

Potential Outcomes and Causal Inference

Gov 51: Data Analysis and Politics



Gov 51 Lecture

Harvard University

Week 11

April 7, 2026



**Two Experiments,
One Lesson**

April 26, 1954: 1.8 million children lined up to stop polio

The Salk Polio Vaccine Trial

- ▷ 1.8 million “Polio Pioneers” enrolled
- ▷ Randomized: vaccine vs. placebo
- ▷ Double-blind: children, parents, doctors all unaware
- ▷ Result: 71% reduction in paralytic polio

Largest field trial in history

Polio cases

Placebo group: 57 per 100k

Vaccine group: 16 per 100k

$$p < 0.0001$$

Lanarkshire, 1930: 20,000 children, four months, no answer

The Lanarkshire Milk Experiment

- ▷ 20,000 Scottish schoolchildren
- ▷ Question: does supplemental milk improve growth?
- ▷ Treatment: free daily milk for 4 months
- ▷ Selection: teachers chose who got milk

Teachers gave milk to the **neediest** children — thinner, malnourished, poorer

The problem

Treated group: worse baseline
Control group: healthier baseline

Comparison is contaminated

“Student” (W.S. Gosset) showed why Lanarkshire was useless

What the data showed:

- ▷ Milk group gained *less* weight than control
- ▷ Did milk hurt children? No.
- ▷ Milk group started *shorter and lighter*
- ▷ Selection of the treated group poisoned every comparison

The treated would have had *worse* outcomes even without milk



Design matters more
than sample size



What Is a Cause?

Hume (1748): to find a cause, imagine the world without it

David Hume, *An Enquiry Concerning Human Understanding* (1748):

- ▷ Causality requires imagining the world *without* the cause
- ▷ Did the match cause the fire? Only if, without the match, there would have been no fire
- ▷ First clue: a cause is defined by its **absence**

“Had it been absent, its effects would have been absent as well”

Mill (1843): a cause is what separates the world where you lived from the one where you didn't

John Stuart Mill, *A System of Logic*
(1843):

- ▷ A man eats a meal and dies. Had he not eaten it, he would not have died.
- ▷ Causation = comparing **two states of the world**
- ▷ The tension: this definition requires measuring the unmeasurable — the same person in both states
- ▷ Second clue: causes are *counterfactual contrasts*

To find the cause:
compare what happened
to what *would*
have happened

Lewis (1973): causation is dependence across possible worlds

David Lewis, *Counterfactuals* (Harvard University Press, 1973):

- ▷ Event c causes event e iff: in the *nearest possible world* where c doesn't occur, e doesn't occur
- ▷ Possible worlds make the counterfactual precise
- ▷ Third clue: causation lives in the **gap between worlds**

All three philosophers are pointing at the same thing: the counterfactual comparison

The cause is what changed
between the world
that happened
and the world that didn't

Fisher arrived at Rothamsted in 1919 to answer: which fertilizer actually works?

Rothamsted Experimental Station (Hertfordshire, UK)

- ▷ Founded 1843 — world's oldest agricultural research station
- ▷ Continuous crop trials since 1843 (Broadbalk wheat plot, still running)
- ▷ Fisher hired 1919 to make sense of decades of messy field data
- ▷ Problem: plots differ in soil, drainage, sunlight

His solution: randomize which plots get which treatment

Fisher's insight

Randomize treatment to plots

⇒ confounding averages out

⇒ inference is valid

Neyman (1923): the first formal statement of potential outcomes

Jerzy Neyman (1923), written in Polish:
“On the Application of Probability Theory to Agricultural Experiments”

- ▷ Comparing crop varieties on randomized plots
- ▷ Introduced the notation $Y_i(1)$, $Y_i(0)$ for potential yields
- ▷ Derived unbiasedness of the difference in means estimator
- ▷ Published in English *67 years later* in 1990

Both Fisher and
Neyman were
answering one question:

*Which seed
grows more food?*

Broockman & Kalla (2016): can a conversation change minds about trans rights?

The experiment:

- ▷ Canvassers visited voters' doors in Florida
- ▷ $D_i = 1$: conversation about transgender rights
- ▷ $D_i = 0$: canvasser discussed recycling (placebo)
- ▷ Outcome: transgender feeling thermometer (0–100)
- ▷ Effects persisted **3 months** later (+6.2 points)

**Broockman
& Kalla**

Science, 2016

$N \approx 1,000$ voters

Durable attitude change



The Potential Outcomes Framework

Every unit has two potential outcomes: one for each treatment state

Notation

- ▷ i indexes units (people, countries, households)
- ▷ $D_i \in \{0, 1\}$: treatment received
- ▷ Y_i^1 : outcome *if* treated
- ▷ Y_i^0 : outcome *if* not treated

Potential outcomes exist *before* treatment is assigned

Y_i^1 : outcome if treated

Y_i^0 : outcome if not treated

The individual treatment effect is the difference — and it is never observed

Individual Treatment Effect

- ▷ $\delta_i = Y_i^1 - Y_i^0$
- ▷ How much did treatment change *this person's* outcome?
- ▷ Requires observing the same person under both states
- ▷ **Impossible:** a person either receives treatment or they don't

$$\delta_i = Y_i^1 - Y_i^0$$

The Fundamental Problem of Causal Inference:
we observe at most one potential outcome per unit

What we observe is determined by treatment received

$$Y_i^{\text{obs}} = D_i \cdot Y_i^1 + (1 - D_i) \cdot Y_i^0$$

If $D_i = 1$ (treated):

$$Y_i^{\text{obs}} = Y_i^1$$

Y_i^0 is the **counterfactual** —
unobserved

If $D_i = 0$ (control):

$$Y_i^{\text{obs}} = Y_i^0$$

Y_i^1 is the **counterfactual** —
unobserved

The counterfactual for every unit is missing data

ATE, ATT, ATU: three ways to average the same unit-level quantities

These are population characteristics:

- ▷ $ATE = E[\delta_i]$: average over the *full* population
- ▷ $ATT = E[\delta_i | D_i = 1]$: average over those who were treated
- ▷ $ATU = E[\delta_i | D_i = 0]$: average over those not treated

None can be computed directly — every $\delta_i = Y_i^1 - Y_i^0$ requires both POs per person

$$ATE = E[Y_i^1 - Y_i^0]$$

ATE, ATT, and ATU differ only in *whose* δ_i we average over



Why Naive Comparisons Fail

Do hospitals make people sicker?

People who go to hospitals are already sick — that's the selection

Naive comparison:

- ▷ Hospital patients have worse outcomes than non-patients
- ▷ Hospital mortality rate $>$ community mortality rate
- ▷ **Conclusion?** Hospitals kill people

What's wrong:

- ▷ Hospital patients were already sicker
- ▷ Their Y_i^0 (without hospital) is already worse
- ▷ The two groups are not comparable

Treated and untreated groups differ in ways *correlated* with potential outcomes

A numerical example: job training with self-selection

Worker	Y_i^1	Y_i^0	δ_i	D_i	Y_i^{obs}
Anna	\$40K	\$30K	\$10K	1	\$40K
Bob	\$50K	\$40K	\$10K	1	\$50K
Carlos	\$20K	\$10K	\$10K	0	\$10K
Diana	\$30K	\$20K	\$10K	0	\$20K

True ATE: $\delta_i = \$10K$ for everyone

Naive SDO:

$$\frac{40+50}{2} - \frac{10+20}{2} = \$45K - \$15K = \mathbf{\$30K}$$

The simple difference in outcomes decomposes into ATE plus selection bias

$$\text{SDO} = \underbrace{E[Y_i^1 - Y_i^0]}_{\text{ATE}} + \underbrace{E[Y_i^0 | D_i = 1] - E[Y_i^0 | D_i = 0]}_{\text{Selection Bias}}$$

In the example:

ATE = \$10K

Selection bias = (35 - 15) = \$20K

SDO = 10 + 20 = \$30K ✓

Selection bias is the gap in baseline outcomes — how much better-off the treated were even *before* treatment

Selection bias: the treated would have differed even without treatment

What selection bias measures:

- ▷ $E[Y_i^0 | D_i = 1]$: what the treated *would have* earned without training
- ▷ $E[Y_i^0 | D_i = 0]$: what the untreated actually earned
- ▷ If these differ, the comparison is contaminated

Positive selection bias

$$E[Y_i^0 | D_i = 1] > E[Y_i^0 | D_i = 0]$$
$$\text{SDO} > \text{ATE}$$

Negative selection bias

$$E[Y_i^0 | D_i = 1] < E[Y_i^0 | D_i = 0]$$
$$\text{SDO} < \text{ATE}$$



**Randomization as
the Solution**

Randomization makes potential outcomes independent of treatment

Random assignment means:

- ▷ Who gets treated is determined by a coin flip
- ▷ Not by earnings potential, health, or anything else
- ▷ Potential outcomes $\{Y_i^0, Y_i^1\}$ are *independent* of D_i

Consequence:

- ▷ $E[Y_i^0 \mid D_i = 1] = E[Y_i^0 \mid D_i = 0] = E[Y_i^0]$
- ▷ Selection bias = 0

$$\{Y_i^0, Y_i^1\} \perp\!\!\!\perp D_i$$

“Independence”: treatment assignment carries no information about potential outcomes

Under independence, the SDO is an unbiased estimator of the ATE

The algebra:

$$\begin{aligned}\text{SDO} &= E[Y_i^1 \mid D_i = 1] - E[Y_i^0 \mid D_i = 0] \\ &= E[Y_i^1] - E[Y_i^0] \quad (\text{by independence}) \\ &= E[Y_i^1 - Y_i^0] \\ &= \text{ATE}\end{aligned}$$

Under randomization:

$$\text{SDO} = \text{ATE}$$

The naive comparison
is the causal effect

Randomization doesn't solve the Fundamental Problem — it averages around it

We still cannot observe $\delta_i = Y_i^1 - Y_i^0$

- ▷ Each person is still only seen in one state
- ▷ Individual treatment effects remain unidentified
- ▷ What randomization gives us: **average** effects

Randomization creates *exchangeable* groups—
neither group is special

What randomization actually does:

- ▷ Creates two groups with identical distributions of Y^0 (in expectation)
- ▷ Any difference in outcomes is therefore due to treatment



From Theory to Data

Step 1: load the BK data and check covariate balance

```
library(tidyverse)
bk <- read_csv("bk2016.csv")

# Check balance: key covariates by treatment group
bk |>
  group_by(treatment) |>
  summarize(pct_female = mean(female),
            mean_age    = mean(age),
            mean_therm0 = mean(trans_therm_t0, na.rm=TRUE),
            n = n())
```

##	treatment	pct_female	mean_age	mean_therm0	n
##	0	0.593	46.3	52.95	913
##	1	0.582	47.7	53.61	912

Randomization worked: all pre-treatment covariates are nearly identical across groups

A large sample doesn't fix a bad design — it makes it worse

The CLT/LLN paradox:

- ▷ LLN: sample means converge to population means as $n \rightarrow \infty$
- ▷ CLT: sampling variance shrinks to zero
- ▷ But what do they converge *to*?

If selection bias exists in the population:

- ▷ More data \Rightarrow more precise estimate of ATE + Selection Bias
- ▷ You converge to the *wrong answer*, with certainty

Precision \neq Validity

With infinite data and
no randomization,
 $SDO \xrightarrow{p} ATE + SB$

Step 2: estimate the ATE as a difference in means

```
# Outcome: trans_therm_t3 (feeling thermometer, 0-100)
# Restrict to 3-month follow-up respondents
bk_t3 <- bk |> filter(responded_t3 == 1)
bk_t3 |>
  group_by(treatment) |>
  summarize(mean_therm = mean(trans_therm_t3), n = n())
## treatment mean_therm n
##          0         52.51 287
##          1         58.74 280
# ATE estimate: 58.74 - 52.51 = 6.23
```

$$\widehat{\text{ATE}} = \bar{Y}_{D=1}^{\text{obs}} - \bar{Y}_{D=0}^{\text{obs}} = 58.74 - 52.51 = 6.23$$

Step 3: OLS with a binary treatment gives the same estimate

```
# OLS regression: treatment as binary variable
fit <- lm(trans_therm_t3 ~ treatment, data = bk_t3)
coef(fit)
## (Intercept)      treatment
##      52.505          6.230

# Intercept = E[Y | treatment=0] = 52.5
# Coefficient = difference in means = 6.2 pts (same)
```

The Neyman variance estimator is the right standard error for a randomized experiment

$$\widehat{\text{SE}}_{\text{Neyman}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$$

- ▷ s_1^2, s_0^2 : sample variances in treated and control groups
- ▷ n_1, n_0 : group sample sizes
- ▷ Does *not* assume equal variances across groups

OLS with HC2

heteroskedasticity-robust standard errors gives *exactly* the Neyman SE. Classical (homoskedastic) OLS SE does not.

The OLS coefficient on a binary treatment equals the difference in means

$$\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$$

Why this is true:

- ▷ With one binary regressor, OLS fits two group means
- ▷ Slope = difference in group means
- ▷ Intercept = mean of the $D = 0$ group

OLS is the natural estimator. Adding covariates increases precision, never changes what it estimates under randomization.



A coin flip makes
groups comparable.
That's why it's the most powerful
tool in social science.