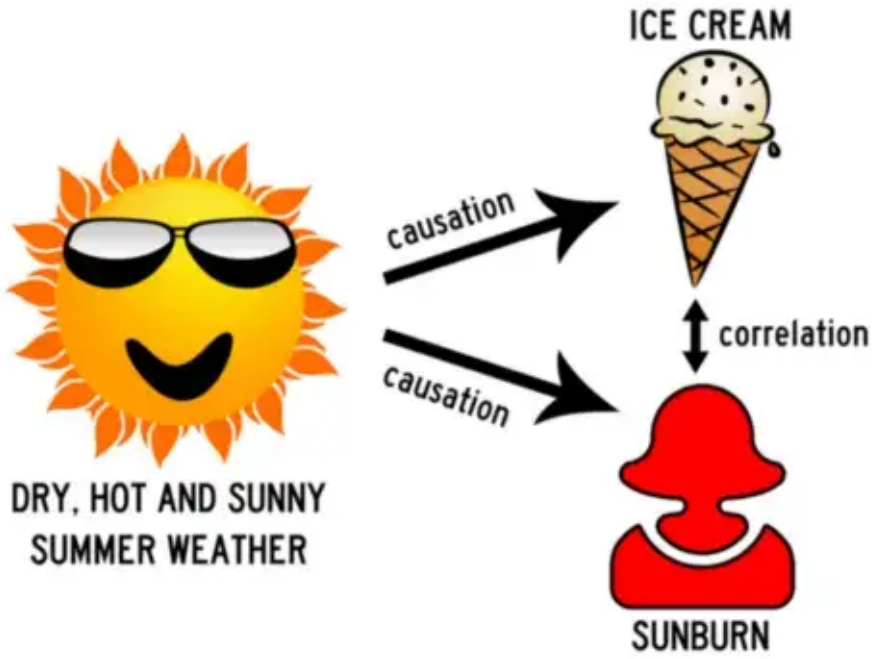
Published in Towards Data Science

Anthony Figueroa Follow

Aug 13, 2019 · 8 min read · · Listen



Correlation is not causation

Why the confusion of these concepts has profound implications, from healthcare to business management

Introduction

In correlated data, a pair of variables are related in that one thing is likely to change when the other does. This relationship might lead us to assume that a change to one thing causes the change in the other. This article clarifies that kind of faulty thinking by explaining correlation, causation, and the bias that often lumps the two together.

The human brain simplifies incoming information, so we can make sense of it. Our brains often do that by making assumptions about things based on slight relationships, or bias. But that thinking process isn't foolproof. An example is when we mistake correlation for causation. Bias can make us conclude that one thing must cause another if both change in the same way at the same time. This article clears up the misconception that correlation equals causation by exploring both of those subjects and the human brain's tendency toward bias.

About correlation and causation

Correlation is a relationship or connection between two variables where whenever one changes, the other is likely to also change. But a change in one variable doesn't cause the other to change. That's a correlation, but it's not causation. Your growth from a child to an adult is an example. When your height increased, your mass increased too. Getting taller didn't make you also get wider. Instead, maturing to adulthood caused both variables to increase — that's causation.

Causation in business

Let's say that we want to offer a promotion or discount to some of our customers. Our marketing department wants to maximize the delta, in other words, the increase in sales as a result of the promotion. So we need to decide which customers will give us the best return on our investment in the promotion or discount. Do we want to offer it

of which customers to offer the promotion to might be totally different. In the absence of valid experimentation or analytics, you don't have accurate answers to those questions.

Cognitive bias

There are many forms of cognitive bias or irrational thinking patterns that often lead to faulty conclusions and economic decisions. These types of cognitive bias are some reasons why people assume false causations in business and marketing:

- **Confirmation bias.** People want to be right. They often can't admit or accept that they're wrong about something, even if that attitude causes eventual harm and loss.
- **The illusion of causality.** Putting too much weight on your own personal beliefs, over-confidence, and other unproven sources of information often produce an illusion of casualty. An economic example is the recent U.S. housing bubble. Millions of people believed that buying a home for much more than its actual value would continue to result in a return on the investment just because that happened in the past.
- **Money.** You want to sell your product. You might spend more than your return on investment (ROI) on marketing and other business expenses if the desire to make money clouds your logic.
- **Major marketing implications.** Marketing statistics and data are often complicated and confusing. It can be easy to see relationships between changing sales numbers and the many other variables in your business when no causation exists.

Experimentation

To know that something is valuable takes experimentation. Experimentation helps you understand if you're making the right choices. But it has a cost. If you hold a workgroup back by not giving them things in value, you'll lose money. But you'll learn the importance of that feature.

The value of an experiment lies in the accomplishment of these two things:

- Decide between different choices.
- Quantify the value of the best choice.

Experimental variables

A scientifically valid experiment needs to have three types of variables: controlled, independent, and dependent:

- A **controlled** variable is kept constant, so other variables that change in relation to each other can be measured in a static environment.
- An experiment's **independent** variable is the only one that can be changed.
- **Dependent** variables are the results that are observed when changes are made to independent variables.

Any **uncontrolled** variables, or mediator variables, can cloud an experiment's accuracy. So they need to be identified and eliminated in order to properly assess the experiment's results. Differences in uncontrolled variables can also impact the relationship between independent and dependent variables.

Uncontrolled variables add the influence of unrelated factors to an experiment's results. Correlations might be assumed, and an hypothesis might be formed where none exist. Accurate analysis becomes difficult or impossible. Examples of conclusions draw from uncontrolled variables are shown in the children's music lessons and mobile phone cancer examples that follow.

How our brain tricks us

It's easy to watch correlated data change in tandem and assume that one thing causes the other. That's because our brains are wired for cause-relation cognitive bias. We need to make sense of large amounts of incoming data, so our brain simplifies it. This process is called heuristics, and it's often useful and accurate. But not always. An example of where heuristics goes wrong is whenever you believe that correlation implies causation.

Spurious correlations

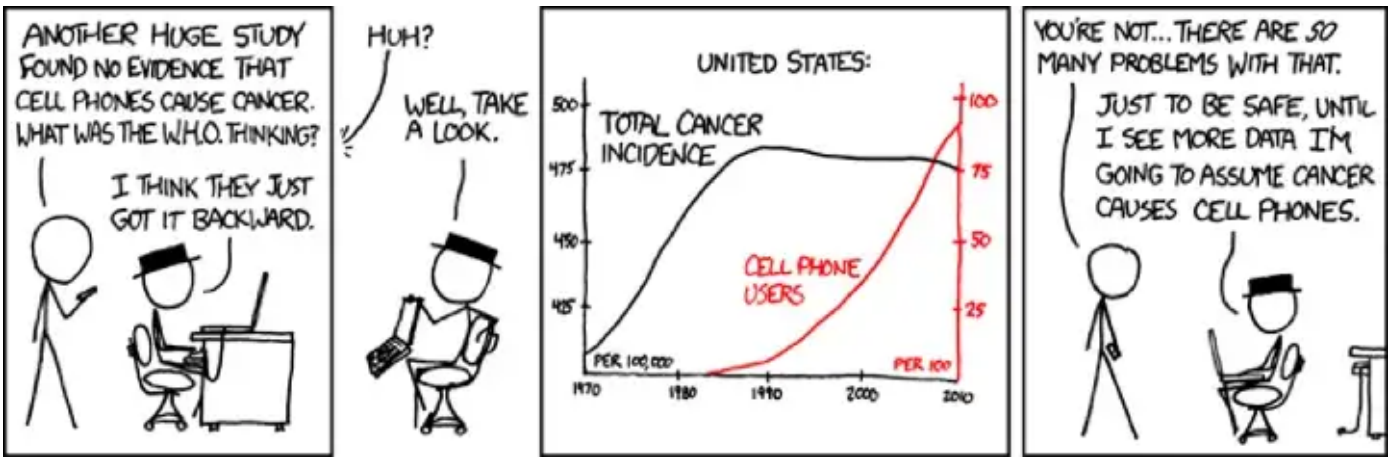
It is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor

Children and music lessons

After a study of human brain development, researchers concluded that kids between 4 and 6 years old who took music lessons showed evidence of boosted brain development in the areas related to memory and attention. Based on this study, our biased brain might connect the dots quickly and conclude that music lessons improve brain development. But there are other variables to consider. The fact that the children took music lessons is an indicator of wealth. So they probably had access to other resources that are known to boost brain development like good nutrition.

The point of this example is that researchers can't assume from only this much data that music lessons affect brain development. Yes, there's clearly a correlation, but there's no actual evidence of causation. We need more data to get a true causal explanation.

Cancer and mobile phones



If you study a chart that shows both the number of cancer cases and the number of mobile phones, you'll notice that both numbers went up in the last 20 years. If your brain processes this information with cause-relation cognitive bias, you might decide that mobile phones cause cancer. But that's ridiculous. There's no proof other than both datapoints happening to increase. A lot of other things have also increased in the past 20 years, and they can't all-cause cancer or be caused by mobile phone use.

Explainability

To find causation, we need explainability. In the era of artificial intelligence and big data analysis, this topic becomes increasingly more important. AIs make data-based recommendations. Sometimes, humans can't see any reason for those recommendations except that an AI made them. In other words, they lack explainability.

Explainability in medicine

The FDA won't approve cancer treatments that lack explainability. Think about this situation for a minute. Do you want the best possible treatment for your cancer, based on an AI's analysis of your genomes, your cancer DNA, millions of other cases, and more data, even if you can't explain how the computer's neural network came up with that exact treatment? Or would you rather have a suboptimal treatment that you can explain the reasoning for?

Medical explainability will be probably one of the biggest topics of this century.

One way versus two way

Correlations go both ways. We can say that mobile phone usage correlates to increased cancer risk and that cancer cases correlate to the number of mobile phones. Basically, you can swap the correlation. In causation relationships, we can say that a new marketing campaign caused an increase in sales. But saying that the increase in sales (after the campaign ran) caused the marketing campaign doesn't make any sense.

Any causal statement, by definition, is one way. That's a big clue about whether you're dealing with correlation or causation.

The big dilemma

In "The causal effect of education on earnings," David Card says that better education is correlated to higher earnings. But the most important thing he says is that if we can't do an experiment, with all our

variables constant, we can't infer causation from a correlation. We can always bring explainability to the table. But in real life and with big enough problems, causations based on explainability are hard to prove. From a scientific viewpoint, they can't be called anything more than a theory.

> In the absence of experimental evidence, it is very difficult to know whether the higher earnings observed better-educated workers are caused by their higher education, or whether individuals with greater earning capacity have chosen to acquire more schooling.
>
> — David Card, The causal effect of education in earnings

Does higher-earning cause higher education? Does higher education cause higher earning potential? We don't know. However, we can make predictions. We can use this correlation to predict the earning potential of an individual based on his education. We can also predict his education based on his earnings.

Good predictions are based on correlations

It sounds like a contradiction, given the context of this article. Correlation is about analyzing static historical datasets and considering the correlations that might exist between observations and outcomes. However, predictions don't change a system. That's decision making. To make software development decisions, we need to understand the difference it would make in how a system evolves if you take an action or don't take action. Decision making requires a casual understanding of the impact of an action.

What are predictions?

We don't make better predictions by developing a better casual understanding. Instead, we need to know the precise limits of the techniques we use to make predictions and what each method can do for us.

References

Lovestats (2019). "Cartoons." *The LoveStats Blog*. Retrieved from lovestats.wordpress.com.

Card, D.. (1999). "The causal effect of education on earnings." *Handbook of Labor Economics*, vol 3.

Machine Learning | Data Science | AI | Correlation Vs Causation | Explainability