The Challenge of Causal Inference in the Social Sciences

Justin Murphy, PhD
jmrphy.net
@jmrphy

What is Inference?

- The challenge of **inference** is to use available information to make the best possible conclusions about what we don't know but would like to know.
    - **Descriptive inference** seeks to describe the existence of something.
        - Example: The number of people who participate in a riot.
- **Causal inference** seeks to understand the effect of some variable(s) on some other variable(s)
    - Example: The *causal effect* of unemployment on the probability a riot will occur.
    - Example: The *causal effect* of a riot on next year's government spending.

Some Key Terms

- A **unit of analysis** is simply the object of study.
    - E.g., the individual human being, the constituency, the country, etc.
- A **variable** is the measurement of some concept that varies across a set of units.
    - E.g. unemployment rate across EU countries.
        - An **observation** is one realisation of a variable for one unit.
        - E.g., UK unemployment is equal to 6.0% in 2014.
- Our **sample** is the set of observations we gather to make inferences about the world *outside the sample.*
    - I.e. a quantitative dataset or the cases you select to investigate.
- The **population** is what we call the world outside the sample we want to make generalisations about.

Descriptive Inference

Let's say we want to know how much of the British population supports the current government.

1. Take a *random, representative sample* of, say, 5,000 Brits.

2. Ask them if they support the government.

3. The sample mean can be used to *infer* the population mean.

4. Statistical theory provides rigorous rules for this inference, accounting for sample size, variance, and random error.

Potential sources of error

in estimating a population distribution using a sample

Sampling error

Non-sampling error

Because the sample is not the whole population

Poor sampling method

Questionnaire or measurement error

Behavioural effects

Source: https://learnandteachstatistics.wordpress.com/2014/09/04/sampling-and-non-sampling-error/

Causal Inference

- A **causal inference** is a statement about why something happens.
    - A causal inference therefore states the existence of a relationship between at least two variables.
- The **dependent variable** measures that variation which we would like to explain (find a cause for).
    - Also called ***Y***, or the “outcome” or “response” or variable.
- The **independent variable** measures that variation which we think explains variation in the dependent variable.
    - Also called ***X***, or the “treatment” or “study” variable.

What is Causation?

- What does it mean to say “X causes Y” and how are we able to know this?

- This is more complicated than it seems and there are many philosophies of causation.

- We’ll use the “counterfactual” framework.

  - AKA: “potential outcomes” or “Neyman-Rubin” framework.

  - Dominant framework in the social sciences today.

The causal effect of a treatment is the difference between what happens to a unit after that treatment and what would have happened had the unit not been treated.

The Consistency Assumption

- AKA the “SUTVA”: The Stable Unit Treatment Value Assumption

- "the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox 1958)

- $Y_i(x) = Y_i \quad \text{if} \quad X_i = x$

- Very important/tricky in social research (hint: strategic interactions, time, etc.)

The Fundamental Problem of Causal Inference

For any unit, we only ever observe one potential outcome.

- In other words, to directly calculate a causal effect would require us to rewind the world and re-run it with a different value on the independent variable.

- In other words, causality *cannot* be directly and certainly observed.

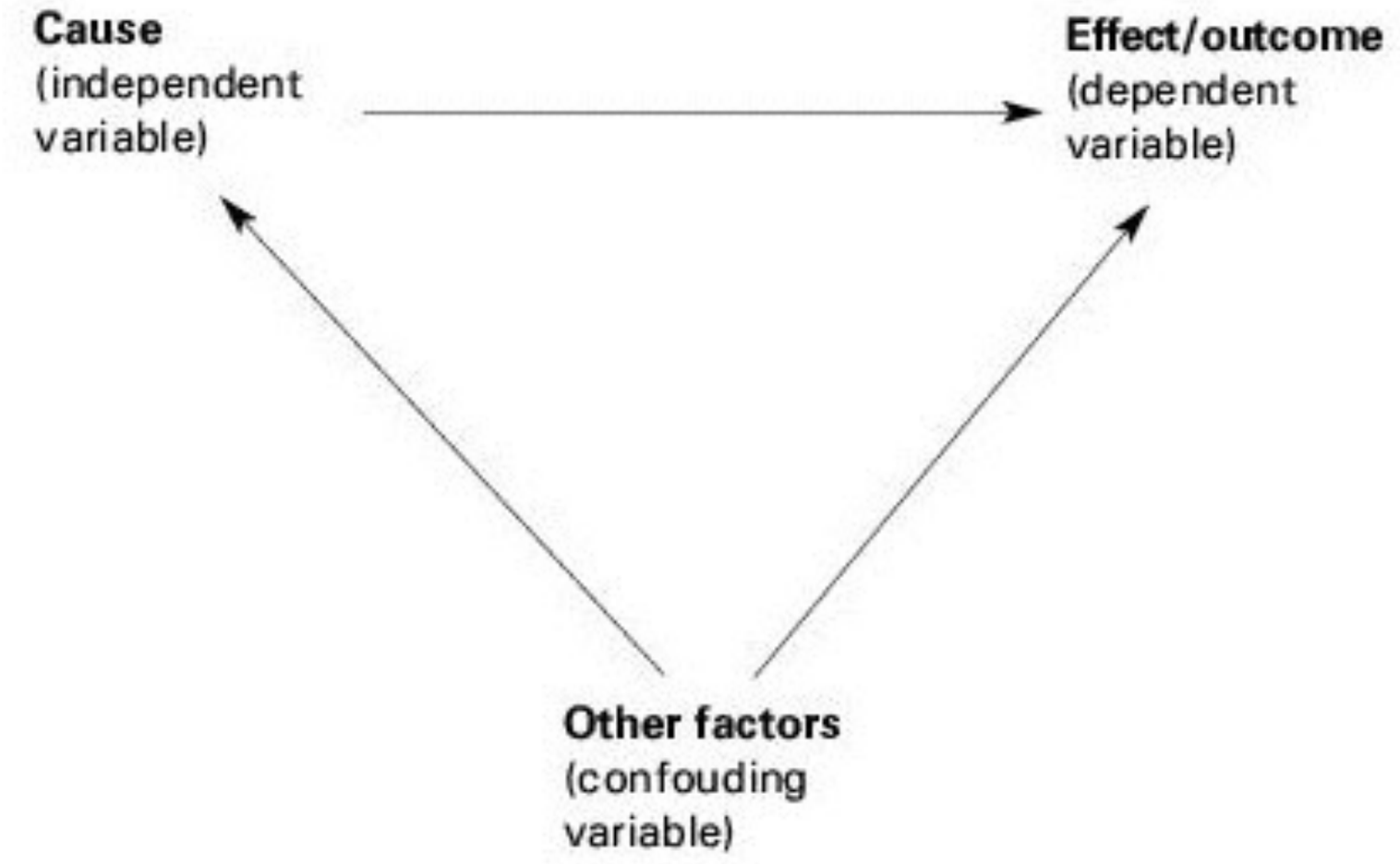
The Experiment as an Imperfect Solution to the FPCI

- Suppose some units $i = 1, ..., N$
- A dependent variable Y_i
- An independent variable X_i

- The value of ***Y*** given some treatment $Y_i(x = 1)$
- The value of ***Y*** given no treatment is $Y_i(x = 0)$
- A basic formal statement of the causal effect is

$$\frac{1}{N}\Sigma_{i=1}^{N} Y_i(x = 1) - Y_i(x = 0)$$

Identifying causal effects in observational research is very hard.



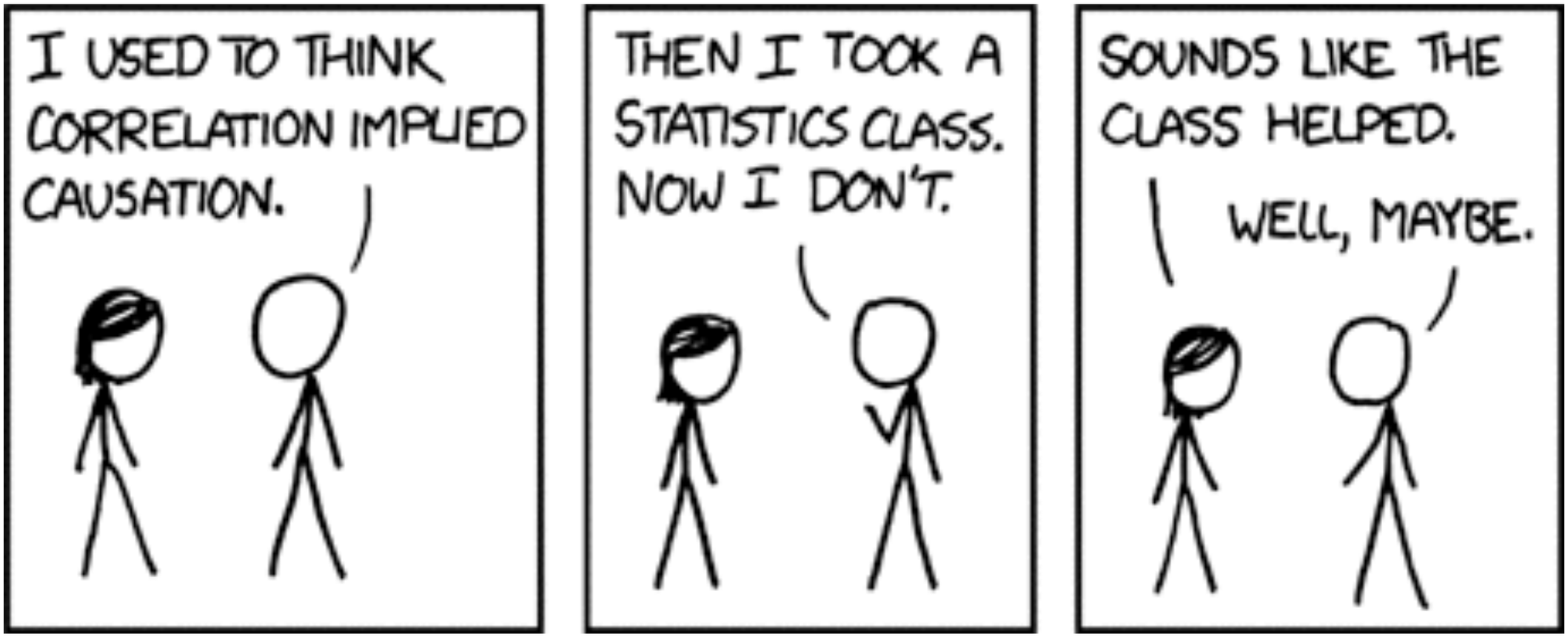
Designing observational research is about collecting and analysing information in a way that mimics experiments.

2. If doing case studies, we select cases strategically to maximize causal leverage.

E.g., two countries that are as similar as possible but different on the independent variable.

3. If quantitative data is available, we can use statistical models to mathematically isolate correlation between an independent and dependent variable.

E.g., regression analysis.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.