MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 2: Randomized Experiments

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2026

(Randomized) Experiments

Experiment:
A research design where the assignment mechanism is individualistic, probabilistic, uncounfounded, and controlled by the researcher.

Randomization:
Treatment values are assigned to N units at random, with known and positive assignment probabilities for each treatment to each unit (often called a 'randomized controlled trial' or RCT).

We consider the 'completely randomized experiment': a random subset of N_1 units assigned to treatment ($D = 1$) and remaining $N_0 = N - N_1$ to control.

- Note the slight difference to simple randomization (Bernoulli trials).
- Extension to cases with more than two levels is reasonably straightforward.
- Other randomized designs are introduced briefly at the end of this lecture.

"It's an illusion, Michael"

The Problem

Recall our basic problem:

$$E[Y|D=1] - E[Y|D=0] = E[Y_1|D=1] - E[Y_0|D=0]$$
$$= \underbrace{E[Y_1|D=1] - E[Y_0|D=1]}_{\text{ATT}} + \underbrace{\{E[Y_0|D=1] - E[Y_0|D=0]\}}_{\text{Selection bias}}$$

Randomized Experiment: Identification Assumption

Our goal is to find conditions under which we can identify our unobservable causal estimand with only observed data.

Randomization implies that assignment probabilities do not depend on potential outcomes (in expectation):

$$P(D|Y_0, Y_1) \quad = \quad P(D)$$

This is often called independence or unconfoundedness:

$$(Y_1, Y_0) \perp\!\!\!\perp D$$

(Note: $\perp\!\!\!\perp$ means "is independent of".)

To check understanding, does randomization imply $Y \perp\!\!\!\perp D$? No!

$(Y_1, Y_0) \perp\!\!\!\perp D$ means Y_0 is (in expectation) the same for those with $D = 1$ and $D = 0$ (similarly for Y_1). Says nothing about equivalence of Y between groups.

Randomized Experiment: Key Identification Result

Under independence from randomization:

$$E[Y_0|D=1] = E[Y_0|D=0] = E[Y_0]$$
$$\therefore E[Y_0|D=1] - E[Y_0|D=0] = 0$$

Read: Selection bias is, in expectation, equal to zero.

Returning to the problem at hand:

$$E[Y|D=1] - E[Y|D=0] = E[Y_1|D=1] - E[Y_0|D=0]$$
$$= \underbrace{E[Y_1|D=1] - E[Y_0|D=1]}_{\text{ATT}} + \underbrace{\{E[Y_0|D=1] - E[Y_0|D=0]\}}_{\text{Selection bias = 0}}$$
$$= \underbrace{E[Y_1|D=1] - E[Y_0|D=1]}_{\text{ATT}}$$

We can prove that our estimator equals our estimand $\rightarrow$ identification.

Randomized Experiment: Equivalence of Estimands

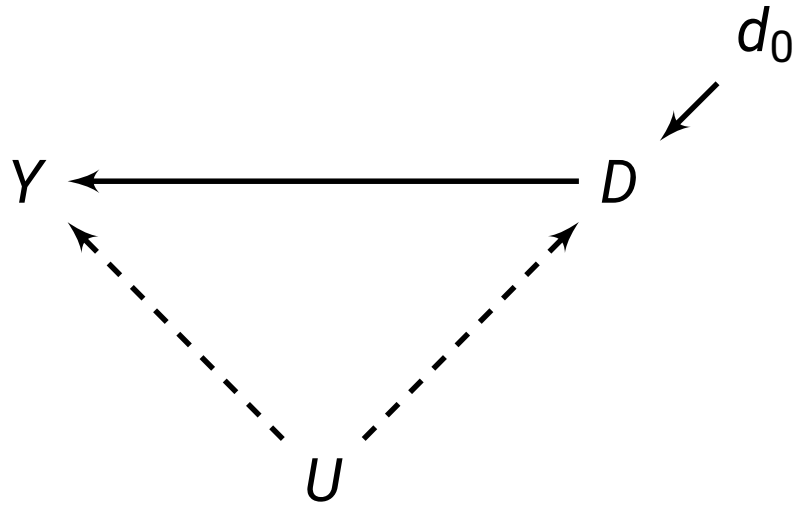
Independence tells us that $E[Y_1|D=1] = E[Y_1|D=0] = E[Y_1]$ (and for Y_0) , thus:

$$\begin{aligned}\tau_{ATT} &= E[Y_1|D=1] - E[Y_0|D=1] = E[Y_1|D=0] - E[Y_0|D=0] \\ &= \tau_{ATU} = E[Y_1] - E[Y_0] \\ &= \tau_{ATE}\end{aligned}$$

Read: Under independence, the ATE, ATT, and ATU are equal, and are thus simultaneously identified by the observed difference-in-means.

Note: We can also identify most other population-level causal effects, since they are comparisons of some features of the distributions of Y_0 and Y_1 and we can now estimate both of these distributions.

Graphical Representation



Consider a setting in which $D \leftarrow U \rightarrow Y$ is a back-door path connecting D and Y through unobserved U.

This is canonical confounding with the unobserved U confounding $D \rightarrow Y$

Randomization is equivalent to imposing $do(d_0)$ or $do(d_1)$, eliminating $U \rightarrow D$

There are now no back-door paths, so $D \rightarrow Y$ is identified.

Randomization and the Balancing Property

In expectation, complete randomization balances all observed and unobserved pre-treatment characteristics between treatment and control.

Why? For units with the same probability of treatment, X_i is independent of treatment assignment $\rightsquigarrow$ the balancing property.

(Note: We will dive deeper into this next week, when we introduce propensity scores.)

In a given experimental sample, we can empirically check for balance in observed pre-treatment covariate X using so called 'balance tests' (e.g., t-tests or equivalence tests) to see if the distributions $p(X|D = 1)$ and $p(X|D = 0)$ are not meaningfully different:

- In any one sample and treatment regime we might expect some chance imbalance.
- You could 'control' for imbalanced covariates, but don't have to (more later).
- Stratified randomization can guarantee exact balance in some observed X.
- Even more aggressive randomization procedures exist (e.g. pair-matching).

Complications and Limitations in Randomized Experiments

Randomization (and thus internal validity) can be complicated by:

- Missing data (e.g. dropout/attrition) – outcome is unobserved for some units in a way that is associated with D or potential outcomes.
- Measurement problems – Hawthorne effects etc.
- Non-compliance – some units receive a different treatment than the one they were assigned to.

Randomization does not help with external validity: Do causal effects in this sample apply to a broader population, or other populations?

- Can differentiate Sample ATE (SATE) from Population ATE (PATE) – randomization identifies SATE, but PATE also requires random sampling.
- Moving to a different population entirely would require other (often heroic) assumptions.

Randomized experiments can be weak in construct validity: How well do treatment and outcome in the experiment match the concept we are substantively interested in?

Estimation vs. Inference

Estimation:

- Choosing the right function to apply to our observed data.
- We can use the distributions $p(Y|D = 1)$ and $p(Y|D = 0)$ in the observed data to estimate the distributions of Y_1 and Y_0 in the population, and thus population ('group level')causal effects.
- Typically quite simple and familiar methods are sufficient for experiments.

Statistical inference:

- Characterizing uncertainty around our estimates and testing statistical hypotheses.
- Hypothesis tests and confidence intervals tend to be based on the source of identifying variation (i.e., what is 'random?')
- See the discussion in Chapters 5–8 of Imbens & Rubin for more on this, if you are interested.

Estimating ATE

$$\tau_{ATE} = E[Y_1] - E[Y_0]$$

An obvious estimator we have already seen is the sample difference-in-means:

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0$$

where

$$\bar{Y}_1 = \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i$$

$$\bar{Y}_0 = \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i$$

with $N_1 = \sum_i D_i$
and $N_0 = \sum_i (1 - D_i) = N - N_1$

Have already proven that $\widehat{\tau}$ is an unbiased estimator of τ_{ATE} under randomization!

Estimating ATE: Regression

The same τ_{ATE} can also be estimated using a linear regression model

$$Y_i = \hat{\gamma} + \hat{\tau} D_i + \hat{\varepsilon}_i$$

(Recall: $\hat{\tau}$ from a bivariate regression with a binary independent variable is equivalent to the diff-in-means.)

It is not necessary to include covariates X in this model. Why?

But pre-treatment covariates are sometimes included:

- Can increase precision (reduce standard error) by modeling residual variation in Y
- Control for observable imbalance (generated by random chance)
- Allow for estimation of heterogeneous treatment effects by X (by including interactions in the model)
- There is a risk of inducing small-sample bias (Freedman, 2008) – more in a few weeks when we introduce the 'fully-interacted estimator' (Lin, 2013)
- Note: you will almost never want to include post-treatment covariates. (See Montgomery et al., 2018; Cinelli et al., 2022)

Asymptotic Inference

When using either a simple difference-in-means or a linear regression, inference can be performed with a t-test:

1. Estimate the parameter of interest (τ_{ATE}) and variance
2. State hypotheses of interest, typically: H_0: $\tau_{ATE} = 0$ and H_1: $\tau_{ATE} \neq 0$
3. Calculate the relevant t-statistic:
   a. For a difference-in-means, a two-sample t-test:

$$t = \frac{\widehat{\tau}}{\sqrt{\frac{\widehat{\sigma}_1^2}{N_1} + \frac{\widehat{\sigma}_0^2}{N_0}}} \xrightarrow{d} N(0,1),$$

$$\text{where} \quad \widehat{\sigma}_d^2 = \sum_{D_i=d} (Y_i - \bar{Y}_d)^2 / N_d \quad \text{for} \quad d \in \{0,1\}.$$

   b. For regression, estimate robust standard errors and calculate t-statistic
4. We reject the null hypothesis H_0: $\tau_{ATE} = 0$ at the asymptotic $\alpha = 5\%$ significance level if $|t| > 1.96$. (The choice of α is arbitrary – more later.)

With more complex randomization schemes (e.g. cluster randomization), adjust standard error estimation ('analyze as you randomize').

Randomization Inference

For our *t*-tests, the null hypothesis was that the average treatment effect τ_{ATE} is zero, i.e.

$$H_0 : E[Y_1] = E[Y_0], \quad H_A : E[Y_1] \neq E[Y_0]$$

Consider now instead the sharp null hypothesis (and alternative)

$$H_0^s : Y_1 = Y_0, \quad H_A^s : Y_1 \neq Y_0$$

i.e. that all individual causal effects are zero.

Assuming H_0^s, then $Y_i = Y_{0i} = Y_{1i}$ for every unit. We can thus construct the full population distributions of Y_{0i} and Y_{1i}, under the null hypothesis!

Why? Under the sharp null the observed data Y_i for every unit would have been exactly the same, no matter the value of D_i

This is called randomization inference, permutation test, or Fisher's exact test

Randomization Inference

Procedure for randomization inference with complete randomization:

1. Permute the values of D_i (N_1 1s and N_0 0s) differently across the N units, keeping Y_i unchanged.
2. Calculate and store the value of $\widehat{\tau}_j$ (or any other appropriate statistic, such as the t-test statistic) for each of these permuted datasets j.
3. Calculate p-value as the proportion of $\widehat{\tau}_j$ that are as or more extreme than the actually observed $\widehat{\tau}$

With small N, we can consider *all* the permutations of D_i

- There are $\binom{N}{N_1} = N!/(N_1!N_0!)$ of them
- With larger N, use a random sample of all the permutations

Randomization Inference Example

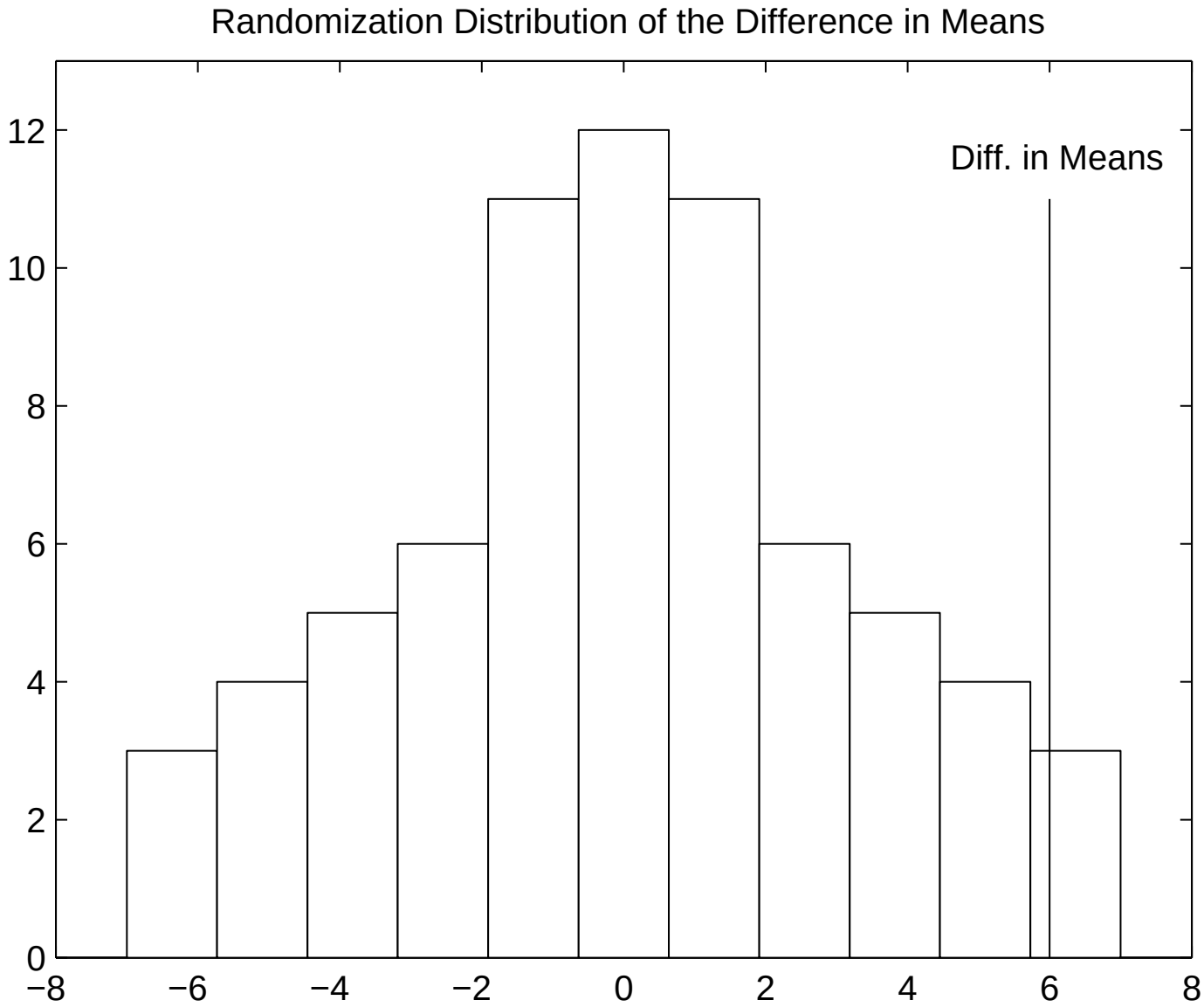
Consider an experiment with 8 units, 4 randomly assigned to treatment.

We can permute all $\binom{8}{4} = 70$ possible assignments.

We can then calculate the sample mean differences that would have been obtained for each of them if the sharp null hypothesis were true.

Y_i	12	4	6	10	6	0	1	1	
D_i	1	1	1	1	0	0	0	0	$\hat{\tau} = 6$
									$\hat{\tau}_j$
$j = 1$	1	1	1	1	0	0	0	0	6
$j = 2$	1	1	1	0	1	0	0	0	4
$j = 3$	1	1	1	0	0	1	0	0	1
$j = 4$	1	1	1	0	0	0	1	0	1.5
				...					
$j = 70$	0	0	0	0	1	1	1	1	-6

Randomization Inference Example



$$p = \Pr(|\hat{\tau}_j| \geq 6) = 0.0857$$

The Bootstrap

Another common method for uncertainty estimation is bootstrapping

The basic idea: Simulate the sampling distribution of a statistic via resampling with replacement

Useful when:

- Statistic is so complicated that analytically deriving its sampling variance is too difficult or cumbersome
- Data are so skewed that inference based on asymptotic normality is unlikely to perform well
- Statistic is of a form that makes CLT kick in only slowly, so normal approximation does not work well

Weakness: Computationally costly, sometimes prohibitively so.

Not a general solution for small samples (a common misunderstanding!)

Nonparametric Bootstrap and Parametric Bootstrap

Nonparametric bootstrap:

1. Draw B resamples of size n from X with replacement
2. For each X_b^*, compute $\hat{\theta}_b^*$, where $b = 1, ..., B$

3a. To estimate s.e. of $\hat{\theta}$, use the sample standard deviation of $\hat{\theta}^* = \{\hat{\theta}_1^*, ..., \hat{\theta}_B^*\}$ (bootstrap standard errors)

3b. To compute 95% CI, use 2.5/97.5 percentiles of $\hat{\theta}^* = \{\hat{\theta}_1^*, ..., \hat{\theta}_B^*\}$ as the lower/upper bounds (bootstrap percentile CI)

3c. If you know that $\hat{\theta} \overset{\text{approx.}}{\sim} \text{N}$, you can use 3a. and compute the bootstrap normal CI

Not only can you do this without any assumption about P, you can use this for any function of data $\hat{\theta} = f(X)$

Block bootstrap: When observations are clustered, resample clusters with replacement instead of individual units

Example: Labor market information

Can improving information about workseekers' skills improve outcomes?

Carranza et al (2022) study Johannesburg, South Africa, a context with high unemployment.

Conduct field experiment with unemployment organization.

D: Workseekers randomly assigned one of:

- shareable report of their skills assessment (treatment)
- no information about their performance (control)
- some other conditions (which we will ignore for now)

Y: employment outcomes, hours worked, earnings, wages, formality

X: economic/demographic characteristics measured before assignment

Example: Treatment

THE WORLD BANK

harambee
YOUTH EMPLOYMENT ACCELERATOR
WORK FOR WORK

REPORT ON CANDIDATE COMPETENCIES

name.. surname..
ID No. id..

This report provides information on assessments conducted by Harambee Youth Employment Accelerator (harambee.co.za), a South African organisation that connects employers looking for entry-level talent to young, high-potential work-seekers with a matric or equivalent. Harambee has conducted more than 1 million assessments and placed candidates with over 250 top companies in retail, hospitality, financial services and other sectors. Assessments are designed by psychologists and predict candidates' productivity and success in the workplace. This report was designed and funded in collaboration with the World Bank. You can find more information about this report, the assessments and contact details at www.assessmentreport.info. «name» was assessed at Harambee on 13 September, 2016.

«name» completed assessments on English Communication (listening, reading, comprehension), Numeracy, and Concept Formation:

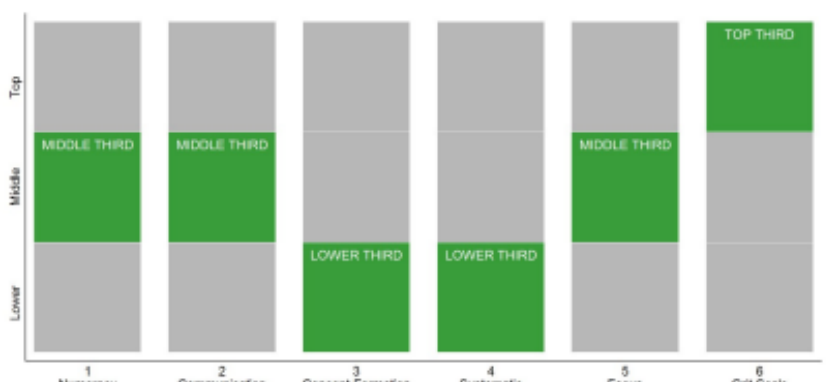
1. The Numeracy tests measure candidates' ability to apply numerical concepts at a National Qualifications Framework (NQF) level, such as working with fractions, ratios, money, percentages and units, and performing calculations with time and area. This score is an average of two numeracy tests the candidate completed.
2. The Communication test measures a candidate's grasp of the English language through listening, reading and comprehension. It assesses at an NQF level, for example measuring the ability to recognise and recall literal and non-literal text.
3. The Concept Formation Test is a non-verbal measure that evaluates candidates' ability to understand and solve problems. Those with high scores are generally able to solve complex problems, while lower scores indicate an ability to solve less complex problems.

«name» also completed tasks and questionnaires to assess their soft skills:

4. The Planning Ability Test measures how candidates plan their actions in multi-step problems. Candidates with high scores generally plan one or more steps ahead in solving complex problems.
5. The Focus Test assesses a candidate's ability to distinguish relevant from irrelevant information in potentially confusing environments. Candidates with high scores are generally able to focus on tasks in distracting surroundings, while candidates with lower scores are more easily distracted by irrelevant information.
6. The Grit Scale measures whether candidates show determination when working on challenging problems. Those with high scores generally spend more time working on challenging problems, while those with low scores choose to pursue different problems.

«name»'s results have been compared to a large benchmark group of young (age 18-34) South Africans assessed by Harambee. All candidates have a matric certificate and are from socially disadvantaged backgrounds. The benchmark group is 5,000 for cognitive skills and 400 for soft skills.

«name» scored in the «tercile_num» THIRD of candidates assessed by Harambee for Numeracy, «tercile_lit» THIRD for Communication, «tercile_cft» THIRD for Concept Formation, «tercile_tol» THIRD for Planning Ability, «tercile_troop» THIRD for Focus and «tercile_grit» THIRD for the Grit Scale.



Example: Balance

Table D.3: Balance Tests in Baseline and Endline Samples

	Baseline Sample Means				
Variable	Control	Private	Public	p:equal	Control
Age	23.5	23.8	23.7	0.583	23.5
Male	0.389	0.365	0.387	0.267	0.386
University degree / diploma	0.158	0.178	0.171	0.889	0.151
Any other post-secondary qualification	0.214	0.223	0.202	0.642	0.217
Completed secondary education only	0.617	0.593	0.612	0.794	0.620
Panel B: Assessment Results					
Numeracy score	-0.002	-0.018	0.024	0.523	-0.007
Communication score	0.038	-0.002	-0.029	0.206	0.031
Concept formation score	0.020	-0.012	-0.005	0.764	0.017
Grit score	-0.045	0.026	0.018	0.089	-0.042
Other scores	0.020	-0.010	-0.003	0.851	0.024
Panel C: Labor Market Measures					
Employed	0.364	0.386	0.387	0.468	0.362
Earnings	609	584	517	0.083	607
Ever worked	0.693	0.716	0.703	0.418	0.690
Ever held a long-term job	0.095	0.090	0.086	0.696	0.095
Panel D: Job Search Measures					
Searched	0.967	0.975	0.960	0.058	0.967
Applications submitted[a]	9.9	10.1	9.6	0.809	9.6
Search cost	205	240	280	0.276	205
Search hours	17.6	17.0	16.4	0.231	17.5
Offers received[a]	1.00	1.41	1.12	0.280	1.02
Panel E: Belief Measures					
Planned applications[a]	19.7	22.4	107.0	0.252	19.6

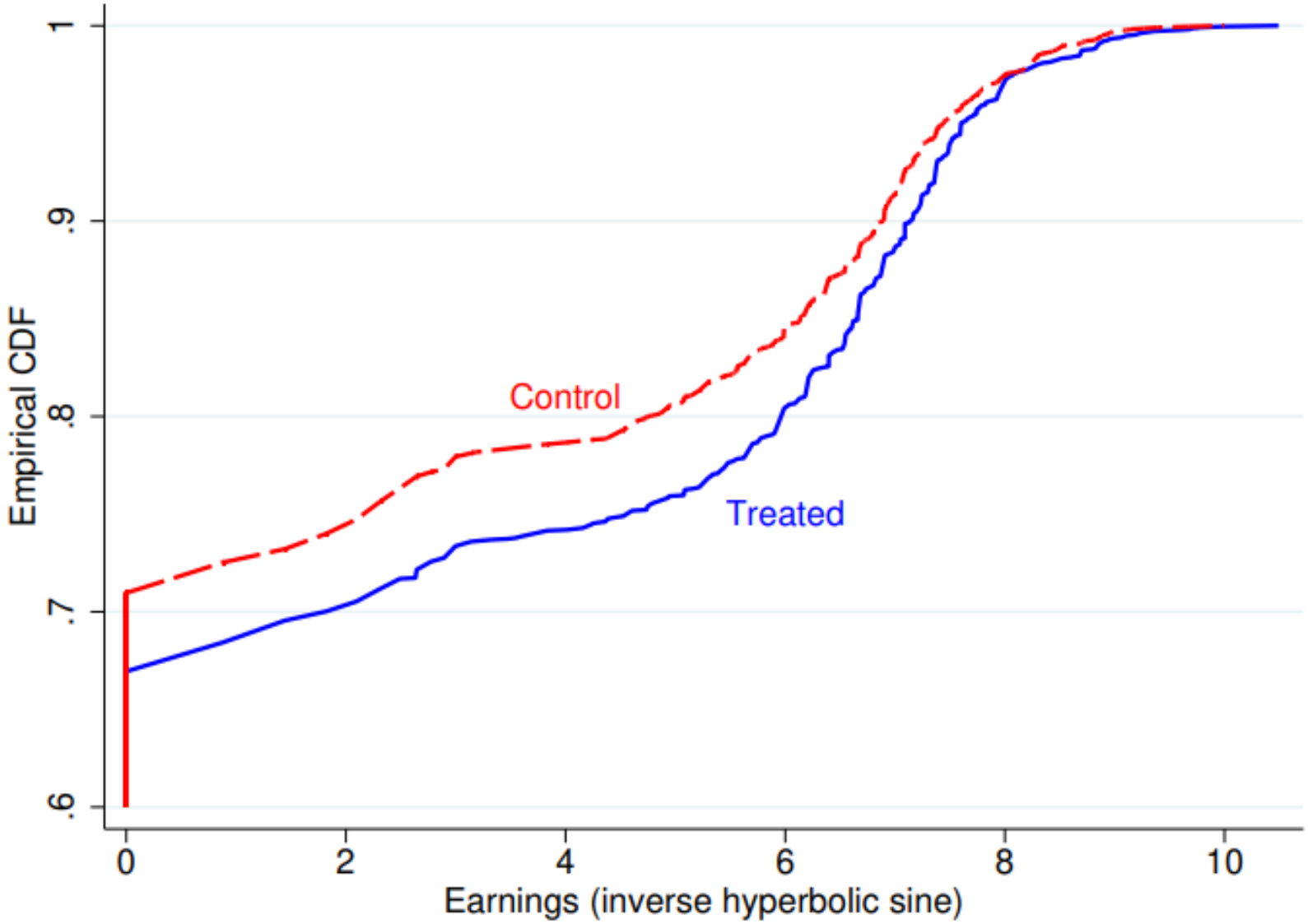
Example: Treatment Effect Estimates

Table 1: Treatment Effects on Labor Market Outcomes

	(1) Employed	(2) Hours[c]	(3) Earnings[c]	(4) Hourly wage[c]	(5) Written contract
Treatment	0.052	0.201	0.337	0.197	0.020
	(0.012)	(0.052)	(0.074)	(0.039)	(0.010)
Mean outcome	0.309	8.848	159.291	9.840	0.120
Mean outcome for employed		28.847	518.291	32.283	0.392
# observations	6607	6598	6589	6574	6575
# clusters	84	84	84	84	84

Coefficients are from regressing each outcome on a vector of treatment assignments, randomization block fixed effects, and prespecified baseline covariates (measured skills, self-reported skills, education, age, gender, employment, discount rate, risk aversion). Heteroskedasticity-robust standard errors shown in parentheses, clustering by treatment date. Mean outcomes are for the control group. All outcomes use a 7-day recall period. Outcomes marked with [c] use the inverse hyperbolic sine transformation for the treatment effects but the control group means are reported in levels. All monetary figures are reported in South Africa Rands. 1 Rand $\approx$ USD 0.167 in purchasing power parity terms. The sample sizes differ across columns due to item non-response, mostly from respondents reporting that they don't know the answer.

Example: Quantile Analysis



Common Experiment Types

Modern social/data science experiments tend to come in four broad flavors:

1. Lab experiments
2. Policy experiments
3. Field experiments
4. Survey experiments

Each type has different design considerations and traditions worth paying close attention to.

Think carefully about what you are randomizing and why... The act of randomization itself is not a sufficient condition for a good study!

For example, are you randomizing to estimate an effect or to measure?

Other Randomization Schemes

The completely randomized design is only one option:

1. Stratified (conditional, blocked) randomized experiment are randomized separately within levels of some covariate(s) X
   - An extreme version is a pairwise randomized experiment: Each stratum (block) contains 2 units, one assigned to treatment, the other to control.
   - Note: Stratification will be an important concept when we move on to observational assignment mechanisms.

2. Cluster randomized experiments randomize units in groups. Every unit within a group (called a cluster) gets the same treatment level.
   - E.g. randomizing whole villages of people or whole classrooms of pupils.

3. Cross-over experiments have units switch treatment status over time.
   - E.g. varying treatments for sick patients over time.

4. Adaptive experiments where randomization updates as results emerge.
   - E.g. 'multi-arm bandits' that allocate subjects to most effective treatment.

Statistical Power: Foundations

Telescope analogy: Is our telescope (experiment) powerful enough to see astronomical bodies (effects) that really exist?

We want to test hypotheses about the true average treatment effect:

$$H_0 : \tau_{ATE} = 0 \quad , \quad H_A : \tau_{ATE} \neq 0$$

Power is the probability of rejecting H_0 when H_A is true:

$$\text{Power} = 1 - \psi = \Pr\left(|t| > z_{1-\alpha/2} \mid \tau_{ATE} \neq 0\right)$$

The power of any given study depends on:

- true effect size τ_{ATE}
- true outcome variability (σ_1^2, σ_0^2)
- sample sizes N_1, N_0
- false-positive tolerance α
- false-negative tolerance ψ

Statistical Power: Implications

Assuming fixed tolerances (α and ψ):

	Sample size (N)	Min. Det. Effect (MDE)	Outcome variability
Smaller	$\downarrow$ power	$\uparrow$ power	$\downarrow$ power
Larger	$\uparrow$ power	$\downarrow$ power	$\uparrow$ power

Some implications:

1. Smaller true effects or noisier outcomes require larger N
2. Larger true effects or less variable data allow smaller N
3. Experimental designs that reduce variability in Y can be more useful than larger N
3. Large estimated effects in small samples may in fact indicate a power problem with a given study... yikes!

Power Calculations in Practice

How can we use these correspondences to design better experiments?

Two typical prospective approaches:

1. When N is fixed, specify explicitly the minimum detectable effect (MDE or MDES) that can reliably be detected for a given α and ψ... ask yourself if this is a "meaningful" magnitude.
2. When we can control N, use hypothesised MDE/MDES to pick N, given tolerances... ask yourself if N is "in budget."

Alternatively, one can perform retrospective power analyses:

- Using the estimated effect and the observed variability in Y, calculate $1 - \psi$... but this exercise is circular!
- Instead, meta-analytical approaches exist to do this across many studies, given that observed effects may not be representative of (unobserved) population-level effects. For more, see Ioannidis et al (2017), Gelman (2019), and Arel-Bundock et al (2025).