Correlation coefficient and *p*-values: what they are and why you need to be very wary of them

(From Chapter 1 of “Risk Assessment and Decision Analysis with Bayesian Networks”, by Norman Fenton and Martin Neil, CRC Press, 2012)

The correlation coefficient is a number between –1 and 1 that determines whether two paired sets of data (such as those for *height* and *intelligence* of a group of people) are related. The closer to 1 the more ‘confident’ we are of a positive linear correlation and the closer to –1 the more confident we are of a negative linear correlation (which happens when, for example one set of numbers tends to decrease when the other set increases as you might expect if you plotted a person’s age against the number of toys they possess).

When the correlation coefficient is close to zero there is no evidence of any relationship.

Confidence in a relationship is formally determined not just by the correlation coefficient but also by the number of pairs in your data. If there are very few pairs then the coefficient needs to be very close to 1 or –1 for it to be deemed ‘statistically significant’, but if there are many pairs then a coefficient closer to 0 can still be considered ‘highly significant’.

The standard method that statisticians use to measure the ‘significance’ of their empirical analyses is the ***p*-value**. Suppose we are trying to determine if the relationship between height and intelligence of people is significant; then we start with the ‘null hypothesis’ which, in this case is the statement ‘height and intelligence of people are unrelated’. The *p*-value is a number between 0 and 1 representing the probability that this data would have arisen if the null hypothesis were true. In medical trials the null hypothesis is typically of the form that ‘the use of drug X to treat disease Y is no better than not using any drug’.

The calculation of the *p*-value is based on a number of assumptions that are beyond the scope of this discussion, but people who need *p*-values can simply look them up in standard statistical tables (they are also computed automatically in Excel when you run Excel’s regression tool). The tables (or Excel) will tell you, for example, that if there are 100 pairs of data whose correlation coefficient is 0.254, then the *p*-value is 0.01. This means that there is a 1 in 100 chance that we would have seen these observations if the variables were unrelated.

A low *p*-value (such as 0.01) is taken as evidence that the null hypothesis can be ‘rejected’. Statisticians say that a *p*-value of 0.01 is ‘highly significant’ or say that ‘the data is significant at the 0.01 level’

A competent researcher investigating a hypothesized relationship will set a *p*-value in advance of the empirical study. Typically, values of either 0.01 or 0.05 are used. If the data from the study results in a *p*-value of less than that specified in advance, the researcher will claim that their study is significant and it enables them to reject the null hypothesis and conclude that a relationship really exists.

In their book *The Cult of Statistical Significance* Ziliak and McCloskey expose a number of serious problems in the way *p*-values have been used across many disciplines. Above all, their main arguments can be summarised as:

- Statistical significance (i.e. the *p*-value) is arbitrarily set and generally has no bearing on what we are really interested in, namely impact or magnitude of the effect of one or more variables on another.
- By focusing on a null hypothesis all that we are ever considering are existential questions, the answers to which are normally not interesting. So, for example, we might produce a very low *p*-value and conclude that road deaths and temperature are not unrelated. But the *p*-value tells us nothing about what we are really interested in, namely the nature and size of the relationship.
- Statisticians often wrongly assume that the *p*-value (which remember is chance of observing the data if the null hypothesis is true) is equivalent to the chance that the null hypothesis is true given the data. So, for example, if they see a low *p*-value of say 0.01 they might conclude that there is a 1 in a 100 chance of no relationship (which is the same as a 99% chance that there is a relationship). This is, in fact, demonstrably false (we will show this in Chapter 5); it is an example of one of the most pernicious and fundamental fallacies of probability theory that permeates many walks of life (called the *fallacy of the transposed conditional*).
- In those many studies (notably medical trials) where the null hypothesis is one of ‘no change’ for some treatment or drug, the hypothesis comes down to determining whether the arithmetic mean of a set of data (from those individuals taking the treatment/drug) is equal to zero (supposedly representing status quo). In such cases, we have the paradox that, as we substantially increase the sample size, we will inevitably find that the mean of the sample, although approximately close to and converging to zero, will be significantly different from zero, even when the treatment genuinely has no effect (this is covered in Chapter 10 on hypothesis testing and is known as Meehl’s conjecture).
- The choice of what constitutes a valid *p*-value is arbitrary. Is 0.04 radically different from 0.05? A treatment or putative improvement that yields a *p*-value that just misses the 0.05 target may be completely rejected and one that meets the target may be adopted.

Ziliak and McCloskey cite hundreds of examples of studies (all published in highly respected scientific journals) that contain flawed analyses or conclusions arising from the above misunderstandings. They give the following powerful hypothetical example of a fundamental weakness of using *p*-values:

> Suppose we are interested in new drugs for reducing weight in humans. Two candidate drugs (called *Precision* and *Oomph* respectively) are considered. Neither has shown any side-effects and their cost is the same. For each drug we conduct a study to test the null hypothesis ‘taking the drug leads to no weight loss’. The results are:
>
> - For drug Precision the mean weight loss is 5 lbs and every one of the 100 subjects in the study loses between 4.5 lb and 5.5 lb.

- For drug *Oomph* the mean weight loss is 20 lbs and every one of the 100 subjects in the study loses between 10 lb and 30 lb.

Since the objective of weight loss drugs is to lose as much weight as possible, any rational, intuitive review of these results would lead us to recommend drug *Oomph* over *Precision*. Yet the *p*-value test provides the opposite recommendation. For drug *Precision* the *p*-value is much lower (i.e. more significant) than the *p*-value for drug *Oomph*. This is because *p*-values inevitably ‘reward’ low variance more than magnitude of impact.