the result:

Height cm	151	154	155	156	156	158	158	163	165	169	169	170	171	172	174	175	177	179	180	182
Weight kg	45	47	60	51	54	50	62	59	60	61	56	63	60	58	72	63	68	66	70	77

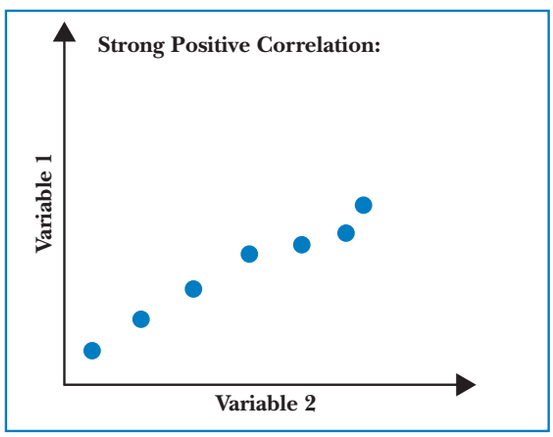
Scatter Graph Plot of Weight vs. Height (for Transition Year Boys).

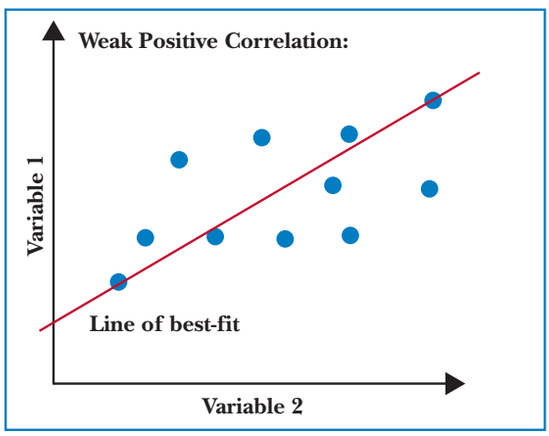


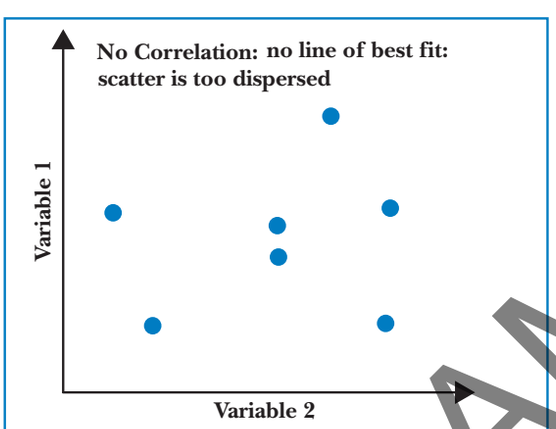
Comment: The mean weight can be calculated (like Example 2) to be 60·1 kg, and the mean height to be 166·7 cm. But when we draw a Scatter Graph, we are interested in the connection or relationship between the two sets of data. It is clear from the graph that there is a diagonal band of points across the graph, from the bottom left to the top right. This is called a **Positive Correlation**. Generally speaking, the taller boys are heavier, and the shorter boys are lighter. This is what we would have expected.

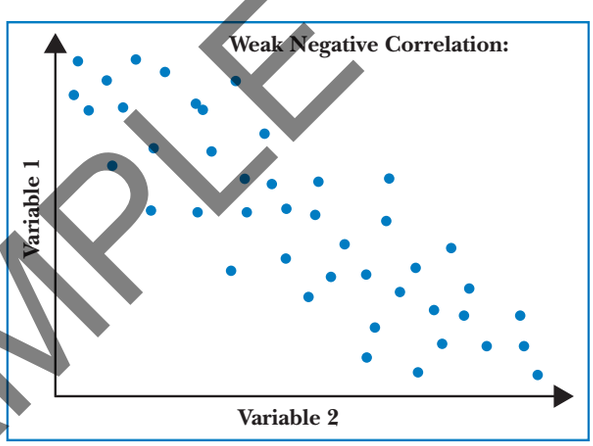
Correlation:

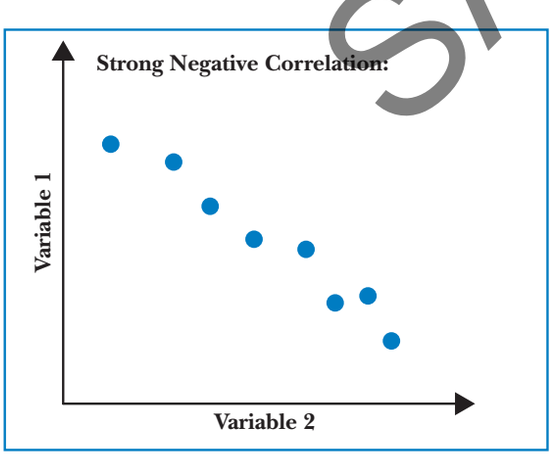
This is the relationship between the two sets of data. We will distinguish between five different relationships. Most sets of data will fit into one of these five categories.











The correlation between weight and height in Example 7 is a weak positive correlation.

PROJECT 2.5 **Height and Shoe Size**

Work in pairs with a measuring tape or two metre sticks to collect the data. Measure the height of everyone in the class and ask them for their shoe size. Record the pairs of data and then plot them on a Scatter Graph. Put shoe size on the x-axis and height on the y-axis. When you have completed the plot, comment on the correlation.

The data you collect in this project is called **Primary Data**, because you collected it yourself. If you run out of time, you may need to get some sets of data from other pupils. This is known as **Secondary Data** as you did not collect it yourself.

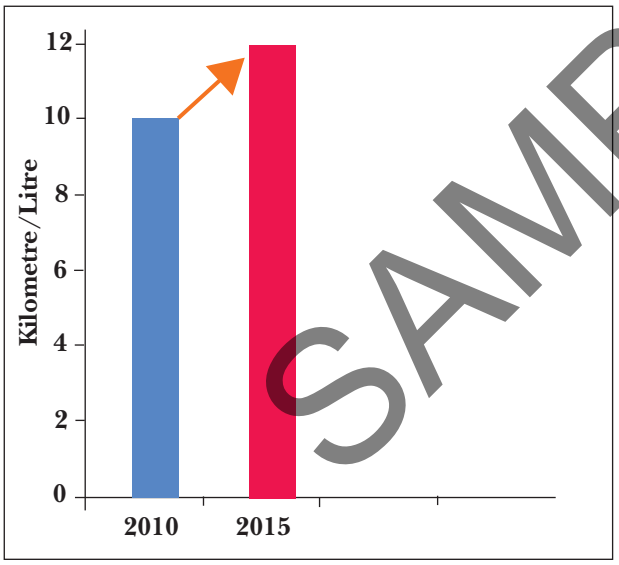
The Lie Factor (use of statistics to distort or exaggerate the truth)

$$\text{The Lie Factor} = \frac{\text{size of the effect shown in graph}}{\text{size of the effect in the data}}$$

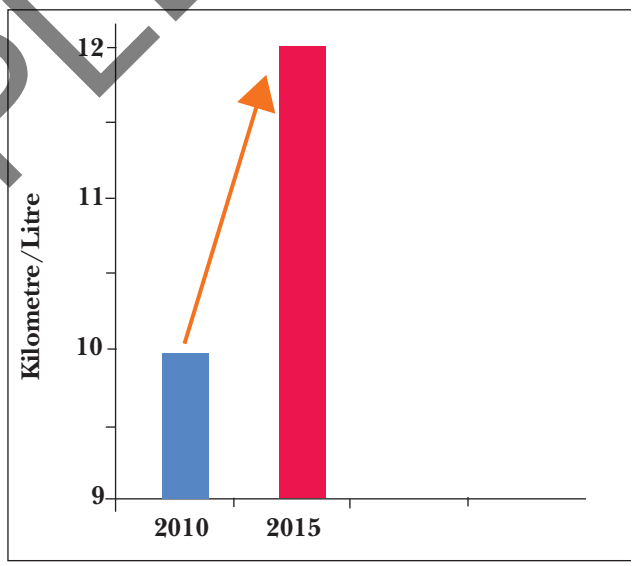
Statistics and charts have been used to present information falsely or to exaggerate the information since they were invented. The most common ways are to change the scale to exaggerate differences or to omit some data.

We have already noticed how the average can be distorted depending on whether we use the mean, mode or median.

Example 8: A car company notices that the fuel efficiency of one of its car models has improved from 10 kilometres per litre in 2010 to 12 kilometres per litre in 2015. Plot (i) a Fair Chart to advertise this improvement and (ii) a Distorted Chart to exaggerate the fuel efficiency improvement. (iii) Calculate the Lie Factor of the exaggeration and comment on it.



Car Fuel Efficiency Improvement 2010 to 2015



Car Fuel Efficiency Improvement 2010 to 2015

(i) The first chart gives a realistic representation of the improvement in fuel efficiency from 10 km/l to 12 km/l. It is a 20% increase in efficiency

(ii) The second chart distorts the improvement by changing the vertical scale. It appears as though the cars are now 200% more efficient

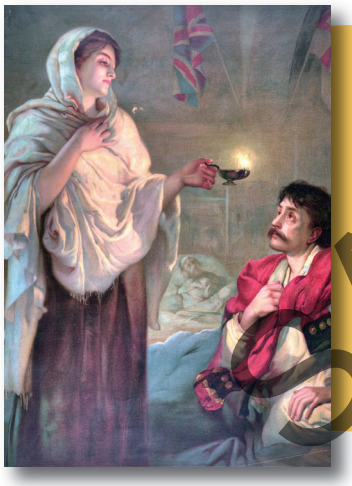
(iii) $\text{The Lie Factor} = \dfrac{\text{the \% change in the size of the chart}}{\text{the \% change in the size of the data}}$

$$\text{Chart 1: Lie Factor} = \frac{\%\text{ change in chart}}{\%\text{ change in data}} = \frac{\left(\frac{6\text{cm}-5\text{cm}}{5\text{cm}}\right)\times 100 = 20\%}{\left(\frac{12-10\text{km}/1}{10\text{km}/1}\right)\times 100 = 20\%} = 1 \text{ (no distortion)}$$

$$\text{Chart 2: Lie Factor} = \frac{\%\text{ change in chart}}{\%\text{ change in data}} = \frac{\left(\frac{6\text{cm}-2\text{cm}}{2\text{cm}}\right)\times 100 = 200\%}{\left(\frac{12-10\text{km}/1}{10\text{km}/1}\right)\times 100 = 20\%} = 10 \text{ (large distortion)}$$

The true improvement in efficiency is 20%; the second chart gives the impression that the improvement is 200%. The Lie Factor of 1 implies the first chart gives an accurate impression. The Lie Factor of 10 in chart 2 indicates a 10-fold distortion.

There is nothing inaccurate with Chart 2; it just exaggerates visually the improvement in efficiency.



Florence Nightingale:
English nurse and mathematician, (1820–1910)

Florence Nightingale is best remembered for her work as a nurse, particularly during the Crimean War, and for setting up a school of nursing. However, she also had a gift for mathematics and was a pioneer in the visual presentation of information and statistical graphs. She was an early user of pie charts and circular histograms.