Experiments and Causal Inference

Ec626 | Spring 2020 | University of San Francisco
Kyle Carlson

Lesson 1 - Causality fundamentals

Table of contents

The importance of causality

In the context of business decisions and actions, people often speak of “pulling levers”. This idiom is a colloquial way of refering to causality. The people in the business want to take some action that will improve the business’s growth, efficiency, security, etc. The idiom highlights that people think causality is important but often do not have a rigorous, formal way to think about it. In this class we will learn how to think rigorously and practically about questions like these:

- If our social media site buys Google search ads, how much will it increase our monthly visitors?
- If our insurance company switches our actuarial model from a logistic regression to gradient boosted decision trees, how much will our annual claims payouts decrease?
- If our courier company gives a 25% wage increase to bike messengers, how many more hours will they work per week?

- If our online banking site implements a CAPTCHA for new accounts, will our conversion rate fall by less than 1 percent?
- If our hotel increases prices by 10 percent, how much will monthly booking revenue increase?

What is causality?

The definition of causality is debated by statisticians, economists, computer scientistis, and philosophers. We are in an economics course, so we will follow the conceptual framework most popular with economists, the potential outcomes model, also known as the (Neyman-)Rubin causal model. This framework underlies thousands of observational studies and experiments in economics, political science, and sociology.

Potential outcomes model

The main idea of the potential outcomes model is that each individual has two potential outcomes, i.e., outcomes that would obtain under different states of the treatment of interest. For example, suppose we are looking at individuals using a search engine and that the treatment is seeing a display ad for Doctors Without Borders (DWB). We are interested in whether the ad causes individuals to donate money to DWB. In the potential outcomes model each individual has two donation outcomes: one in the theoretical world in which they saw the ad and another in the world in which they did not. The ad exposure corresponds to a “treatment” state and the lack of ad to a “control” state.

To analyze this problem rigorously we can use a formal, mathematical model. We will index individuals by i. A given individual i has the outcomes y_i^0 and y_i^1, corresponding to the control state and treatment state, respectively. The y values represent how much money i donates:

- y_i^0 = donation amount in the control state (no ad),
- y_i^1 = donation amount in the treatment state (saw ad).

For each individual i we will suppose that there is only one instant in which they can be exposed to the ad and make a donation decision. Therefore, we can only observe, in our data, one of y_i^0 or y_i^1, not both. If the actual treatment state is $d_i \in 0, 1$, then the observed outcome is $y_i = y_i^1 d_i + y_1^0(1 - d_i) = y_i^{d_i}$.

We want to estimate causal effects. The basic units of causality are the individual-level treatment effects $y_i^1 - y_i^0$, which represent the difference in donation amount caused by exposure to the ad. However, it is impossible to directly calculate this treatment effect from data because one of the values is always missing from out data! This limitation is the *fundamental problem of causal inference*. We will return to this problem, but we are not yet done building up our framework. It needs to be embedded in a probability model.

Example data from DWB ads problem

The table below shows the potential donation amounts for five individuals. Such a data set can only be obtained by an oracle that knows what each individual will do under both conditions. In practice, we non-superhuman economists can only obtain the first 3 columns of data.

Individual (i)	**Actually exposed to ad?** (d_i)	**Actual donation amount** (y_i)	**Potential donation amount: No ad** (y_i^0)	**Potential donation amount: Exposed to ad** (y_i^1)	**Individual treatment effect** (δ_i)
1	1 (Yes)	$0	$0	$0	0
2	0 (No)	$0	$0	$5	5
3	1 (Yes)	$10	$10	$10	0
4	1 (Yes)	$10	$5	$10	5
5	0 (No)	$5	$5	$0	-5

One could calculate the average actual donation amount for the exposed (~$6.67) and non-exposed ($2.50) groups. The difference between these is ~$4.17. Is this number a good representation of the causual effect of the ads? One might say “no” and note that a sample size of 5 is very small. Although sample size is a concern, this problem is not the fundamental one at the moment. Imagine the data set is expanded to an abritrarily large sample size: millions, billions, or more. Now we can put aside our concerns about sample size. This is the study of *identification*, what we can learn from an arbitarily large quantity of data. If we imagine our data as realizations of random variables, then the average of any variable in the sample will be close to the *expectation* of the random variable. We know this because of the law of large numbers. Now we can work in terms of the random variables, which will make our analysis tractable and lets us focus on the essential issues in causality and experimentation.[1]

A probability model for potential outcomes

To make progress, let us be specific about building probability into the potential outcomes model. Suppose that y_i^0 and y_i^1 are realizations of the random variables Y^0 and Y^1. The treatment state is commonly representated by a random variable D. The observed outcome is $Y_i = Y_i^1 D_i + Y_1^0(1 - D_i)$. (Note: This follows the convention of using upper case characters to represent random variables and lower case to represent realizations of the variables.)

Framing this as a probability model gives us all the tools of probability and statistics. (This abstraction also reduces the number of objects to work with. Instead of $O(N)$ data points there are a handful of distributions.) While the individual treatments effects $y_i^1 - y_i^0$ cannot be calculated in practice, we do have hope of calculating the *average treatment effect* (ATE) $E[\Delta] = E[Y^1 - Y^0]$, where E is the expectation operator. Our hope comes from statistics, which gives us tools for relating data to probability models. In causal inference we typically want to make computations on our data $(d_i, y_i) : i = 1 ... N$ to estimate the parameter ATE of our probability model.

This particular construction of the probability model warrants a few comments about why it is used and what it implies for our analysis. Other constructions exist, but this one is common in econometrics and many other disciplines. This approach is also implicit in many industrial applications of causal inference, for example, conventional A/B testing approaches.

Our framework has some unintuitive interpretations. A typical introductory statistics course will use a framework with a “population” and a “sample”. For example, a population might be the 327 million individuals in the United States. A sample might be a random selection of 1000 of those individuals whose data is collected as part of a research survey. A statistical problem would then be using the average age among the sample (“sample mean”) to estimate the average age in the population (“population mean”). However, in our construction there is no requirement that our data “sample” actually be a subset of any well-defined population. In applications we may have *all* of the data. Many studies have all of the data in a relevant class, for example, all births in the United States (CDC Vital Statistics birth certificates data), all deaths in the United States (CDC death certificate data), or all people in the United States (U.S. Census data). Similarly, an internet business will typically have data about *all* of the people that visited the web site, not a sample. In these cases, the conventional approach is to adopt the useful fiction that the data were drawn from an imaginary “superpopulation”. The variables (Y, Y^0, Y^1, D) characterize the superpopulation. Why is it useful to learn about an imaginary population? The answer is fairly simple in the case of industrial applications. Suppose our search engine conducted an A/B test on all of the visitors in the past month and estimated the effect of the DWB advertisement. If we imagine each future month as a new random sample drawn from the superpopulation, we can generalize our learnings from the A/B test sample to future visitors to the search engine. Without assuming some commonality between past and future our analysis of the data will be useless for future decisions. The assumption of a shared superpopulation is perhaps the simplest useful form of commonality.[2]

Variations of average treatment effects

The table below shows key treatment effects of interest. The first three (ATE, ATT, ATC) are average treatment effects are different subsets of the population. The ATE is the treatment effect averaged across the entire population. The ATT is the treatment effect but averaged across only the individuals *that actually received treatment*. Many policy researchers focus on the ATT because it is the effect of a program, e.g., education or training, on those who received it. Finally, the ATC averaged across only the individuals that did *not* receive the treatment. It tells us how much those individuals would be affected in the counterfactual world where they were subject to the treatment.

Effect name	**Quantity**
Average Treatment Effect (ATE)	$E[Y^1 - Y^0]$
Average Treatment Effect on the Treated (ATT)	$E[Y^1 - Y^0 \vert D = 1]$
Average Treatment Effect on the Control (ATC)	$E[Y^1 - Y^0 \vert D = 0]$
Naive Average Treatment Effect (NATE)	$E[Y^1 \vert D = 1] - E[Y^0 \vert D = 0]$

Bias of the naive average treatment effect

Suppose we have a sample of data and are interested in the ATE. We naively contrast the average outcomes in the two treatment groups: $\frac{1}{N_1}\sum_{i:d_i=1} Y_i - \frac{1}{N_0}\sum_{i:d_i=0} Y_i$. The expectation of this quantity is the Naive Average Treatment Effect (NATE). The equation below shows how the NATE can be decomposed into the ATE plus two bias terms: selection bias and differential effect bias.[3] The two bias terms show that we cannot necessarily expect our estimate to be close to the true ATE. Later we will see how randomized experiments eliminate these bias terms.

$$E\left[\frac{1}{N_1}\sum_{i:d_i=1} Y_i - \frac{1}{N_0}\sum_{i:d_i=0} Y_i\right]$$

$$= \underbrace{E[Y^1|D=1] - E[Y^0|D=0]}_{\text{NATE}}$$

$$= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}} + \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}} + \underbrace{E[D] * (E[\Delta|D=1] - E[\Delta|D=0])}_{\text{Differential effect bias}}$$

Below is the same statement in a less precise but more readable form:

$$E\left[\bar{Y}_{\text{Treatment}} - \bar{Y}_{\text{Control}}\right]$$

$$= \underbrace{E[Y^1|\text{Treatment}] - E[Y^0|\text{Control}]}_{\text{NATE}}$$

$$= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}} + \underbrace{E[Y^0|\text{Treatment}] - E[Y^0|\text{Control}]}_{\text{Selection bias}} + \underbrace{P(\text{Treatment}) * (ATT - ATC)}_{\text{Differential effect bias}}$$

Selection bias

Selection bias is a systematic difference in Y^0 between the treatment and control groups. Selection bias occurs when the two groups are systematically different even in the absence of an actual treatment. In the DWB ads example this would happen if the ad is more likely to be shown to users that searched for “Doctors Without Borders”. Even in a world where there are no DWB ads, the people who searched for DWB are much more likely to donate than those who did not.

Differential effect bias

Differential effect bias is a systematic difference between how the treatment affects the treated group and how the treatment would affect the control group. This type of bias is typical in cases where some form of agency is involved:

- The units opt into the treatment based on expectations about its individual effects. Example: Suppose there is a worker training program that people can opt into. Those who do opt in to the program are likely to be those who expect the gain from it. That is, a reasonable hypothesis is $E[Y^1 - Y^0 \mid \text{Enrolled}] > E[Y^1 - Y^0 \mid \text{Not Enrolled}]$.
- Treatments are assigned to units based on expectations about the individual effects. Example: In the DWB example, suppose that people who search for “Doctors Without Borders” are more easily persuaded by the ads than are others who do not search for DWB. The marketing agency of DWB knows this and bids specifically on the “Doctors

Without Borders” term to target their ad budget in the most effective way possible. This will generate $E[Y^1 - Y^0 \mid \text{Exposed to ad}] > E[Y^1 - Y^0 \mid \text{Not exposed}]$.

Randomized experiments remove bias by enforcing independence

In the result above we calculated the naive difference between the treatment and control groups and showed that it has two bias terms. For both bias terms to be zero, we need the following conditions to be true:

$$\begin{aligned} &E[Y^0|D=1] = E[Y^0|D=0], \text{and} \\ &E[\Delta|D=1] = E[\Delta|D=0]. \end{aligned}$$

The key assumption for supporting these conditions is the *independence assumption*:

$$(Y^0, Y^1) \perp D.$$

The condition means that the potential outcomes are jointly independent of treatment assignment. Properly randomized experiments implement the independence assumption. Randomization should assign treatment or control to each unit independently of every other factor, for example, by a coin flip for each unit. This eliminates any relationship between (Y^1, Y^0) and D. In particular, this means that:

$$\begin{aligned} E[Y^0 \mid D=d] &= E[Y^0], d = 0, 1, \\ E[Y^1 \mid D=d] &= E[Y^1]; d = 0, 1. \end{aligned}$$

First, the independence assumption eliminates *selection bias*:

$$\begin{aligned} (Y^0, Y^1) \perp D &\rightarrow E[Y^0|D=d] = E[Y^0], d = 0, 1 \\ &\rightarrow E[Y^0|D=1] - E[Y^0|D=0] = 0. \end{aligned}$$

Second, the independence assumption eliminates *differential effect bias*. The argument is similar to the one above.

Conditional independence

The independence assumption very strong and unlikely to hold true except in the most strictly controlled experiments. Economists and other social scientists often work with an alternative version called the *conditional independence assumption* (CIA):

$$(Y^0, Y^1) \perp D|X.$$

This condition means that the potential outcomes are independent of treatment assignment *conditional on a vector of covariates* X. Typical covariates in social science include age, race, location, and family background. In many settings these variables will influence both the potential outcomes and treatment assignment. A classic example is the effect of a college degree on earnings. We know there are strong relationships between both of those variables and the covariates listed above. In common practice the researcher will include these variables in a regression as “control variables” and argue, with varying success, that treatment is “as good as randomly assigned” conditional on those variables.

The CIA is also called “selection on observables” (especially in econometrics), which means that we can account for selection bias using observable variables. “Observable” means we have data on those variables. The CIA is also called “unconfoundedness” or “ignorability”.[4]

Notes

1. This argument has parallels in *Learning From Data* (2012), the book on machine learning fundamentals by Abu-Mostafa, Magdon-Ismail, and Lin. Section 1.3 “Is Learning Feasible?” discusses the possibility of learning about an *unknown* target function from data. The way to make this feasible is to use probability. ↩
2. For a classic philosophical take on the problem of learning from past experience to understand the future, start with “Kant and Hume on Causality” in the Standford Encyclopedia of Philosphy. This topic is out of scope for this course. ↩
3. The derivation is available in *Causal Inference: The Mixtape* by Scott Cunningham. ↩
4. “Confounding” means to mix things together in a way that makes them difficult to distinguish. In our case this means mixing together how the potential outcomes and treatment assignment are determined. ↩