Potential Outcomes

Nick Eubank

Potential Outcomes

For a unit of analysis i, we WANT to compare:

- outcome y_i under treatment $t = 1$ (denoted $y_{i,t=1}$) to
- outcome y_i under no treatment $t = 0$ (denoted $y_{i,t=0}$).

We call these the potential outcomes for i under different treatments.
In an ideal world, we'd call $\delta = y_{i,t=1} - y_{i,t=0}$ our causal estimate.

- *Counter-factual model of causality*

Potential Outcomes Framework

… but we can't see both $y_{i,t=1}$ and $y_{i,t=0}$. Each person can only experience one outcome.

So we'll do two things. First, let's move to populations. Ideally we want:

$$\begin{aligned} E(\delta) &= E(Y_{T=1} - y_{T=0}) \\ &= E(Y_{T=1}) - E(y_{T=0}) \end{aligned}$$

Called *Average Treatment Effect*, or *ATE*

Potential Outcomes Framework

But we *still* can't actually see ATE. What we *can* see is:

$$\widehat{ATE} = E(Y_{T=1}|D=1) - E(Y_{T=0}|D=0)$$

where $D \in \{0,1\}$ tell us whether a given observation *actually* experienced the treatment or not.

Two concepts:

- $T \in 0, 1$: *Potential* states of the world.
- $D \in 0, 1$: *Actual* populations of people.

Potential Outcomes Framework

What we *want* is for $\widehat{ATE} = ATE$. When is that true?

$$\begin{aligned}
\widehat{ATE} &= E(Y_{T=1}|D=1) - E(Y_{T=0}|D=0) \\
&= E(Y_{T=1}|D=1) - E(Y_{T=0}|D=0) + \\
&\quad E(Y_{T=0}|D=1) - E(Y_{T=0}|D=1) \\
&= \underbrace{E(Y_{T=1}|D=1) - E(Y_{T=0}|D=1)}_{\text{Avg Treatment on the Treated}} + \\
&\quad \underbrace{E(Y_{T=0}|D=1) - E(Y_{T=0}|D=0)}_{\text{Baseline Difference}}
\end{aligned}$$

Potential Outcomes

$$\underbrace{E(Y_{T=1}|D=1) - E(Y_{T=0}|D=1)}_{\text{Treatment on the Treated}} + \underbrace{E(Y_{T=0}|D=1) - E(Y_{T=0}|D=0)}_{\text{Baseline Difference}}$$

Baseline Difference: Absent treatment, would those who actually got treatment have turned out the same as those who hadn't received treatment.

Potential Outcomes Framework

$$\underbrace{E(Y_{T=1}|D=1) - E(Y_{T=0}|D=1)}_{\text{Avg Treatment on the Treated}} + \underbrace{E(Y_{T=0}|D=1) - E(Y_{T=0}|D=0)}_{\text{Baseline Differences}}$$

Treatment on the Treated: What we measure. This is equal to Average Treatment effect iff

$$E(Y_{T=1}|D=1) - E(Y_{T=0}|D=1) \quad = \quad E(Y_{T=1}|D=0) - E(Y_{T=0}|D=0)$$

in which case

$$E(Y_{T=1}|D=1) - E(Y_{T=0}|D=1) = E(Y_{T=1}) - E(Y_{T=0}) \tag{1}$$

In other words, $ATT = ATE$ if the response to treatment of people for whom $D = 1$ is the same as that of those for whom $D = 0$.

Potential Outcomes Framework

What we estimate is equivalent to $ATE = E(Y_{T=1}) - E(Y_{T=0})$ if:

1. No baseline difference (absent treatment, same outcomes)
2. Same treatment response (no difference in how treated and untreated would respond if treated)

$\Rightarrow$ Both groups to have same potential outcomes

Potential Outcomes Framework

Suppose we measured the effect of an exercise program on health by just comparing health of people in an exercise class with health of people not in an exercise class.
How might these two things have been violated:

1. No baseline difference
2. Same treatment response

Potential Outcomes Framework

Suppose we measured the effect of advertising on sales by correlating sales with advertising expenditures.
How might these two things have been violated:

1. No baseline difference
2. Same treatment response