# MS&E 228 (Winter 2026) — Lecture 1 Notes (Student Handout)<br>Foundations via Potential Outcomes

Vasilis Syrgkanis
Stanford University

January 22, 2026

**Readings.**

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

These notes are written in a "chapter" style but follow the lecture flow closely. The goal is to give you something you can read after class that reproduces the narrative, definitions, and examples we covered.

## 1 What is a causal question?

Many questions in data science and policy have the form "What is the effect of doing $D$ on an outcome $Y$?" For example: in tech, does showing feature $X$ increase retention; in healthcare, does a new protocol reduce readmissions; in social science, do tutoring programs increase test scores; in operations, do preventive maintenance policies reduce downtime. A unifying theme is that these are not just questions about association; they ask what would happen *under a change in the world*. The central difficulty is that for each unit (a user, a patient, a school, a machine) we only see what actually happened, but causal reasoning requires comparing that to what would have happened otherwise.

### Goals for today

By the end of this lecture, you should be able to:

1. define potential outcomes (counterfactuals) $Y(0), Y(1)$;
2. define the average treatment effect (ATE) as a *causal estimand* $\mathbb{E}[Y(1) - Y(0)]$;
3. understand confounding both formally and via examples;
4. see why randomized experiments (A/B tests, randomized controlled trials) are the "ideal" for causal identification.

## 2 The unit, the treatment, and the outcome

We index units by $i = 1, \ldots, n$ (users, patients, schools, machines, …). Each unit has:

- a *binary* treatment indicator $D_i \in \{0, 1\}$ (e.g., receives a feature, a protocol, a program, a policy);
- an outcome $Y_i$ of interest (retention, readmission, test score, downtime, …).

To keep notation light we will typically drop the unit index $i$, but it is always there in the background.

**Key conceptual gap.** We only observe each unit under one treatment state: either treated ($D = 1$) or untreated ($D = 0$). Causal questions require a principled way to compare *two* states for the *same* unit.

## 3 Potential outcomes: defining causal effects

The potential outcomes framework (Neyman and Fisher in the 1920s; Rubin 1974) makes the above gap explicit by introducing two random variables for each unit:

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

where $Y(1)$ is the outcome the unit would have under treatment and $Y(0)$ is the outcome the same unit would have under no treatment.

**Individual causal effect.** The most direct causal object is the unit-level effect $Y(1) - Y(0)$. However, this immediately surfaces the *fundamental problem of causal inference*: for any given unit, we never observe both $Y(1)$ and $Y(0)$.

## 4 From potential outcomes to observed data: consistency and SUTVA

To connect the counterfactual notation to the data we observe, we use:

**Consistency.** The observed outcome equals the potential outcome corresponding to the realized treatment:

$$Y = Y(D).$$

So if $D = 1$ then $Y = Y(1)$, and if $D = 0$ then $Y = Y(0)$.

**SUTVA (Stable Unit Treatment Value Assumption).** Our notation $Y(1), Y(0)$ implicitly assumes: (i) *no interference* between units (my outcome does not depend on your treatment), and (ii) *no hidden versions* of treatment ("treated" means the same intervention for everyone). If interference exists, then unit $i$ might need potential outcomes of the form $Y_i(d_1, \ldots, d_n)$, depending on everyone else's treatment assignment.

**Discussion**

When might SUTVA fail in the motivating examples? For instance, what kinds of product features, health interventions, education programs, or operational policies create spillovers?

# 5 Causal estimands: what do we want to learn?

Because unit-level effects are typically not identifiable, we often focus on average causal effects.

## 5.1 Average treatment effect (ATE)

The average treatment effect is the population average causal effect:

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

This estimand matches questions like "What would happen if we rolled out a feature to everyone?"

## 5.2 Average treatment effect on the treated (ATT)

Sometimes we care about the effect for those who were actually treated:

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

This is natural for evaluation questions like "Among adopters/participants, what was the program's effect?"

## 5.3 Conditional average treatment effect (CATE)

For targeting and heterogeneity, we use covariates $X$ and define:

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

This estimand supports decisions like "For which subpopulations should we deploy an intervention?"

**Quick Check**

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? ATE, ATT, CATE, or none of the above.

# 6 Correlation is not causation: the identification problem

A causal estimand is defined in terms of $Y(1)$ and $Y(0)$, which we never jointly observe. The **identification problem** asks: *Can we rewrite a causal estimand as a function of the observed data distribution?*

## 6.1 The naive comparison

A common (but typically wrong) approach is the difference in observed group means:

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

It is tempting because it is easy to compute and easy to explain. It is usually wrong because treated and untreated groups can differ in many ways besides treatment.

### 6.2 Decomposition: causal effect + confounding bias

Using consistency $Y = Y(D)$, we can decompose:

$$\begin{aligned}\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] &= \mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0] \\ &= \mathbb{E}[Y(1) - Y(0) \mid D = 1] + \underbrace{\Big(\mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]\Big)}_{\text{confounding bias}}.\end{aligned}$$

The first term is the $ATT$. The second term is the difference in untreated potential outcomes between treated and untreated groups: it is nonzero when treatment selection is related to what would have happened without treatment.

**Remark**

Key message: even if treatment helps (think $Y(1) > Y(0)$), confounding bias can mask or exaggerate the effect in the naive comparison.

## 7 Confounding

We say there is **confounding** when treatment status is predictive of the potential outcomes:

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

Equivalently, treated and untreated units are not comparable "as if" randomized.

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

$$X \to D, \qquad X \to Y.$$

Examples from the lecture: power users may receive feature rollouts earlier; sicker patients may receive aggressive treatment; motivated workers may self-select into training; high-performing sites may adopt new processes sooner.

### 7.1 Confounding "in the wild": chocolate and Nobel laureates

An iconic example: countries with more chocolate consumption also tend to have more Nobel laureates. A tempting (wrong) conclusion is "chocolate causes Nobel prizes." A plausible confounder is country wealth: wealth increases chocolate consumption and also increases investments in education and science. The association can be explained by a "backdoor" path $D \leftarrow X \to Y$ even when $D$ does not causally affect $Y$.

**A warning sign.** In observational data you can find many "predictors" of Nobel counts (wine, tea, IKEA stores, ...). The takeaway is that strong correlation can be produced by shared causes, not by a causal effect of $D$ on $Y$.

**Quick Check**

Suppose a wellness app is adopted more by already health-conscious users. If we compare adopters ($D = 1$) vs non-adopters ($D = 0$), the estimated effect on health outcomes is likely:
(A) upward biased (B) downward biased (C) unbiased (D) cannot tell sign.

### 7.2 Mechanistic view: why a common cause creates bias

One way to internalize confounding is to imagine two “counterfactual worlds”: a world where everyone is untreated (revealing $Y(0)$) and a world where everyone is treated (revealing $Y(1)$). In the observed world, we see $Y(0)$ for some units (those with $D = 0$) and $Y(1)$ for others (those with $D = 1$). If treatment selection is related to a common cause $X$, then the subset of units for which we observe $Y(0)$ is not representative of the population distribution of $Y(0)$, and similarly for $Y(1)$. The “missing” counterfactuals are missing in a structured (non-random) way, which is precisely what creates confounding bias.

## 8 When does correlation equal causation? Randomized assignment

Randomized assignment is the clean case where design eliminates confounding. If treatment is randomized, then

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

Under ignorability and consistency,

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

**Design beats analysis:** randomization makes treated and control groups comparable in expectation.

## 9 Empirical example: confounding bias in the LaLonde/NSW job training data

To make the abstraction concrete, we turn to the classic “LaLonde” dataset, which is interesting because it includes both: (i) an experimental sample (random assignment) and (ii) a widely used observational control group.

### 9.1 Where the data come from

The National Supported Work (NSW) program evaluation (mid-1970s) studied a federally funded program providing 12–18 months of work experience to disadvantaged participants (jobs ranged from restaurant work to construction). Among eligible applicants, access to NSW was determined by random assignment: some were offered the program, others were assigned to a control group and not offered it. Pre-intervention variables were collected from surveys and Social Security Administration records. The main outcome is earnings in 1978; assignment occurred over 51 months (Mar 1975–Jun 1977).

### 9.2 How we construct the experimental and observational datasets

We will use two datasets:

- `lalonde_exp`: NSW treated ∪ NSW randomized controls.
- `lalonde_obs`: `lalonde_exp` ∪ *survey-based observational controls.*

The observational controls come from two large U.S. surveys: the Panel Study of Income Dynamics (PSID; a long-running household panel survey) and the Current Population Survey (CPS; a large repeated labor-force survey).

**Quick Check**

Given the structure of the two datasets, the experimental estimate best corresponds to which causal estimand in the observational setting? (A) ATE (B) ATT.

**In-class activity**

What we compute live in the notebook: (1) the experimental estimate: difference in mean 1978 earnings between treated and randomized controls; (2) a naive observational estimate: treated vs pooled controls (experimental + non-experimental), illustrating large confounding bias; (3) a covariate balance check: compare pre-treatment variables across groups.

### Results from running the notebook

When we run the notebook on the Dehejia–Wahba subset of the NSW experiment (treated + randomized controls) and then append the PSID/CPS non-experimental controls, we see the central phenomenon of *confounding bias* very starkly.

**Experimental benchmark (random assignment).** The treated group has mean 1978 earnings $\widehat{\mathbb{E}}[Y \mid D = 1] \approx 6,349.14$ while the randomized controls have mean $\widehat{\mathbb{E}}[Y \mid D = 0] \approx 4,554.80$, giving the difference-in-means estimate:[1]

$$\widehat{\delta}_{\text{exp}} \;=\; 1,794.34, \qquad \widehat{\text{SE}}(\widehat{\delta}_{\text{exp}}) \approx 671.00, \qquad 95\% \text{ CI} \approx [479.19,\, 3,109.50].$$

This is the estimate we treat as a benchmark for the ATE in this experimental sample.

**Naive observational comparison (treated vs. pooled non-experimental controls).** In the observational dataset, the controls have much higher mean earnings (about 14,715.46) than the treated (about 6,349.14), producing a *negative* naive difference:

$$\widehat{\delta}_{\text{naive}} \;=\; -8,366.32, \qquad \widehat{\text{SE}}(\widehat{\delta}_{\text{naive}}) \approx 583.05, \qquad 95\% \text{ CI} \approx [-9,509.09,\, -7,223.54].$$

The sign flip relative to the experiment is not a paradox: it is exactly what we expect when treated units are systematically more disadvantaged than the non-experimental controls.

**Where the confounding bias shows up: covariate imbalance.** A quick balance check shows that the treated group looks very different from the (mostly CPS/PSID) control population in pre-treatment variables. Table 1 summarizes several large imbalances via standardized differences.

[1] we report also here the standard error of the estimate, though we will get into standard errors more formally in subsequent slides; for now just think of them as the typical deviation around the point estimate

Table 1: Covariate imbalance in the observational dataset (treated vs. controls).

| Covariate | Treated mean | Control mean |
|---|---:|---:|
| age | 25.82 | 32.74 |
| education | 10.35 | 11.87 |
| black | 0.84 | 0.12 |
| married | 0.19 | 0.69 |
| no_degree | 0.71 | 0.33 |
| re74 | 2,095.57 | 13,670.53 |
| re75 | 1,532.06 | 13,089.37 |

**Regression "adjustment" moves the estimate but does not solve the problem.** If we run a simple linear regression of $Y = \texttt{re78}$ on the treatment and the pre-treatment covariates, the coefficient on `training` becomes positive:

$$\widehat{\delta}_{\text{reg}} \approx 945.18, \qquad \widehat{\text{SE}} \approx 569.15, \qquad 95\% \text{ CI} \approx [-170.38,\ 2,060.75].$$

This moves in the direction of the experimental benchmark, but it remains sensitive to functional-form assumptions and extrapolation when overlap is poor (as the balance table suggests). This is exactly why, later in the course, we will develop more advanced adjustment methods (doubly robust estimators with ML-based approaches). We will also formally see under which assumptions does this simple linear regression recover the correct ATE.

## Key code snippets

Below are the core blocks from the notebook (trimmed for readability).

**Loading and assembling the experimental and observational datasets**

```python
import pandas as pd

cols = ["training","age","education","black","hispanic","married",
        "no_degree","re74","re75","re78"]

exp_files = ["nswre74_treated.txt", "nswre74_control.txt"]
ctrl_files = ["psid_controls.txt", "psid2_controls.txt", "psid3_controls.txt",
              "cps_controls.txt", "cps2_controls.txt", "cps3_controls.txt"]

lalonde_exp = pd.concat(
    [pd.read_csv(f, sep=r"\s+", header=None, names=cols) for f in exp_files],
    ignore_index=True
)

lalonde_obs = pd.concat(
    [pd.read_csv(f, sep=r"\s+", header=None, names=cols) for f in exp_files + ctrl_files],

    ignore_index=True
)
```

**Difference-in-means (experimental or observational)**

```
import numpy as np

def diff_in_means(df):
    treated = df[df.training == 1]["re78"]
    control = df[df.training == 0]["re78"]
    ate = treated.mean() - control.mean()
    se = np.sqrt(treated.var(ddof=1)/len(treated) + control.var(ddof=1)/len(control))
    return ate, se

ate_exp, se_exp = diff_in_means(lalonde_exp)
ate_obs, se_obs = diff_in_means(lalonde_obs)
```

**A simple balance check (means and standard deviations)**

```
pretreat = ["age","education","black","hispanic","married","no_degree","re74","re75"]
balance = lalonde_obs.groupby("training")[pretreat].agg(["mean","std"])
```

**Linear regression adjustment (illustrative)**

```
import statsmodels.formula.api as smf

formula = "re78 ~ training + age + education + black + hispanic + married + no_degree +
    re74 + re75"
res = smf.ols(formula, data=lalonde_obs).fit()
adj_ate = res.params["training"]
```

## 10 Preview: adjusting for confounders

Later we will formalize identification and estimation in observational studies. For now, two preview ideas:

- **Stratify by $X$.** If treated and control groups are comparable within strata of $X$, then within-stratum differences can be closer to causal effects.
- **Re-weight controls.** Weight control units so that, in terms of $X$, the weighted control population resembles the treated population; this simulates a better comparison group.

## Wrap-up

Four takeaways from the lecture:

1. Potential outcomes provide a precise language for causal questions.
2. The 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 misleading.
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).