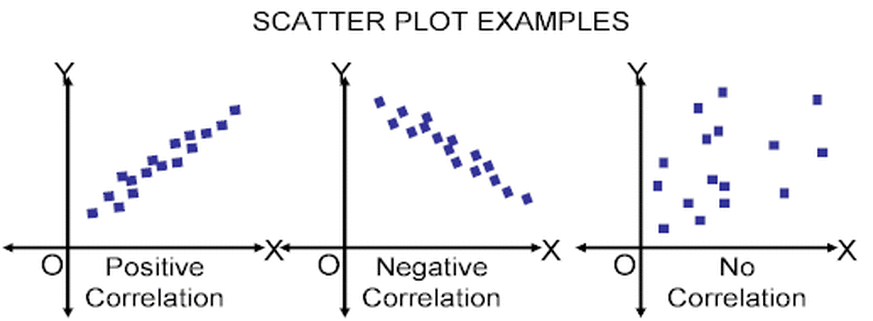
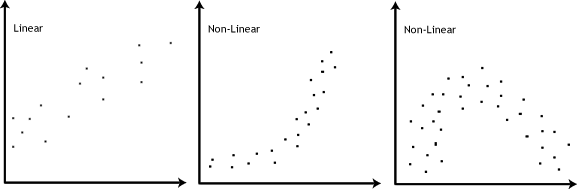
**Correlation – Measuring the relationship between bivariate data**

At GCSE we became familiar with the use of the term correlation when describing points plotted on a scatter-graph. Such points are based on the values of two variables, such as the monetary value compared to the mileage of a second hand car, or the marks awarded in Paper 1 and Paper 2 of a Maths exam. We would then describe these relationships as having ‘positive’, ‘negative’, or ‘no’ correlation.



Where there was evidence of correlation we learned how to draw and use a line of best fit on the plot.

In statistics there are different ways to **quantify** the strength of correlation exhibited between a pair of variables. These are known as correlation coefficients. When it comes to linear correlation (other forms of correlation exist, see below), the two measures used are **Pearson’s Product Moment Correlation Coefficient (PMCC)** and **Spearman’s Coefficient of Rank Correlation**.



Pearson’s coefficient is calculated on raw bivariate data and measures how close the data lie to a straight line. A value of +1 would indicate that all the data lie on a straight line with a positive gradient (perfect positive correlation) and a correlation coefficient of -1 would indicate perfect negative correlation, with all the data laying on a straight line with a negative gradient.

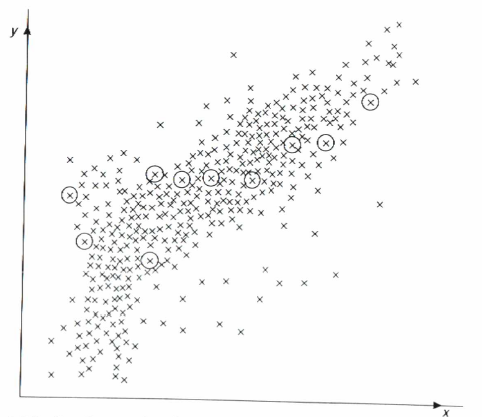
Spearman’s coefficient is used when the data has been ranked in some way. For example if two judges in a bake off rank cakes baked by 10 different people. If both judges agreed exactly on the ranking of the cakes this would produce a coefficient of +1, a value of -1 would occur if the two judges had the ranking in exactly reverse order, 1st place for Judge 1 = 10th place for Judge 2, etc.

In Unit 4, we are not required to actually calculate the coefficients, just interpret them.

**The meaning of a correlation coefficient**

As described on the previous page, if the value of a correlation coefficient for a sample (denoted by the letter r), is close to +1 or -1, we can be satisfied that there is linear correlation. The question we are looking to address in this work is what happens in a case such as r = 0.6?

When calculating a value of r, we are actually using a sample of bivariate data from a much larger parent population. For example, in the scatter diagram below we may have sampled the ten circled points to calculate r.



There will be a level of correlation within the parent population and this is denoted by the Greek letter ρ (pronounced rho).

The calculated value of r, based on the sample can be used as an estimate for ρ. It can also be used to carry out an hypothesis test on the value of ρ, the parent population correlation coefficient. Used in this way it is described as a *test statistic*.

The simplest hypothesis test which you can carry out is that there is no correlation within the parent population. This gives rise to the null hypothesis:

H0: ρ = 0 ‘There is no correlation between the two variables’

There are three possible alternative hypotheses, according to the sense of the situation being investigated. These are:

H1: ρ ≠ 0 ‘There is correlation between the variables’ (2-tail test)

or H1: ρ > 0 ‘There is positive correlation between the variables’ (1-tail test)

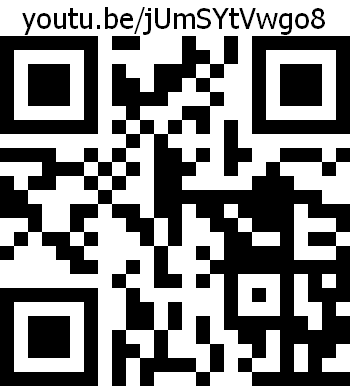
or H1: ρ < 0 ‘There is negative correlation between the variables’ (1-tail test)

The test is carried out by comparing the calculated value of r with the appropriate entry in a table of critical values (actually easier to use than our calculator for this!). This will depend on the size of the sample, the significance level at which we are testing and whether the test is a 1-tail or a 2-tail test.

Eg7 A language teacher wished to test whether there is any correlation between students’ ability in their own and a foreign language. Accordingly she collects the marks of 8 students, all native English speakers, in their end of year examinations in English and French.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Candidate | A | B | C | D | E | F | G | H |
| English | 65 | 34 | 48 | 72 | 58 | 63 | 26 | 80 |
| French | 74 | 49 | 45 | 80 | 63 | 72 | 12 | 75 |

1. Use your calculator to determine the PMCC
2. State the null and alternative hypotheses
3. Using the value of r as the test statistic, carry out the test at the 5% significance level.



Eg8 “You can’t win without scoring goals”. So says the coach of a netball team. Jamila, who believes in solid defensive play, disagrees and sets out to prove that there is no correlation between scoring goals and winning matches. She collects the following data for the goals scored and the points scored by 12 teams in a netball league.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Goals Scored | 41 | 50 | 54 | 47 | 47 | 49 | 52 | 61 | 50 | 29 | 47 | 35 |
| Points Scored | 21 | 20 | 19 | 18 | 16 | 14 | 12 | 11 | 11 | 7 | 5 | 2 |

1. Use your calculator to determine the PMCC
2. State suitable null and alternative hypotheses, indicating whose position each represents
3. Carry out the hypothesis test at the 5% significance level and comment on the result



Eg9 Charlotte is a campaigner for temperance, believing that drinking alcohol is an evil habit. Michael, a representative of a wine company, presents her with figures which he claims show that wine drinking is good for marriages.

|  |  |  |
| --- | --- | --- |
| Country | Wine Consumption  (kg/person/year) | Divorce Rate  (/1000 inhabitants) |
| Belgium | **20** | **2.0** |
| Denmark | **20** | **2.7** |
| Germany | **26** | **2.2** |
| Greece | **33** | **0.6** |
| Italy | **63** | **0.4** |
| Portugal | **54** | **0.9** |
| Spain | **41** | **0.6** |
| UK | **13** | **2.9** |

1. Write Michael’s claim in the form of an hypothesis test and carry it out at the 5% significance level.
2. Charlotte claims that Michael is ‘indulging in pseudo-statistics’. What arguments could she use to support this point of view?

**Interpreting Correlation**

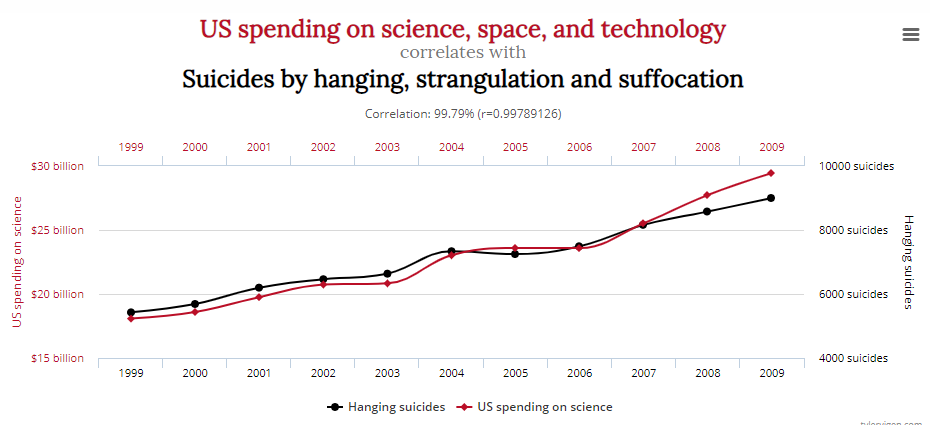
**Correlation does not imply causation**

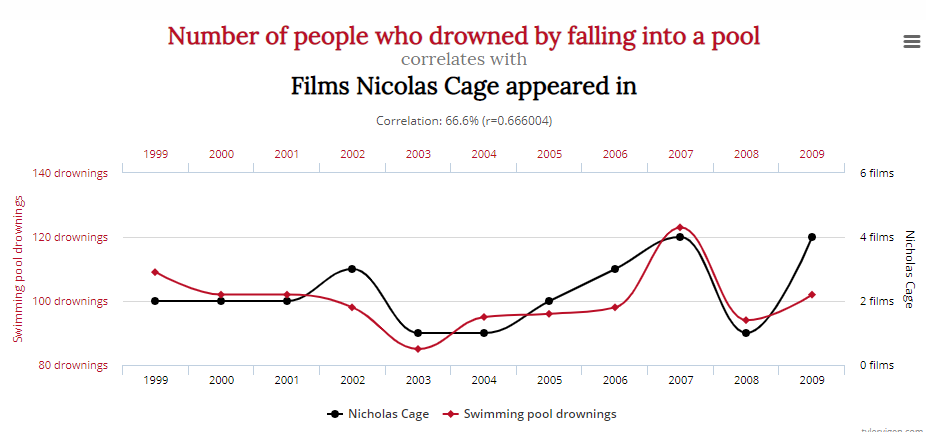
The above example demonstrates a situation where a high level of correlation does not necessarily mean that there is a direct connection between the two variables.

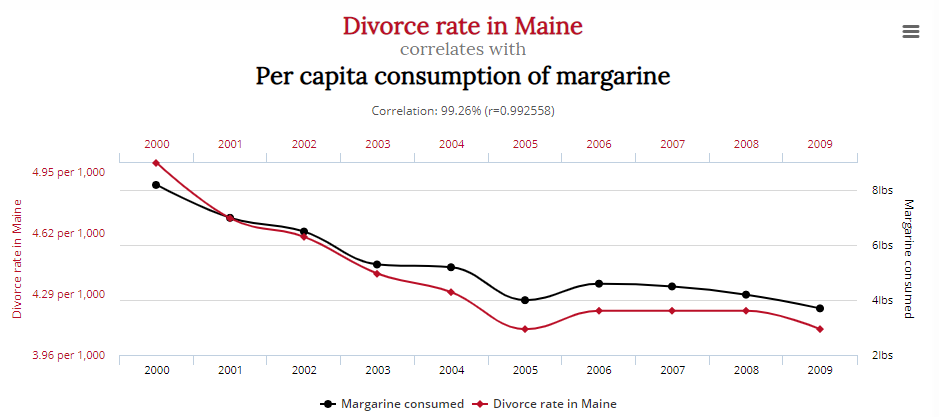
Although there may be a high level of correlation between variables A and B it does not mean that A → B or that

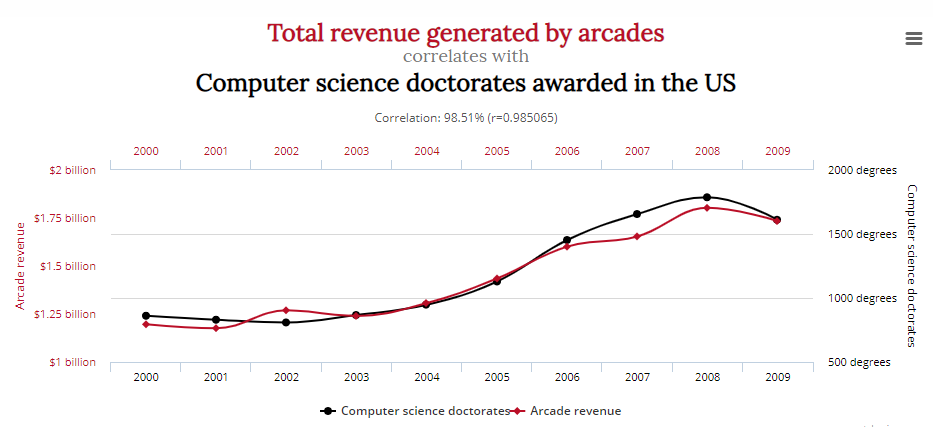
A ← B. It may well be that a third variable C causes both A and B or it may be a more complicated set of relationships.

Here are some examples taken from ‘Spurious Correlations’, [www.tylervigen.com](http://www.tylervigen.com)



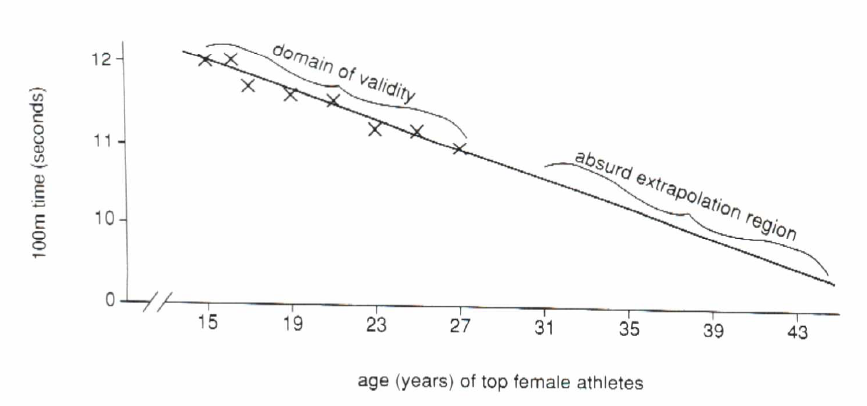






**Extrapolation**

A linear relationship established over a particular domain should not be assumed to hold outside of this range. For instance, there is a strong correlation between the age in years and 100 metre times of female athletes between the ages of 8 and 20 years. To extend the connection, in the plot below, would suggest that veteran athletes are quicker than athletes in their prime, and if they live long enough can even run 100m in no time at all!



**Using a p-value**

Sometimes you may be given a p-value instead of a value for the correlation coefficient. In this context, the p-value is the probability that, if there is no correlation, a random sample gives the given value for the correlation coefficient – in other words, the probability that this correlation coefficient could have been obtained by chance.

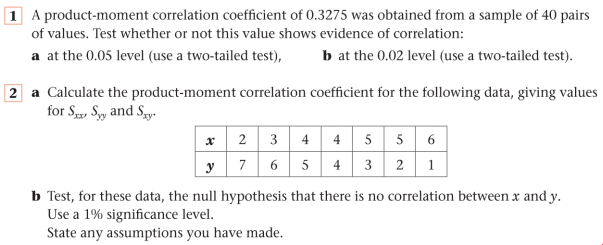
As with the hypothesis testing for sample means, we compare the p-value with the significance level. If the p-value is less than the significance level, this means that it is very unlikely that this value for the correlation coefficient could have risen by chance, so the null hypothesis is rejected.

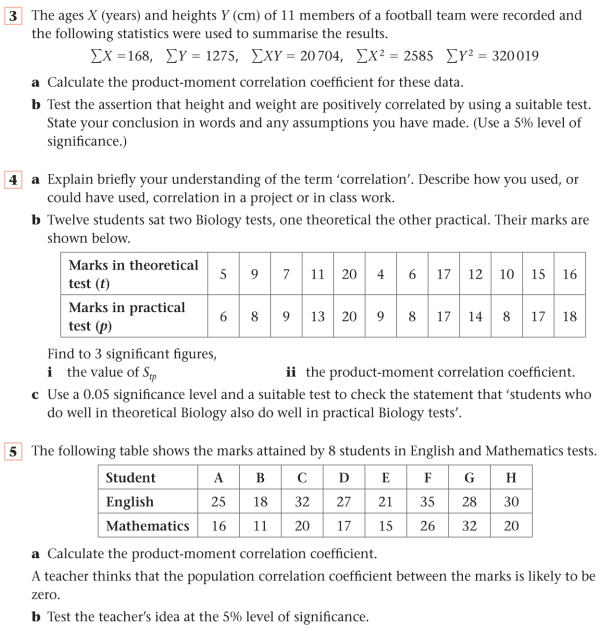
Eg10 A random sample of 50 pairs of bivariate data (x, y) produce a product moment correlation coefficient of -0.3. This gives a 1-tail p-value of 0.0171.

What is the 2-tail p-value?

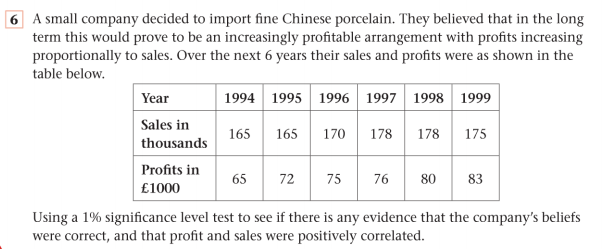
 Stating your hypotheses clearly, test at the 5% significance level whether there is negative correlation between x and y within the parent population from which the values are drawn.

**Exercise 3.2 (Edexcel S3 Ex 5B)**





**r = 0.677**



**Answers**

