Experiments: Part I

PS200C Quant Methods in Politics, Spring 2026

Chad Hazlett

UCLA

chazlett@ucla.edu

Randomization Solves the Selection Problem

Recall this formula for the bias of the DIM ($\tilde{\tau}$) relative to ATT:

$$\begin{aligned}
\tilde{\tau} &= \mathbb{E}[Y_i|D_i=1] - \mathbb{E}[Y_i|D_i=0] \quad \text{(obs. diff. in means)} \\
&= \mathbb{E}[Y_{1i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=0] \\
&= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i=1]}_{\tau_{ATT}} + \underbrace{\mathbb{E}[Y_{0i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=0]}_{\text{Bias}}
\end{aligned}$$

Stare at the bias term and ask: What could ensure it is zero?

Random assignment of D_i will make the treated and untreated units identical on average, such that

$$\mathbb{E}[Y_{0i} \mid D_i = 1] = \mathbb{E}[Y_{0i} \mid D_i = 0]$$

Historical Context

From “Experimentation and social interventions: a forgotten but important history” (Ann Oakley, 1998, *BMJ*). From the summary points:

- Many social scientists argue that randomised controlled trials are inappropriate for evaluating social interventions, but they ignore a considerable history, mainly in the United States, of the use of randomised controlled trials to assess different approaches to public policy and health promotion
- A tradition of experimental sociology was well established by the 1930s, built on the early use of controlled experiments in psychology and education
- From the early 1960s to early 1980s randomised experiments were considered the optimal design for evaluating public policy interventions in the United States, and major evaluations using this design were carried out.[1]
- This approach became less popular as policy makers reacted negatively to evidence of “near zero” effects
- Lessons to be learnt about implementing randomised controlled trials in real life settings include the difficulty of assessing complex multi-level interventions and the challenge of integrating qualitative data

[1] Also note Campbell & Stanley published “Experimental and Quasi-experimental designs for research” in 1966

Re-rise of experiments

Though much of the history forgotten, renewed enthusiasm and "rediscovery" of experiments for social sciences early this century

- *Abbreviated* list of examples (from Green 2008):
  - Program evaluation: development programs, education programs, SAT prep classes, weight loss programs, diversity training, deliberative polls, advertising campaigns, website designs...
  - Public policy evaluation: teacher pay, student incentives, class size, speed traps, vouchers, alternative sentencing, job training, health insurance subsidies, tax compliance, public housing
  - Behavioral research: persuasion, mobilization, education, income, interpersonal influence, conscientious health behaviors, media exposure, deliberation, discrimination
  - Research on institutions: transparency, corruption, electoral systems, information
- Also 2019 Nobel prize in economics.

Identification vs. Estimation

Goal of causal inference: Learn about an *unobservable, counterfactual-dependent* quantity of interest (QoI) using *observed* data.

Two inferential hurdles:

1. Identification: If you can observe data from an entire *population*, can you learn about your QoI?
2. Estimation: Given your finite amount of data on a *sample*, how well can you learn about your QoI?

Golden rule of inference: **Identification precedes estimation**

The Template

Before we say more about experiments, take note of the general template for talking about an identification strategy; we will repeat it through the remainder of the course.

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

Setup:

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

Notes:

- Random assignment can take one of several forms:
  - 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):

$$\{Y_{1i},\ Y_{0i}\} \perp\!\!\!\perp D_i$$

Quantity of Interest:

$$\tau_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

How does our ID assumption lead to identification of τ_{ATE}?

$$\begin{aligned}
\mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \quad \text{(why?)} \\
&= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \quad \text{(why?)} \\
&= \text{DIM Estimand} \\
&\approx \overline{Y}_1 - \overline{Y}_0 \quad \text{(the DIM estimator)}
\end{aligned}$$

SATE and PATE

- Often we focus on estimating the average causal effect in a particular sample: **S**ample **A**verage **T**reatment **E**ffect (SATE)
  - uncertainty arises only from hypothetical randomizations.
  - inferences are limited to the sample in our study.
- Compare to the **P**opulation **A**verage **T**reatment **E**ffect (PATE)
  - Imagines uncertainty in both the sample drawn from the (larger) population, and in treatment assignment.
- If your sampling was random from a given population, then $\mathbb{E}[SATE] = PATE$. For this reason, we're often uninterested in distinguishing the two quantities.

Standard Errors for the ATE

Variance of the difference in means estimator:

$$\mathbb{V}(\overline{Y}_1 - \overline{Y}_0) = \frac{\sigma^2_{Y1}}{N_1} + \frac{\sigma^2_{Y0}}{N_0}$$

where σ^2_{Y1} and σ^2_{Y0} are the variance of the Y_{1i} and Y_{0i} in the population.

The Neyman estimator:

$$\widehat{SE}_{ATE} = \sqrt{\frac{\hat{\sigma}^2_{Y1}}{N_1} + \frac{\hat{\sigma}^2_{Y0}}{N_0}}$$

Key facts:

- This is the correct SE for the **PATE**
- It is what you get from t-tests (unequal variance), or regression with robust SE
- It is **conservative** for the **SATE** (proof in appendix)

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16,199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17,038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15,040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16,180
- Estimated average effect of training:
  - $\widehat{\tau}_{ATE} = \bar{Y}_1 - \bar{Y}_0 = 16,199 - 15,040 = \$1,159$
- Estimated standard error for effect of training:

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{\sigma}^2_{Y|D_i=1}}{N_1} + \frac{\widehat{\sigma}^2_{Y|D_i=0}}{N_0}} = \sqrt{\frac{17,038^2}{7,487} + \frac{16,180^2}{3,717}} \approx \$330$$

- Is this consistent with a zero average treatment effect $\alpha_{ATE} = 0$?

Testing the Null Hypothesis of Zero Average Effect

Null hypothesis, H_0: $\tau_{ATE} = 0$

We observe a difference in mean earnings of $\hat{\tau}_{ATE} = 1,159$

Would it be surprising to see such an estimate had the null been true?

- you likely know already:
- 1,159 is more than 2 SEs from 0, so you know t-stat/ z-score would be over 2 and p-value would be less than 0.05.
- to recall the formalities,

Testing the Null Hypothesis of Zero Average Effect

Use a two-sample t (or z) test with unequal variances:

$$t = \frac{\widehat{\tau}}{\sqrt{\frac{\widehat{\sigma}^2_{Y_i|D_i=1}}{N_1} + \frac{\widehat{\sigma}^2_{Y_i|D_i=0}}{N_0}}} = \frac{\$1,159}{\sqrt{\frac{\$17,038^2}{7,487} + \frac{\$16,180^2}{3,717}}} \approx 3.5$$

- For large enough N, $t \xrightarrow{d} N(0,1)$,
- Getting a 3.5 or above from a normal distribution would be quite a shock:

```
> 2*(1-pnorm(abs(3.5)))
[1] 0.000465
```

- Inverting the test statistic we can construct a 95% confidence interval

$$\widehat{\tau}_{ATE} \pm 1.96 \cdot \widehat{SE}_{\widehat{ATE}}$$

- What assumptions did we need along the way?

Testing the Null Hypothesis of Zero Average Effect

```
> d <- read.dta("jtpa.dta")
> t.test(earnings~assignmt,data=d,var.equal=FALSE)

        Welch Two Sample t-test

data:  earnings by assignmt
t = -3.5084, df = 7765.599, p-value = 0.0004533
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1807.2427  -511.6239
sample estimates:
mean in group 0 mean in group 1
       15040.50        16199.94
```

Regression to Estimate ATE in Experiments

Estimator (Regression with Experiments)

The ATE can be expressed as a regression coefficient:

$$
\begin{aligned}
Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
&= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\varepsilon} \\
&= \alpha + \tau_{Reg} D_i + \varepsilon_i
\end{aligned}
$$

- Does this assume constant treatment effects?
- Our SE estimator allows different variance for $D = 1$ and $D = 0$.
  - Implies heteroskedasticity
  - Use "HC2" heteroskedasticity-robust variance:

$$\hat{\sigma}^2_{HC2} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = \tilde{\mathbb{V}}(\tilde{\tau})$$

Regression to Estimate the Average Treatment Effect

```
> library(sandwich)
> library(lmtest)
>
> lout <- lm(earnings~assignmt,data=d)
> coeftest(lout,vcov = vcovHC(lout, type = "HC2"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 15040.50     265.38 56.6752 < 2.2e-16 ***
assignmt     1159.43     330.46  3.5085 0.0004524 ***
--
```

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_1] = \mathbb{E}[Y_0], \quad H_1 : \mathbb{E}[Y_1] \neq \mathbb{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test:

$$H_0 : Y_{1i} = Y_{0i}, \quad H_1 : Y_1 \neq Y_0 \ \ \forall i, \qquad \text{(sharp null of no effect)}$$

Key idea: Under the sharp null, we "observe" all potential outcomes – can compute what ATE would be under alternate randomizations

...which allows you to see many effect estimates all under this null.

Testing in Small Samples: Fisher's Exact Test

Let Ω be the set of all possible ways to assign treatments.

Fisher's exact test procedure:

1. Calculate a statistic $\hat{\theta}_{true}$ (e.g. difference in means) from original treatment assignment data
2. Obtain the null distribution of the statistic by calculating the same statistic $\hat{\theta}(\omega)$ under the sharp null for every possible (or many) ω in Ω
3. Compare $\hat{\theta}_{true}$ to the null distribution of $\hat{\theta}(\omega)$'s to see how "extreme" it is

Fisher's Exact Test: Setup

i	Y_{1i}	Y_{0i}	D_i
1	3	?	1
2	1	?	1
3	?	0	0
4	?	1	0
$\widehat{\tau}_{\text{ATE}}$			1.5

What do we know given the sharp null $H_0 : Y_{1i} = Y_{0i}$?

Fisher's Exact Test: Sharp Null Fills In POs

i	Y_{1i}	Y_{0i}	D_i
1	3	3	1
2	1	1	1
3	0	0	0
4	1	1	0
$\widehat{\tau}_{\text{ATE}}$			1.5
$\hat{\tau}(\omega)$			1.5

Given the full schedule of potential outcomes under the sharp null, we can compute the null distribution of the ATE across all possible randomizations.

Fisher's Exact Test: First Alternative Randomization

i	Y_{1i}	Y_{0i}	D_i	ω_1
1	3	3	1	1
2	1	1	1	0
3	0	0	0	1
4	1	1	0	0
$\widehat{\tau}_{\text{ATE}}$			1.5	
$\hat{\tau}(\omega)$			1.5	0.5

Fisher's Exact Test: Adding Randomizations

i	Y_{1i}	Y_{0i}	D_i	ω_1	ω_2
1	3	3	1	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	0	0	1
$\widehat{\tau}_{\text{ATE}}$			1.5		
$\hat{\tau}(\omega)$			1.5	0.5	1.5

Fisher's Exact Test: Adding Randomizations

i	Y_{1i}	Y_{0i}	D_i	ω_1	ω_2	ω_3
1	3	3	1	1	1	0
2	1	1	1	0	0	1
3	0	0	0	1	0	1
4	1	1	0	0	1	0
$\widehat{\tau}_{\text{ATE}}$			1.5			
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5

Fisher's Exact Test: Adding Randomizations

i	Y_{1i}	Y_{0i}	D_i	ω_1	ω_2	ω_3	ω_4
1	3	3	1	1	1	0	0
2	1	1	1	0	0	1	1
3	0	0	0	1	0	1	0
4	1	1	0	0	1	0	1
$\widehat{\tau}_{\text{ATE}}$			1.5				
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-0.5

Fisher's Exact Test: All Randomizations

i	Y_{1i}	Y_{0i}	D_i	ω_1	ω_2	ω_3	ω_4	ω_5
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{\text{ATE}}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-0.5	-1.5

So *under null*: $\Pr\left(|\hat{\tau}(\omega)| \geq |\widehat{\tau}_{\text{ATE}}|\right) = 4/6 \approx .67$.

Scaling Up

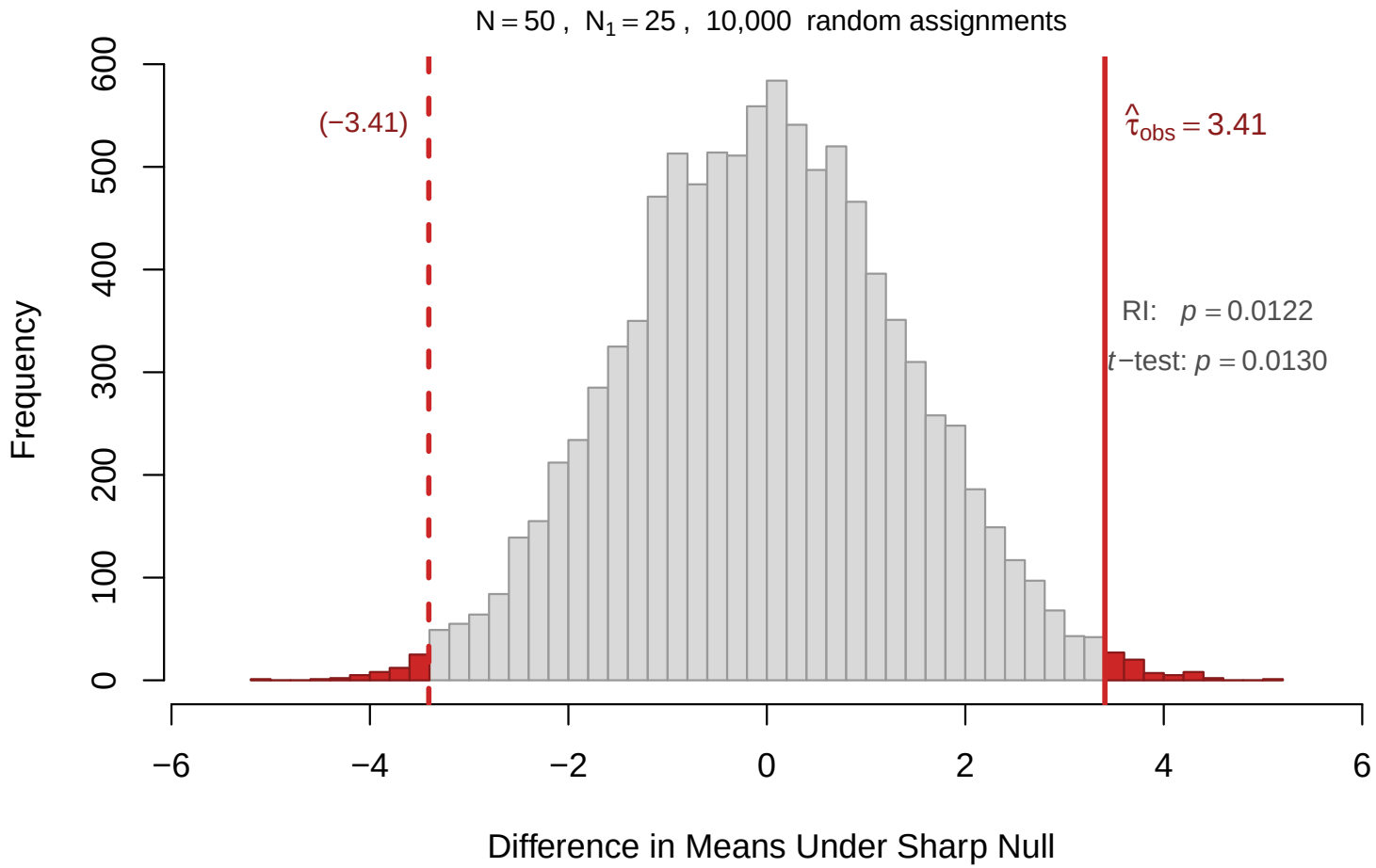
With 4 units and 2 treated, there are only $\binom{4}{2} = 6$ possible assignments — we can enumerate them all.

With 50 units and 25 treated, there are $\binom{50}{25} \approx 1.26 \times 10^{14}$ possible assignments.

We can't enumerate — but we can sample from Ω.

- Randomly re-assign treatment many times (e.g. 10,000)
- Compute $\hat{\tau}(\omega)$ each time under the sharp null
- This gives us the **null distribution** of the test statistic
- Compare our observed $\hat{\tau}$ to this distribution

Randomization Inference in Practice



Histogram: 10,000 sampled re-randomizations under sharp null.
Solid line: observed $\hat{\tau}$. Shaded regions: values at least as extreme $\Rightarrow$ p-value.

How do these inferences differ?

Finite-sample exactness.

- The Fisher/RI p-value is exact under the sharp null — no appeal to large-N approximations (CLT).
- When N is large enough for the CLT, the null distribution looks normal and you get nearly the same answer as a t-test (as we saw)

The sharp null is a strong hypothesis: $Y_{1i} = Y_{0i}$ for *every* unit.

- But any summary test statistic (DIM, rank statistic, etc.) is only sensitive to effects that accumulate — not to effects that cancel. In practice, rejecting the sharp null via the DIM tells "effects exist *and* don't average to zero", similar to weak null.

Compare to the bootstrap

- The bootstrap resamples *units* — simulating "what if we'd drawn different people from the population?" Reflects super-population uncertainty.
- RI is "design-based" – re-randomizes *treatment assignment* within the fixed sample — "what same people received different treatment assignment?"

Bottom line: RI valuable in small samples where normality is a worry.

Appendix

The exact SE for SATE

In reality we can only assign treatment within a finite sample so expect $cov(\overline{Y}_1, \overline{Y}_0) \neq 0$.

Considering this, under complete randomization the true standard error of the sample ATE is actually

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{N - N_1}{N - 1}\right)\frac{\sigma^2_{Y1}}{N_1} + \left(\frac{N - N_0}{N - 1}\right)\frac{\sigma^2_{Y0}}{N_0} + \left(\frac{1}{N - 1}\right)2Cov[Y_1, Y_0]}$$

with population variances and covariances σ^2_{Yd}, $Cov(Y_1, Y_0)$.

We don't use this—can you see why?

Proof that Neyman SE is conservative for SATE

Two preliminaries:

- $Cov(Y_1, Y_0) \leq SD(Y_1)SD(Y_0)$
- $Var(Y_1) + Var(Y_0) \geq 2\sqrt{Var(Y_1)Var(Y_0)}$

With that,

$$
\begin{aligned}
SE_{\widehat{ATE}} &= \sqrt{\left(\frac{N - N_1}{N - 1}\right)\frac{Var[Y_1]}{N_1} + \left(\frac{N - N_0}{N - 1}\right)\frac{Var[Y_0]}{N_0} + \left(\frac{1}{N - 1}\right)2Cov[Y_1, Y_0]} \\
&= \sqrt{\frac{1}{N - 1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + 2Cov[Y_1, Y_0]\right)} \\
&\leq \sqrt{\frac{1}{N - 1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + 2\sqrt{Var[Y_1]Var[Y_0]}\right)} \\
&\leq \sqrt{\frac{1}{N - 1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + Var[Y_1] + Var[Y_0]\right)}
\end{aligned}
$$

$SE_{\widehat{ATE}} \leq \widehat{SE}_{\widehat{ATE}}$, continued.

$$
\begin{aligned}
SE_{\widehat{ATE}} &\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + Var[Y_1] + Var[Y_0]\right)} \\
&\leq \sqrt{\frac{N_0^2 Var[Y_1] + N_1^2 Var[Y_0] + N_1 N_0 (Var[Y_1] + Var[Y_0])}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{(N_0^2 + N_1 N_0)Var[Y_1] + (N_1^2 + N_1 N_0)Var[Y_0]}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{(N_0 + N_1)N_0 Var[Y_1]}{(N-1)N_1 N_0} + \frac{(N_1 + N_0)N_1 Var[Y_0]}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{N\,Var[Y_1]}{(N-1)N_1} + \frac{N\,Var[Y_0]}{(N-1)N_0}} \\
&\leq \sqrt{\frac{N}{N-1}\left(\frac{Var[Y_1]}{N_1} + \frac{Var[Y_0]}{(N_0)}\right)} \\
&\leq \sqrt{\left(\frac{Var[Y_1]}{N_1} + \frac{Var[Y_0]}{(N_0)}\right)}
\end{aligned}
$$

$\therefore$ our estimator is conservative.