

MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 2: Randomized Experiments

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2024

Lecture Roadmap

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments

(Randomized) Experiments

Definition (Randomized Experiment)

An **experiment** is a **research design** where the **assignment mechanism** is individualistic, probabilistic, uncounfounded, and **controlled** by the researcher.

In a (classical) **randomized experiment** ('randomized controlled trial' or RCT) treatment values are assigned to N units **at random**, with known and positive assignment probabilities for each treatment to each unit.

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 treatment levels is reasonably straightforward.
- Other randomized designs are introduced briefly at the end of this lecture.

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments



G.O.B.
GEORGE OSCAR BLUTH II
Magician (part-time)

"It's an illusion, Michael"

The Problem

Recall our basic problem:

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

Randomization

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

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

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

Or, said another way:

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

(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 that (in expectation) Y_0 is the same for those with $D = 1$ and for those with $D = 0$ (and similarly for Y_1), says nothing about equivalence of Y between these groups.

Randomization Eliminates Selection Bias

Back to the problem at hand:

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

Under independence from randomized treatment assignment, we have

$$E[Y_0|D = 1] = E[Y_0|D = 0] = E[Y_0]$$

thus selection bias equals zero (in expectation).

We also have $E[Y_1|D = 1] = E[Y_1|D = 0] = E[Y_1]$, thus

$$\tau_{ATT} = E[Y_1|D = 1] - E[Y_0|D = 1] = E[Y_1] - E[Y_0] = \tau_{ATE}$$

Randomization: Key Identification Result

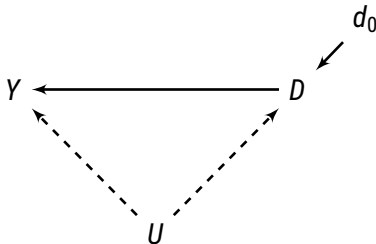
Under complete randomization we can assume:

$$E[Y|D = 1] - E[Y|D = 0] = E[Y_1] - E[Y_0] = \tau_{ATE}$$

The observed means difference between the treatment group **identifies** the causal average treatment effect ATE (as well as ATT and ATU, which both equal to ATE in this case).

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.
- Non-compliance – some units receive a **different treatment** than the one they were assigned to.

Randomization does not help with **external validity**: How well do causal effects for this sample apply to 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?

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments

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 causal effects.
- Typically quite simple and familiar methods are sufficient for experiments.

Statistical inference:

- Characterizing uncertainty around our estimates.
- 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 of this is the sample difference-in-means:

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

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 $\hat{\tau}$ 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: **do not** include post-treatment covariates. (Montgomery et al., 2018; Cinelli et al., 2022)

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference**
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments

Two Sample T-Test for Inference

From statistical theory, we know that under $H_0: \tau_{ATE} = 0$,

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

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

We reject the null hypothesis $H_0: \tau_{ATE} = 0$ against the alternative $H_1: \tau_{ATE} \neq 0$ at the asymptotic 5% significance level if $|t| > 1.96$.

Randomization Inference

For the two-sample t-test, 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

Randomization Inference

Procedure for randomization inference (permutation test, Fisher's exact test):

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 $\hat{\tau}_j$ (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 $\hat{\tau}_j$ that are as or more extreme than the actually observed $\hat{\tau}$

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

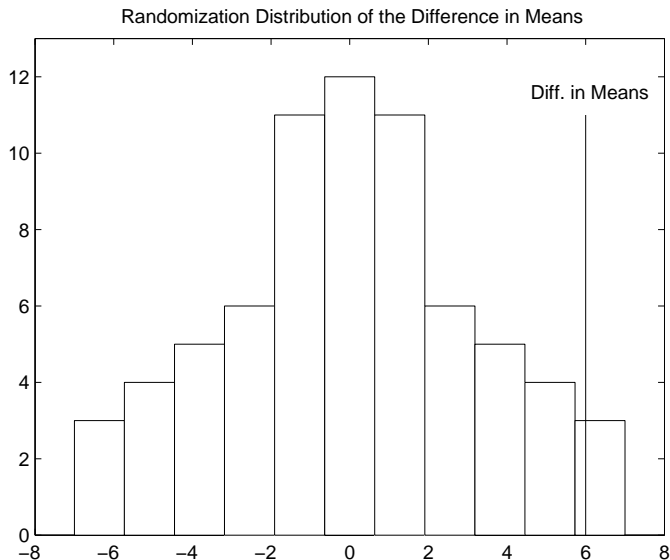
Consider an experiment with 8 units, 4 randomly assigned to treatment.

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

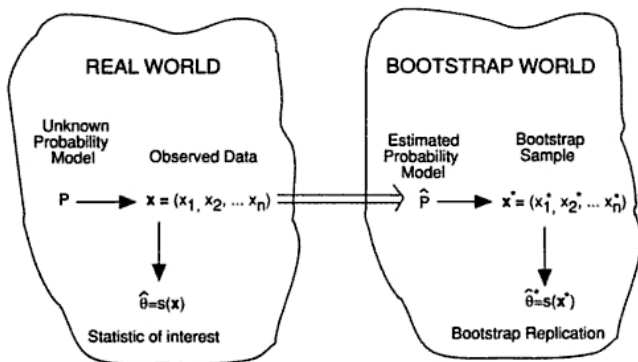
Useful when:

- 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!)

Bootstrap World



Goal: Estimate uncertainty in $\hat{\theta}$ (any statistic or parameter of interest) without making any assumption about P

Idea: If n is sufficiently large, the sample \mathbf{X} should be a good approximation of P

→ Think of \mathbf{X} as an estimated population probability model P , and just like \mathbf{X} is a realization from P , let's draw a **resample** \mathbf{X}^* from \mathbf{X}

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 $b = 1, \dots, B$
- 3a. To estimate s.e. of $\hat{\theta}$, use the sample standard deviation of $\hat{\theta}^* = \{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ (**bootstrap standard errors**)
- 3b. To compute 95% CI, use 2.5/97.5 percentiles of $\hat{\theta}^* = \{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ as the lower/upper bounds (**bootstrap percentile CI**)
- 3c. If you know that $\hat{\theta} \overset{\text{approx.}}{\sim} N$, 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

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments



Google images result for “stock photo of people upskilling in a business setting”

Example: Job Training Partnership Act (JTPA)

Largest randomized training evaluation ever undertaken in the U.S.; started in 1983 at 649 sites throughout the country

Sample: “Underskilled” and “economically disadvantaged” persons in the labor market (previously unemployed or low earnings)

D: (Invitation) to one of three general service strategies:

- classroom training in occupational skills
- on-the-job training and/or job search assistance
- other services (eg. probationary employment)

Y: earnings 30 month following assignment

X: Characteristics measured before assignment (age, gender, previous earnings, race, etc.)

Means and Standard Deviations for JTPA Experiment

	Entire Sample	Assignment		Difference (t-stat.)
		Treatment	Control	
A. Men				
Number of observations	5,102	3,399	1,703	
<i>Treatment</i>				
Training	.42 [.49]	.62 [.48]	.01 [.11]	.61 (70.34)
<i>Outcome variable</i>				
30 month earnings	19,147 [19,540]	19,520 [19,912]	18,404 [18,760]	1,116 (1.96)
<i>Baseline Characteristics</i>				
Age	32.91 [9.46]	32.85 [9.46]	33.04 [9.45]	-.19 (-.67)
High school or GED	.69 [.45]	.69 [.45]	.69 [.45]	-.00 (-.12)
Married	.35 [.47]	.36 [.47]	.34 [.46]	.02 (1.64)
Black	.25 [.44]	.25 [.44]	.25 [.44]	.00 (.04)
Hispanic	.10 [.30]	.10 [.30]	.09 [.29]	.01 (.70)
Worked less than 13 weeks in past year	.40 [.47]	.40 [.47]	.40 [.47]	.00 (.56)

JTPA Experiment: Estimated effects separately by group

Exhibit 5 Impacts on Total 30-Month Earnings: Assignees and Enrollees, by Target Group

	<i>Mean earnings</i>		<i>Impact per assignee</i>		
	<i>Treatment group (1)</i>	<i>Control group (2)</i>	<i>In dollars (3)</i>	<i>As a percent of (2)</i>	<i>Impact per enrollee in dollars</i>
Adult women	\$ 13,417	\$ 12,241	\$ 1,176***	9.6%	\$ 1,837***
Adult men	19,474	18,496	978*	5.3	1,599*
Female youths	10,241	10,106	135	1.3	210
Male youth non-arrestees	15,786	16,375	-589	-3.6	-868
Male youth arrestees					
Using survey data	14,633	18,842	-4,209**	-22.3	-6,804**
Using scaled UI data	14,148	14,152	-4	0.0	-6

- 1 The experimental ideal
- 2 The 'magic' of randomization
- 3 Estimation
- 4 Inference
- 5 Example experiment: JTPA
- 6 Beyond simple randomized experiments

Some Other Randomization Schemes

The completely randomized design is only one option:

- **Stratified** (conditional, blocked) randomized experiments are randomized separately within levels of some covariate(s) X
 - e.g. separately for men and women
 - An extreme version is a *pairwise randomized experiment*: Each stratum (block) contains 2 units, one assigned to treatment, the other to control.
 - Stratification will be very important from next week, when we move on to observational assignment mechanisms.
- **Cluster randomized** experiments randomize units in **clusters**. Every unit within a cluster gets the same treatment level.
 - e.g. randomizing whole villages of people or whole classrooms of pupils.
- **Cross-over** experiments have units switch treatment status over time.
 - e.g. varying treatments for sick patients over time