2. Causal Inference: Experimental and Observational Designs

Ryan T. Moore

2023-09-11

1 The Goal of Causal Inference

Recall that the two potential outcomes in a simple experiment, $Y_i(1)$ and $Y_i(0)$, are *potential*. They represent (for unit i) what would happen under treatment, and what would happen under control, respectively. Since we can never observe both potential outcomes for a given unit in our study,[1] we rarely focus on estimating the individual-level true treatment effect, $\tau_i = Y_i(1) - Y_i(0)$.

[1] What is this fact called?

Instead, we focus on estimating an average effect across many units. That is, we are usually interested in an *average treatment effect* (ATE). Specifically, we may be interested in an estimate of the average of the treatment effects for the individuals in our sample, the *sample average treatment effect*, or SATE.[2] To help express the average, we'll use the symbol $\sum$ ("sigma"), which just means "add up the values". For example, $\sum_{1}^{10}$ means "add up the first 10 observations"; more explicitly, $\sum_{i=5}^{8} x_i$ means "add up the 5th through the 8th values of x". To express, "add up the x values of units in the treatment group", we'll use $\sum_{i \in T} x_i$.

[2] The average (the "mean") takes all the quantities, adds them up, and divides by how many you have.

$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

One last bit of notation: we will let the subscript $\_T$ represent the units that actually receive treatment, and the subscript $\_C$ represent those that actually receive control. So, $Y_{iT}(1)$ is the potential outcome under treatment for a unit that actually **did** receive treatment. That's something we observe: what would that unit have done, if it received treatment – which it did. On the other hand, $Y_{iT}(0)$ is something we cannot observe: what the unit would have done under control (but it was in the treatment group).

$Y_{iT}(0)$ = the outcome under control for a treated unit (not observable)

The **true** SATE is the average of the individual treatment effects for the n units in our sample:

$$SATE = \frac{1}{n} \sum_{i=1}^{n} [Y_i(1) - Y_i(0)]$$

We'll simplify the notation a bit, and use the $\overline{overline}$ to represent the average of a quantity. So, the SATE is also

$$SATE = \overline{Y_i(1) - Y_i(0)}$$

Since this quantity relies on a set of individual effects $Y_i(1) - Y_i(0)$,

it cannot be directly observed.[3] The average is just made up of additions and a multiplication by $\frac{1}{n}$, so we can equivalently write

$$SATE = \underbrace{\overline{Y_i(1)}}_{\text{Avg if all Tr}} - \underbrace{\overline{Y_i(0)}}_{\text{Avg if all Co}}$$

That is, the SATE is the difference between the average outcome of all units, if all units were subject to treatment, minus the average outcome of all units, if all units were subject to control.

Let's consider $\overline{Y_i(1)}$. This is the average outcome if *everyone* received treatment, and $\overline{Y_i(1)} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1)$. We can break this up into two groups:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(1) = \frac{1}{n}\left(\sum_{i\in T} Y_i(1) + \sum_{i\in C} Y_i(1)\right)$$

This consists of the outcomes under treatment for those who *actually* received treatment, and the outcomes under treatment for those who *actually* received control.[4] There's a similar quantity for the $\overline{Y_i(0)}$. So, overall, we can write the true SATE using four quantities:

$$\begin{aligned}
SATE &= \overline{Y_i(1) - Y_i(0)} \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\right] \\
&= \frac{1}{n}\left[\left(\sum_{i=1}^{n} Y_i(1)\right) - \left(\sum_{i=1}^{n} Y_i(0)\right)\right] \\
&= \frac{1}{n}\left[\left(\sum_{i\in T} Y_i(1) + \sum_{i\in C} Y_i(1)\right) - \left(\sum_{i\in T} Y_i(0) + \sum_{i\in C} Y_i(0)\right)\right]
\end{aligned}$$

In order to estimate the SATE, we need an estimate of $\sum_{i\in C} Y_i(1)$ and the similar quantity $\sum_{i\in T} Y_i(0)$. If the treated outcomes give us a perfect estimate of what would have happened to the controls, and vice-versa, we can simply substitute in the values we observe for those we don't. We then estimate the true SATE using the simple observed "difference in means" estimator between the treated and control groups:

$$\begin{aligned}
\widehat{SATE} &= \frac{1}{n_T}\left[\sum_{i\in T} Y_i(1)\right] - \frac{1}{n_C}\left[\sum_{i\in C} Y_i(0)\right] \\
&= \overline{Y_T(1)} - \overline{Y_C(0)}
\end{aligned}$$

[3] What is this fact called?

[4] Which one of these can we observe? Which can we never observe?

2 Why Experiments?

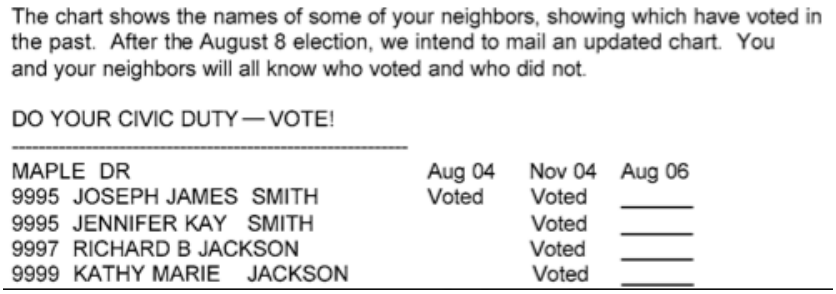
Experiments generate the best observable estimates of the quantities we can't observe. An experiment is far more likely to generate treatment and control groups that are similar than is an observational design.

If we, the researchers, randomized whether someone gets treatment or control, then we know that being assigned treatment or control is not related to your outcomes, on average. If the outcomes under treatment, the $Y_i(1)$, differ by whether or not you are in the treatment or control group, then $\overline{Y_T(1)} \neq \overline{Y_C(1)}$. In that case, we cannot use $\overline{Y_T(1)}$ as a good estimate of the unobservable $\overline{Y_C(1)}$. A randomized experiment gives us the best chance of the outcome being unrelated to whether you received treatment or control, and thus, the best chance of $\overline{Y_T(1)}$ being a good estimate of $\overline{Y_C(1)}$; similarly, an experiment gives us the best chance of the observable $\overline{Y_C(0)}$ being a good estimate of the unobservable $\overline{Y_T(0)}$.

2.1 The Social Pressure GOTV Experiment

Gerber et al. [2008] describe an experiment in which voters are randomly assigned to four different conditions. In the control condition, there is no contact. In the other 3, the voter gets a mailer with one of three messages:

- Civic duty: "DO YOUR CIVIC DUTY – VOTE!"
- Hawthorne: "YOU ARE BEING STUDIED! …VOTE!"
- Naming-and-shaming:



The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

Why randomly assign these messages, instead of just asking people in a survey "Has anyone encouraged you to vote by telling you to do your civic duty?", and then comparing turnout among those who say "yes" or "no"?

3 Observational Studies and Confounding

In *observational studies*, as opposed to randomized experiments, the researchers do not decide which units get treatment and which get control. The difficulty with doing causal inference in observational studies is that the units may not be put into treatment groups randomly; the groups may differ systematically. If the potential outcomes differ by whether you receive treatment or control, then we can't use one group to estimate the other. Formally, $\overline{Y_T(1)}$ won't be a good estimate of $\overline{Y_C(1)}$, and $\overline{Y_C(0)}$ won't be a good estimate of $\overline{Y_T(0)}$. In this case, something else, not the treatment, may cause a difference between the treatment and control groups. This is a *confounding* variable. Confounders are pre-treatment variables that partly determine (a) the outcome (or else they are irrelevant), and (b) the treatment (which group a unit is sorted into).

Confounders influence Y and T, the outcome and the treatment assignment.

Unknown or unmeasurable confounders are the biggest threat to valid causal inference in observational studies. Confounders prevent us from being able to distinguish *causation* from *association* or *correlation*.

Note that the *observed* outcomes may differ between treatment and control – that would be consistent with an actual effect! However, the underlying, only-half-observed *potential* outcomes should not vary between the two groups.

3.1 Addressing Confounding: Subclassification

Some observational designs are better at yielding valid causal inferences than other designs. In observational designs, sometimes unsuccessfully, we try to adjust our naive estimate, the treatment-control difference, statistically. One strategy for this is to *subclassify*, and estimate the treatment-control difference within classes defined by the confounder.

Substantively, suppose the social pressure experiment had only been an observational design. We could ask "Did turnout differ between those who had happened to encouter Civic Duty messages in their daily lives in 2006, versus those who did not encounter those messages?" Suppose we find, yes, those who heard Civic Duty messages were more likely to turn out. Was it the messages that *caused* them to turn out more? Probably not. The people who encountered Civic Duty messages were probably more likley to turn out anyway. Maybe they encountered those messages because they worked on campaigns, or they paid attention to politics, or, especially, because they voted last time. The two groups likely have different fractions that turned out in 2004.

Subclassification would say, "Don't just compare Civic Duty to No

Civic Duty turnout. Compare turnout just *in the subset of those who voted in 2004*, and then just *in the subset that didn't vote in 2004*". Then, we know that *within* those groups, prior voting is controlled. Within the subsets, prior turnout is equal, so prior turnout can't explain differences we observe between Civic Duty vs. No Civic Duty. The Civic Duty group that turned out in 2004 can best be compared to the No Civic Duty that also turned out in 2004. The Civic Duty group that did not turn out in 2004 can best be compared to the No Civic Duty that also did not turn out in 2004.

3.2 Addressing Confounding: Difference-in-Differences Designs

Before-After designs ask the simple question "Did the outcome go up or down?" In the NJ minimum wage study (Card and Krueger [1994]), a Before-After design would ask "After the new minimum wage, did employment rise or fall in NJ?"[5]

[5] What's weak about this question? What is being ignored?

Difference-in-differences (DiD) designs improve on Before-After designs. Where a Before-After design only uses 2 pieces of information, the DiD design uses 4. The Before-After estimate for NJ would be

$$\text{Before-After estimate} = \overline{\text{Employment}}_T^{\text{After}} - \overline{\text{Employment}}_T^{\text{Before}}$$

The DiD estimate is just the difference between two Before-After estimates. The DiD is the Before-After estimate for NJ, but we subtract off the change that occurred in the control state, PA. This second part removes the part of the change in employment that would have occurred anyway, in the absence of the law change. (E.g., perhaps the economy was growing, or fast food restaurants had been hit by health scares and had to lay off workers.)

$$\begin{aligned}\text{DiD estimate} &= (\overline{Y}_T^{\text{After}} - \overline{Y}_T^{\text{Before}}) - (\overline{Y}_C^{\text{After}} - \overline{Y}_C^{\text{Before}}) \\ &= (\underbrace{\overline{Y}_{NJ}^{\text{After}} - \overline{Y}_{NJ}^{\text{Before}}}_{\text{Change in Treated}}) - (\underbrace{\overline{Y}_{PA}^{\text{After}} - \overline{Y}_{PA}^{\text{Before}}}_{\text{Change in Control}})\end{aligned}$$

The "Change in Control" is our estimate of what would have happened in the treated unit (NJ), in absence of treatment (i.e., without the minimum wage change). The key assumption is "parallel trends" – the change in PA is what the change in NJ would have been, without the new minimum wage. Figure 1 shows this assumption graphically.

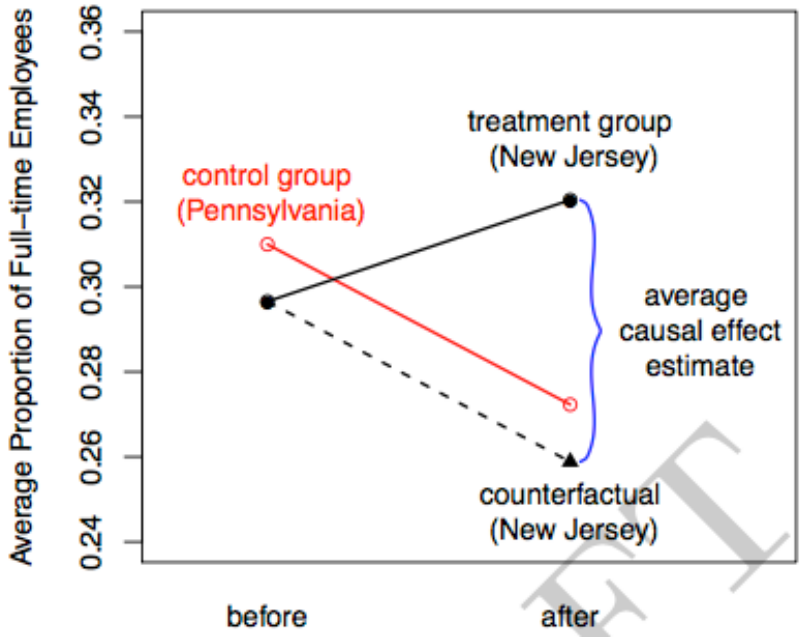


Figure 1: The change in PA (control, in red) is what we assume would have happened in NJ (treated, dotted −−), in the absence of the law change. This is an assumption of "parallel trends". The black segment for NJ shows what actually happened.

References

David Card and Alan B. Krueger. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, 90(5):1397–1420, 1994.

Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1):33–48, 2008.