1. Causal Inference: The Potential Outcomes Framework

Ryan T. Moore

2023-09-07

Causal inference, as we will use it, is estimating the *effects* of *causes*, rather than the causes of effects. That is, we will primarily consider questions like "What was the effect of the new minimum wage on employment?", "What was the effect of the health insurance program on household finances?", and "How much more likely is a resume with a 'white-sounding' name to get a job interview than a resume with a 'black-sounding' name?".

Questions like "What caused World War II?" and "Why did Donald Trump win the Republican primary?" tend to be less well-posed. Specifically, they tend to be over-determined. There are likely *many* factors which contribute to an outcome. We will seek to isolate factors and attempt to estimate their effects.

Causal inference in political methodology, as in medicine, statistics, economics, and other fields, refers to the causal factor of interest as the *treatment*, or the *intervention*.

1 Potential Outcomes

At the core of causal inference in the Rubin Causal Model (Rubin [1974], Holland [1986]) are *potential outcomes*. In order to estimate the effect of a factor, we would like to know what happens if that factor is present, and what happens if it is absent. These two states of the world are called *potential outcomes*. In Bertrand and Mullainathan [2004], for example, the treatment is receiving a resume with a 'black-sounding' name. The *control* condition is receiving a resume with a 'white-sounding' name. When the researchers themselves decide randomly which units of study get which condition, we describe the study as an *experiment*.[1]

In this experiment, the researchers randomly assigned some employers to receive black-sounding names, others to receive white-sounding names. However, the researchers could have assigned a given employer to the other condition – e.g., a white-sounding name instead of a black-sounding one. (For our causal inferences to make sense, it is important that each unit of study have the *possibility* of receiving either treatment.) We'll let T stand for the treatment, and $T = 1$ denote that the employer received a resume with a black-sounding name, and $T = 0$ denote that the employer received a resume with a white-sounding name.

$$T = \begin{cases} 1 & \text{emplr. gets "black" resume} \\ 0 & \text{emplr. gets "white" resume} \end{cases}$$

[1] We will use "experiment" in this sense — to describe randomized assignments; we will avoid using "experiment" more colloquially to mean "trying something".

The outcome of interest, which we'll denote Y, is whether the employer called back the fake applicant. $Y(1)$ is the outcome under treatment, and $Y(0)$ is the outcome under control.[2] For individual employer i, the true causal effect of the race of the name is τ_i (pronounced "tau sub-i"), and

$$\tau_i = Y_i(1) - Y_i(0)$$

So, for each employer, a row in Table 1, we can think about whether they *would have* called if they received a resume with a white- or a black-sounding name.

Table 1: Potential Outcomes for Employers under Different Resume Race-Sounding Names

Employer	T	Y(1)	Y(0)
1	?	Call	Call
2	?	Call	No Call
3	?	No Call	Call
4	?	No Call	No Call
⋮	⋮	⋮	⋮

Let's describe what each row of Table 1 means, substantively.

Now, let's let $Y(1)$ and $Y(0)$ be numeric, such that $Y(T) = 1$ for a given treatment if the employer calls back, and $Y(T) = 0$ if the employer does not call back.

$$Y(T) = \begin{cases} 1 & \text{employer calls back} \\ 0 & \text{employer does not call back} \end{cases}$$

Then, Table 1 becomes Table 2.

Table 2: Potential Outcomes for Employers under Different Resume Race-Sounding Names

Employer	T	Y(1)	Y(0)
1	?	1	1
2	?	1	0
3	?	0	1
4	?	0	0
⋮	⋮	⋮	⋮

2 The True Treatment Effect

If we assume that we know all the information in Table 2, then we can calculate the **true** causal effect for each employer, τ_i. That is, for each employer, we can calculate the effect of getting a black-sounding

[2] Put this in substantive terms.

versus a white-sounding name. For each of the four employers in Table 3, calculate the true effect of receiving a black-sounding resume. In other words, for each employer, how many more or fewer calls would the black-sounding name get than the white-sounding one? Then, calculate the true average treatment effect, the ATE, across the four employers.

Table 3: Potential Outcomes for Employers under Different Resume Race-Sounding Names

Employer	T	Y(1)	Y(0)	$\tau_i = Y_i(1) - Y_i(0)$
1	?	1	1	
2	?	1	0	
3	?	0	1	
4	?	0	0	
Avg				

The same sort of logic applies when the outcome is not binary. For example, in Table 4 below, the units are precincts, the treatment is whether a candidate's campaign bought TV ads or not (1 = TV ads, 0 = no TV ads), and the outcome is the number of votes the candidate received in that precinct. For each precinct, calculate the true effect of the TV ads on votes; then calculate the average across the precincts.

Table 4: Potential Outcomes for Precincts under Different Advertising Strategies

Precinct	T	Y(1)	Y(0)	$\tau_i = Y_i(1) - Y_i(0)$
W	?	500	200	
X	?	400	300	
Y	?	100	300	
Z	?	750	750	
Avg				

Unfortunately, in applied research, there is a problem. For a given employer, and a given resume, at a given moment in time, we can never know both a) what would happen if we sent a black-sounding name, and b) what would happen if we sent a white-sounding name. We can send one, or we can send the other, and we see what happens. "Why can't we send both, and see what happens to each?" That's a logical question. However, you've changed the treatment! You've suggested we send a *combination* of two identical resumes with different names. That's a different intervention, and it would likely lead to very different outcomes.[3]

This problem is so fundamental to causal inference, we call it the **fundamental problem of causal inference**: we can never observe both potential outcomes for a given unit of observation. To give a canonical example, suppose you have a headache and we are interested in whether aspirin removes the headache. You cannot, right now, both take the aspirin and not take the aspirin. You can do one, and we can see if your headache improves. Say you take the aspirin, and your headache gets better. We can never know for certain what would have happened had you *not* taken the aspirin at that same moment.

3 Treatment Effects We Can Estimate

In an actual experiment, we can only observe one potential outcome for each unit. We let Y_i represent the outcome we actually observe; sometimes we use Y_i^{obs} or $Y_i(T_i)$. In Table 5, fill in the value of Y_i^{obs} for the treatment vector T given here. That is, suppose that the first three employers received black-sounding names ($T = 1$) and the fourth received a white-sounding name. Which potential outcome would we observe for each employer?

Table 5: One Experimental Data Generating Process.

Employer	T	Y(1)	Y(0)	Y^{obs}
1	1	1	1	
2	1	1	0	
3	1	0	1	
4	0	0	0	

Avg for $T = 1$ Group:
Avg for $T = 0$ Group:
Difference :

Now, calculate the difference in the observed outcomes between the treatment group average and the control group average. From these data, what is our estimate of the effect of a black-sounding name? How much were we off from the **true** ATE?

Now, similarly, suppose that we instead observe the experiment in Table 6. In this version of the world, the first two employers received white-sounding names, the third and fourth received black-sounding names. Note that the potential outcomes are all the same – the only difference between the two tables is which potential outcome is revealed as Y^{obs}. Substantively, whether each employer will call back a resume with a black- or white-sounding name is fixed – either the employer will or won't. Through the assignment of employers to treatment or control, one of these becomes observable.

[3] If you received two identical resumes with different names, would you call either one back? If you had received a single strong resume with either name, you might call the applicant. But if you receive both, something strange is going on here ...

For Table 6, calculate the effect of a black-sounding name, and how much were we off of the **true** ATE by.

Table 6: Another Experimental Data Generating Process.

Employer	T	Y(1)	Y(0)	Y^{obs}
1	0	1	1	
2	0	1	0	
3	1	0	1	
4	1	0	0	

Avg for $T = 1$ Group:
Avg for $T = 0$ Group:
Difference :

4 *What Empirical Data Look Like*

Since we can never observe both potential outcomes, in a real study, our data will come in a form like Table 7.

Table 7: The Data from Table 6 As We Would Encounter It

Employer	T	Y^{obs}
1	0	1
2	0	0
3	1	0
4	1	0

Note that Table 7 is perfectly consistent with Table 6. However, it's also consistent with many other tables of potential outcomes. Imagine you conduct a real study, and you observe Table 7. This information is repeated in Table 8. Complete the parts of Table 8 that you can, using **only** this information.

Table 8: Using *only* the data here, fill in what we know.

Employer	T	Y^{obs}	Y(1)	Y(0)
1	0	1	_	_
2	0	0	_	_
3	1	0	_	_
4	1	0	_	_

In a real experiment, we will only know half the data for certain. At most. If there are more conditions, there will be more unobserved potential outcomes. If we have treatment A, treatment B, and control ($T = C$), then our data might appear as

Table 9: Some data from a 3-condition experiment.

Employer	T	Y^{obs}
1	C	1
2	C	0
3	A	0
4	A	0
5	B	1
6	B	0
7	A	1
8	B	1
	⋮	

Using the data from Table 9 (repeated below), fill in **only** what you can below.

Employer	T	Y^{obs}	Y(A)	Y(B)	Y(C)
1	C	1	_	_	_
2	C	0	_	_	_
3	A	0	_	_	_
4	A	0	_	_	_
5	B	1	_	_	_
6	B	0	_	_	_
7	A	1	_	_	_
8	B	1	_	_	_

References

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4): 991–1013, 2004.

Paul Holland. Statistics and causal inference. *The Journal of the American Statistical Association*, 81(396):945–960, 1986.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.