Potential Outcomes

Nick Eubank

Mastering ’Metrics Take-Aways

If we want to know the causal effect of some treatment D, we can’t just compare people who got the treatment with those who did not.

Why not?

Mastering 'Metrics Take-Aways

There may be unobserved differences between people who got the treatment ($D = 1$) and those that did not ($D = 0$) besides differences caused by the treatment that affect our outcome.

- *Selection Effect*

$\Rightarrow$ Comparing these two groups would conflate effects of treatment with other unobserved differences.

Mastering ’Metrics Take-Aways

Why might it be a problem to...

- Estimate effect of Diet Coke on weight by comparing Diet Coke drinkers to non-Diet Coke drinkers?
- Estimate effect of buying TV ads by comparing revenues of companies that buy ads with those that don’t?
- Estimate effect of cholesterol meds on heart attack risk by comparing heart attack rates among those who take cholesterol meds to those that don’t?

Mastering ’Metrics Take-Aways

Experiments fix this. How? *On average*, random assignment

ensures that in large enough samples, treated and control subjects will be the same as one another.

Study Validity

- **Internal Validity:** Are we accurately estimating the quantity of interest in study?
- **External Validity:** Do we think results will generalize to other contexts?

RAND Study:

- High internal validity (randomized)
- Mixed external validity: study population was average people, but most uninsured today are young, poor, less educated, so may not speak to results of real policies to expand insurance.

Study Validity

Often (though not always) tension between internal and external validity:

- **Internal Validity:** Maximized by controlling the environment
- **External Validity:** About “realism” of study

Study Validity

Often (though not always) tension between internal and external validity:

- **Internal Validity:** Maximized by controlling the environment
- **External Validity:** About “realism” of study

Study Validity

Which of the following has higher internal validity, which has higher external validity, and why?

1. Psychology experiment where undergrads are put in a lab and randomly assigned to solve puzzles, some with emotionally disturbing imagery and some with happy imagery to test effects of emotional stress on problem solving.
2. Political scientists interested in how social pressure effects whether people vote mail fliers to a random set of voters that includes data on what elections they've voted in in the past several years so they'll know whether they vote is public. They then see if they turnout at higher rates than a control group.

Potential Outcomes Framework

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 assignment of treatment.

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

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

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