

Correlation is not causation

Reference: Nathan Green's S word

<https://www.theguardian.com/science/series/nathan-green-statistics>

Correlation is not causation. Correlation is not causation. Correlation is not causation ... " At times during my statistics studies I felt like Jack Nicholson in the film *The Shining*, in which we witness his descent into madness as he types the same sentence over and over again, "All work and no play makes Jack a dull boy. All work and no play makes Jack a dull boy ... "

"Correlation is not causation" is a statistics mantra. It is drilled, military school-style, into every budding statistician. But what does it actually mean? Well, correlation is a measure of how closely related two things are. Think of it as a number describing the relative change in one thing when there is a change in the other, with 1 being a strong positive relationship between two sets of numbers, -1 being a strong negative relationship and 0 being no relationship whatsoever.

"Correlation is not causation" means that just because two things correlate does not necessarily mean that one causes the other. As a seasonal example, just because people in the UK tend to spend more in the shops when it's cold and less when it's hot doesn't mean cold weather causes frenzied high-street spending. A more plausible explanation would be that cold weather tends to coincide with Christmas and the new year sales.

Despite embodying an important truth, the phrase has not caught on in the wider world. It's easy to see why. Our preconceptions and suspicions about the way things work tempt us to make the leap from correlation to causation without any hard evidence.

Correlations between two things can be caused by a third factor that affects both of them. This sneaky, hidden third wheel is called a confounder.

Arguably the most well known and important example of a correlation being clear but causation being in doubt concerned smoking and lung cancer in the 1950s. There had been a sixfold increase in the rate of lung cancer in the preceding two decades. Nobody disputed that there was a correlation between lung cancer and smoking, but to prove that one caused the other would be no mean feat.

There might be a confounder that was responsible for the correlation between smoking and lung cancer. The increased rate could have been the result of better diagnosis, more industrial pollution or more cars on the roads belching noxious fumes. Perhaps people who were more genetically predisposed to want to smoke were also more susceptible to getting cancer?

It took a study involving more than 40,000 doctors in the UK to show conclusively that smoking really does cause cancer.