# The Potential Outcome Framework

## PS200C, Causal Inference, Spring 2026

Chad Hazlett

UCLA

chazlett@ucla.edu

# What is causality?

What do we mean when we say $X$ causes $Y$, or $Y$ was caused by $X$?

Thoughts?

Reading: Chapers 1 and 2 from “What If" (Hernan & Robins)

# Flavors of causality

**Effect of causes.** One kind of causal question the most common type of causal question in applied areas-regards the *effect* of some potential cause, e.g. what is the effect of:

- political institutions on corruption?
- effect of a drug or other therapy on a health outcome
- an ad on purchasing behavior
- peace-keeping missions on peace?
- body cameras on police behavior?

**Causes of effects/ causal attribution**. Another set of questions asks "what led to an observed outcome".

Other/more general/complicated concepts can be embedded in a deep enough framework (namely structural causal models)

# Two languages

Most work on causality now employs one of two languages:

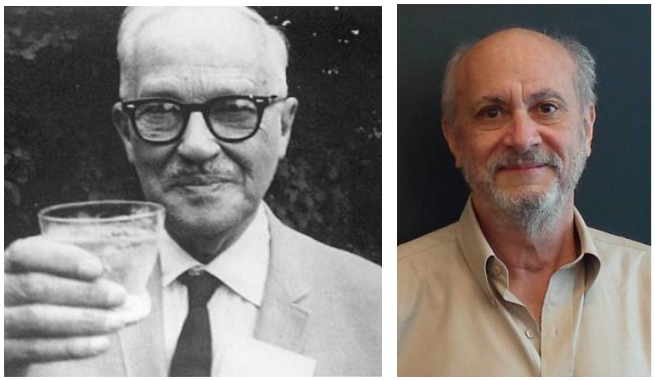


Figure: Neyman and Rubin

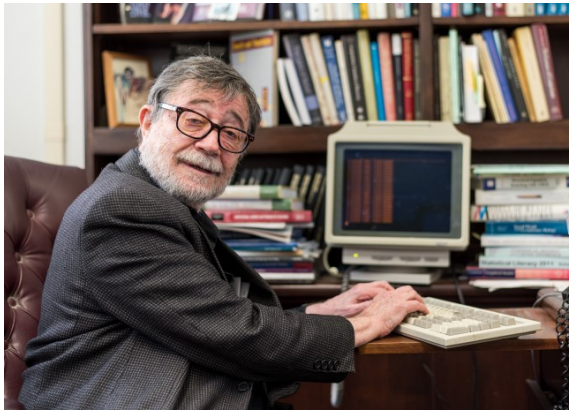


Figure: Judea Pearl

- Neyman-Rubin causal model, aka the Potential Outcomes Model (POM).
- Pearl associated with structural causal models (SCMs), graphs (directed acyclic graphs, DAGs)

We will use both extensively, but start with potential outcomes.

# A Running Example

**Research question**: Does getting a college degree make people more politically liberal?

- **Treatment** ($D_i$): Completing a college degree (vs. not)
- **Outcome** ($Y_i$): Liberal attitudes (e.g. a policy liberalism scale)
- **Confounders to imagine**:
  - Parental income and education
  - Urban vs. rural upbringing
  - Pre-existing political attitudes

We'll use this example throughout to make the abstract machinery concrete.

# What Causal Inference Actually Asks

**What we do naturally:**

- Split people into a treated group and a control group
- Compare average outcomes across groups
- "College grads are more liberal than non-grads"

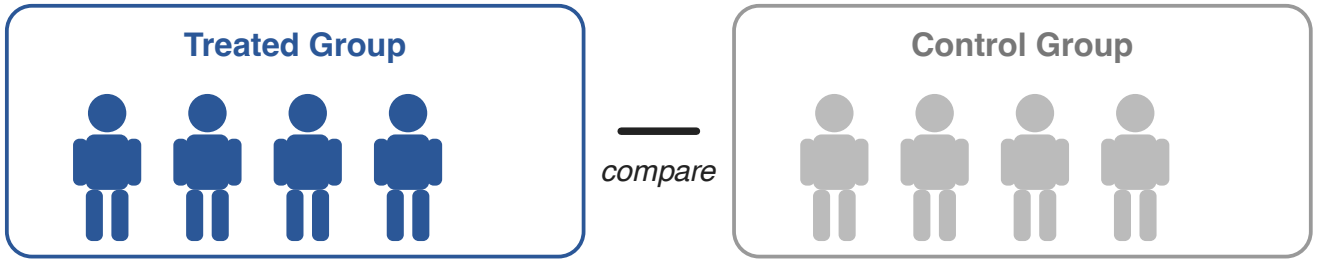
**What causal inference asks:**

- For each person, 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 people. The second is a comparison within each person, between two versions of their life.

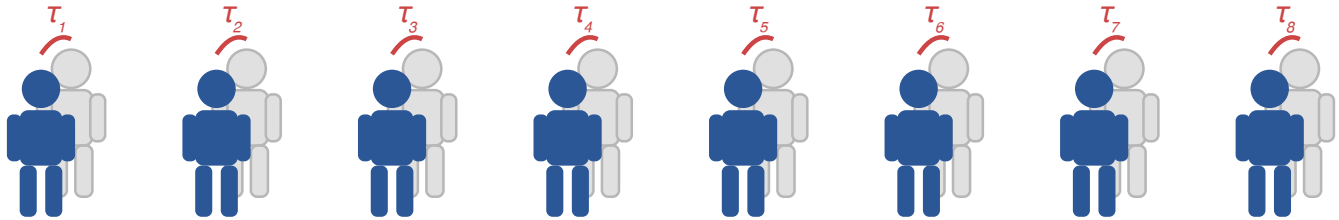
# What Causal Inference Actually Asks

**Group comparison**

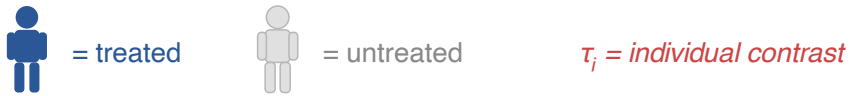



*Contrasting two groups*

**Causal (counterfactual) comparison**




*Contrasting two (potential) outcomes for everybody*



# Potential Outcomes Notation


- $Y_i(1)$: the outcome unit $i$ *would have* under treatment
- $Y_i(0)$: the outcome unit $i$ *would have* under control
- $\tau_i = Y_{1i} - Y_{0i}$: the causal effect for unit $i$
- Sample average causal effect: $\frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(Y_{1i} - Y_{0i})$

The fundamental problem of causal inference: you only get to see *one* potential outcome per unit. We need assumptions to somehow "fill in the missing" potential outcome.

# From Potential Outcomes to Observed Data

Now add the treatment indicator $D_i$ (1 if treated, 0 if not). Key insight: *treatment doesn't change the potential outcomes—it changes which one you see*:

$$Y_i = D_i\, Y_i(1) + (1 - D_i)\, Y_i(0)$$

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

1. $Y_i(1)$ and $Y_i(0)$ (at the same time, for the same unit)
2. $D_i$
3. $Y_i$
4. $\tau_i$

# Quantities of Interest

- Average Treatment Effect (ATE):

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau_i]$$

- Average treatment effect on the treated (ATT):

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

- Average treatment effect on the controls (ATC):

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

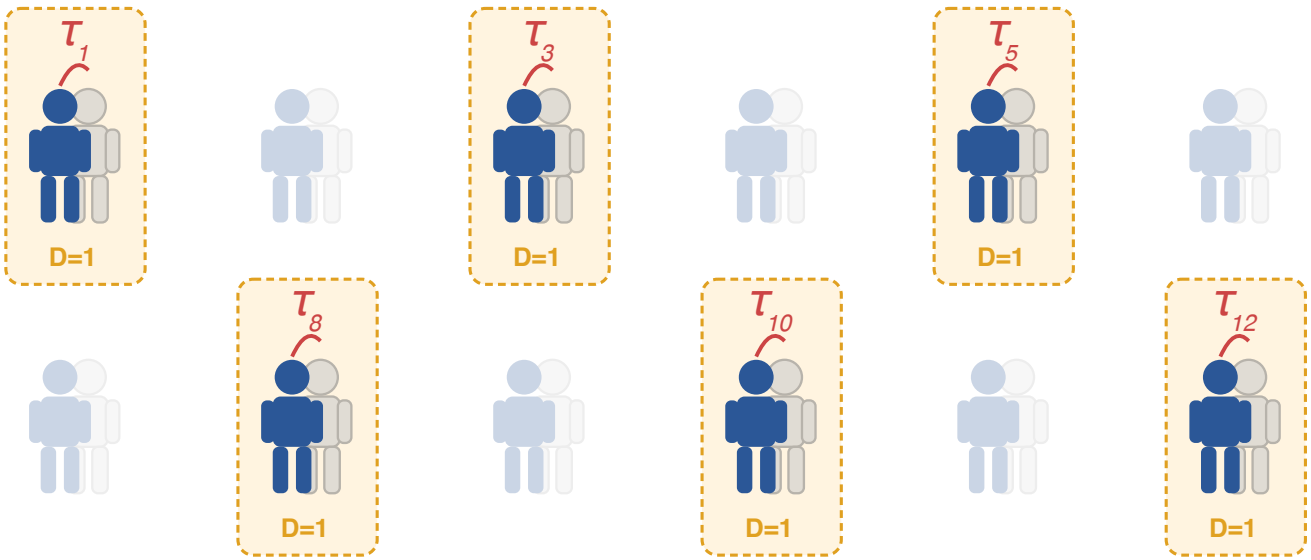
- Average treatment effect for sub-groups ($ATE(X)$):

$$ATE(X) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$$

# Visualizing the ATT

The ATT asks: among units who were actually treated ($D_i = 1$), what is the average of their individual causal effects?

Not a group comparison—imagine each treated unit's two potential lives:



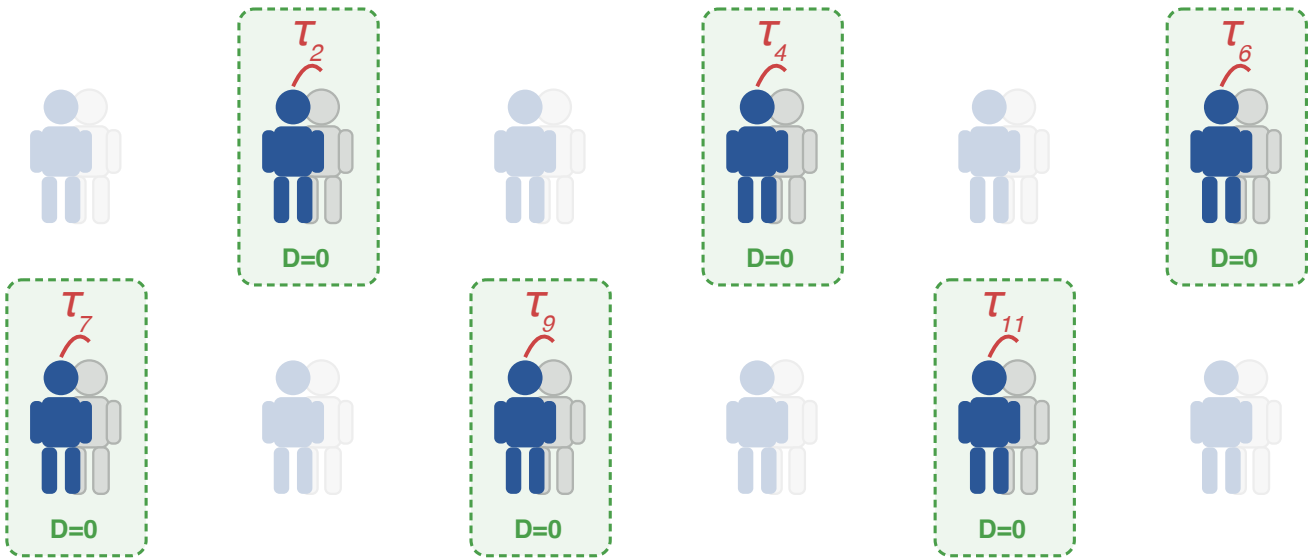

ATT = average of

$\tau_1, \tau_3, \tau_5, \tau_8, \tau_{10}, \tau_{12}$

# 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 $i$ received treatment $d$.

Worrying about consistency:

- **Interference**: other units' treatment affects $i$'s outcome, so $Y_i(d)$ is not well-defined without specifying everyone else's status. Problematic.

- **Treatment variation**: treatment may differ across units (e.g. which project a village picks in a CDD program). This is not usually a problem—variation only threatens consistency if it differs systematically by assignment, which is handled by randomization or identification assumptions later.

You will encounter **SUTVA** (Rubin 1980) = no interference + "one version of treatment."

- These are threats to consistency, *not additional assumptions*, don't go listing assumptions for no reason.

# Problematic POs

Just because you can write a PO doesn't mean you should.

- With a PO we imagine changing the treatment status of a unit. If you cannot imagine how you would intervene, it may signal ill-definedness.

Example: what is the effect of leader gender on some outcome?

- For US presidents, this involves $Y_{trump}(female)$, i.e. "What would Donald Trump have done if he was female?"
- "Changing Trump to female" would change countless subsequent experiences (and the probability of being in the sample of presidents anyway).

Similarly, a concept may seem modifiable yet we may be unclear about what exactly it would change (e.g. the effect of obesity on health).

# Train your brain

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

What licenses it?

# Train your brain

**Test 2**: Someone is assigning $D_i$ at random.

- What can we say about $\mathbb{E}[Y_{1i}|D_i = 1]$ compared to $\mathbb{E}[Y_{1i}|D_i = 0]$ and $\mathbb{E}[Y_{1i}]$?

- What would $\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$ tell us?

# Train your brain

**Test 3**: Suppose $Y_{1i} = Y_{0i}$ for all $i$ (no effect). But someone assigns $D_i = 1$ more often when $Y_{1i}$ is higher.

- Does $\mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 0]$?
- If you estimated $\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$, do you expect zero, positive, or negative?
- What is the direction of bias, and why?

# Train your brain

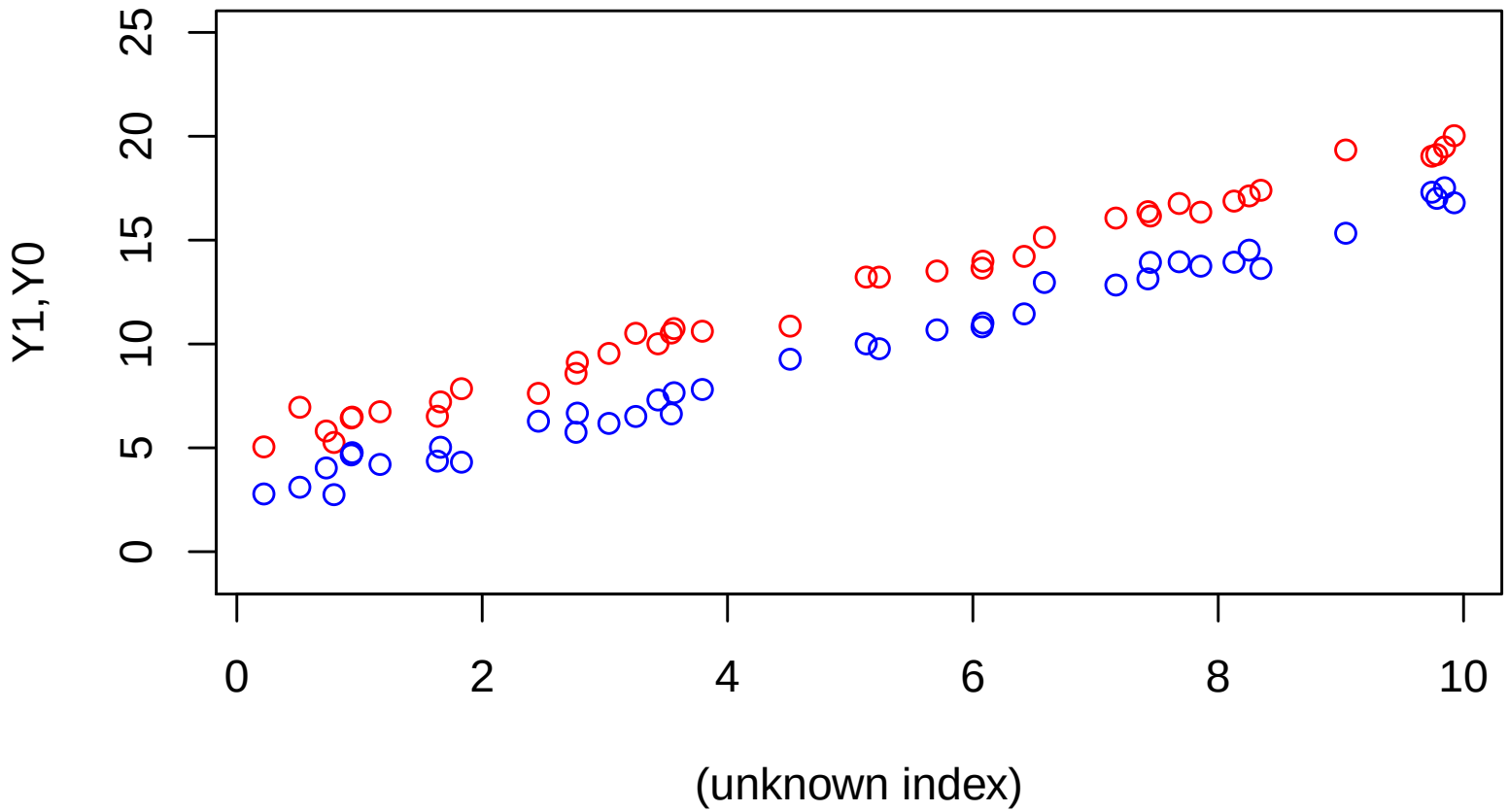
**Test 4**: Suppose you don't know anything about how $D_i$ is assigned w.r.t. $Y_{1i}$ and $Y_{0i}$.

- Can we say anything about how $\mathbb{E}[Y_1 i | D = 1]$ compares to $\mathbb{E}[Y_1 i]$?
- Can we say anything about how $\mathbb{E}[Y_0 i | D = 0]$ compares to $\mathbb{E}[Y_0 i]$?
- So can we say anything about how $\mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0]$ compares to $\mathbb{E}[Y_{1i} - Y_{0i}]$?

Which of these scenarios are we usually in with observational data?

# Visual Practice

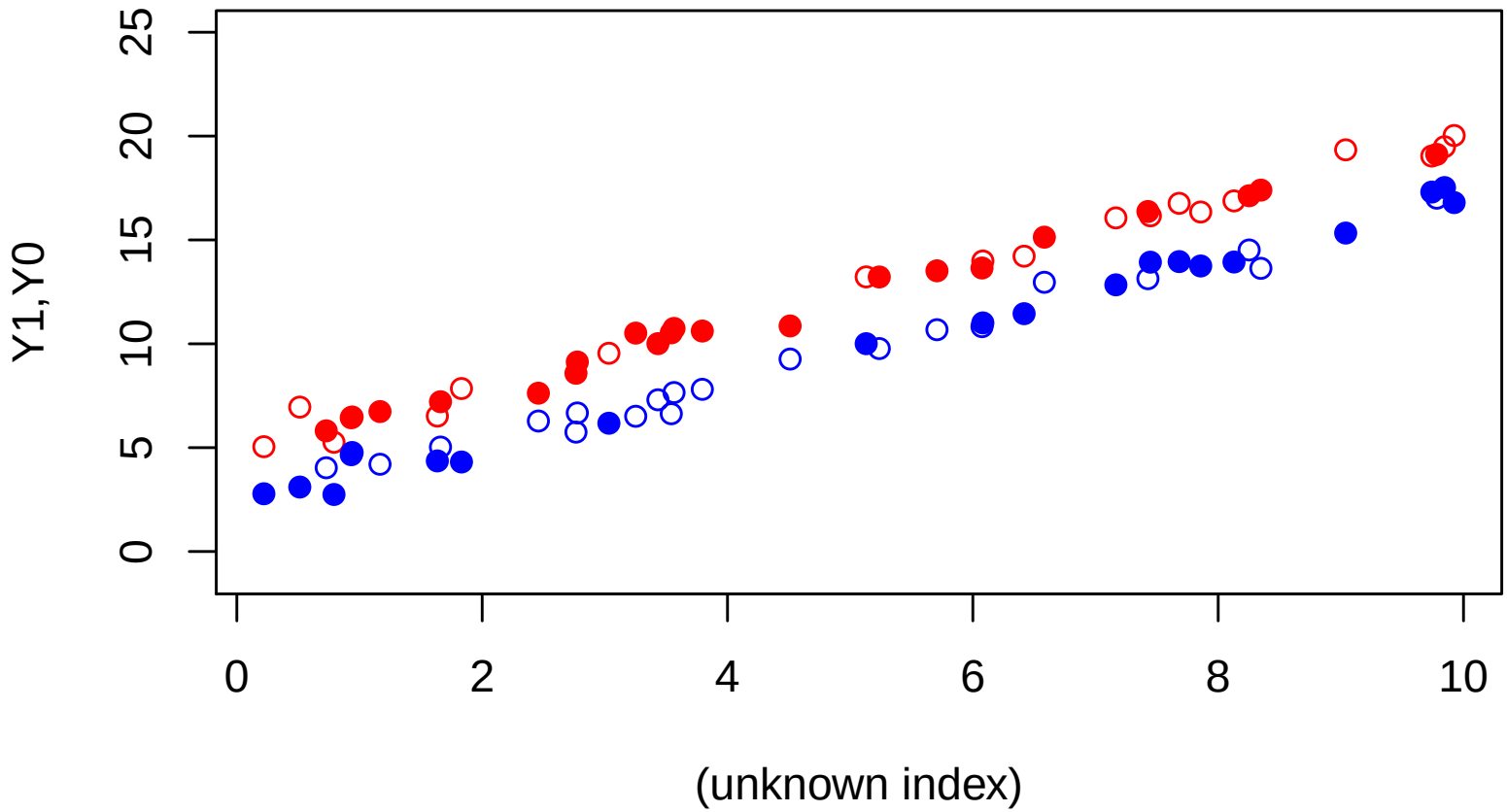
Let's looks at $Y_{1i}$ and $Y_{0i}$ alone first.




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

# Random Assignment of Treatment

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

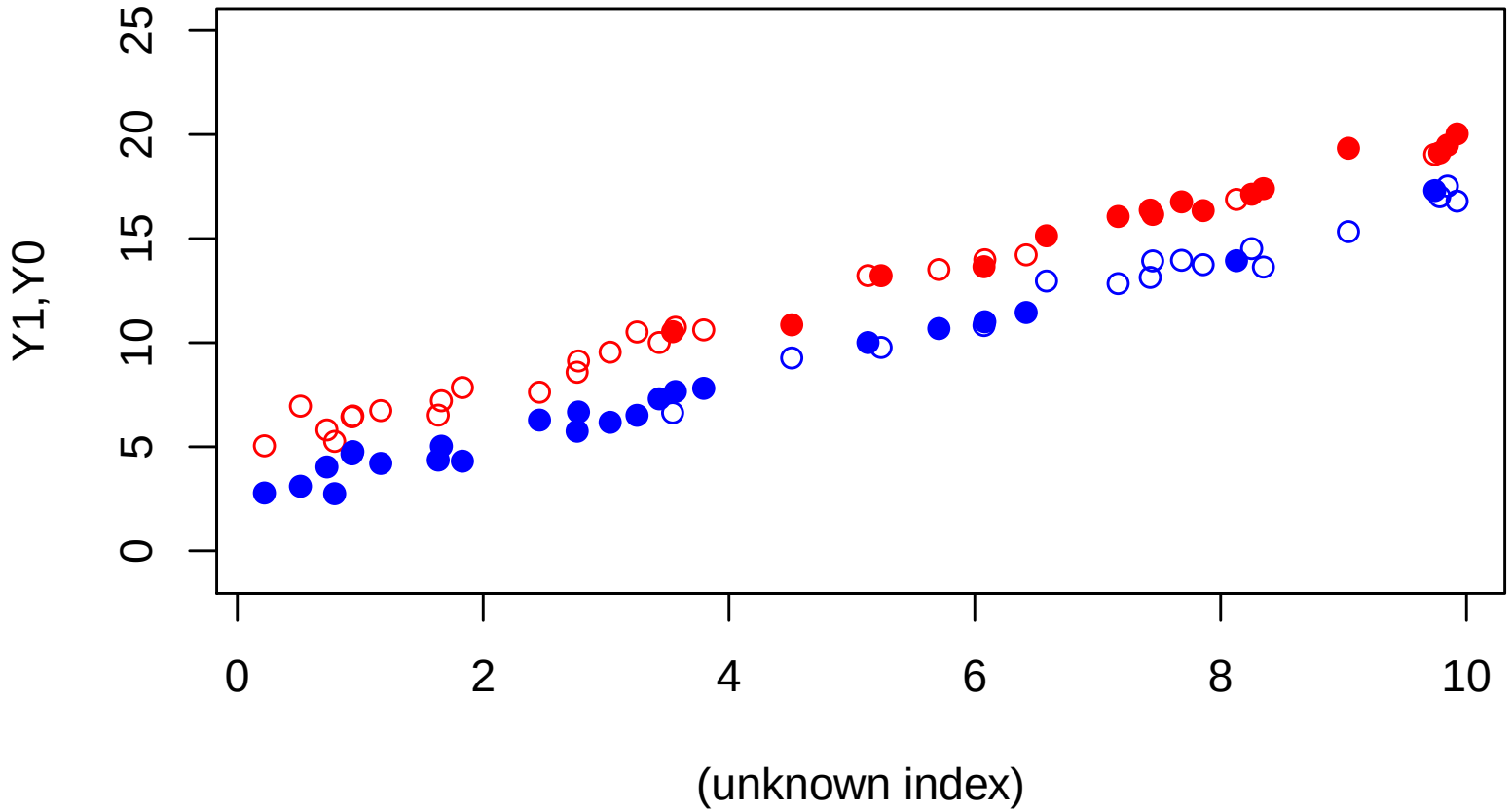



We get

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

# Non-Random Assignment of Treatment 1

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

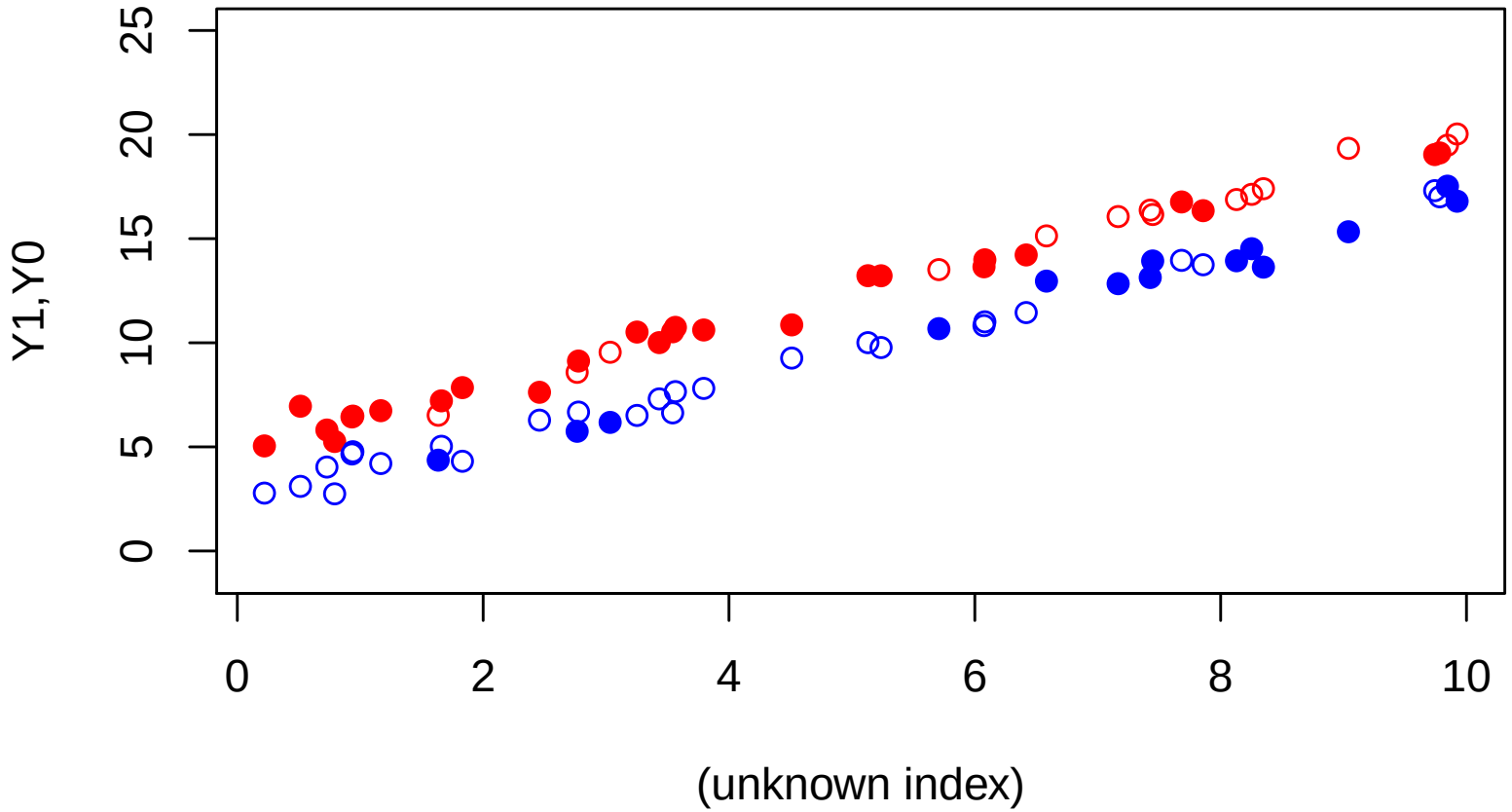



What difference in means do you expect?

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

# Non-Random Assignment of Treatment 2

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



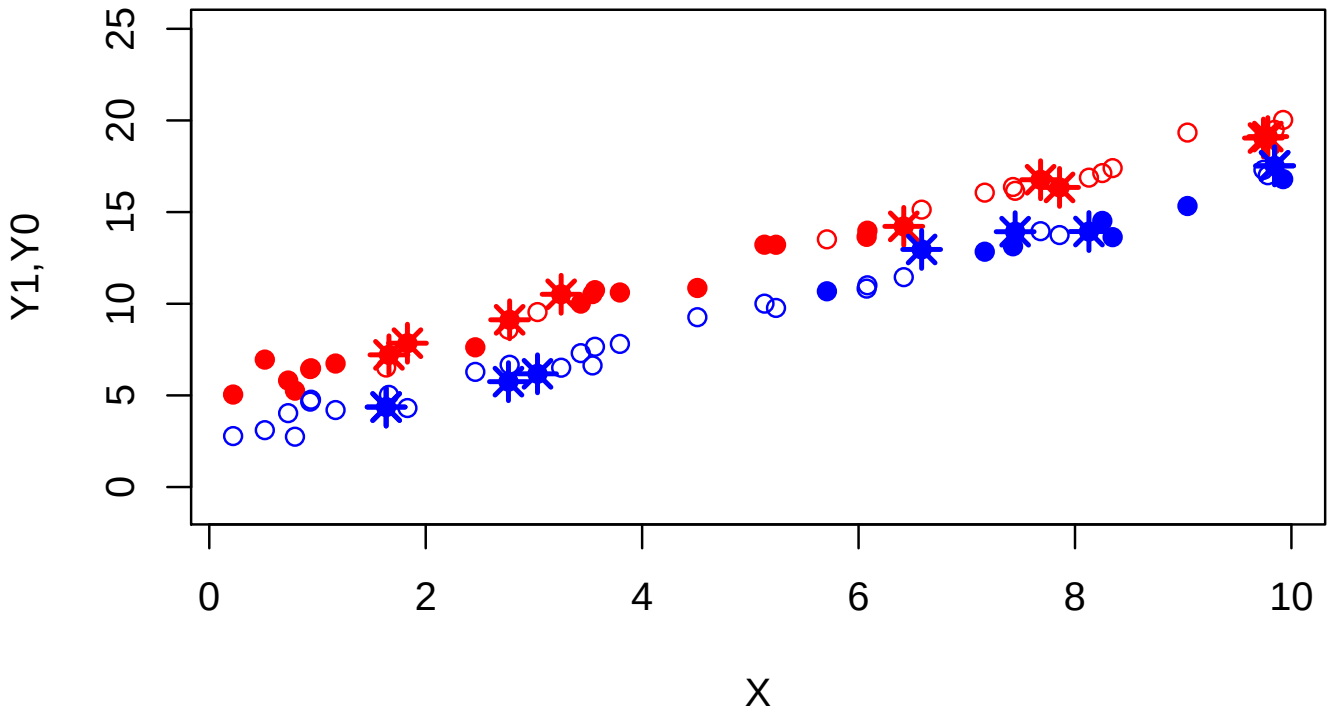

What difference in means do you expect now?

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

What can we do about this? Nothing, yet

# Preview: what if we observe a confounder?

Again $Pr(D_i = 1)$ decreases to the right, but now suppose $X$ is observable and $D$ is random conditional on $X$. Compare treated to control at the same $X$:




$$[\overline{Y}_i|D = 1] - [\overline{Y}_i|D = 0] = \widehat{ATE} = 2.63$$

# More PO practice: the "science table"

Imagine a study population with 4 units:

| $i$ | $D_i$ | $Y_{1i}$ | $Y_{0i}$ | $\tau_i$ |
|---|---|---|---|---|
| 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 $\tau_i$ are)

2. What are the ATT and ATC?

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

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

# Naive Comparison: 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 is

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

POM helps us compare this to QOIs. One terrific 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?

# Selection bias: examples

Recall: DIM $=$ ATT $+$ selection bias, where
bias $= \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$.

Example: $D =$ Church Attendance, $Y =$ Political Participation

- Church goers may already differ from non-goers (e.g. civic duty). So bias $\neq 0$.

Example: Human rights treaties

- Countries willing to sign may already have better human rights. So bias $\neq 0$.

Our running example: $D =$ College degree, $Y =$ Liberal attitudes

- People who get degrees may already be more liberal (parental education, urban upbringing). So bias $\neq 0$.

# Assignment mechanism

We often start an analysis by thinking about how units are assigned to treatment.

We can classify many important examples (similar to the graphical examples we tried) by what we know of the assignment mechanism (for $D_i$):

- When we know it was random, we know those with $D = 1$ and $D = 0$ are "comparable", meaning same average POs.
- We previewed that if $D$ is random conditionally on observed $X$, we can manage.
- When we don't know anything about assignment, the data say nothing about the causal effects without importing some other assumption.

Those represent important assignment mechanisms with hallowed names in varied fields:

- random assignment, ignorability, or exchangeability
- selection on observables, conditional ignorability
- bad news

# What's next

From here: random assignment, then random-conditional-on-observables.