Data dredging

Significance is abolished by data dredging

Data dredging (also known as p-hacking) is a term that refers to testing a large number of hypotheses concerning a single data set in order to find a correlation (any correlation). It is an improper and potentially unethical procedure when applied after the fact in order to find some kind of correlation.

Conventional tests of statistical significance are based on the probability that a particular result would arise if chance alone were at work. In any significance test one necessarily accepts some risk of mistaken rejection of the null hypothesis or alternatively mistaken confirmation of the null hypothesis.

If you look for a correlation, you may well find one (even by random chance). But, it is not significant. To pretend that it is significant can be unethical.

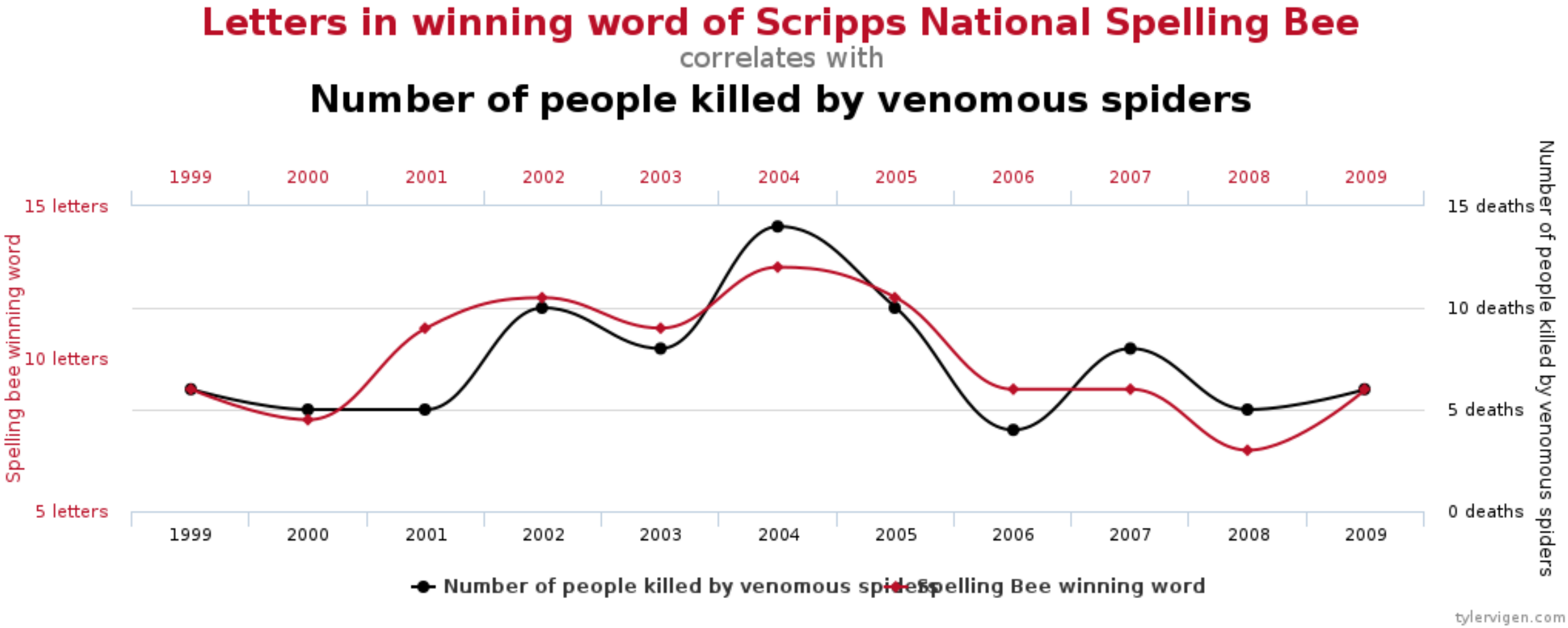
Looking for a correlation may give one by chance

If a large number of tests is conducted, then some of them will produce false results that appear to be significant. For example, by chance alone, 5% of randomly chosen hypotheses turn out to be significant at the 5% level ($p < 0.05$) and 1% turn out to be significant at the 1% significance level ($p < 0.01$), etc.

If enough hypotheses are tested, it is virtually certain that some will appear to be statistically significant. These are misleading, since almost every data set with any degree of randomness is likely to contain some spurious correlations.

This type of approach has been used in epidemiological studies to give the appearance that a particular therapy has efficacy, when in reality it is no better than placebo.

A correlation that lacks significance



Tyler Viglen – Spurious correlations