Potential Outcomes and Causal Inference

Gov 51 Section — Week 11

George

Harvard University

April 8, 2026

The Problem with Self-Selection

Harvard students choose their concentrations — are those groups comparable?

Two Harvard concentrations:

- Government: ~120 students
- Computer Science: ~80 students
- Students choose their own concentration

Question: Is the gender composition similar across concentrations?

- Gov: $\approx 62\%$ female
- CS: $\approx 22\%$ female
- **Not even close to balanced**

Self-selection problem

The groups differ on a pre-existing characteristic

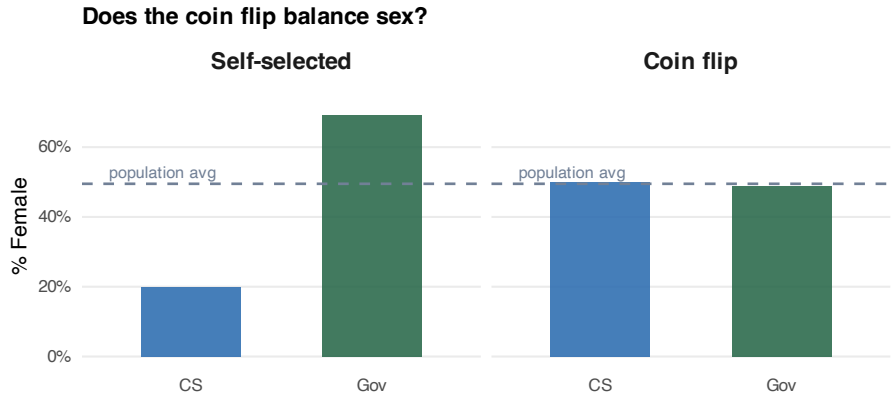
Any Gov vs. CS comparison is contaminated

If we flipped a coin to force students into concentrations, would the sex ratios differ?

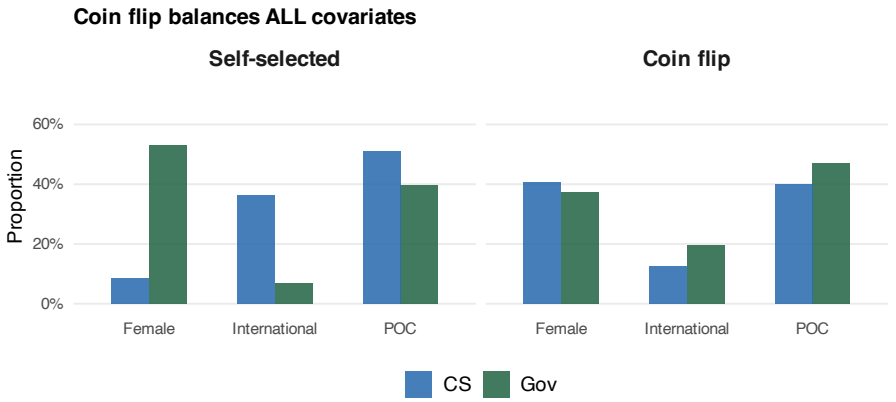
Code: self-selection creates imbalance — coin flip destroys it

```
set.seed(200)
harvard <- tibble(
  female = c(rbinom(120, 1, 0.62), rbinom(80, 1, 0.22)),
  choice = c(rep("Gov", 120), rep("CS", 80)),
  coin   = rbinom(200, 1, 0.5)    # random assignment
)
harvard |> group_by(choice) |> summarize(pct_female = mean(
    female))
## Gov: 69%     CS: 20%    <-- very different
harvard |> group_by(coin)  |> summarize(pct_female = mean(
    female))
## coin=0: 49%    coin=1: 50%    <-- balanced
```

The coin flip balances sex across concentration groups



The coin flip balances *every* pre-existing characteristic simultaneously



Why randomization balances everything: independence

What a coin flip does:

- Treatment assignment D_i is determined by chance
- D_i carries no information about any pre-existing attribute
- Sex, race, age, ambition, GPA — all unrelated to D_i

Result:

- Both groups have the same distribution of every covariate *in expectation*
- Not just one covariate — **all of them at once**

Randomization $\Rightarrow$

All pre-existing attributes have equal distribution across groups

Even potential outcomes Y_i^1 and Y_i^0 balance under randomization

```r
set.seed(51)
po_sim <- tibble(
  Y1 = rnorm(1000, mean=75, sd=15),   # PO if treated
  Y0 = rnorm(1000, mean=60, sd=15),   # PO if not treated
  D  = rbinom(1000, 1, 0.5)           # coin flip
)

po_sim |> group_by(D) |>
  summarize(mean_Y1=mean(Y1), mean_Y0=mean(Y0))
## D  mean_Y1  mean_Y0
## 0     74.8     59.7
## 1     75.2     60.3
```

Y^1 and Y^0 are nearly identical across treatment groups — the coin flip is independent of both

Randomized Experiments

A randomized experiment makes the treatment group comparable to control

Key ingredients:

- A **treatment** $D_i \in \{0, 1\}$ assigned by chance
- An **outcome** Y_i we observe afterward
- Groups comparable *at baseline* by design

What randomization buys:

- No systematic differences between groups
- Any difference in Y must be due to D

Randomization $\Rightarrow$

treated and control groups are *exchangeable at baseline*

Today's experiment: race and hiring in the labor market

Bertrand & Mullainathan (2004)

- Sent 4,870 fictitious resumes to real job ads
- Randomly assigned white- or Black-sounding names
- Everything else held identical
- Outcome: did the employer call back?

Treatment: `white` $= 1$ (white-sounding name)
Outcome: `call` $= 1$ (employer called back)

The key insight

Name is randomly assigned $\Rightarrow$ no selection bias $\Rightarrow$ diff in call-backs is causal

Estimator 1: difference in means

$$\widehat{\text{ATE}} \quad = \quad \bar{Y}_{D=1} \; - \; \bar{Y}_{D=0}$$

```
library(tidyverse)
resume <- read_csv("resume.csv") |>
  mutate(white = as.integer(race == "white"))
resume |>
  group_by(white) |>
  summarize(callback = mean(call))
## white  callback
##     0  0.03448  # Black-sounding names
##     1  0.06652  # White-sounding names
0.06652 - 0.03448   # ATE estimate = 0.032
```

3.2 percentage points more callbacks for white-sounding names

Estimator 2: OLS regression with a binary treatment

$$Y_i = \alpha + \beta D_i + \varepsilon_i \quad \Rightarrow \quad \hat{\beta} = \bar{Y}_{D=1} - \bar{Y}_{D=0}$$

```
# OLS regression
fit <- lm(call ~ white, data = resume)
coef(fit)
## (Intercept)         white
##   0.03448276   0.03203285

# Intercept = mean(call | white = 0) = 3.45%
# Slope     = diff in means          = 3.20 pp
```

The OLS coefficient on `white` is identical to the difference in means

OLS with a binary regressor is just a difference in means

Why they're the same:

- OLS minimizes SSR by fitting two group means
- With one binary regressor:
  - $\hat{\alpha} = \bar{Y}_{D=0}$ (control mean)
  - $\hat{\beta} = \bar{Y}_{D=1} - \bar{Y}_{D=0}$ (gap)
- The "regression" is just connecting two points

Difference in means: 0.032

OLS coefficient: 0.032

Identical — always

Adding covariates to OLS increases *precision* but doesn't change what it estimates under randomization

Potential Outcomes

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

Notation:

- Y_i^1: outcome unit i would get *if treated*
- Y_i^0: outcome unit i would get *if not treated*
- $D_i \in \{0, 1\}$: treatment actually received

Observed outcome:

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

Y_i^1 exists for *every* unit
Y_i^0 exists for *every* unit
But we only *observe* one
The other is the counterfactual

The individual treatment effect is the difference — never observed

Individual treatment effect:

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

- How much did treatment change *this person's* outcome?
- Requires observing same person under both states
- **Impossible**: Holland's (1986) Fundamental Problem of Causal Inference

The Fundamental Problem

We observe Y_i^1 *or* Y_i^0
— never both —

δ_i is always missing data

The ATE is the average we can identify with randomization

Average Treatment Effect (ATE):

$$\text{ATE} = E[Y_i^1 - Y_i^0]$$

- Average δ_i across the full population
- We can't observe individual effects — but we can average
- Under randomization: SDO = ATE

Why randomization works:

- D_i assigned by chance $\Rightarrow$
- $\{Y_i^0, Y_i^1\} \perp\!\!\!\perp D_i$
- Treated and control groups have identical *baseline* outcomes in expectation
- Any difference in observed Y = causal effect

Harvard vs. Dartmouth

Does attending Harvard raise your salary?

Setup: 6 students, two universities, wages at age 30

The scenario:

- 3 students attend Harvard ($D_i = 1$)
- 3 students attend Dartmouth ($D_i = 0$)
- Outcome: annual salary at age 30 ($K)
- **Key:** Harvard students *chose* to attend Harvard
- Assumption: Harvard raises every student's salary by $30K

Constant treatment effect

$\delta_i = 30$ for every student

Harvard adds $30K to *whatever* you would have earned

The full potential outcomes table

Student	Y_i^1 (Harvard)	Y_i^0 (Dartmouth)	δ_i	D_i	Y_i^{obs}
Aria	\$160K	\$130K	\$30K	1	\$160K
Ben	\$180K	\$150K	\$30K	1	\$180K
Chen	\$170K	\$140K	\$30K	1	\$170K
Dana	\$90K	\$60K	\$30K	0	\$60K
Eli	\$110K	\$80K	\$30K	0	\$80K
Fiona	\$100K	\$70K	\$30K	0	\$70K

True ATE = \$30K (constant for everyone) | Shaded rows: we observe only Y_i^{obs}

The naive comparison massively overstates Harvard's effect

What we can observe:

- Harvard students' avg salary: $\frac{160+180+170}{3} = \$170K$
- Dartmouth students' avg salary: $\frac{60+80+70}{3} = \$70K$

The simple difference:

$$\text{SDO} = \$170K - \$70K = \mathbf{\$100K}$$

True ATE = \$30K

Naive comparison = \$100K

The SDO is *3× larger* than the true effect

Harvard students would have earned more even at Dartmouth

The baseline potential outcomes:

- Harvard students' Y_i^0: $\frac{130+150+140}{3} = \$140K$
- Dartmouth students' Y_i^0: $\frac{60+80+70}{3} = \$70K$

Selection bias:

$$E[Y_i^0 \mid D_i = 1] - E[Y_i^0 \mid D_i = 0] = \$140K - \$70K = \mathbf{\$70K}$$

Harvard students are more ambitious, more talented, from wealthier backgrounds. They would have outearned Dartmouth students *regardless* of which school they attended. **This is selection bias.**

The decomposition: SDO = ATE + Selection Bias

$$\underbrace{\text{SDO}}_{\$100K} = \underbrace{\text{ATE}}_{\$30K} + \underbrace{E[Y_i^0 \mid D_i = 1] - E[Y_i^0 \mid D_i = 0]}_{\text{Selection Bias}=\$70K}$$

Breaking it down:

- \$30K of the gap = Harvard's causal effect
- \$70K of the gap = Harvard admits better students
- We *cannot* tell them apart without randomization

$$\$100K = \$30K + \$70K$$

✓

Selection bias: the treated would have been better off anyway

Definition:

$$\text{Selection Bias} = E[Y_i^0 \mid D_i = 1] - E[Y_i^0 \mid D_i = 0]$$

- Compares baseline potential outcomes across treatment groups
- Positive selection: treated group would have done better *even without treatment*
- Exists whenever treatment is correlated with pre-existing characteristics

In our example:

- Harvard admits ambitious, talented students
- These students would have earned \$140K at Dartmouth
- Dartmouth students would have earned \$70K
- The \$70K gap is the selection, not Harvard

The ATE: what Harvard *actually* does to wages

ATE = $30K in our example

- Every student gets $30K more from attending Harvard vs. Dartmouth
- This is the causal effect — what would change if you *randomly assigned* students to schools
- Aria: $160K − $130K = 30K$
- Ben: $180K − $150K = 30K$
- Dana: $90K − $60K = 30K$

$$\text{ATE} = E[Y_i^1 - Y_i^0] = \$30K$$

The ATE is what a *lottery* for Harvard admissions would reveal. Without randomization, selection bias hides it.

Randomization would make the Harvard comparison valid

What randomization would do:

- Randomly assign students to Harvard or Dartmouth
- Treatment group: random mix of Aria, Dana, Fiona, etc.
- Control group: random mix of the same types
- In expectation: $E[Y_i^0 \mid D = 1] = E[Y_i^0 \mid D = 0]$
- Selection bias $= 0$
- SDO $=$ ATE

Under randomization:

SDO $=$ ATE $= \$30K$

No selection bias

Dale & Krueger (2002) tried to approximate this by comparing students who were admitted to the same set of schools but chose differently

Simulation:
The Weight Loss Pill

Step 1: generate 10,000 people with realistic demographics

```
set.seed(51); N <- 10000
sim <- tibble(
  age      = round(rnorm(N, 45, 12)),
  female   = rbinom(N, 1, 0.52),
  poc      = rbinom(N, 1, 0.35),
  college  = rbinom(N, 1, 0.45),
  # Males 18 lbs heavier on average; slight age gradient
  Y0 = round(185 + 18*(1-female) + 0.3*(age-45) + rnorm(N
     ,0,28))
)
```

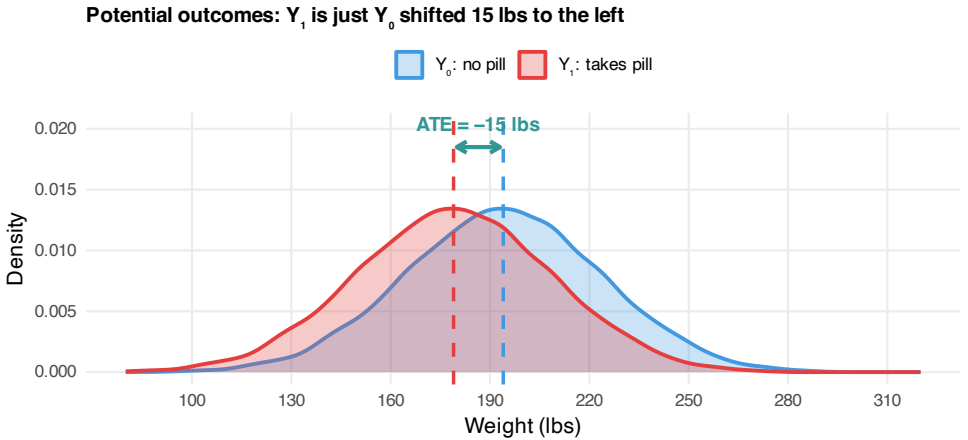
Mean weight: **194 lbs**. 52% female. 35% POC. 45% college.

Step 2: the pill reduces weight by 15 lbs — for everyone

```
sim <- sim |>
  mutate(
    delta = -15,          # constant treatment effect: -15 lbs
    Y1    = Y0 + delta    # every person's treated weight = Y0
        - 15
  )

mean(sim$delta)   # ATE = -15
mean(sim$Y0)      # 194.0 lbs
mean(sim$Y1)      # 179.0 lbs
```

The two potential outcomes are the same distribution, shifted left 15 lbs



If we gave the pill to everyone, average weight would drop from 194 to 179 lbs.

But we can't observe both Y_i^0 and Y_i^1 for the same person.

So how do we learn about the ATE from data where each person gets one pill or no pill?

Step 3: the switching equation — we only observe one outcome per person

```
# Assign treatment (two scenarios below)
sim <- sim |>
  mutate(
    Y_obs = D * Y1 + (1 - D) * Y0
  )
# If D=1: we see Y1 (treated weight)
# If D=0: we see Y0 (untreated weight)
# The other potential outcome is forever unobserved
```

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

Scenario A:
Heavy People Choose the Pill

Scenario A: only above-average weight people take the pill

```
mean_Y0 <- mean(sim$Y0)    # 194 lbs

sim <- sim |>
  mutate(
    D_A = as.integer(Y0 > mean_Y0),    # 1 if heavier than
        average
    Y_A = D_A * Y1 + (1 - D_A) * Y0   # switching equation
  )

sim |> group_by(D_A) |>
  summarize(mean_Y0 = mean(Y0), n = n())
## D_A   mean_Y0     n
##   0     169.9  4920     (lighter people: no pill)
##   1     217.3  5080     (heavier people: took pill)
```

Before we compute anything: the treated group starts 47 lbs heavier than the control group.

What do you expect the naive comparison $(\bar{Y}_{D=1} - \bar{Y}_{D=0})$ to show?

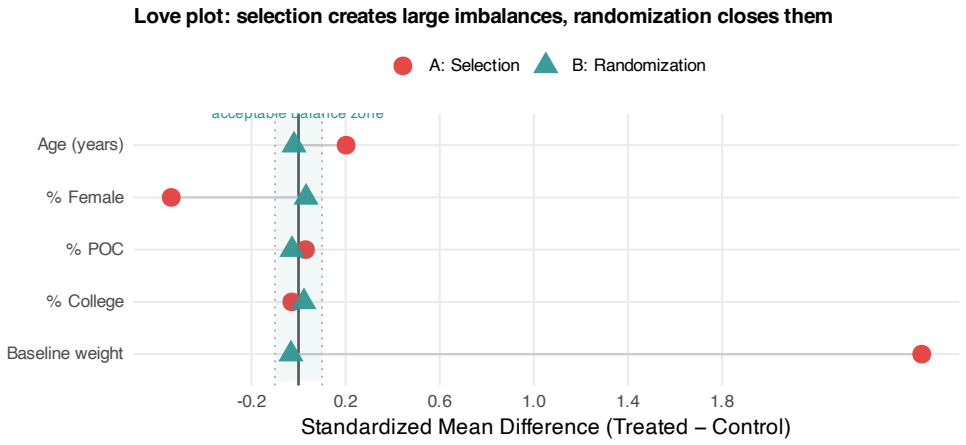
Will it look like the pill works?

Scenario A: covariate balance check — the groups are systematically different

```
sim |> group_by(D_A) |>
  summarize(
    mean_age    = mean(age),
    pct_female  = mean(female)*100,
    pct_poc     = mean(poc)*100,
    mean_weight = mean(Y0)
  )
## D_A  mean_age  pct_female  pct_poc  mean_weight
##   0      43.4        63.1     35.5        169.9
##   1      46.4        40.7     34.4        217.3
```

Treated group: older, **more male**, much heavier — selection is pervasive

Love plot: selection creates large imbalances — randomization closes them



Scenario A: computing the SDO

```r
# Mean observed outcome by treatment group
mean(sim$Y_A[sim$D_A == 1])    # 202.3 lbs  (heavy people
    AFTER pill)
mean(sim$Y_A[sim$D_A == 0])    # 169.9 lbs  (light people, no
     pill)

# Simple difference in outcomes
SDO_A <- mean(sim$Y_A[sim$D_A==1]) - mean(sim$Y_A[sim$D_A
    ==0])
SDO_A    # +32.4 lbs  <--- pill appears to make people
    HEAVIER!
```

The pill reduces weight by 15 lbs but
the naive comparison says +32 lbs

Scenario A: the decomposition reveals why

$$\underbrace{\text{SDO}}_{+32.4} = \underbrace{\text{ATE}}_{-15} + \underbrace{E[Y_i^0 \mid D=1] - E[Y_i^0 \mid D=0]}_{\text{Selection Bias}=+47.4}$$

The math:

- ATE $= -15$ lbs (pill works!)
- Selection bias $= 217.3 - 169.9 = +47.4$ lbs
- Heavy people chose the pill — they would have been heavier *anyway*

$-15 + 47.4 = +32.4$

Selection bias is so large it **flips the sign**

OLS confirms: the pill looks like it adds 32 lbs

```r
# Regression of Y_obs on treatment indicator
fit_A <- lm(Y_A ~ D_A, data = sim)
coef(fit_A)
## (Intercept)         D_A
##      169.87       32.37     <-- pill appears to add 32 lbs!

# Intercept = E[Y | D=0] = 169.9
# Coefficient = SDO = 32.4    (OLS = diff-in-means, as always
    )
```

The regression is not wrong — it correctly estimates the SDO. The SDO is wrong.

Scenario B:
The Pill is Randomly Assigned

Scenario B: assign the pill by coin flip

```
sim <- sim |>
  mutate(
    D_B = rbinom(N, 1, 0.5),            # coin flip
    Y_B = D_B * Y1 + (1 - D_B) * Y0   # switching equation
  )

sim |> group_by(D_B) |>
  summarize(mean_Y0 = mean(Y0), n = n())
## D_B   mean_Y0     n
##   0     194.5  4940    (control group)
##   1     193.5  5060    (treated group)
```

Treated and control groups start at **the same weight**. Selection bias $\approx$ 0.

The treated and control groups now have the same average baseline weight — and also the same average age, sex ratio, and education level.

Why does the coin flip balance all of these at once?

We never explicitly controlled for any of them.

Scenario B: computing the SDO

```r
mean(sim$Y_B[sim$D_B == 1])    # 178.5 lbs  (treated group)
mean(sim$Y_B[sim$D_B == 0])    # 194.5 lbs  (control group)

SDO_B <- mean(sim$Y_B[sim$D_B==1]) - mean(sim$Y_B[sim$D_B
   ==0])
SDO_B    # -15.95 lbs  <--- approximately equal to ATE!

# Verify the decomposition
SB_B <- mean(sim$Y0[sim$D_B==1]) - mean(sim$Y0[sim$D_B==0])
SB_B     #  -0.95 lbs  (approximately zero)
# ATE + SB = -15 + (-0.95) = -15.95 = SDO  [check]
```

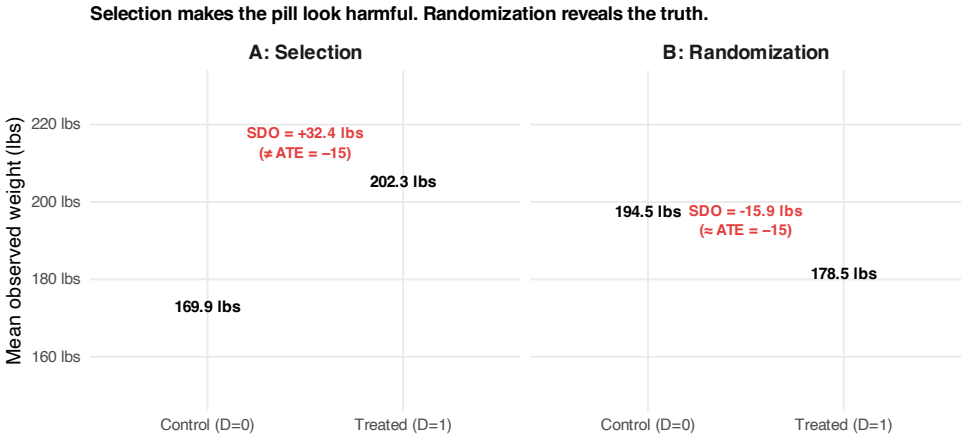
OLS gives the causal estimate under randomization

```r
fit_B <- lm(Y_B ~ D_B, data = sim)
coef(fit_B)
## (Intercept)         D_B
##      194.47      -15.95    <-- pill reduces weight ~15 lbs

# Intercept = E[Y | D=0] = 194.5
# Coefficient = SDO = -15.95  (~ATE = -15)
```

Comparing the Two Scenarios

The same pill, the same world — only the assignment changed



Summary: what changed from A to B?

	True ATE	SDO	Selection Bias
Scenario A (selection)	-15 lbs	$+32.4$ lbs	$+47.4$ lbs
Scenario B (randomization)	-15 lbs	-15.9 lbs	-0.9 lbs

What the decomposition says:

- SDO = ATE + Selection Bias — always
- In A: SB $= +47.4$ dominates and flips the sign
- In B: SB ≈ 0, so SDO $\approx$ ATE

Same data-generating process.
Same true ATE.

Only the assignment rule changed.

A pharmaceutical company runs a trial where patients choose whether to take the drug.

The drug group has worse health outcomes than the control group.

Does the drug make people sicker?

What does this simulation tell us about how to interpret that result?

Putting It Together

The SDO decomposition is a mechanical identity — always true

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

This is algebra, not an assumption:

- Add and subtract $E[Y_i^0 \mid D_i = 1]$ from the SDO
- The decomposition is *always* true — with or without randomization
- It tells us what we're missing when we compare groups naively

With our coin flip:

$E[Y_i^0|D = 1] = E[Y_i^0|D = 0]$

$\Rightarrow$ Selection Bias $= 0$

$\Rightarrow$ SDO $=$ ATE

Randomization zeroes out the selection bias term

Why the coin flip works:

- Coin flip $\Rightarrow \{Y_i^0, Y_i^1\} \perp\!\!\!\perp D_i$
- Both treated and control groups are a random sample from the same population
- Their baseline outcomes Y_i^0 have the same distribution
- So $E[Y_i^0 \mid D_i = 1] = E[Y_i^0 \mid D_i = 0]$
- Selection bias $= 0$, always

Under randomization:

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

The naive comparison *is* the causal effect

The SDO decomposition is always true.

Randomization makes the selection bias term zero.

That's the whole idea.