

The Potential Outcome Framework, Part 1

Stat 256, Causality

Chad Hazlett

UCLA

`chazlett@ucla.edu`

Spring 2026

Chapter 1-3 from Hernan & Robins, Causal Inference: What If.

The Neyman-Rubin Potential Outcome Model

Much of the applied causal inference literature speaks the language of the **Neyman-Rubin causal model**, aka the **Potential Outcomes Model (POM)**.

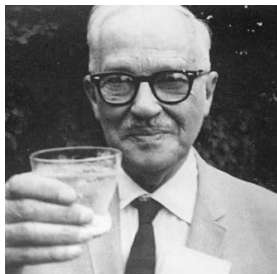


Figure: Neyman



Figure: Rubin

What causal inference actually asks

What we do naturally:

- Split units into a treated group and a control group
- Compare average outcomes *across groups*
- “College grads are more liberal than non-grads.”

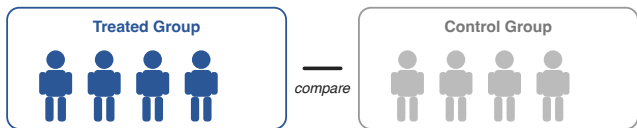
What causal inference asks:

- For each unit, compare how they would do *if treated* to how they would do *if untreated*
- “Would *this person* be more liberal *if they got a degree* than *if they didn't*?”

The first is a comparison *between groups of different units*. The second is a comparison *within each unit*, between two versions of its world.

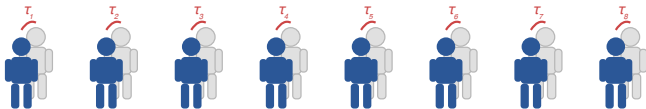
Group comparison vs. within-unit comparison

Group comparison



Contrasting two groups

Causal (counterfactual) comparison



Contrasting two (potential) outcomes for everybody



= treated



= untreated

τ_i = individual contrast

Potential Outcomes, Labeled



- Y_{1i} : the outcome unit i *would have* under treatment
- Y_{0i} : the outcome unit i *would have* under control
- $\tau_i = Y_{1i} - Y_{0i}$: the causal effect for unit i

Fundamental problem: we only ever see *one* potential outcome per unit.

From potential to observed

Adding the treatment indicator $D_i \in \{0, 1\}$, the observed outcome is

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

Key idea: D_i doesn't change the potential outcomes — it just picks which one you see.

Which of these are *always*, *sometimes*, or *never* observable?

- D_i Y_i Y_{1i} Y_{0i} τ_i

POs and QOIs

Some quantities of interest will be:

- Average treatment effect (ATE): $\mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[\tau]$
- Average treatment effect on the treated (ATT)

$$\mathbb{E}[\tau_i | D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$$

- Average treatment effect on the controls (ATC)

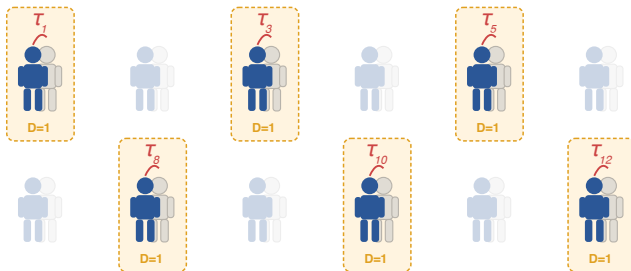
$$\mathbb{E}[\tau_i | D_i = 0] = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 0]$$

- Conditional average treatment effect (CATE):

$$\mathbb{E}[\tau_i | X = x] = \mathbb{E}[Y_{1i} - Y_{0i} | X = x]$$

Visualizing the ATT

The ATT asks: among units actually treated ($D_i = 1$), what's the average of their individual causal effects? Not a group comparison — imagine each treated unit's two potential lives:

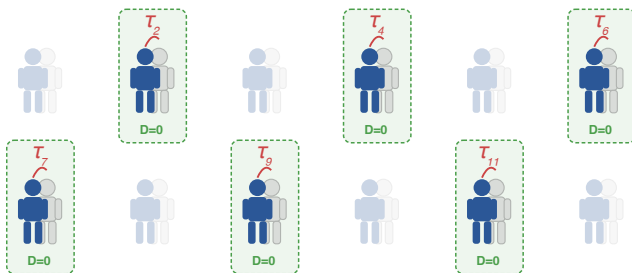


ATT = average of

$T_1, T_3, T_5, T_8, T_{10}, T_{12}$

Visualizing the ATC

The ATC asks the same question for the untreated ($D_i = 0$). These units were *not* treated, but we still ask what *would have happened* if they had been:



ATC = average of

$\tau_2, \tau_4, \tau_6, \tau_7, \tau_9, \tau_{11}$

The one embodied assumption: Consistency

By writing $Y_i(d)$ we assert **consistency**: $Y_i(d)$ is the Y_i you'd see if i actually received treatment d . We name it so we can reference it later and notice where it breaks.

Main threat: **interference/spillovers**. If unit j 's treatment affects unit i 's outcome, then $Y_i(d)$ isn't well-defined without also specifying everyone else's status. (Think vaccines, education, monitoring.)

You'll encounter **SUTVA** (Rubin 1980) = no interference + "one version of treatment." Essentially this is just consistency when d has a single well-defined value — a re-statement, not a new assumption.

("One version of treatment" is easy to overread; we'll save that for later.)

Not every PO is a good PO

You can write $Y_i(d)$ for anything. That doesn't mean it's well-defined or that it's what you want.

Requires clarification — what exactly are you changing?

- Race, gender: when do you “change” it? In whose eyes?
- Obesity: which feature of it?

Arguably ill-defined — the PO may not exist:

- “the size of the tumor in the brain of a patient who in fact died” (truncation by death).

A common guideline: a cause should be **manipulable** — you should be able to describe an ideal experiment that would set $D = d$.

- Helpful prodding, not a hard rule: non-manipulable things can have effects too, and the SCM can define those cleanly.

Brain Training

Test 1: Is $\mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_i|D_i = 1]$?

What licenses this?

Brain Training

Test 2: Suppose D_i randomly assigned.

- How would the Y_{1i} you observed (i.e. when $D_i = 1$) compare to the Y_{1i} you don't observe? And for Y_{0i} ?
- What can we say about $\mathbb{E}[Y_{1i}|D_i = 1]$ compared to $\mathbb{E}[Y_{1i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}]$?

(While we're here: this is ignorability, aka exchangeability!)

- So can we estimate $\mathbb{E}[Y_{1i}]$? And $\mathbb{E}[Y_{0i}]$? How? What licenses do you use?

Brain Training

Test 3: Suppose $Y_{1i} = Y_{0i}$ for all i . But now, someone magically sees Y_{1i} , and chooses $D_i = 1$ more often when Y_{1i} is higher.

- How would the Y_{1i} you see compare to Y_{1i} you don't see?
- So does $\mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 0] = \mathbb{E}[Y_{1i}]$?
- How would the Y_{0i} you see compare to the ones you don't?
- If you estimate $\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$, do you expect something near zero, positive, or negative?
- Does this tell us anything about $\mathbb{E}[Y_{1i} - Y_{0i}]$?
- Can you say anything about expected direction of bias?

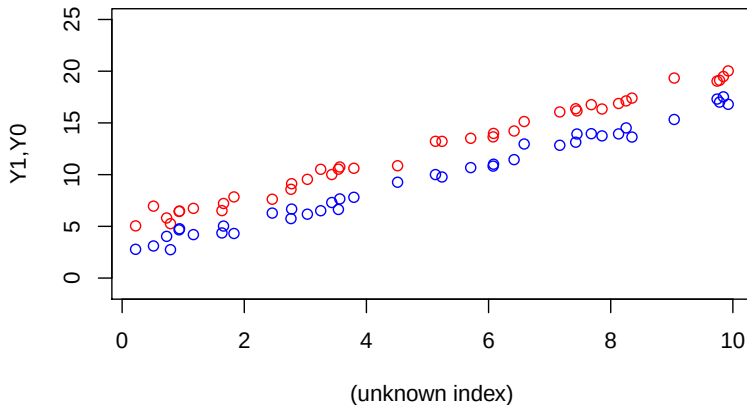
Brain Training

Test 4: Suppose you don't know anything about how D_i is assigned w.r.t. Y_{1i} and Y_{0i} .

- Can we say anything about how $\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$ compares to $\mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$?
- Unfortunately, this is often the scenario we are in with observational data.

Brain Training: Visual Practice

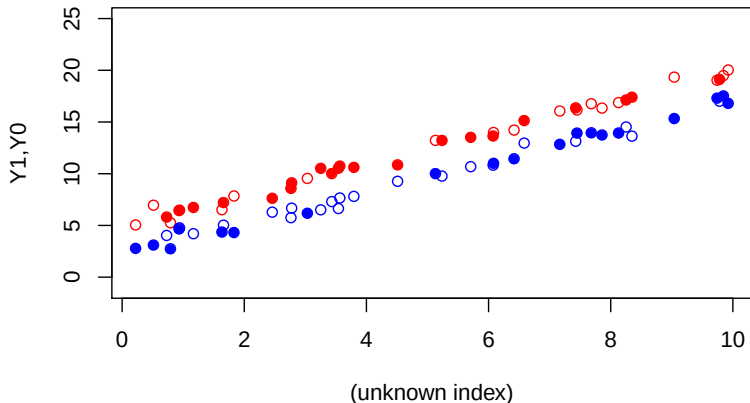
Let's look at Y_{1i} and Y_{0i} alone first.



True $\mathbb{E}[Y_{1i} - Y_{0i}] = \text{ATE} = 3$

Random Assignment of Treatment

Now suppose $D_i = 1$ is randomly assigned. We see only the filled dots:

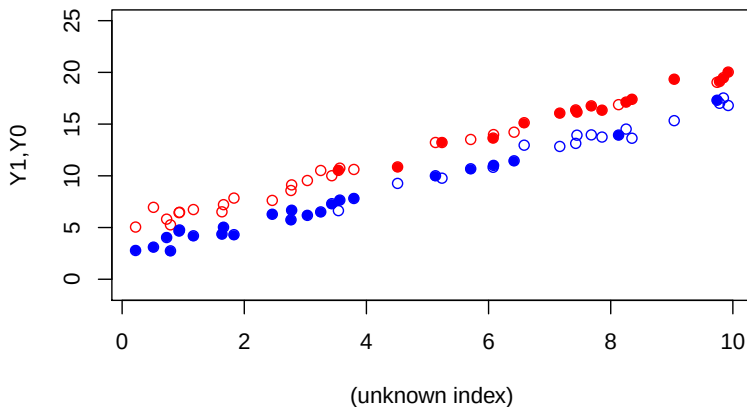


We get

$$\hat{\mathbb{E}}[Y_i | D_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0] = \widehat{\text{DIM}} = 3.01 \approx \text{ATE}$$

Selection into treatment, case 1

Suppose $Pr(D_i = 1)$ increases to the right. Now we observe (filled dots):

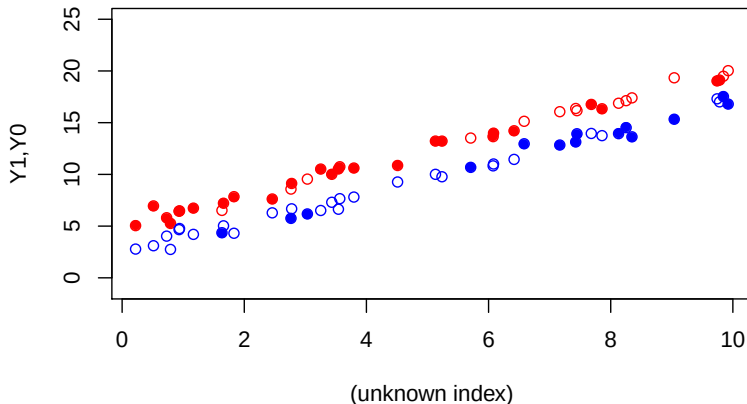


What difference in means do you expect?

$$\hat{\mathbb{E}}[Y_i | D_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0] = \widehat{DIM} = 9.97 \neq ATE = 3$$

Selection into treatment, case 2

Now suppose $Pr(D_i = 1)$ decreases to the right. We observe (filled dots):



What difference in means do you expect now?

$$\hat{\mathbb{E}}[Y_i | D_i = 1] - \hat{\mathbb{E}}[Y_i | D_i = 0] = \widehat{DIM} = -1.59 \neq ATE = 3$$

Yet More Brain Training: “science” tables

Imagine a study population with 4 units:

i	D_i	Y_{1i}	Y_{0i}	τ_i
1	1	10	4	6
2	1	1	2	-1
3	0	3	3	0
4	0	5	2	3

1. What is the ATE? $\mathbb{E}[Y_{1i} - Y_{0i}] = 1/4 \times (6 - 1 + 0 + 3) = 2$

(Note: average effect is positive, but not all τ_i are)

2. What are the ATT and ATC?

$$\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1] = .5(6 - 1) = 2.5$$

$$\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 0] = .5(0 + 3) = 1.5$$

Difference in Means

You saw earlier how simple comparison of observed outcomes can be misleading. Let's use POM to see this more rigorously.

The **Difference in Means** estimand:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

What does the POM tell us about this quantity? One decomposition:

$$\begin{aligned}\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + (\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0])\end{aligned}$$

What are the terms in **blue** & **green**?

Summing Up: The Potential Outcomes Model

- **Causality is a within-unit comparison.** $\tau_i = Y_i(1) - Y_i(0)$ — not a contrast between a treated group and a control group, but between two versions of the same unit's world, then averaged.
- **The fundamental problem.** You only ever see one of $Y_i(1)$, $Y_i(0)$. Every causal question is a missing-data problem.
- **One embodied assumption: consistency.** Writing $Y_i(d)$ already asserts it. Interference is the main way it breaks; SUTVA is just consistency when d is well-defined.
- **POs are a language, not a free lunch.** You can write $Y_i(d)$ for anything, but some POs are ill-defined (truncation by death) or not what you actually want. The notation doesn't rescue a bad question.
- **Next:** what has to be true about how D is assigned for the observed difference to recover a causal quantity? → randomization, selection-on-observables, and the rest of the course.