John Myles White
"He who refuses to do arithmetic is doomed to talk nonsense."

Three-Quarter Truths: Correlation Is Not Causation

By John Myles White on 10.1.2010

Other than our culture's implicit association between lies, damned lies and statistics, I think no idea has stifled the growth of statistical literacy as much as the endless repetition of the words *correlation is not causation*. This phrase seems to be primarily used to suppress intellectual inquiry by encouraging the unspoken assumption that correlational knowledge is somehow an inferior form of knowledge.

I'd like to defend correlation for a bit. Here are four reasons why I think we should learn to love correlation and stop worrying so much about causation.

Claim 1: Most Knowledge is Correlational Knowledge

The majority of reliable human knowledge is already correlational. Spend a few days making a list of things that you know for certain about the world. I claim that you will find that a solid majority of them will be correlational statements rather than causal statements. For example, you might notice that you know that teenagers who own skateboards generally like punk music more than world music, though you are certainly aware that listening to the Sex Pistols isn't the cause of their desire to learn how to ollie. And you almost certainly know that 's' is followed by 't' more often than 's' is followed by 'r' in English, though you would never claim that an 's' causes an 'r'. [1]

Hopefully those two examples are enough to make you suspect that you have an enormous quantity of correlational information stored inside your head. I'd like to further suggest that, despite its low status in our scientific culture, this sort of correlational knowledge has enormous practical value to you, because it allows you to make sense of a world in which you have incomplete information and are constantly required to fill in the blanks. For example, if you're out at night in the deep South and suddenly see someone charging towards you dressed in white sheets, you'll almost certainly run away, even though you don't believe that white sheets cause lynchings. [2] Correlational knowledge can keep you alive when worrying about causality would get you killed.

Claim 2: The Value of Information is More Complex than the Opposition of Causation and Correlation Would Suggest

Taking this point a step further, it's worth noting that assessing the value of information is a far more difficult problem than one might think. In practice, you always need to ask yourself what you're trying to do with information. In many cases, you aren't trying to control things, which I would claim is the only scenario in which causal knowledge could not be replaced with correlational knowledge in principle. In most real world problems, correlational knowledge is enough to make predictions with very high accuracy. For example, imagine that you run a bank and want to predict whether a person will default on their loan. You find that

their zipcode predicts their rate of default quite well. You know full well that a zipcode cannot possibly cause a person to default on their loan, because it's just a number based on a fairly arbitrary way of cutting up neighborhoods. But the absence of a causal relationship is completely irrelevant to you as a banker, since your interest lies in making money — and not in learning something about the hidden causes of human behavior.

If you want to predict something, rather than control it, the most important thing to ask is how well the information you can acquire will allow you to make predictions. After addressing this problem, you will also need to consider the relative costs of acquiring different sorts of information. For example, suppose that you want to predict a person's height. Most of us accept that our genes are the ultimate cause of our height, barring serious illness or malnutrition as children. That's why the heights of identical twins are so similar, while the heights of fraternal twins can be quite different. Focusing on causal pathway from genes to phenotype might suggest that you should try to measure someone's genes to predict their height. People have done this and it doesn't work very well. More importantly, it provides mediocre results at a fairly high cost. Acquiring a genotype is constantly going down in price, but it still costs a few hundred dollars.

Another approach comes from the inventor of the concept of correlation: Francis Galton. Galton's method simply takes your parents' heights and uses a correlational model to predict your height. This approach is correlational because no one believes that your parents' heights cause your height: your parents' genes caused their heights, then their genes caused your genes, and finally your genes caused your height. This is a perfect example of the way in which two things can be correlated because they share a common cause.

By making clever use of correlational information, Galton's method only requires data that is available at almost zero cost, and yet it is more than ten times as accurate as the genetic screening method described above. Sometimes cheap correlational information provides high predictive accuracy, while costly causal information provides almost no predictive power. If you want to do something with information, you should always consider the possibility that a correlational pathway may be cheaper to observe than a causal one — at the same time that it provides comparable predictive power or even greater predictive power.

Claim 3: Causation is a Moving Target

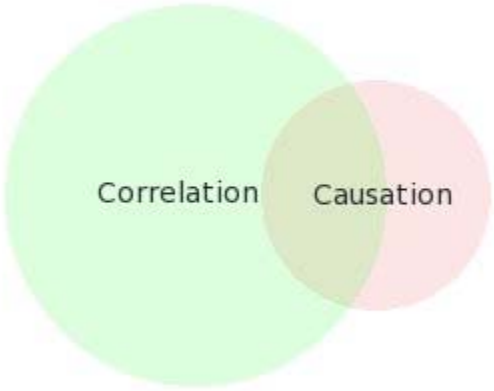
Causation is not an entirely well-defined concept. It is an intuitive notion like justice or intelligence, and therefore may not have any definition that corresponds to all of the ways in which the word "cause" is used in normal language. Despite considerable work by philosophers and mathematicians, our accumulated understanding of what causation means is still very weak.

This vagueness works in causation's favor. Because correlation is so much more precise as a concept than causation, it's easier to come up with examples in which correlation doesn't provide us with useful information than it is to come up with examples of the irrelevance of causal knowledge. This discrepancy in falsifiability is really a general property of mathematical models when compared with intuitive arguments: the precision of mathematical models makes them much more vulnerable to attack than vague ideas. But this brittleness is really a unrecognized virtue, because it is inseparable from the exactness that makes mathematical models directly comparable, precisely communicable and easily modified and extended. Despite their intuitive appeal, ideas whose true or falsehood is hard to assess are less amenable to the incremental improvements that has made scientific knowledge so valuable to humanity.

Claim 4: Correlation and Causation are Related

Last, but not least, I think correlation and causation are themselves correlated. By this I mean that if you were to list pairs of related things like height and weight, ethnicity and voting preferences, or zipcodes and mortgage default rates; and then classified each relationship as correlational and causal, you'd find that many instances of correlation were accompanied by causation. And you'd find that even more instances of

causation were accompanied by correlation. Following Drew Conway's lead, I'll draw a Venn diagram of the relationship that I believe holds between correlation and causation:



This claim is incredibly hard to test: it is merely meant to remind us how wasteful it can be to focus exclusively on the differences between correlation and causation when they also have important similarities. It is true that correlation is not causation. But it is also true that human beings are not chimpanzees. And yet, in spite of that, we've been able to learn a lot about the human brain from studying the brains of chimpanzees, because there are many cases in which the similarities between humans and chimps are more important than the differences. Similarly, studying correlations can give us valuable information, including information about where to start looking for causal relationships. And even when it can't do this, there is nothing wrong with correlational knowledge that is not also causal knowledge. Knowledge of causation is only necessary when we want to control the world. But there are many aspects of the world that we are largely unable to control, even in principle. In those cases, we simply need to have accurate predictions, because prediction without causation is enough for us to make the best of what is going to happen in the future. Assessing our ability to make predictions is vitally important, and it is the habit of making testable and precise predictions that an education in statistics can give to us. So let's embrace a world with rich data sets that can provide us with formal, testable knowledge based on unambiguous, formal models — even if those models won't ultimately provide us with causal mechanisms.

With all of that said, if you really want to understand the distinctions between correlation and causality, there is a rich academic literature that is far subtler and more interesting than the folk philosophy of science that I've been attacking. The current classic is Judea Pearl's masterwork, entitled simply "Causality". It is very challenging material, but well worth the effort. And understanding it will require you to master so much of the machinery of prediction that you'll walk away enlightened even if you decide in the end that causality doesn't really interest you.

For most people, though, I have a different closing message. Please don't allow the absence of causation to be used as a justification for remaining ignorant about the correlational structure of our world. Though there are cases in which knowing that A is related to B is much less useful than knowing that A causes B, knowing that A and B are related at all is still far better than knowing nothing at all — and we currently know nothing about many things. We should stop focusing on the ways in which correlation is not causation and instead follow Voltaire's advice: do not allow the perfect to become the enemy of the good. [3]

1. You can quickly check this with the following shell script on OS X:

```
grep 'sr' /usr/share/dict/words | wc -l
grep 'st' /usr/share/dict/words | wc -l
```

Running those commands should show you that there are 156 examples of ‘sr’ and 21,407 examples of ‘st’ in the standard UNIX dictionary.

2. Especially not if you’ve seen Santa Semana celebrations in Spain.
3. These ideas came up during a recent planning session for O’Reilly’s upcoming Strata Conference, during which Chris Wiggins said that he thought the distinction between correlation and causation was a red herring. My desire to expand on the reasons why I agreed with him inspired me to write my own ideas down.]

Posted in Statistics | 18 Responses

18 responses to “Three-Quarter Truths: Correlation Is Not Causation”

1. 

Chris Diehl
10.1.2010 at 2:14 pm | Permalink

Excellent post John!

2. 

Stephen
10.1.2010 at 2:15 pm | Permalink

Good post. While I respect the value of correlational knowledge, I know of very few instances in which pure prediction without the need to control something exist outside of academia. In many instances, particularly in finance, the focus on improving prediction accuracy results in less focus on causation and statements like “no one could have predicted…” are used as justification for a lack of a responsible decision making process. Voltaire’s quote can work both ways if correlation based knowledge provides better prediction than causal models. In the larger realm of decision making, improvement in prediction accuracy is not necessarily positive.

3. 

zyxo
10.1.2010 at 3:23 pm | Permalink

Like someone said: if you want to know for sure whether it is causation or merely correlation then you have to experiment !

4. 
Antonio
10.1.2010 at 6:19 pm | Permalink

Why “merely correlation”? That’s very point the of the post!

5. 
James
10.1.2010 at 6:31 pm | Permalink

If you haven’t already, I would recommend reading Paul Holland’s classic 1986 article “Statistics and causal inference,” published in the Journal of the American Statistical Association, 81 (396). It discusses the Neymann-Rubin Causal Model, a formal model for defining cause and effect that is a broad counterpoint to Pearl’s approach.

6. 
John Myles White
10.1.2010 at 9:12 pm | Permalink

@Everyone,

Thanks for all the feedback. I’ve really appreciated the conversation this post started.

@Stephen,

My, perhaps quite mistaken, impression is that the dream of finance has been precisely prediction without control. For example, portfolio diversification seems to depend exclusively on correlations between stock prices rather than causal relationships. And one of the great problems, it seems, for financial models is that they fail to consider the effect that large-scale algorithmic trading has on the patterns of trading that emerge: the model is affecting the system it tries to predict.

That said, my knowledge of finance is very, very slim. I’d be interested in discussing this topic more with you at some point since I am starting to do some research work in behavioral finance.

@James,

I haven’t read that paper. Thanks for the pointing it out. I knew about the approach through Andrew Gelman, but had never sat down to read anything. I’ll fix that this weekend.

7. 

John
10.2.2010 at 5:47 pm | Permalink

The Galton height example would have worked better had you used it to predict parents' height from child's height, as it is then even clearer that the child's height did not *cause* in any the parents' height.

8. 
Farrel
10.4.2010 at 2:21 pm | Permalink

I believe the point of previous criticisms may have been missed. Causality is not superior to correlation (or association). The article does a fine job of pointing out instances where correlation would be far superior to causality. Rather the issue is that many people falsely conclude that causality has been proven (or at least strongly suggested) when correlation has been demonstrated. If A is associated with B and one suspects that A causes B then one may point to the association to confirm the suspicion. FAIL! Rather if A was NOT associated with B then one may have grounds to weaken the suspicion that A causes B. If A is associated with B then all of the following have an equal probability of being true: A causes (directly or indirectly) B, B causes (directly or indirectly) A, Y causes (directly or indirectly) both A and B. In medicine it is humongously easier (from an ethics and/or a resource point of view) to design a study that would detect if association does or does not exist that to detect if causality exists or does not exist. So the association studies are done by the bucket load and then used to confirm biases or to provide the illusion that causality has been demonstrated. The illusion can result in people fearing phenomenon they do not need to fear or it can result in people undergoing treatment that does not affect the outcome over and above the natural course.

9. 
anon
10.8.2010 at 8:31 pm | Permalink

Shouldn't causation be a subset of correlation?

10. 
John Myles White
10.8.2010 at 8:33 pm | Permalink

That depends on your definition of correlation. Given a standard definition, like the Pearson product moment correlation, there are many causal systems with zero correlation.

11. 

Antonio
10.8.2010 at 8:46 pm | Permalink

No. There could be causation without correlation. Just think about a situation in which two causal effects cancel each other, because the run in opposite directions. There will be no correlation in the data but there is still causation. An example is campaign contribution and votes in US congressional elections. In this setting the correlation is typically negative! The more you spend the the smaller the number of votes you get!!! Why? The common explanation is that the more competitive is the election (i.e. the more difficult is to get votes) the higher is the spending. Yet given the high level of competition you get less votes per unit of money. This real word example is of negative correlation, not the lack of thereof. But hope I made the point somewhat clear.

12. 

John
10.9.2010 at 12:01 am | Permalink

C'mon all; you know what @anon meant: not literally linear correlation, per se, but something like relations, a "goes-togetherness"; in which case, despite the fact that science has yet to define in any clear mathematical way what that means (but see Judea Pearl's book Causality for a seemingly good stab at it), we all know that what was meant: there are relations, and then there (as a subset), causal relations. Beyond that, I am not sure @anon's point was, though.

13. 

anon
10.9.2010 at 1:14 am | Permalink

@john Yes, something like that. Look, I'm just trying to understand what everyone is getting at. I mean, I don't know what a causal effect means other than a conditionally independent effect of X on Y. For instance, in the elections example, the theoretical effect of spending is to raise turnout of voters, but the confounding factor is the competitiveness of the election. It would be wrong to conclude that less spending gets you more votes. Perhaps related, but I'm not sure what causality means independent of the notion of it we derive from a proposed explanation (i.e. theory). I mean, imagine we can identify an effect of X on Y when controlling for all relevant factors (i.e. a "perfect experiment" or the "perfect" subset of control variables). This doesn't magically turn the effect parameters into a causal effect except to the extent we can make that inference because we understand the system through a theory that has a priori identified the relevant factors.

@ John could you provide examples of systems in which X causes Y and has no correlation?

14. 

John
10.9.2010 at 1:36 am | Permalink

@anon
Assuming, again, that by no correlation you mean something like no relation (i.e., under all circumstances, x is independent of y, such that p(x|y) = p(x) for all values of y and x), then no, I can't see how X *in that situation* could in anyway be considered a cause of Y. But, that was why I was defending your stance: real experiments assume they have met this requirement via random assignment.

15. 

John Myles White
10.9.2010 at 8:15 am | Permalink

Thanks for all the smart comments, @John and @anon. Here's a quick pass at a reply. I am not even close to being an expert in this area, please so be vigilant for mistakes in my reasoning.

Pedantic Component
I can easily give examples of systems in which X causes Y, but there is no correlation under specific definitions of correlation. The Yerkes-Dodson Law is an example for which there is zero Pearson correlation and zero Spearman correlation as well. In general, symmetric parabolic equations will break any definition of correlation that requires monotonic relationships between variables.

If you say that correlation simply means non-independence, then I think it may be true that causation is a subset of non-independence, but I'm not quite sure yet. Unfortunately, I'm even less sure this knowledge could be practically valuable, because a universal test for non-independence may be impossible to construct, much like a universal algorithm for solving Diophantine equations cannot exist. Also, see the points below for how complex a notion independence is in practice, especially when conditioning on other variables.

Substantive Component
I would really encourage reading something by Pearl, like this review: (http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf) My knowledge base is enormously smaller than Pearl's. Here are some issues raised during my reading of his work, though.

First, conditional independence is problematic, because it only means something given the variables you're conditioning on. If you don't have access to the relevant variables (e.g. because you didn't collect them during your observational study), you can't even start to test notions of conditional independence. As Pearl puts it, given a graphical model M, the information in M is not sufficient to make causal inferences about M. How, for example, are you supposed to deal with the relationships between symptoms and diseases using only their joint probability distribution?

Second, conditional (in)dependence doesn't satisfy my intuitions about causality at all. Let's consider a specific graphical model in which there are directed edges, X -> Y and Y -> Z. In this model, conditional on Y, X is independent of Z. But I most definitely believe that X is a cause of Z in some instantiations of this model, so your claim, at least as I understand it (and I may be confused), seems

not to work in this case. Causality exists simultaneously with conditional independence. Since you can translate that graphical model into examples involving billiard balls quite easily, I think it's natural to generalize and conclude that studying conditional independence will not be sufficient to address concerns of causality.

More generally, Pearl wants us to remember that probabilistic calculations on a set of variables can never be sufficient to determine that we have data from a "perfect experiment". (The random assignment theory is nice, but in practice is clearly based on a hope that your groups do not differ on unmeasured confounding variables — a hope that you can easily convince yourself is not always satisfied: you only need to enumerate enough variables before you'll find one that was not properly matched across groups by random assignment. The hope aspect comes in because you simply assume these other variables don't matter rather than verify your claim.) Pearl's point implies that you will never have done enough to verify this claim if you only use probabilistic calculations, because causality is something external to statements that can be derived from joint probability distributions.

16. 
Andrew L.
10.10.2010 at 7:44 pm | Permalink

Devil's advocate, but essentially it is. Not because of lower inherent value, but because it's so often misused.

As a result, people weight correlations more lightly because according to their experience correlational data is less dependable than causative.

It's not an unreasonable response, and however unfortunate it's probably generally beneficial.

17. 
Steven D'Aprano
10.17.2010 at 7:43 pm | Permalink

Nice article, but I question the very first paragraph: is there *any* evidence that statistical literacy is suffering because of overuse of the phrase "correlation is not causation"? It seems to me that you have jumped from the correlation between "people get told correlation does not imply causality a lot" and "people find statistics difficult", and imagined a causal relationship where there is none.

Or that correlational knowledge is being undervalued? I'd say the opposite — your post demonstrates quite well that correlational knowledge isn't undervalued at all, it is quite higher valued.

So much so that when correlation and causality point in opposite directions — say, A causes B but due to some other uncorrected factor(s) the correlation between A and B is low, people in general are extremely resistant to the idea that A causes B. We call such pairs of causal relationships "unintuitive", and as difficult as it is to discover them, that's nothing compared to the difficulty of convincing people of that relationship.

I think it is quite absurd to suggest that the reason — or even *a* reason — why most people are functionally illiterate when it comes to statistics is because they're being "stifled" by overuse of

"correlation is not causality". I can't imagine that there is anyone who would say they didn't understand why sample standard deviation divides by n-1 instead of n because they heard "correlation does not imply causality" too many times. Or that they couldn't quite grasp when to use mean and when to use median.

Statistics is complex and complicated and highly mathematical. We as a species don't do well at *any* of those. Our brains evolved in an environment where there were no tools, and no time, for carefully teasing out deeply hidden causal relationships, where if you didn't have *fast* heuristics for deciding how to behave you would die, where false positives were less harmful than false negatives (better to wrongly avoid eating the good fruit than to wrongly eat the poison one). It is no surprise that we find correlation easy to detect (even when there is none — correlation is the basis of superstitious behaviour) and causality hard. The problem isn't encouraging people to look for correlation, but to stop them jumping from correlation to causality inadvisablely.

This post was an excellent defence of correlation, marred only by the fact that correlation needs no defence.

18. 

John Myles White
10.19.2010 at 4:23 pm | Permalink

@Steven,

You make some very valid points, but I'd caution you against appealing to evolutionary psychology too readily. Such reasoning seems prima facie plausible, but it doesn't take into account the contemporary research on human causal learning. I'd recommend starting with Josh Tenenbaum's research and then moving on to what's been well known for decades. The high-level summary of the literature I'll offer is that you need to make the words "complex" and "complicated" essentially meaningless before your claim stands up against the empirical data on causal learning. Humans are capable of extraordinarily subtle causal reasoning: they simply do it in an seemingly odd fashion.

Leave a Reply

Name *

Email *

Website

Comment

Submit

☐ Notify me of followup comments via e-mail

my del.icio.us

- More Women Without Children - Pew Research Center 2010/07/26
- TinEye Image Search Engine 2010/07/26
- HPCwire: Data Mining Made Faster 2010/07/26
- Five Myths about the Death Penalty 2010/07/26
- Microsoft Learning to Rank Datasets 2010/07/26

@johnmyleswhite

- Could not connect to Twitter

« Previous Next »

Powered by WordPress and Hybrid.