Experiments and Causal Inference

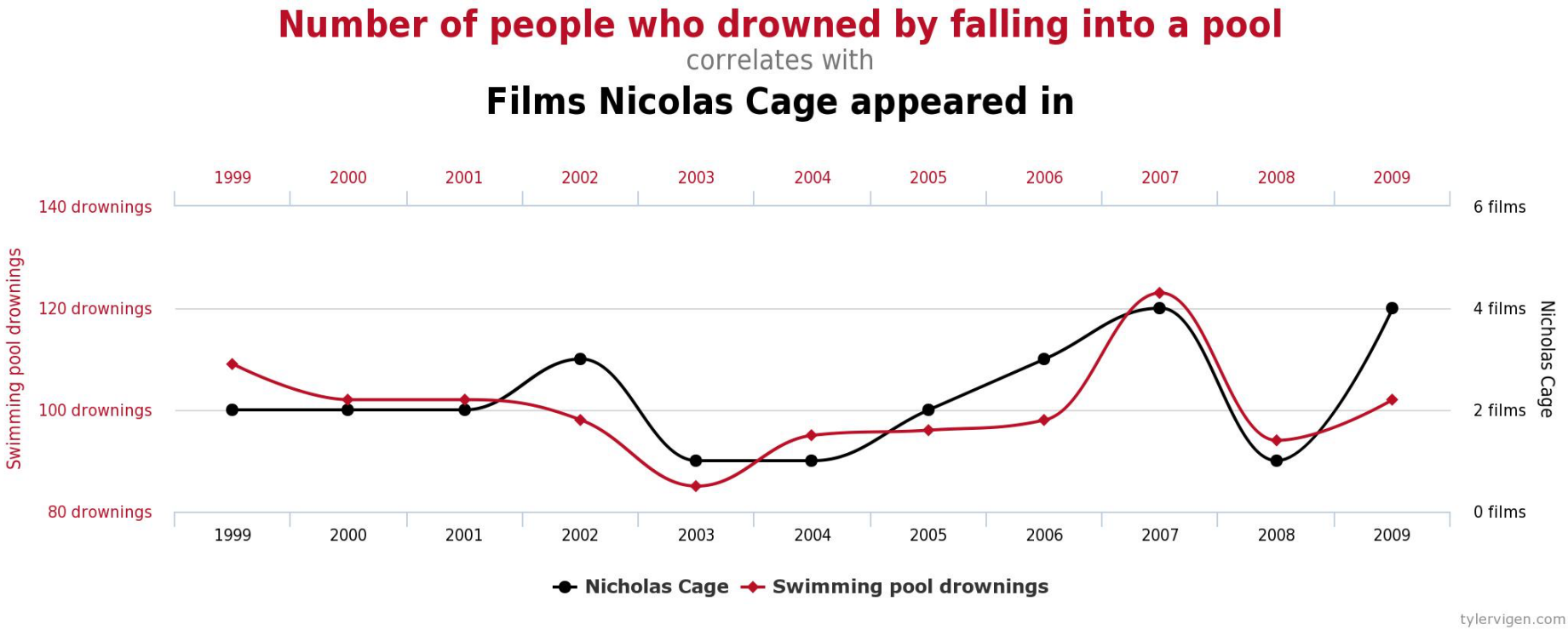
DATA 301

Dr. Stephen Pettigrew

"Correlation does not imply causation"

It's good to be skeptical of causal arguments based entirely on correlations...

It's good to be skeptical of causal arguments based entirely on correlations...



But on the other hand...

But on the other hand...

John Horton 
@johnjhorton

Follow 

researcher writes 60 page paper, 55 of which are on how to infer causality from the observational data Twitter: Yeah, but what about causality...

3:12 PM - 6 Mar 2018 from Greenwich, CT

23 Retweets **176** Likes 

3 23 176

Why are we talking about this in a data science class?

Lots of data science work today is focused on using big data and machine learning

- This is just a fancy way of saying we'll take a bunch of data and look for correlations in it

Why are we talking about this in a data science class?

Lots of data science work today is focused on using big data and machine learning

- This is just a fancy way of saying we'll take a bunch of data and look for correlations in it

This type of analysis is all about *prediction*:

- How can we use data from the 2016 election to predict who will turn out to vote in 2020?
- How can we use data about basketball players' career trajectories to predict how well a 2nd year player will perform in his 5th year?
- How can we use data to predict which companies' stocks will rise or fall over the next year?

Why are we talking about this in a data science class?

Prediction tasks tell us *how* X and Y are related to each other

But often, we want to know *why* X and Y are related

Why are we talking about this in a data science class?

Prediction tasks tell us *how* X and Y are related to each other

But often, we want to know *why* X and Y are related

Many of the most interesting questions we have are *causal* in nature. Did X cause Y? And if so, by how much?

As a result, experiments (sometimes called A/B testing) have become more common in business analytics and data science

Reasons why correlation may not imply causation

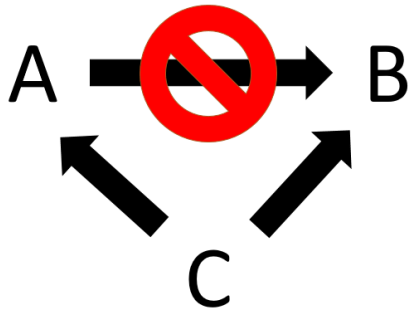
Reason 1: confounding (spurious relationship)

Something else (C) is causing both A and B, which makes A correlate with B, even if A isn't causing B

Reasons why correlation may not imply causation

Reason 1: confounding (spurious relationship)

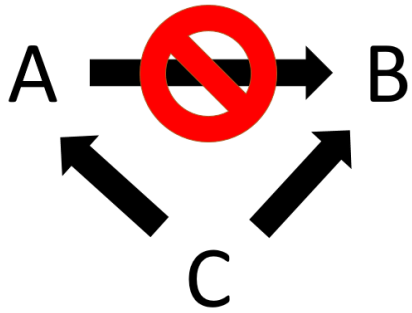
Something else (C) is causing both A and B, which makes A correlate with B, even if A isn't causing B



Reasons why correlation may not imply causation

Reason 1: confounding (spurious relationship)

Something else (C) is causing both A and B, which makes A correlate with B, even if A isn't causing B

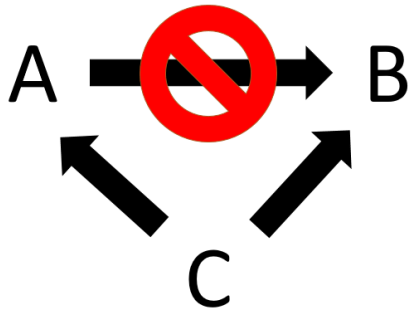


Example: Spending a night in a hospital (A) makes you more likely to die (B)

Reasons why correlation may not imply causation

Reason 1: confounding (spurious relationship)

Something else (C) is causing both A and B, which makes A correlate with B, even if A isn't causing B



Example: Spending a night in a hospital (A) makes you more likely to die (B)

Confounder (C): You got hit by a bus. You got Ebola.

Reasons why correlation may not imply causation

Reason 2: reverse causation

We think that A is causing B, but it's actually B that's causing A.

Reasons why correlation may not imply causation

Reason 2: reverse causation

We think that A is causing B, but it's actually B that's causing A.

Reasons why correlation may not imply causation

Reason 2: reverse causation

We think that A is causing B, but it's actually B that's causing A.

Example: Windmills rotating (A) don't cause the wind to blow (B). Instead, the wind blowing (B) causes windmills to rotate (A)

Reasons why correlation may not imply causation

Reason 3: Bidirectional/reciprocal causation

A is causing B, but B is also causing A

Reasons why correlation may not imply causation

Reason 3: Bidirectional/reciprocal causation

A is causing B, but B is also causing A

Reasons why correlation may not imply causation

Reason 3: Bidirectional/reciprocal causation

A is causing B, but B is also causing A

Example: Predators and prey. The number of predators (A) affect the numbers of prey (B), but the abundance (or scarcity) of prey (B) can affect the number of predators (A).

Reasons why correlation may not imply causation

Reason 4: Relationship is coincidental

A and B are correlated, but neither is causing the other

Reasons why correlation may not imply causation

Reason 4: Relationship is coincidental

A and B are correlated, but neither is causing the other

Reasons why correlation may not imply causation

Reason 4: Relationship is coincidental

A and B are correlated, but neither is causing the other

Example: Every president elected in a year ending in 0 (A) between 1840 and 1980 either died in office or had an assassination attempt against them (B)
Example: Every bald or balding Russian leader since 1825 (A) has been succeeded by a hairy Russian leader (B)

What does it mean for something to have *caused* something else?

What does it mean for something to have *caused* something else?

Causal inference is all about *counterfactuals*

What does it mean for something to have *caused* something else?

Causal inference is all about *counterfactuals*

Imagine that Y represents whether or not somebody dies of a disease

And X represents whether or not somebody is given a particular prescription drug to treat that disease

What does it mean for something to have *caused* something else?

Causal inference is all about *counterfactuals*

Imagine that Y represents whether or not somebody dies of a disease

And X represents whether or not somebody is given a particular prescription drug to treat that disease

If we want to know the effect of X on Y, we need to know whether somebody would die if they were given the drug (call this $Y_{treatment}$), and whether they would die if they weren't ($Y_{control}$)

What does it mean for something to have *caused* something else?

Causal inference is all about *counterfactuals*

Imagine that Y represents whether or not somebody dies of a disease

And X represents whether or not somebody is given a particular prescription drug to treat that disease

If we want to know the effect of X on Y, we need to know whether somebody would die if they were given the drug (call this $Y_{treatment}$), and whether they would die if they weren't ($Y_{control}$)

The effect of X is the difference between $Y_{treatment}$ and $Y_{control}$

$Y_{treatment}$ and $Y_{control}$ are often referred to as *potential outcomes*

Potential outcomes framework

Most often, we want to know what the average effect of X on Y (usually called the Average Treatment Effect or ATE)

Potential outcomes framework

Most often, we want to know what the average effect of X on Y (usually called the Average Treatment Effect or ATE)

But every person in our study is either assigned to the *treatment* group or the *control* group. Never both.

This problem is often referred to as *the fundamental problem of causal inference*

So how can we average the treatment effects across everybody in our study?

Causal inference as a problem of missing data

Treatment group	$Y_{observed}$
Treated	\$84,000
Treated	\$70,000
Control	\$44,000
Treated	\$56,000
Control	\$59,000
Control	\$53,000
Control	\$61,000
Treated	\$61,000
Treated	\$64,000
Control	\$54,000
	\$60,600

Causal inference as a problem of missing data

Treatment group	$Y_{observed}$	$Y_{treated}$	$Y_{control}$	$Y_{treated} - Y_{control}$
Treated	\$84,000	\$84,000	?	?
Treated	\$70,000	\$70,000	?	?
Control	\$44,000	?	\$44,000	?
Treated	\$56,000	\$56,000	?	?
Control	\$59,000	?	\$59,000	?
Control	\$53,000	?	\$53,000	?
Control	\$61,000	?	\$61,000	?
Treated	\$61,000	\$61,000	?	?
Treated	\$64,000	\$64,000	?	?
Control	\$54,000	?	\$54,000	?
	\$60,600	\$67,000	\$54,200	?

How could we fill in all the missing values in the table?

Randomized experiments

Making causal inferences is often contingent on some form of *randomization* or *random* process

In this context, a treatment is *random* if nothing can predict whether or not a person ends up in the *treatment group* or the *control group*

- In other words, we could have flipped a coin to determine who ended up in which group

Randomized experiments

Making causal inferences is often contingent on some form of *randomization* or *random* process

In this context, a treatment is *random* if nothing can predict whether or not a person ends up in the *treatment group* or the *control group*

- In other words, we could have flipped a coin to determine who ended up in which group

Another way to think of randomization is that the treatment group is a random sample drawn from a population, and the control group is another random sample drawn from the same population

Randomized experiments

The key thing we have to assume to do causal inference is that the treatment group is (on average) identical to the control group, except that one group received the treatment and one did not

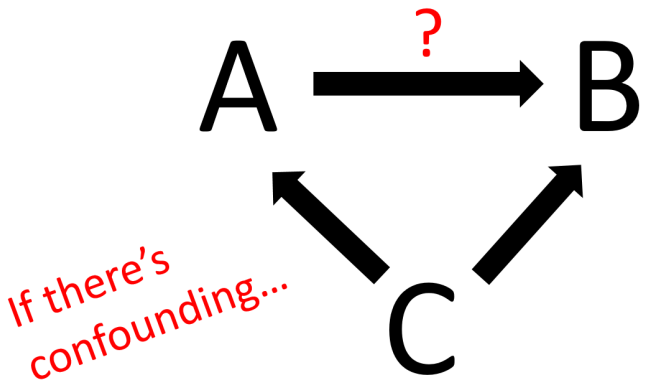
- In practice this will never be exactly true, but with a large enough sample or if we run the same experiment enough times it will be

Randomized experiments

The key thing we have to assume to do causal inference is that the treatment group is (on average) identical to the control group, except that one group received the treatment and one did not

- In practice this will never be exactly true, but with a large enough sample or if we run the same experiment enough times it will be

This solves the problem of confounding, since no *pre-treatment variables* (i.e. C) are correlated with the treatment anymore

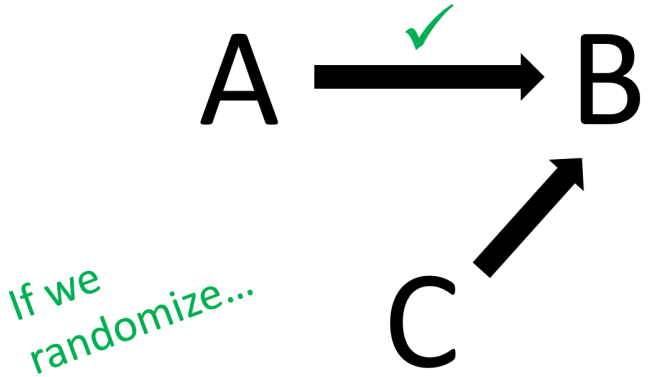


Randomized experiments

The key thing we have to assume to do causal inference is that the treatment group is (on average) identical to the control group, except that one group received the treatment and one did not

- In practice this will never be exactly true, but with a large enough sample or if we run the same experiment enough times it will be

This solves the problem of confounding, since no *pre-treatment variables* (i.e. C) are correlated with the treatment anymore



If treatment is randomized...

Treatment group	$Y_{observed}$	$Y_{treated}$	$Y_{control}$	$Y_{treated} - Y_{control}$
Treated	\$84,000	\$84,000	?	?
Treated	\$70,000	\$70,000	?	?
Control	\$44,000	?	\$44,000	?
Treated	\$56,000	\$56,000	?	?
Control	\$59,000	?	\$59,000	?
Control	\$53,000	?	\$53,000	?
Control	\$61,000	?	\$61,000	?
Treated	\$61,000	\$61,000	?	?
Treated	\$64,000	\$64,000	?	?
Control	\$54,000	?	\$54,000	?
	\$60,600	\$67,000	\$54,200	?

If treatment is randomized...

Treatment group	$Y_{observed}$	$Y_{treated}$	$Y_{control}$	$Y_{treated} - Y_{control}$
Treated	\$84,000	\$84,000	?	?
Treated	\$70,000	\$70,000	?	?
Control	\$44,000	\$67,000	\$44,000	?
Treated	\$56,000	\$56,000	?	?
Control	\$59,000	\$67,000	\$59,000	?
Control	\$53,000	\$67,000	\$53,000	?
Control	\$61,000	\$67,000	\$61,000	?
Treated	\$61,000	\$61,000	?	?
Treated	\$64,000	\$64,000	?	?
Control	\$54,000	\$67,000	\$54,000	?
	\$60,600	\$67,000	\$54,200	?

If treatment is randomized...

Treatment group	$Y_{observed}$	$Y_{treated}$	$Y_{control}$	$Y_{treated} - Y_{control}$
Treated	\$84,000	\$84,000	\$54,200	?
Treated	\$70,000	\$70,000	\$54,200	?
Control	\$44,000	\$67,000	\$44,000	?
Treated	\$56,000	\$56,000	\$54,200	?
Control	\$59,000	\$67,000	\$59,000	?
Control	\$53,000	\$67,000	\$53,000	?
Control	\$61,000	\$67,000	\$61,000	?
Treated	\$61,000	\$61,000	\$54,200	?
Treated	\$64,000	\$64,000	\$54,200	?
Control	\$54,000	\$67,000	\$54,000	?
	\$60,600	\$67,000	\$54,200	?

If treatment is randomized...

Treatment group	$Y_{observed}$	$Y_{treated}$	$Y_{control}$	$Y_{treated} - Y_{control}$
Treated	\$84,000	\$84,000	\$54,200	\$29,800
Treated	\$70,000	\$70,000	\$54,200	\$15,800
Control	\$44,000	\$67,000	\$44,000	\$23,000
Treated	\$56,000	\$56,000	\$54,200	\$1,800
Control	\$59,000	\$67,000	\$59,000	\$8,000
Control	\$53,000	\$67,000	\$53,000	\$14,000
Control	\$61,000	\$67,000	\$61,000	\$6,000
Treated	\$61,000	\$61,000	\$54,200	\$6,800
Treated	\$64,000	\$64,000	\$54,200	\$9,800
Control	\$54,000	\$67,000	\$54,000	\$13,000
	\$60,600	\$67,000	\$54,200	\$12,800

Therefore this treatment causes salary to increase, on average, by \$12,800

Causal inference when you can't run an experiment

There's lots of things we might want to know the effect of, but it's not possible/feasible/ethical to randomize them in an experiment

- Luckily, there's a whole arsenal of other causal inference methods that can be used to answer these questions

Causal inference when you can't run an experiment

Method 1: creative experimental designs

Problem: we can't randomize somebody's race or their gender, so how can we know if women or people of color are being discriminated against when they apply for jobs?

Causal inference when you can't run an experiment

Method 1: creative experimental designs

Problem: we can't randomize somebody's race or their gender, so how can we know if women or people of color are being discriminated against when they apply for jobs?

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

Marianne Bertrand

Sendhil Mullainathan

AMERICAN ECONOMIC REVIEW
VOL. 94, NO. 4, SEPTEMBER 2004
(pp. 991-1013)

Causal inference when you can't run an experiment

Method 2: advanced causal inference techniques (instrumental variables, regression discontinuity, matched pairs designs)

There's a huge number of statistical techniques used to create as-if randomness in the world, even if nobody was running an experiment

Causal inference when you can't run an experiment

Method 2: advanced causal inference techniques (instrumental variables, regression discontinuity, matched pairs designs)

There's a huge number of statistical techniques used to create as-if randomness in the world, even if nobody was running an experiment

Problem: how can we know if participating in big political rallies (like the Tea Party marches in 2010 or the Women's Marches in 2017) impacted turnout in subsequent elections?

Causal inference when you can't run an experiment

Method 2: advanced causal inference techniques (instrumental variables, regression discontinuity, matched pairs designs)

There's a huge number of statistical techniques used to create as-if randomness in the world, even if nobody was running an experiment

Problem: how can we know if participating in big political rallies (like the Tea Party marches in 2010 or the Women's Marches in 2017) impacted turnout in subsequent elections?

Solution: could maybe use rainfall as a source of random variation that affects participation in the rallies, but would impact whether somebody turned out to vote 12 months later

Causal inference when you can't run an experiment

Method 3: Sensitivity analysis

Using statistical tests to show that the correlation is so strong that there's no confounding variable that could possibly explain the correlation

Causal inference when you can't run an experiment

Method 3: Sensitivity analysis

Using statistical tests to show that the correlation is so strong that there's no confounding variable that could possibly explain the correlation

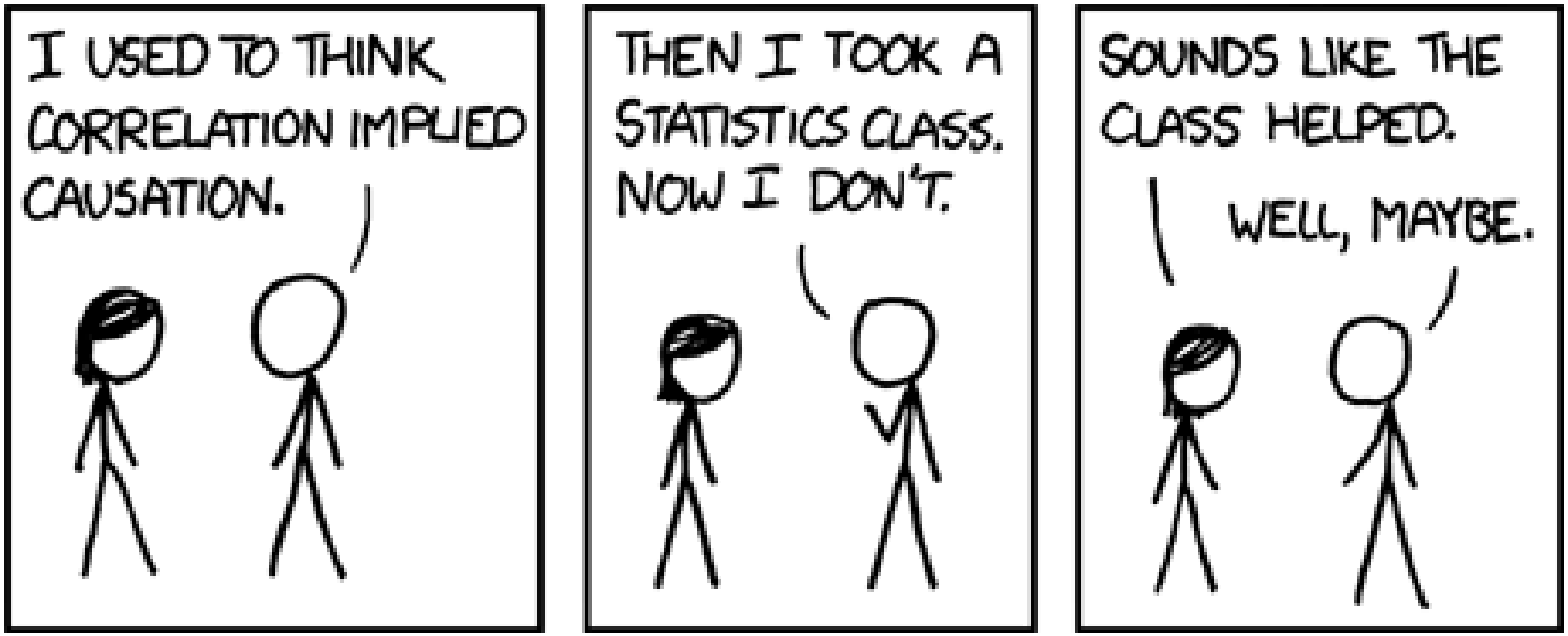
Problem: There has never been a randomized experiment where people were randomly forced to become smokers. So how do we know smoking causes lung cancer?

Causal inference when you can't run an experiment

Method 3: Sensitivity analysis

Using statistical tests to show that the correlation is so strong that there's no confounding variable that could possibly explain the correlation

Solution: The correlation between smoking and lung cancer is so strong that, for it not to be case, there would have to be a confounding factor that was thousands of times better at predicting both lung cancer and whether or not somebody smokes than anything scientists or doctors have ever seen before



source: https://xkcd.com/552/