The Potential Outcomes Framework

Adriane Fresh 2020

The potential outcomes framework provides a mathematical framework for understanding cause and effect within the theoretical framework of the counterfactual theory of causality.

The potential outcomes framework is built around the idea that we all have *potential* outcomes were we to be exposed to treatment and were we to be exposed to a control condition (i.e. not exposed to treatment). These potential outcomes are fixed, defined values within each of us. But the challenge is that we cannot observe counterfactual worlds to see all potential outcomes for individuals. Instead, we observe only realized outcomes, which may muddy or confound our ability to precisely conclude whether a given cause produces a given effect.

The framework allows us to see counterfactuals from a different (mathematical) perspective, to define selection bias, and provides a foundation for formalizing additional research designs and their assumptions that allow us to draw a causal inference in spite of the fundamental problem of causal inference.

Part 1. An Abstract World

1.1 The Potential Outcomes Set Up

Let's begin in the simple world of a binary (potential or theoretical) "treatment" defined below

$$sT \in \{0, 1\}$$

where we call T the treatment, and we say that the treatment can be either $T = 1$, in which case we call it the *treatment condition*, or $T = 0$, in which case we call it the *control condition*. We can think of these as alternative theoretical potential *states of the world*.

Note: Although we use the terminology *treatment* and *control*, there is no, I repeat, **no** notion of randomization here. We can read $T = 1$ as "with the causal factor of interest present" and $T = 0$ as "without the causal factor of interest present," where the reason for the presence or absence of the causal factor of interest is unknown. I suspect this is a source of confusion, so I will reiterate it. We call this a *treatment*, but in doing so we do not imply that the receipt of treatment came from a randomized application. Instead, we are simply saying that $T = 1$, for instance, indicates a state in which the causal factor of interest is present.

For a given "population" – defined in the abstract mathematical sense – we define two population-level random variables as the *potential outcomes* under the states of the world *treatment* and *control* as follows

$$Y_{T=0} \equiv Y_0$$
$$Y_{T=1} \equiv Y_1$$

We use the capital letters to indicate that these are random variables. We read Y_0 as the "random variable

of the potential outcome under the control condition," and Y_1 similarly.[1] For an individual i, we can define the realization of those random variables as

$$y_{i,T=0} \equiv y_{i,0}$$
$$y_{i,T=1} \equiv y_{i,1}$$

where we use lower case letters, and index by i to indicate that these are individually defined potential outcomes. We read $y_{i,0}$ as the "potential outcome value of the random variable for individual i under the control condition," and $y_{i,1}$ similarly. Remember that we are still in the realm of *potential outcomes* here. So, by "realization", I mean the individual-level *value* of the given random variable, *not* the *observed* value.[2]

We can think about these individual-level realizations of the random variable as the potential outcomes in different theoretical states of the world that we all carry within ourselves. We have a potential outcome for if the world is $T = 1$, and we have a potential outcome for if the world is $T = 0$. They both exist simultaneously in us.

1.2 The Observed Outcomes Set Up

Now that we have defined the potential outcomes, let's discuss the set up for observed outcomes. Where T was in effect the theoretical treatment condition that allowed us to define *potential* outcomes for a theoretically applied causal factor, we use D to refer to the *observed* receipt of either treatment or control.

$$D \in \{0, 1\}$$

As with T, whether D is equal to zero or one is not (necessarily) a function of randomization. We should read $D = 1$ as "observed with the causal factor of interest" and read $D = 0$ as "observed without the causal factor of interest."

Because D is observed, not just potential, it is not possible for an individual to be *both simultaneously* be observed in treatment and control. Therefor, for individual i

$$d_i = 0 \text{ or } d_i = 1.$$

Either we have an individual observed with the causal factor of interest present, or without the causal of interest present, but not both.

It follows that, for a given value of D – that is, for the given observed presence of the causal factor of interest or not – we will only ever be able to *observe*

[1] Other ways of saying it: "the potential outcome random variable when the causal condition is not present." Or: "the potential outcome random variable in the absence of treatment."

[2] Random variables are variables that can potentially take on a set of values according to a probability distribution function (pdf). As a random variable, they are imbued with all possibilities according to that function. When we look at the individual value of that variable, we talk about it being "realized."

$$y_{i,0} \text{ if } d_i = 0 \text{ or } y_{i,1} \text{ if } d_i = 1.$$

But for any individual i, we cannot observe both.

1.3 Two Theoretical Quantities of Interest

Given our potential outcomes and our realized application of the causal factor of interest, D, we can define the causal effect for individual i as follows

$$\delta = y_{i,1} - y_{i,0}.$$

That is, for individual i, the causal effect, defined as δ, is the difference between the individual *potential outcome* for the state of the world where the causal factor is present, minus the *potential outcome* for the state of the world where the causal factor is not present.

When we move from the individual to the population, the causal effect needs to be defined as some sort of summary statistic that incorporates the information from the realization of the random variables Y_T for *all* individuals. We oftentimes think *averages* are interesting summary statistics, so we define two different averages for the population.

The first is the Average Treatment Effect, abbreviated as the ATE. Because we have moved from the individual level to the population, and because we are dealing with a full population and not a sample of a population, we use the expectation operator $E()$ to refer to the population average, or *expected value*.

We define the ATE using the expectation operator as follows

$$\begin{aligned} \text{ATE} &= E(\delta) \\ &= E(Y_1 - Y_0) \\ &= E(Y_1) - E(Y_0) \end{aligned}$$

where we can bring the expectation inside the parentheses because the expectation is a linear operator. The equation above is similar in look and interpretation to δ at the individual level further above. And the logic is exactly the same. We have simply moved from assessing the causal effect for an individual to assessing an *average* causal effect for a population, and thus taken the expected value of the individual-level causal effects for our entire population.

In addition to the ATE, we can define another quantity of interest, the ATT. The ATT is the Average Treatment Effect on the Treated. We define this additional quantity because it is helpful mechanically as we manipulate different equations to relate the conditional average to the ATE. We'll see that in a moment.

We define the ATT as follows

$$\begin{aligned} \text{ATT} &= E(\delta|D=1) \\ &= E(Y_1|D=1) - \underbrace{E(Y_0|D=1)}_{\text{Unobserved counterfactual}} . \end{aligned}$$

There are a few things to notice about the above definition for the ATT. First, we have conditioned on *observed exposure* to treatment. But, the ATT is still a theoretical quantity. There are components of this equation above that are *not* directly observable, as noted.

Second, the component of the equation that is not directly observable is the latter part of the equation – $E(Y_0|D = 1)$. We read that part of the equation as saying "the expected value of the *potential outcome* for the state of the world where $T = 0$ for an individual who we observe in the treatment condition (i.e. with the causal factor of interest present)." Individuals have a *potential* outcome for states of the world in which they receive or don't receive treatment. And they have those *potential* outcomes (they have both potential outcomes) regardless of the treatment condition that we *observe* them in. Remember, we have potential outcomes for all possible treatments within us at all times.

Finally, the only difference between the ATE and the ATT is that we are conditioning on a particular part of our population when we define the ATT. Specifically, we are conditioning on the portion of the population that we observe in the treatment condition. Logically, we should not want the ATT and the ATE to differ. Although they are conditioned on different parts of the population, the ideal is that the average causal effect for those actually receiving the treatment would be no different than the average causal effect for the entire population. That is the hope, at least.

1.4 Estimating the ATE

Until now we have written down theoretical quantities of interest. But it has not been the case that we have been able to observe those quantities. Now, we are interested in what happens when we try to estimate the ATE (that is, measure $\widehat{ATE}$).

We return to our definition of the ATE from above. *However*, because we are now in the world of estimating the ATE, we are now in a world in which we can *only* observe the potential outcomes for a given state of the world *when the observed causal state is the same.* That is, we only observe Y_1 when $D = 1$ and we only observe Y_0 when $D = 0$. Therefore, we can effectively add conditionals to our $\widehat{ATE}$ to reflect this required match between the potential outcome we estimate (and observe), and the observed causal state.

We write $\widehat{ATE}$ with these conditionals as follows

$$\widehat{ATE} = E(Y_1|D = 1) - E(Y_0|D = 0)$$

and we do a small mathematical trick of adding zero ($E(Y_0|D = 1) - E(Y_0|D = 1)$) to the equation to better allow us to see the potential bias that might make the ATE *not* equal, as we would hope, to the ATT.

$$\begin{aligned}\widehat{ATE} &= E(Y_1|D = 1) - E(Y_0|D = 0) + \underbrace{E(Y_0|D = 1) - E(Y_0|D = 1))}_{=0} \\ &= \underbrace{E(Y_1|D = 1) - E(Y_0|D = 1)}_{\text{the ATT}} + \underbrace{E(Y_0|D = 1) - E(Y_0|D = 0)}_{\text{possible baseline difference selection bias}}\end{aligned}$$

Rearranging the first line to obtain the second line allows us to see that the first two terms comprise the familiar ATT from above. As noted, we would like $\widehat{ATE}$ and ATE to be equal. But for this to be the case, we need two things: we need our Average Treatment Effect (ATE) to be the same as our Average

Treatment on the Treated (ATT), and we need our baseline difference selection bias to be zero.

Let's start with that baseline difference selection bias term on the right. If – and it is a big if – it is the case that $E(Y_0|D=0) = E(Y_0|D=1)$ then the term at the end will be zero.

So what is this baseline difference selection bias term? It's the difference in the outcome Y for the treated group ($D=1$) and the untreated group ($D=0$) *in a world where neither was treated (T=0).* Or, expressed more succinctly, it's the expected difference in our outcome in a world where we never did anything to anyone. So if our two groups ($D=1$ and $D=0$) would have looked the same in terms of Y absent intervention, then that term is zero.

Now let's look at ATT: $E(Y_1|D=1) - E(Y_0|D=1)$. Clearly, even if our baseline difference selection term is zero, that's not the same as our ATE ($E(Y_1) - E(Y_0)$).

For those to be equal, it must be the case that:

$$\begin{aligned} E(Y_1|D=1) - E(Y_0|D=1) &= E(Y_1|D=0) - E(Y_0|D=0) \\ &= E(Y_1) - E(Y_0) \end{aligned}$$

What would that equality mean intuitively? It means that *the effect of the treatment* is the same for our two groups ($D=1$ and $D=0$).

So taken together, our estimate $\widehat{ATE}$ is the same as the true value we want to estimate (ATE) if and only if

1. our groups would have looked the same absent intervention (no baseline difference selection effect), and
2. the two groups would both respond to treatment in the same way (ATT is the same as ATE).

Or, put differently: $\widehat{ATE} = ATE$ if and only if *all potential outcomes for our two groups are the same.*