Introduction to Causal Inference

Day 1: Potential Outcomes and Ignorability

Chad Hazlett

UCLA

chazlett@ucla.edu

Academia Sinica ISSM, 16 July 2026

The big picture up front

These two days are really about **one idea**: what does it take to make a credible causal claim, and why can't the data do it for us?

1. A causal claim is about a comparison we can never fully observe. So no amount of data settles it by itself.

2. "Identification" comes before estimation. We need assumptions connecting what we want to know to what we can see. Estimation comes after.

3. There are no "causal methods." Credibility comes from the assumptions, and the story behind them, not the technique.

4. Experiments are not a separate kind of knowledge. Every design leans on assumptions; with experiments, the assumption holds by design. For others, they need to be argued.

These may sound odd now, but I will state them and then repeat them at the end, and hopefully you will believe them.

When the data mislead — and when they don't

- **Intergroup contact and prejudice:** For decades, observational studies found that people with more contact across ethnic, religious, or national lines hold *less* prejudice. But field experiments that actually assign contact find modest, conditional, sometimes opposite effects.

- **Microfinance:** Celebrated as a path out of poverty based on observational comparisons of borrowers vs. non-borrowers. Six major RCTs found much more modest effects—helpful in some ways, but not transformative.

- **Smoking and lung cancer:** We will never have an RCT. Yet the causal conclusion is among the most trusted in all of science. Why? Not because of the volume of data, but because researchers could formalize *what it would take* for the result to be wrong—and the confounding required was wildly implausible.

Causal inference $\neq$ data analysis

The shift in thinking needed for these two days begins by **getting away from the idea that data "speak for themselves".** Data $\neq$ evidence.

Rather, there are causal quantities you'd like to know, but *without further assumptions*, data are largely silent on them, no matter how much you have.

Stylized illustration:

- you have data on X and Y
- you want to learn some effect of X on Y
- you compute something with data, say, $coef(Y \sim X)$
- trust me (for now): *without further assumption*, any causal effect can produce any coefficient...and thus any coefficient you get is equally suggestive of any causal effect.

In other words, "looking at the data" is not a sufficient strategy for learning about the world...

So what *does* it take? That's what these two days are about.

Identification: Assumptions + Data $\rightarrow$ Causal Conclusions

Causal inference is about that "further assumption" bit:

- What assumptions let us connect the causal thing we want to know to something in the data?
- Can we identify our quantity of interest with something we can estimate?
- We speak of identification assumptions or identification strategies that provide such bridges.

Key implications:

- Most identifying assumptions cannot be proven correct using data. It's not about "knowing the right test" to run.
- We will always be appealing to information, assumptions, or context from *outside* the data.
- Credibility does NOT derive from "what method you used." There are not "causal methods"! You can't say "it's causal because I did xyz."
- This is one of the hardest things to absorb, in part because we are surrounded by mistakes and false advertising.

Two languages

Most work on causality now employs one of two languages:

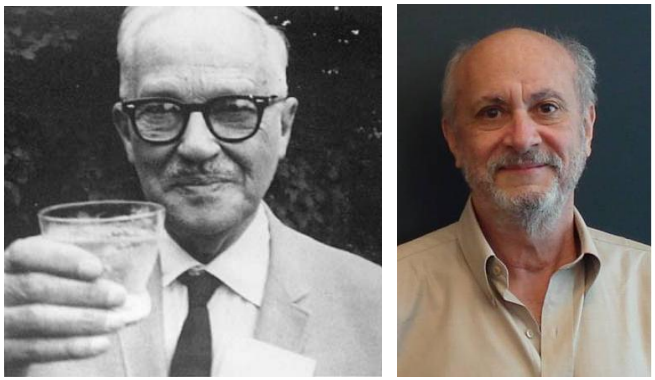


Figure: Neyman and Rubin

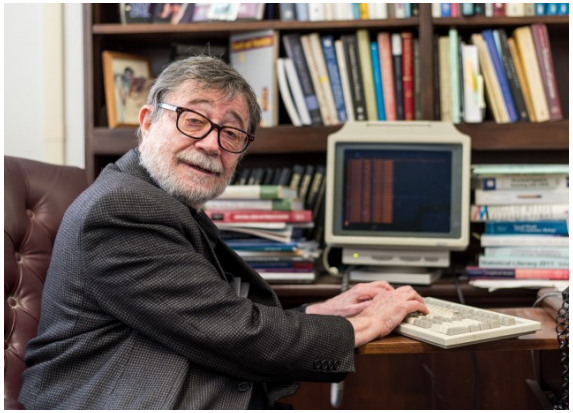


Figure: Judea Pearl

- 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 our brains (problematically) jumpt to:

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

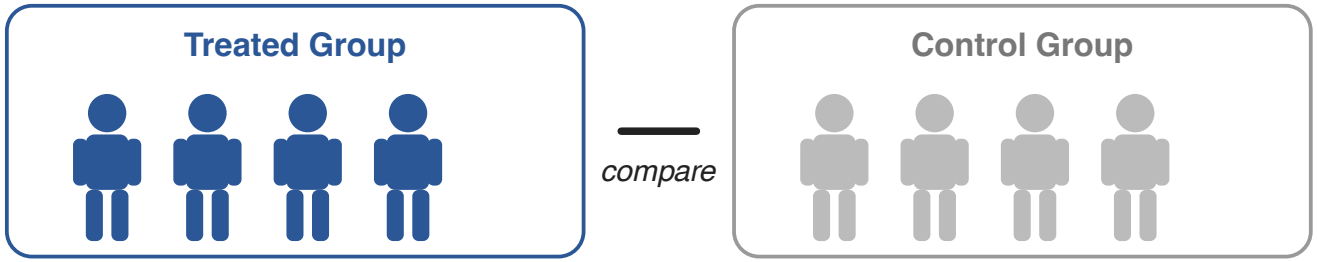
What causal inference asks, by contrast:

- 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 compares groups of *different* people. The second requires a *within person* comparison across counterfactuals.

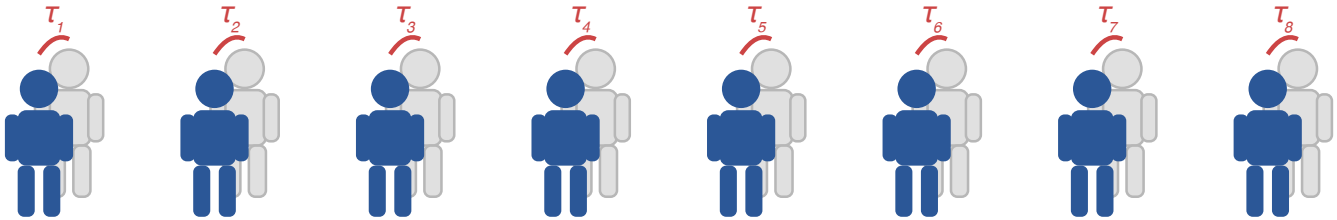
What Causal Inference Actually Asks

Group comparison



Contrasting two groups

Causal (counterfactual) comparison



Contrasting two (potential) outcomes for everybody

= treated = untreated τ_i = *individual contrast*

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
- $\tau_i = Y_i(1) - Y_i(0)$: the causal effect for unit i
- Sample average causal effect: $\frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$

The fundamental problem of causal inference: you only see *one* potential outcome per unit. We rely on assumptions to access the missing outcomes.

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. τ_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 ($ATE(X)$):

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

Reading $\mathbb{E}[\, Y_i(d) \mid D_i = d' \,]$: an outcome, and a filter

These expressions have **two separable pieces**:

$$\mathbb{E}\Big[\quad \underbrace{Y_i(d)}_{\text{the outcome}} \quad \Big|\ \underbrace{D_i = d'}_{\text{the filter}} \quad \Big]$$

The outcome $Y_i(d)$: something every unit has, treated or not — possibly a value we never see.

The filter $\mid D_i = d'$: whom we average over — "in the part of the population with $D_i = d'$", or "among those with $D_i = d'$". It selects units; it does not change what $Y_i(d)$ means.

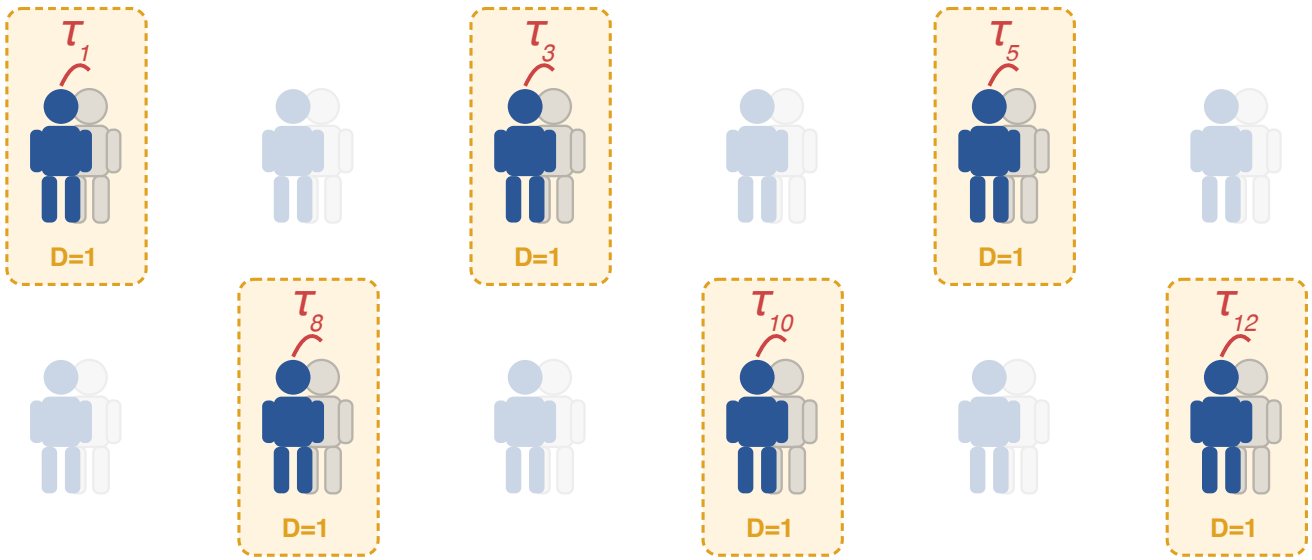
So the outcome and the filter can **match** or **mismatch**:

- $\mathbb{E}[Y_i(1) \mid D_i = 1]$: the treated units' treated outcome — **observed** (and equal to $= \mathbb{E}[Y_i \mid D_i = 1]$).
- $\mathbb{E}[Y_i(0) \mid D_i = 1]$: the non-treatment (control) outcome, among treated units. This is a **counterfactual** average, never seen (but perfectly well-defined).

Visualizing the ATT

The ATT is a good test of undestanding: What does it mean to ask about the treatment effect only among the treated?

- If you are stuck on the group comparison idea, this makes no sense.
- If you get the counterfactual idea, it is easy: everybody has a treatment effect, the ATT is the average of those effects among the treated.



ATT = average of

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

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.

This is definitional to potential outcomes, but we give it a name so we can reference it and worry about it.

What to worry about?

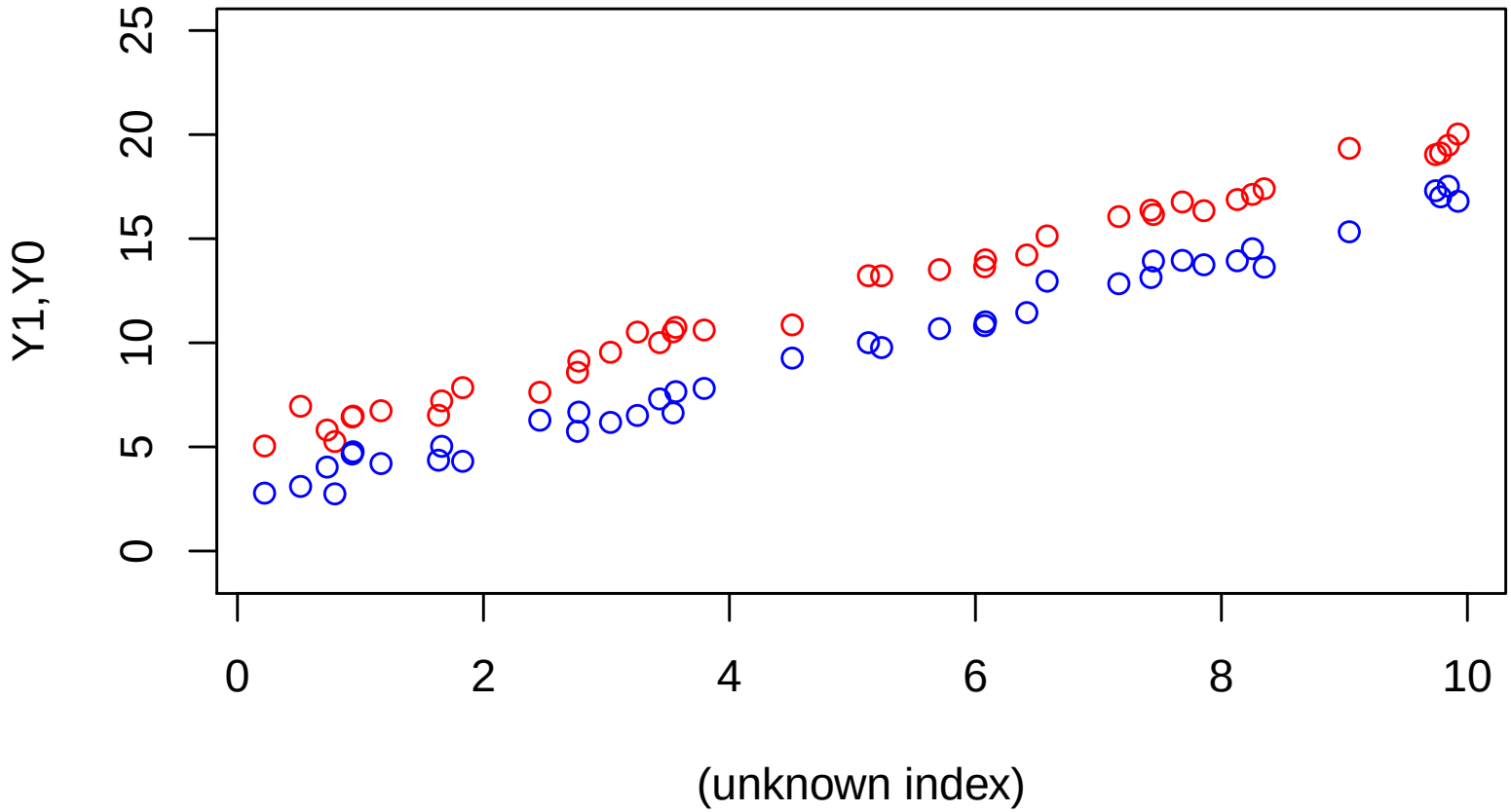
- **Interference/ Spillover**: other units' treatment affects i's outcome, so $Y_i(d)$ is not well-defined without specifying everyone else's status.

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

- Don't worry about "one version of treatment" actually... we will leave it at that.
- These are threats to consistency, *not additional assumptions*, i.e. consistency is enough and all we will use.

The setup: everyone's two potential outcomes

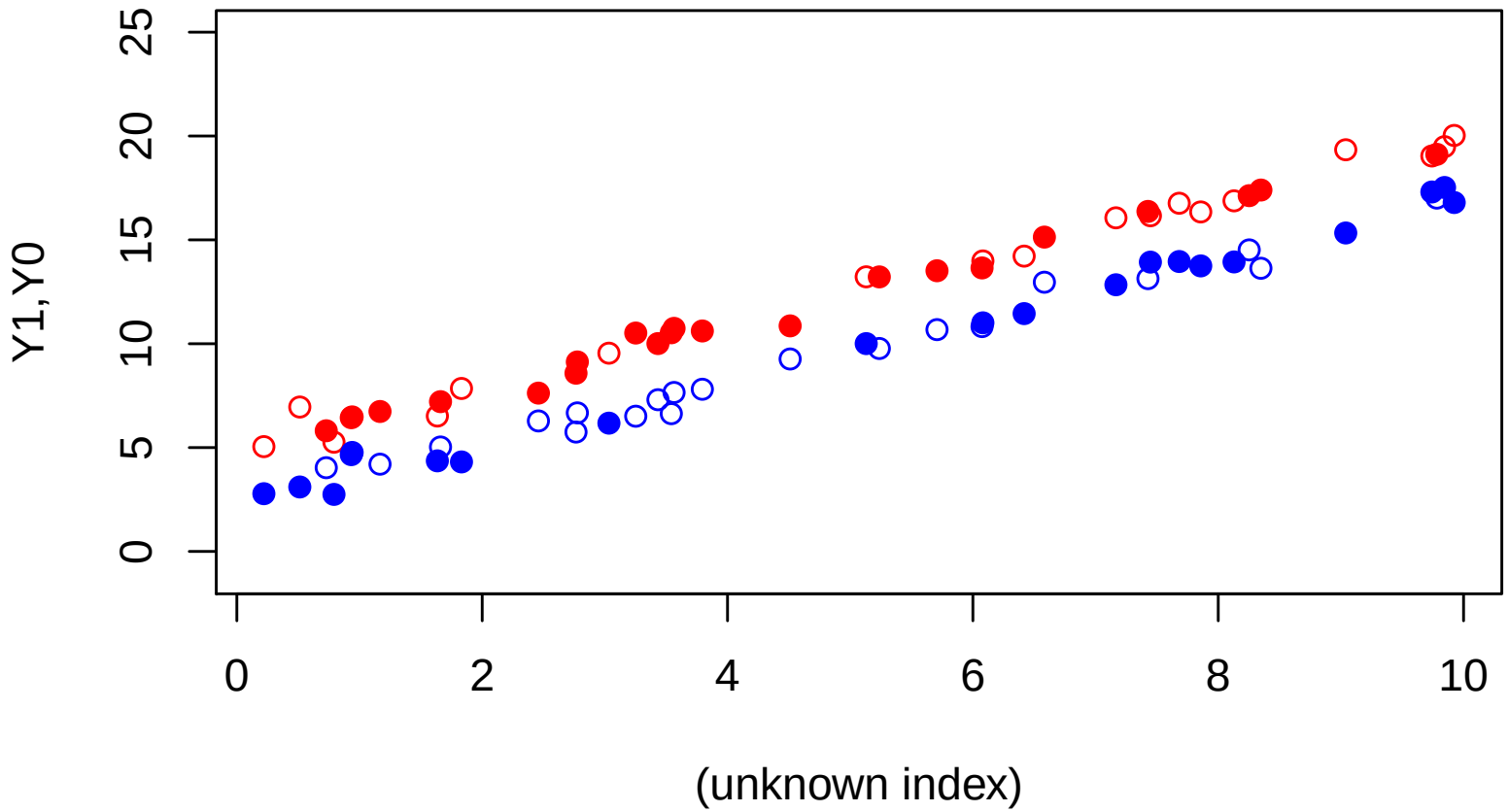
Let's look at $Y_i(1)$ and $Y_i(0)$ for everyone first (we could never see this in reality):



By construction here, the **true** $\mathbb{E}[Y_i(1) - Y_i(0)] = \text{ATE} = 3$.

Random Assignment of Treatment

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

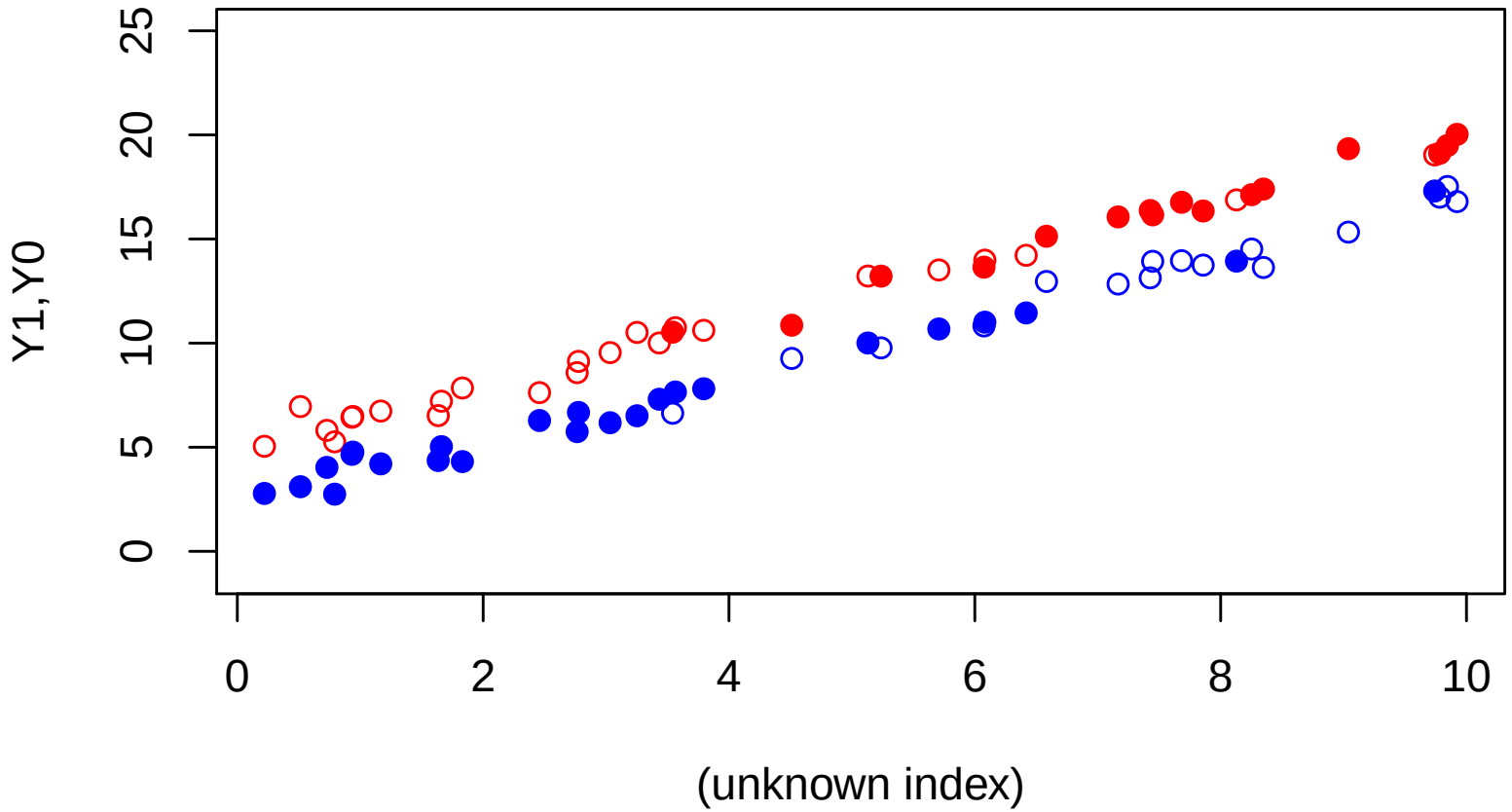


We recover the truth:

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

Non-Random Assignment of Treatment 1

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

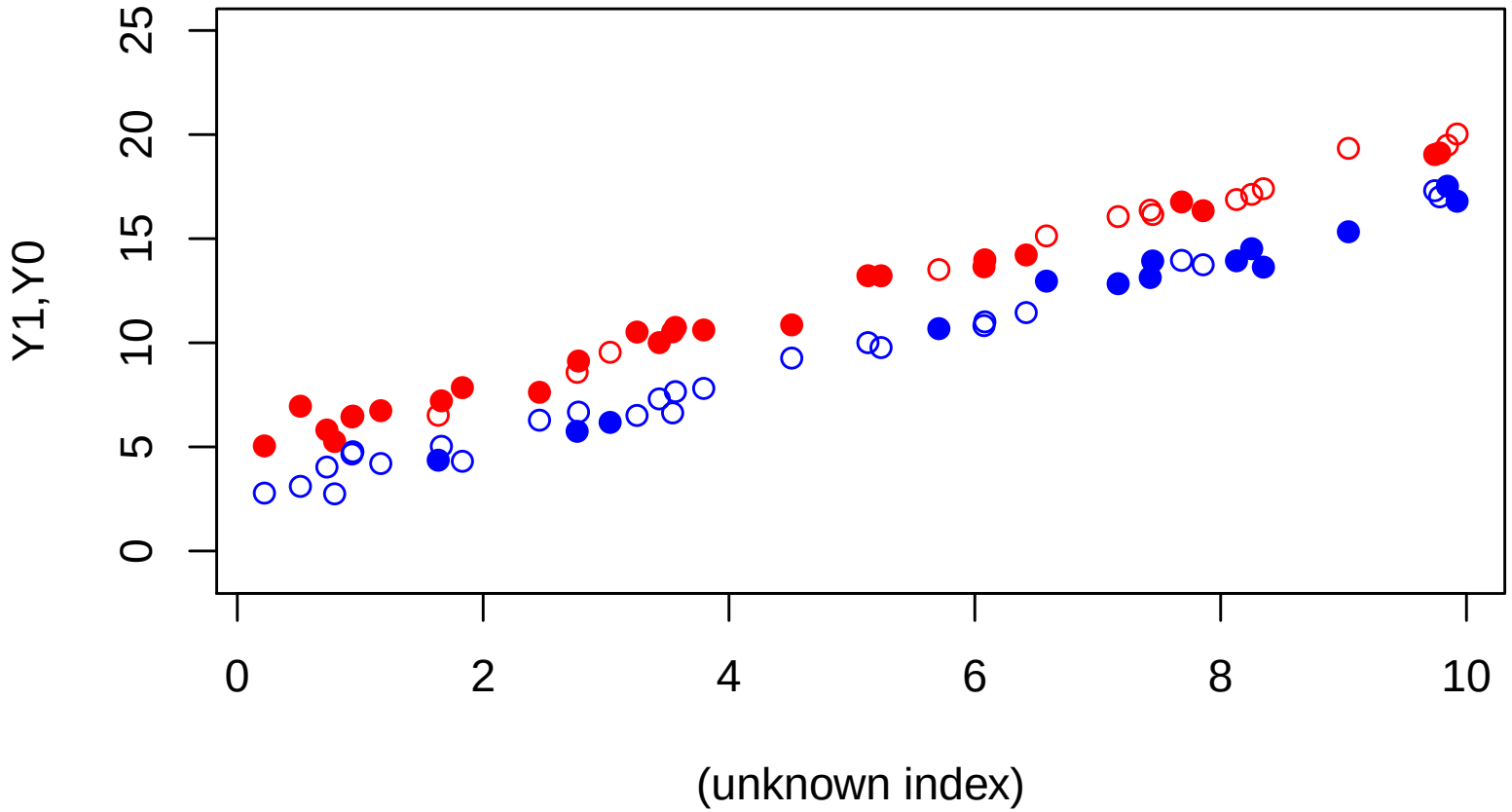


What difference in means do you expect?

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

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?

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

Same true ATE of 3 — selection can even **flip the sign**.

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 (DIM) 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:

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

What are the terms in blue & green?

Selection bias: examples

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

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 $\neq 0$.

Identification vs. Estimation

Hopefully you believe me now: causal inference is an attempt to learn about an *unobservable, counterfactual-dependent* quantity of interest (QoI) using *observed* data.

Two inferential hurdles:

1. Identification: How will we fill in missing potential outcomes?
2. Estimation: Once we can fill in missing potential outcomes, how do we deal with standard statistical inference?

Identification problems persist *no matter how much data you have*. Estimation problems are those that can be solved with more data.

Golden rule of causal inference: **Worry about identification first.**

The Template

Before we say more about experiments, take note of the general template applicable throughout causal inference.

We organize our thinking about how to make causal claims by identification strategy.

For each we discuss:

1. the identification assumption first, which your life will depend upon when using this approach.
2. information from outside the data that help support that assumption
3. the (limited) things you can test in efforts to falsify the assumption
4. estimators
5. ideally, sensitivity to violations of assumptions

Classical Randomized Experiment

The happiest of causal stories: inference depends on an assumption you can defend by design.

Setup:

- Units: $i = 1, ..., N$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: $Y_i(0)$, $Y_i(1)$
- Observed outcome: $Y_i = Y_{D_i i}$
- Number of treated/untreated units: $N_1 = \sum_{i=1}^{N} D_i$ and $N_0 = N - N_1$

Random assignment can take one of several forms, most importantly:

- Complete randomization: Exactly N_1 treated units
- Simple (Bernoulli) randomization: Each unit independently assigned to treatment with probability p

Identification of Average Treatment Effect Under Experiments

Identification Assumption: (guaranteed by random assignment):

$$\{Y_i(1),\ Y_i(0)\} \perp\!\!\!\perp D_i$$

Quantity of Interest:

$$\tau_{ATE} \equiv \mathbb{E}[Y_i(1) - Y_i(0)]$$

How do we actually manage identification of τ_{ATE}?

$$\begin{aligned}
\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] &= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \quad \text{(why?)} \\
&= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \quad \text{(why?)} \\
&= \text{DIM Estimand} \\
&\approx \overline{Y}_{D=1} - \overline{Y}_{D=0} \quad \text{(the DIM estimator)}
\end{aligned}$$

That's a big deal, algebraically shows how we can learn something that is based on half unobservable data, using only observed data!

Worksheet 1

Potential Outcomes

Pen and paper, $\sim$15 min

Can we make causal claims in observational studies?

In an experiment, random assignment guarantees ignorability:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i$$

In other words: treated and control groups are "comparable" on their potential outcomes.

In an observational study, units may select into treatment in ways related to their potential outcomes: the treated and control groups are no longer "comparable" on their potential outcomes.

The usual hope: Perhaps conditionally on X, treated and control are comparable again.

Here we examine what assumptions are needed to make this hope credible, and how to use them to identify causal effects.

The Assumption: Conditional Ignorability

The assumption that would make this work is **conditional ignorability** (CI):

Identification Assumption (Conditional Ignorability / Selection on Observables)

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid X_i$$

This assumption has many names beyond CI:

- Selection on observables (SOO)
- No unmeasured confounders
- Conditional exchangeability
- "Backdoor criterion satisfied" (in graph language, tomorrow)

Conditional ignorability = experiments by strata of X

In an experiment, we had unconditional ignorability:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i.$$

Conditional ignorability just adds "$\mid X_i$":

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid X_i.$$

Read it backwards, from the conditioning inward:

- → Among units with some $X_i = x$ (say, all college-educated 40-year-olds).
- → ... D_i needs to be independent of $\{Y_i(1), Y_i(0)\}$.

So an observational study under CI is a **stack of mini-experiments**: one as-good-as-randomized comparison inside each value of X.

We didn't run the experiment — we are assuming nature ran one within each stratum. Whether that is credible is the whole game.

How could you ever believe CI?

So when is CI defensible? You need to either:

1. **Know the assignment.** You understand why units got treated well enough that every systematic common cause of D_i is included in X_i.
2. **Know the outcome.** If you understand what drives $Y_i(0)$ well enough that, given X_i, nothing left over is associated with D_i.

Both of these are ways of saying **there are no unobserved confounders**. This will be easier to see with graphs tomorrow.

"**Natural experiments**" or "**quasi-experiment**"

- Cases where the nature of treatment assignment is idiosyncratic enough that you argue (conditionally on something) it is basically random.
- E.g. Within regions equally likely to have ben struck by a typhoon, which ones were hit?

CI is untestable! It is not in the data.

Common Support

CI alone is not enough. We also need:

Identification Assumption (Common Support / Overlap)

$$0 < \Pr(D_i = 1 \mid X_i = x) < 1 \quad \textit{for all } x$$

Every type of unit (every value of X) must have a positive probability of receiving either treatment or control.

This can pose trouble when the probability of treatment (or control) is very high or very low for some units — there may be little or no information about the missing counterfactual in that region of X.

Estimation Under CI: The Idea

The identification result says: compute treatment effects *within strata of X*, then average over the population distribution of X.

Subclassification (stratification) does this literally when X is discrete:

1. Divide units into strata defined by X
2. Compute the within-stratum DIM: $\hat{\tau}_j = \bar{Y}_{D=1,j} - \bar{Y}_{D=0,j}$
3. Average: $\hat{\tau}_{ATE} = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\tau}_j$

This is transparent and requires no modeling,

- ...but it breaks down with many covariates or continuous X (curse of dimensionality)
- that's where matching, weighting, and regression come in; there is a vast estimation toolkit we won't dig into here.
- none of that changes or eases the core identification concern.

Example: Smoking and Mortality (Cochran 1968)

Unadjusted death rates per 1,000 person-years:

	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarette only	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

What does this seem to say?

But this is a group comparison, not the causal effect of smoking.

Confounders: start by trying to think of things that are associated with both smoking type and death rate.

A Confounder: Age

Mean ages by smoking group:

	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarette only	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

On average, pipe/cigar smokers are *older* and cigarette smokers are *younger* than non-smokers.

Age is associated with both smoking type and death rate; it is a confounder.

Momentarily, let's pretend we believe CI with age as the confounder:

$$deathrate(smoking) \perp\!\!\!\perp smoking \mid age$$

Subclassification estimator rescues the within-stratum differences

Deaths per 1,000 person-years[1]:

Age group	Rate (Cig)	Rate (Non)	Effect	N (Cig)	N (Non)
< 45	5	3	$+2$	40	10
45–54	10	5	$+5$	30	20
55–64	20	10	$+10$	20	30
65+	45	23	$+22$	10	40
Pooled	13.5	13.5	0	100	100

Within every age group, smokers die at higher rates.

But the pooled difference (0) masks this because smokers are younger.

Reminder: CI claims that "we have an experiment within each stratum of X", so you compute the DIM in each stratum, and average after:

$$\hat{\tau}_{ATE} = \frac{50}{200}\big(2 + 5 + 10 + 22\big) = 9.75$$

[1] Fake data, consistent with the marginal rates above.

Did we just prove smoking kills?

We applied a causal assumption (CI) and an estimator (subclassificaion) and determined that "smoking raises mortality by ≈ 9.75" deaths per 1000 person-years.

No, we didn't!

- Stating an assumption (CI given age) doesn't make it hold.
- Applying an estimator (subclassification) doesn't make the result true.

A result is defensibly causal only if to the degree the assumption is defensible (Thesis 3). The worry here: there are other confounders.

- What else differs between smokers and non-smokers and affects mortality?
- Diet, exercise, alcohol, occupation, income, health conciousness . . . ?

It would not have mattered if we used matching, weighting, regression, double machine learning, etc. *A causal claim is only as good as the assumption is defensible.*

Where we landed today

Today, in the language of **potential outcomes**:

- The causal effect is a comparison of $Y_i(1)$ and $Y_i(0)$ — and we never see both.
- **Experiments** identify the ATE because randomization makes $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i$.
- **Conditional ignorability** is the same idea within strata of X: an experiment we assume nature ran for each value of X.

Bookend: The theses of causal inference

Hopefully the four theses we started with now seem more concrete:

1. A causal claim is about a comparison we can never fully observe. So no amount of data settles it by itself.
2. "Identification" comes before estimation. We need assumptions connecting what we want to know to what we can see. Estimation comes after.
3. There are no "causal methods." Credibility comes from the assumptions, and the story behind them, not the technique.
4. Experiments are not a separate kind of knowledge. Every design leans on assumptions; with experiments, the assumption holds by design. For others, they need to be argued.

Tomorrow: a second language

Potential outcomes are great for precisely saying what we mean by causal effects and what assumptions are required.

But they are clumsy for reasoning about those assumptions, including which variables to condition, why controlling for the wrong thing can create bias, and other opportunities for either mistakes or identification.

Tomorrow we add **causal graphs (DAGs)**: a second language for the same objects. We will see:

- confounding, colliders, and why "control for everything" is problematic
- the **backdoor criterion** — and that it is the same as conditional ignorability, seen in pictures;

Worksheet 2

Experiments & Ignorability

Short, $\sim$10 min — finish with your neighbor if we run short.