MS&E 228: Applied Causal Inference Powered by ML and AI
Lecture 1: Foundations via Potential Outcomes

Vasilis Syrgkanis

Stanford University

Winter 2026

Readings: *Applied Causal Inference Powered by ML and AI*, §2.1; Hernán & Robins, *What If*, Ch. 1.

Goals for Today

1. Define notion of *potential outcomes* (counterfactuals):

$$Y(0), Y(1)$$

2. Mathematically define average treatment effect (ATE) as a *causal estimand*:

$$\mathbb{E}[Y(1) - Y(0)]$$

3. Internalize the concept of *confounding*, formally and through examples
4. Understand why Randomized Control Trials are "the ideal" for *causal identification*

Motivation: why we care

Tech
Does showing feature X increase retention?

Social science
Do tutoring programs increase test scores?

Healthcare
Does a new protocol reduce readmissions?

Operations
Do preventive maintenance policies reduce downtime?

Core issue: For each question, we need a principled way to define and estimate "what would have happened otherwise."

The unit, treatment, and outcome

- Units: $i = 1, \ldots, n$ (users, patients, schools, machines, ...)
- Binary treatment: $D_i \in \{0, 1\}$ (e.g., got feature/protocol/program/policy)
- Outcome: Y_i (retention, readmission, test score, downtime, ...)

Note: typically we will be dropping the unit index i from all random variables

Key conceptual gap

We only observe each unit under *one* treatment state. Causal questions require comparing *two* states.

Potential outcomes (Neyman and Fisher '20s, Rubin'74)

Introduce two random variables for each unit, referred to as *potential outcomes*:

$$Y(1) \quad \text{and} \quad Y(0)$$

- $Y(1)$: outcome if the unit *were treated*
- $Y(0)$: outcome if the unit *were not treated*

Individual causal effect

$$Y(1) - Y(0)$$

Fundamental problem of causal inference: we never observe both $Y(1)$ and $Y(0)$ for the same unit.

From potential outcomes to observed data

Observed outcome is the potential outcome for the assigned treatment:

$$Y = Y(D) \qquad \text{(consistency)}$$

Stable Unit-Treatment Value Assumption (SUTVA)

Implicit Assumption in our Definition:

(i) no interference between units

(ii) no hidden versions of treatment

Otherwise we would need to write $Y_i(d_1, \ldots, d_N)$, if the potential outcome of a unit, depended on treatments received by other units.

Discussion: when might SUTVA fail in the motivating examples?

Causal Estimands: ATE, ATT, and CATE

Individual causal effect almost never identifiable. Typically, we care about averages:

- **Average Treatment Effect (ATE):**

$$\delta = \mathbb{E}[Y(1) - Y(0)]$$

- **Average Treatment Effect on the Treated (ATT):**

$$\delta_1 = \mathbb{E}[Y(1) - Y(0) \mid D = 1]$$

- **Conditional Average Treatment Effect (CATE):**

$$\delta(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

These quantities are called *causal estimands* because they depend on counterfactual random variables (potential outcomes).

Why more than one causal estimand?

Different decisions: product rollouts (ATE), targeted policies (CATE), evaluation among adopters (ATT).

Poll Everywhere

A hospital wants to know: "Among patients who actually received the new discharge protocol, what was its average impact on readmission?"

Which estimand best matches?

1. ATE $\mathbb{E}[Y(1) - Y(0)]$
2. ATT $\mathbb{E}[Y(1) - Y(0) \mid D = 1]$
3. CATE $\mathbb{E}[Y(1) - Y(0) \mid X = x]$
4. None of the above

Correlation is not causation: the naive comparison

The Identification Problem

Express the causal estimand as a quantity that only depends on observed random variables (*statistical estimand*).

A common (but typically wrong) approach:

$$\pi = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0].$$

Why it's tempting

It's easy to compute and easy to explain.

Why it's usually wrong

The treated and untreated groups may differ in many ways besides treatment.

Decomposing the naive difference: effect + selection bias

Start with:

$$\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$$

Using $Y = Y(D)$, we can show:

$$\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] = \mathbb{E}[Y(1) - Y(0) \mid D = 1] + \underbrace{(\mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0])}_{\text{confouning bias}}.$$

Key message

Even if treatment helps ($Y(1) > Y(0)$), selection bias can mask or exaggerate the effect.

Confounding

Confounding is the situation where treatment status D is predictive of the potential outcomes:

$$\mathbb{E}[Y(d) \mid D = 1] \neq \mathbb{E}[Y(d)] \quad \text{for } d \in \{0, 1\}.$$

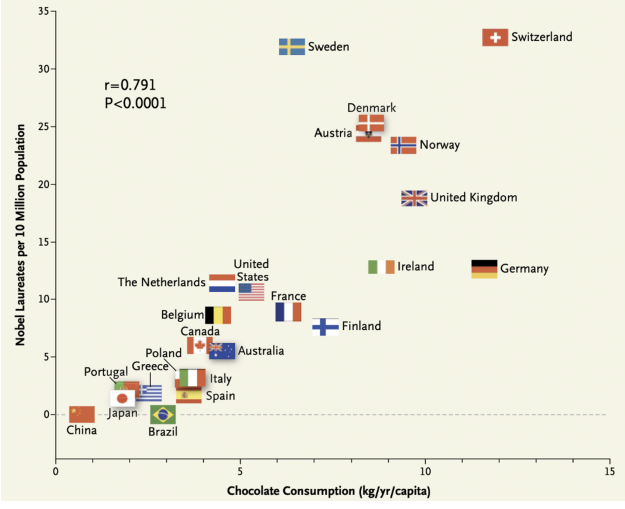
Equivalently, treated and untreated groups are not comparable as-if randomized.

A **confounder** (aka common cause) is a variable X that affects both treatment assignment D and the outcome Y (and therefore also the potential outcomes $Y(d)$).

Examples

- Tech: power users more likely to receive a "new feature" rollout earlier
- Healthcare: sicker patients more likely to receive an aggressive treatment
- Social science: motivated workers self-select into training
- Ops: high-performing sites adopt new process earlier

Confounding in the wild: Chocolate consumption and Nobel laureates



The tempting (most probably wrong) conclusion

"If a country eats more chocolate, it produces more Nobel laureates."

Some studies support the causal story:

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. New England Journal of Medicine. Prinz, A. L. (2020). Chocolate consumption and noble laureates. Social Sciences & Humanities Open.

Where confounding enters: a common cause

Story

A plausible confounder is **country wealth**:

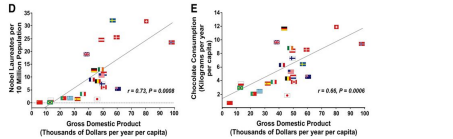
- Wealth $\rightarrow$ more chocolate consumption
- Wealth $\rightarrow$ more investment in education/science

Reminder

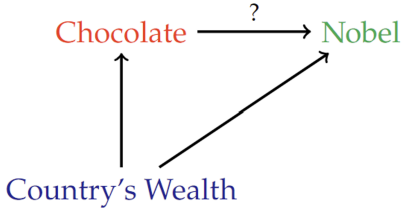
A variable X is a **confounder** if it affects both treatment and outcome:

$$X \rightarrow D \quad \text{and} \quad X \rightarrow Y$$

making D and Y associated even when $D \not\perp\!\!\!\perp Y$.

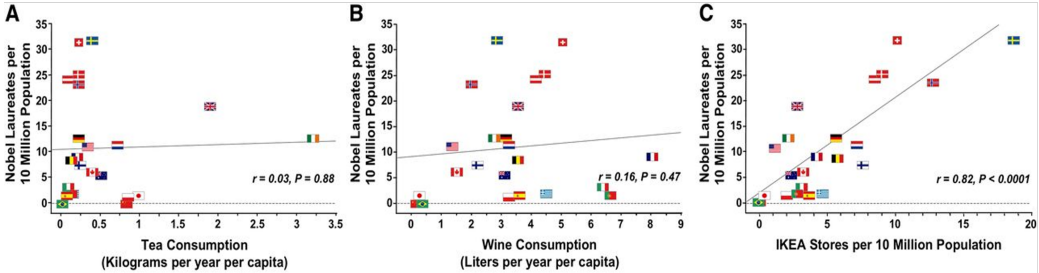


Wealth (GDP) correlates with Nobels and with chocolate.



The "backdoor" path $D \leftarrow X \rightarrow Y$ creates non-causal association.

A warning sign: many things correlate with Nobel counts



You can find many "predictors" of Nobel counts (wine, tea, IKEA stores, . . .).

Takeaway

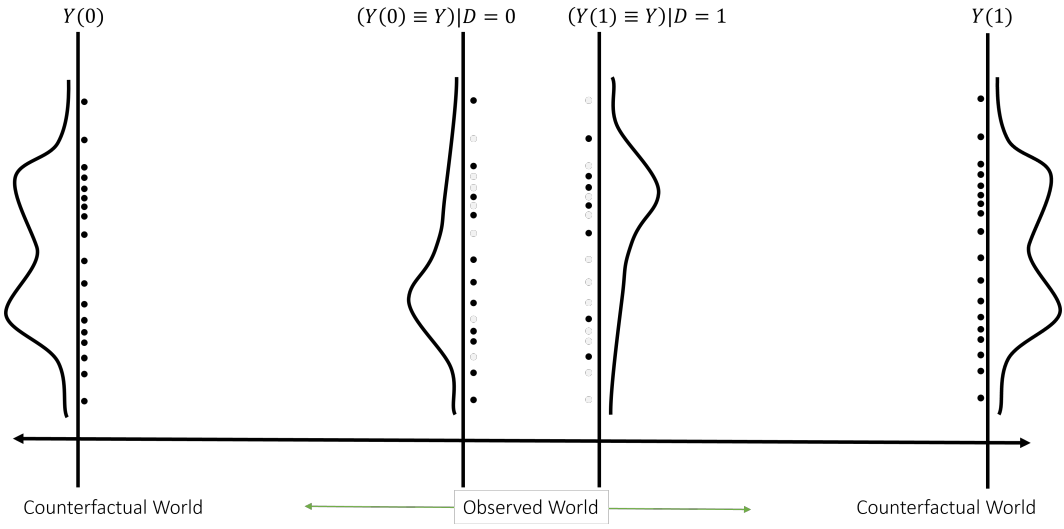
In observational data, a strong correlation can be produced by **shared causes** (confounding), not by a causal effect of D on Y.

Poll Everywhere

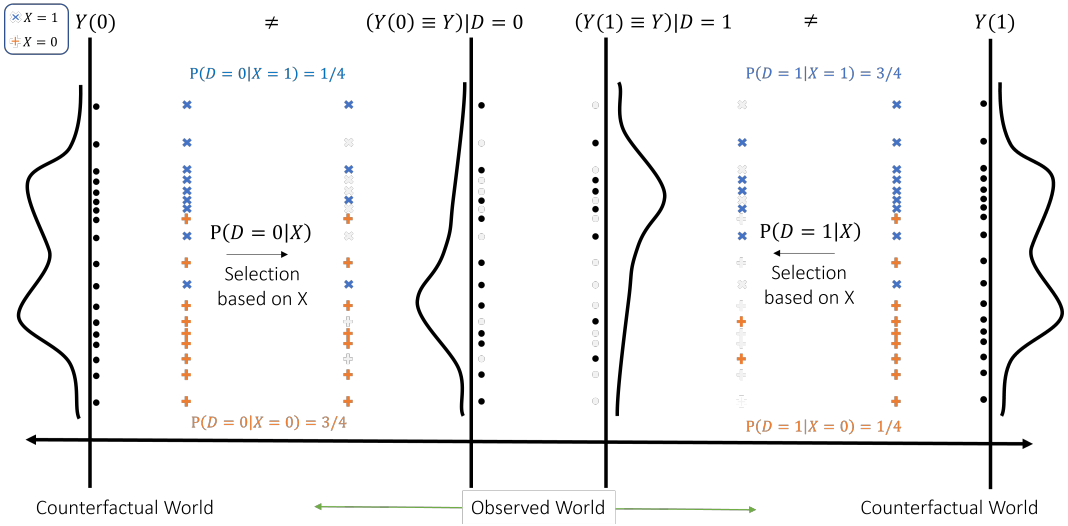
Suppose a wellness app is adopted more by already health-conscious users. If we compare adopters vs non-adopters, the estimated effect on health outcomes is likely:

- A: upward biased
- B: downward biased
- C: unbiased
- D: cannot tell sign

Common Cause Leads to Confounding: a Mechanistic View



Common Cause Leads to Confounding: a Mechanistic View



When does correlation *equal* causation? Randomized assignment

If treatment is randomized,

$$D \perp\!\!\!\perp (Y(1), Y(0)), \qquad \text{(ignorability)}$$

then

$$\begin{aligned}
\mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] &= \mathbb{E}[Y(1) \mid D=1] - \mathbb{E}[Y(0) \mid D=0] && \text{(consistency)} \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] && \text{(ignorability)} \\
&= \mathbb{E}[Y(1) - Y(0)] = \mathsf{ATE}.
\end{aligned}$$

Design beats analysis

Randomization makes treated and control groups comparable *in expectation*.

An Empirical Example

Selection bias and confounding in the Lalonde dataset

Idea

We will look at a dataset (the "Lalonde" dataset) where we can compute:

- an **experimental** (randomized) estimate as a benchmark
- a naive **observational** estimate that can be badly biased

Where does the data come from?

National Supported Work (NSW) Program Evaluation (mid-1970s)

- Federally funded program providing **12–18 months** of work experience to disadvantaged participants
- Among **eligible applicants**, access to NSW was determined by **random assignment** (i.e., some were randomly selected to be *offered* the program, others were assigned to a control group and not offered it).
- Pre-intervention variables collected from **surveys** and **Social Security Administration records**
- Outcome: **1978 earnings**; assignment occurred over **51 months (Mar 1975–Jun 1977)**

LaLonde (1986), *Evaluating the Econometric Evaluations of Training Programs*.

Dehejia & Wahba (1999), *Causal Effects in Nonexperimental Studies: Re-evaluating the Evaluation of Training Programs*.

Empirical example: where the data come from (and how we build it)

- **Experimental sample (random assignment):** treated + randomized controls from NSW program evaluation
- **Observational control samples (non-experimental):**
  - **Panel Study of Income Dynamics** controls (a long-running U.S. household panel survey)
  - **Current Population Survey** controls (a large, repeated U.S. labor-force survey)

We will use two datasets

$$\texttt{lalonde\_exp} = \texttt{NSW treated} \cup \texttt{NSW randomized controls}$$

$$\texttt{lalonde\_obs} = \texttt{lalonde\_exp} \cup \texttt{(survey-based observational controls)}$$

Empirical example: National Supported Work (NSW) job training

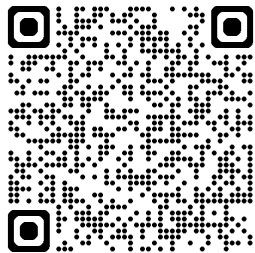
Setup ("Lalonde" dataset)

- Treatment D: offered job training program in 1975
- Outcome Y: earnings in the follow-up year 1978
- Covariates X: age, education, prior earnings ('74,'75), etc.

Why is this dataset interesting?

It contains an experimental sample (random assignment) and a widely used observational control group. It allows us to treat the experimental estimate as a benchmark and demonstrate confounding bias when using the observational controls.

Notebook



Scan to open

Or click here

What we will compute live

1. **Experimental estimate:** difference in mean earnings between treated and randomized controls.
2. **Observational estimate:** treated vs (experimental and non-experimental controls) (large selection bias).
3. **Covariate balance check:** compare pre-treatment variables across groups.

Poll Everywhere

Given the structure of the two datasets, the experimental estimate best corresponds to which causal estimand in the observational setting?

(A) ATE

(B) ATT

Teaser: "adjusting for confounders"

Later in the course we'll formalize identification/estimation in observational studies. For now, a preview of some ideas.

Idea 1: Stratify by X

- If treated and control groups are comparable within strata of X,
- then within-stratum differences can be closer to causal effects.

Idea 2: Re-weight control group

- Put weights on control samples; higher weight on samples that based on X look like they could be coming from treated population,
- then we are "simulating" a control population that looks like the treated population at least in terms of X.

Wrap-up: four takeaways

1. **Potential outcomes** give a precise language for causal questions.
2. **ATE** is a contrast of counterfactual means: $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.
3. **Confounding** is why $\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$ can be wrong.
4. **RCTs/A-B tests** eliminate confounding by design: $D \perp\!\!\!\perp Y(d)$.

Next lecture: identification in observational studies (conditional ignorability; adjusting for confounders).