**If correlation doesn’t imply causation, then what does?**

by Michael Nielsen on January 23, 2012

<http://www.michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/>

It is a commonplace of scientific discussion that correlation does not imply causation. Business Week recently ran an [**spoof article**](http://www.businessweek.com/magazine/correlation-or-causation-12012011-gfx.html) pointing out some amusing examples of the dangers of inferring causation from correlation. For example, the article points out that Facebook’s growth has been strongly correlated with the yield on Greek government bonds: ([**credit**](http://www.businessweek.com/magazine/correlation-or-causation-12012011-gfx.html))



Despite this strong correlation, it would not be wise to conclude that the success of Facebook has somehow *caused* the current (2009-2012) Greek debt crisis, nor that the Greek debt crisis has caused the adoption of Facebook!

Of course, while it’s all very well to piously state that correlation doesn’t imply causation, it does leave us with a conundrum: under what conditions, exactly, can we use experimental data to deduce a causal relationship between two or more variables?

The standard scientific answer to this question is that (with some caveats) we can infer causality from a well designed [**randomized controlled experiment**](http://en.wikipedia.org/wiki/Randomized_controlled_trial). Unfortunately, while this answer is satisfying in principle and sometimes useful in practice, it’s often impractical or impossible to do a randomized controlled experiment. And so we’re left with the question of whether there are other procedures we can use to infer causality from experimental data. And, given that we can find more general procedures for inferring causal relationships, what does causality mean, anyway, for how we reason about a system?

It might seem that the answers to such fundamental questions would have been settled long ago. In fact, they turn out to be surprisingly subtle questions. Over the past few decades, a group of scientists have developed a theory of *causal inference*intended to address these and other related questions. This theory can be thought of as an algebra or language for reasoning about cause and effect. Many elements of the theory have been laid out in a [**famous book**](http://www.amazon.com/Causality-Reasoning-Inference-Judea-Pearl/dp/0521773628) by one of the main contributors to the theory, [**Judea Pearl**](http://en.wikipedia.org/wiki/Judea_Pearl). Although the theory of causal inference is not yet fully formed, and is still undergoing development, what has already been accomplished is interesting and worth understanding.

In this post I will describe one small but important part of the theory of causal inference, a *causal calculus* developed by Pearl. This causal calculus is a set of three simple but powerful algebraic rules which can be used to make inferences about causal relationships. In particular, I’ll explain how the causal calculus can sometimes (but not always!) be used to infer causation from a set of data, even when a randomized controlled experiment is not possible. Also in the post, I’ll describe some of the limits of the causal calculus, and some of my own speculations and questions.

The post is a little technically detailed at points. However, the first three sections of the post are non-technical, and I hope will be of broad interest. Throughout the post I’ve included occasional “Problems for the author”, where I describe problems I’d like to solve, or things I’d like to understand better. Feel free to ignore these if you find them distracting, but I hope they’ll give you some sense of what I find interesting about the subject. Incidentally, I’m sure many of these problems have already been solved by others; I’m not claiming that these are all open research problems, although perhaps some are. They’re simply things I’d like to understand better. Also in the post I’ve included some exercises for the reader, and some slightly harder problems for the reader. You may find it informative to work through these exercises and problems.

Before diving in, one final caveat: I am not an expert on causal inference, nor on statistics. The reason I wrote this post was to help me internalize the ideas of the causal calculus. Occasionally, one finds a presentation of a technical subject which is beautifully clear and illuminating, a presentation where the author has seen right through the subject, and is able to convey that crystalized understanding to others. That’s a great aspirational goal, but I don’t yet have that understanding of causal inference, and these notes don’t meet that standard. Nonetheless, I hope others will find my notes useful, and that experts will speak up to correct any errors or misapprehensions on my part.

**Simpson’s paradox**

Let me start by explaining two example problems to illustrate some of the difficulties we run into when making inferences about causality. The first is known as [**Simpson’s paradox**](http://en.wikipedia.org/wiki/Simpson%27s_paradox). To explain Simpson’s paradox I’ll use a concrete example based on the passage of the Civil Rights Act in the United States in 1964.

In the US House of Representatives, 61 percent of Democrats voted for the Civil Rights Act, while a much higher percentage, 80 percent, of Republicans voted for the Act. You might think that we could conclude from this that being Republican, rather than Democrat, was an important factor in causing someone to vote for the Civil Rights Act. However, the picture changes if we include an additional factor in the analysis, namely, whether a legislator came from a Northern or Southern state. If we include that extra factor, the situation *completely* reverses, in both the North *and* the South. Here’s how it breaks down:

**North:** Democrat (94 percent), Republican (85 percent)

**South:** Democrat (7 percent), Republican (0 percent)

Yes, you read that right: in *both* the North and the South, a larger fraction of Democrats than Republicans voted for the Act, despite the fact that *overall* a larger fraction of Republicans than Democrats voted for the Act.

You might wonder how this can possibly be true. I’ll quickly state the raw voting numbers, so you can check that the arithmetic works out, and then I’ll explain why it’s true. You can skip the numbers if you trust my arithmetic.

**North:** Democrat (145/154, 94 percent), Republican (138/162, 85 percent)

**South:** Democrat (7/94, 7 percent), Republican (0/10, 0 percent)

**Overall:** Democrat (152/248, 61 percent), Republican (138/172, 80 percent)

One way of understanding what’s going on is to note that a far greater proportion of Democrat (as opposed to Republican) legislators were from the South. In fact, at the time the House had 94 Democrats, and only 10 Republicans. Because of this enormous difference, the very low fraction (7 percent) of southern Democrats voting for the Act dragged down the Democrats’ overall percentage much more than did the even lower fraction (0 percent) of southern Republicans who voted for the Act.

(The numbers above are for the House of Congress. The numbers were different in the Senate, but the same overall phenomenon occurred. I’ve taken the numbers from[**Wikipedia’s article about Simpson’s paradox**](http://en.wikipedia.org/wiki/Simpson%27s_paradox#Civil_Rights_Act_of_1964), and there are more details there.)

If we take a naive causal point of view, this result looks like a paradox. As I said above, the overall voting pattern seems to suggest that being Republican, rather than Democrat, was an important causal factor in voting for the Civil Rights Act. Yet if we look at the individual statistics in *both* the North and the South, then we’d come to the exact *opposite* conclusion. To state the same result more abstractly, Simpson’s paradox is the fact that the correlation between two variables can actually be *reversed*when additional factors are considered. So two variables which appear correlated can become anticorrelated when another factor is taken into account.

You might wonder if results like those we saw in voting on the Civil Rights Act are simply an unusual fluke. But, in fact, this is not that uncommon. [**Wikipedia’s page on Simpson’s paradox**](http://en.wikipedia.org/wiki/Simpson%27s_paradox) lists many important and similar real-world examples ranging from understanding whether there is gender-bias in university admissions to which treatment works best for kidney stones. In each case, understanding the causal relationships turns out to be much more complex than one might at first think.

I’ll now go through a second example of Simpson’s paradox, the kidney stone treatment example just mentioned, because it helps drive home just how bad our intuitions about statistics and causality are.

Imagine you suffer from kidney stones, and your Doctor offers you two choices: treatment A or treatment B. Your Doctor tells you that the two treatments have been tested in a trial, and treatment A was effective for a higher percentage of patients than treatment B. If you’re like most people, at this point you’d say “Well, okay, I’ll go with treatment A”.

Here’s the gotcha. Keep in mind that this *really happened*. Suppose you divide patients in the trial up into those with large kidney stones, and those with small kidney stones. Then even though treatment A was effective for a higher overall percentage of patients than treatment B, treatment B was effective for a higher percentage of patients in *both groups*, i.e., for both large and small kidney stones. So your Doctor could just as honestly have said “Well, you have large [or small] kidney stones, and treatment B worked for a higher percentage of patients with large [or small] kidney stones than treatment A”. If your Doctor had made either one of these statements, then if you’re like most people you’d have decided to go with treatment B, i.e., the exact opposite treatment.

The kidney stone example relies, of course, on the same kind of arithmetic as in the Civil Rights Act voting, and it’s worth stopping to figure out for yourself how the claims I made above could possibly be true. If you’re having trouble, you can click through to the [**Wikipedia page**](http://en.wikipedia.org/wiki/Simpson%27s_paradox#Kidney_stone_treatment), which has all the details of the numbers.

Now, I’ll confess that before learning about Simpson’s paradox, I would have unhesitatingly done just as I suggested a naive person would. Indeed, even though I’ve now spent quite a bit of time pondering Simpson’s paradox, I’m not entirely sure I wouldn’t still sometimes make the same kind of mistake. I find it more than a little mind-bending that my heuristics about how to behave on the basis of statistical evidence are obviously not just a little wrong, but utterly, horribly wrong.

Perhaps I’m alone in having terrible intuition about how to interpret statistics. But frankly I wouldn’t be surprised if most people share my confusion. I often wonder how many people with real decision-making power – politicians, judges, and so on – are making decisions based on statistical studies, and yet they don’t understand even basic things like Simpson’s paradox. Or, to put it another way, they have not the first clue about statistics. Partial evidence may be worse than no evidence if it leads to an illusion of knowledge, and so to overconfidence and certainty where none is justified. It’s better to know that you don’t know.

**Correlation, causation, smoking, and lung cancer**

As a second example of the difficulties in establishing causality, consider the relationship between cigarette smoking and lung cancer. In 1964 the United States’ Surgeon General issued a [**report**](http://profiles.nlm.nih.gov/ps/access/NNBBMQ.pdf) claiming that cigarette smoking causes lung cancer. Unfortunately, according to Pearl the evidence in the report was based primarily on correlations between cigarette smoking and lung cancer. As a result the report came under attack not just by tobacco companies, but also by some of the world’s most prominent statisticians, including the great [**Ronald Fisher**](http://en.wikipedia.org/wiki/Ronald_Fisher). They claimed that there could be a hidden factor – maybe some kind of genetic factor – which caused both lung cancer *and* people to want to smoke (i.e., nicotine craving). If that was true, then while smoking and lung cancer would be correlated, the decision to smoke or not smoke would have no impact on whether you got lung cancer.

Now, you might scoff at this notion. But derision isn’t a principled argument. And, as the example of Simpson’s paradox showed, determining causality on the basis of correlations is tricky, at best, and can potentially lead to contradictory conclusions. It’d be much better to have a principled way of using data to conclude that the relationship between smoking and lung cancer is not just a correlation, but rather that there truly is a causal relationship.

One way of demonstrating this kind of causal connection is to do a randomized, controlled experiment. We suppose there is some experimenter who has the power to*intervene* with a person, literally forcing them to either smoke (or not) according to the whim of the experimenter. The experimenter takes a large group of people, and randomly divides them into two halves. One half are forced to smoke, while the other half are forced not to smoke. By doing this the experimenter can *break* the relationship between smoking and any hidden factor causing both smoking and lung cancer. By comparing the cancer rates in the group who were forced to smoke to those who were forced not to smoke, it would then be possible determine whether or not there is truly a causal connection between smoking and lung cancer.

This kind of randomized, controlled experiment is highly desirable when it can be done, but experimenters often don’t have this power. In the case of smoking, this kind of experiment would probably be illegal today, and, I suspect, even decades into the past. And even when it’s legal, in many cases it would be impractical, as in the case of the Civil Rights Act, and for many other important political, legal, medical, and econonomic questions.

**Causal models**

To help address problems like the two example problems just discussed, Pearl introduced a [**causal calculus**](http://ftp.cs.ucla.edu/pub/stat_ser/R212.pdf). In the remainder of this post, I will explain the rules of the causal calculus, and use them to analyse the smoking-cancer connection. We’ll see that even without doing a randomized controlled experiment it’s possible (with the aid of some reasonable assumptions) to *infer* what the outcome of a randomized controlled experiment would have been, using only relatively easily accessible experimental data, data that doesn’t require experimental intervention to force people to smoke or not, but which can be obtained from purely observational studies.

To state the rules of the causal calculus, we’ll need several background ideas. I’ll explain those ideas over the next three sections of this post. The ideas are *causal models* (covered in this section), *causal conditional probabilities*, and *d-separation*, respectively. It’s a lot to swallow, but the ideas are powerful, and worth taking the time to understand. With these notions under our belts, we’ll able to understand the rules of the causal calculus

To understand causal models, consider the following graph of possible causal relationships between smoking, lung cancer, and some unknown hidden factor (say, a hidden genetic factor):



This is a quite general model of causal relationships, in the sense that it includes both the suggestion of the US Surgeon General (smoking causes cancer) and also the suggestion of the tobacco companies (a hidden factor causes both smoking and cancer). Indeed, it also allows a third possibility: that perhaps both smoking and some hidden factor contribute to lung cancer. This combined relationship could potentially be quite complex: it could be, for example, that smoking alone actually reduces the chance of lung cancer, but the hidden factor increases the chance of lung cancer so much that someone who smokes would, on average, see an increased probability of lung cancer. This sounds unlikely, but later we’ll see some toy model data which has exactly this property.

Of course, the model depicted in the graph above is not the most general possible model of causal relationships in this system; it’s easy to imagine much more complex causal models. But at the very least this is an *interesting* causal model, since it encompasses both the US Surgeon General and the tobacco company suggestions. I’ll return later to the possibility of more general causal models, but for now we’ll simply keep this model in mind as a concrete example of a causal model.