Experiments

POLI SCI 210

Introduction to Empirical Methods in Political Science

AI Prompts

- (Fundamental problem of) causal inference
- Potential outcomes framework
- Average treatment effect
- “Help me design a [survey/field/laboratory] experiment about X”
- Internal vs. external validity

Last week

- Surveys and how to make them good
- Importance of random sampling to justify statistical inference via asymptotic properties (CLT)
- **So far:** Mostly about *descriptive statistics*

This week

- Moving from *statistical inference* to *causal inference*
- Experiments as the *gold standard*
- **Tuesday:** Logic of experimentation
- **Thursday:** Learning from experiments

Return to counterfactuals

Study by Tulane researcher suggests marijuana can cause infertility in men

| Lance Sumler lsumler@tulane.edu View PDF

medicine.tulane.edu/news/study-tulane-researcher-suggests-marijuana-can-cause-infertility-men

Making causal statements

- Why were we using cautious language?
- **Ideal study:** Compare the same man with and without smoking
- **Actual study:** Compare two *different* groups of men

What would allow us to make more confident **causal** claims?

Causal inference

Imagine we want to establish whether a medical **treatment** improves people's lives

We want to make sure that the treatment *actually* works

In other words, we want to *attribute* the treatment as the **cause** of health improvement

Can we just show that people who receive the treatment get better?

No! we need some kind of **control**

Potential outcomes framework

Ingredients

- condition (0: control, 1: treatment) $D = \{0, 1\}$
- is the individual **potential outcome** $Y(D)$

Switching equation

$$Y = \begin{cases} Y(0) & \text{if} \quad D = 0 \\ Y(1) & \text{if} \quad D = 1 \end{cases}$$

$$Y = Y(1) \cdot D + Y(0) \cdot (1 - D)$$

Toy example

ID		$Y(1)Y(0)$
1	0	0
2	1	0
3	1	0
4	1	1

is the **individual treatment effect**

$\tau_i = Y(1) - Y(0)$

Toy example

ID		$Y(1)$	$Y(0)$	τ_i
1	0	0	0	
2	1	0	1	
3	1	0	1	
4	1	1	0	

is the **individual treatment effect**

$$\tau_i = Y(1) - Y(0)$$

is the **average treatment effect** (ATE)

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \tau_i$$

Challenge

ID	$Y_i(1)$	$Y_i(0)$	x_i
1	0	0	0
2	1	0	1
3	1	0	1
4	1	1	0

Challenge

	Unobserved		
ID			
	D	$Y(1)$	$Y(0)$	τ_i
1	0	0	0
2	1	0	1
3	1	0	1
4	1	1	0

Assign condition D (0: control, 1: treatment)

Challenge

	Unobserved			Observed	
ID					
1	0$Y(1)$	0$Y(0)$	0τ_i	1D	0Y
2	1	0	1	0	0
3	1	0	1	1	1
4	1	1	0	0	1

To know the ATE we need

$$(Y(1) - Y(0))$$

But we only observe one at a time for each unit

Challenge

	Unobserved			Observed	
ID					
1	${}_0Y(1)$	${}_0Y(0)$	${}_0\tau_i$	${}_1D$	${}_0Y$
2	1	0	1	0	0
3	1	0	1	1	1
4	1	1	0	0	1

This is the **FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE**

The term comes from: Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945-960

What can we do?

We can rewrite the ATE as

$$\tau = E[\tau_i]$$

And then expand since

$$\tau_i = Y(1) - Y(0)$$

$$\tau = E[Y(1) - Y(0)]$$

What can we do?

We can rewrite the ATE as

$$\tau = E[\tau_i]$$

And then expand since

$$\tau_i = Y(1) - Y(0)$$

Which is equivalent to

$$\underbrace{\tau = E[Y(1) - Y(0)]}_{\text{Average individual treatment effect}}$$

$$\tau = E[Y(1)] - E[Y(0)]$$

What can we do?

We can rewrite the ATE as

$$\tau = E[\tau_i]$$

And then expand since

$$\tau_i = Y(1) - Y(0)$$

Which is equivalent to

$$\underbrace{\tau = E[Y(1) - Y(0)]}_{\text{Average individual treatment effect}}$$

(because the sum of the averages = average of the sums)

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

What can we do?

We want to know

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

But we can only calculate

What can we do?

We want to know

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

But we can only calculate

$$\hat{\tau} = E[Y(1)|D = 1] - E[Y(0)|D$$

Aside on notation

Greek

- Letters like μ denote **estimands** (unobserved quantities of interest)
- A hat $\hat{\mu}$ denotes **estimators** (rules to calculate a quantity)

Latin

- Letters like Y denote **actual variables** in our data
- A bar $\overline{Y}$ denotes an **estimate** calculated from our data

$$X \rightarrow \overline{X} \rightarrow \hat{\mu} \xrightarrow{\text{hopefully!}} \mu$$

Aside on notation

Greek

- Letters like μ denote **estimands** (unobserved quantities of interest)
- A hat $\hat{\mu}$ denotes **estimators** (rules to calculate a quantity)

Latin

- Letters like Y denote **actual variables** in our data
- A bar $\bar{Y}$ denotes an **estimate** calculated from our data

$$\text{Data} \rightarrow \text{Estimate} \rightarrow \text{Estimator} -$$

Aside on notation

Greek conceptual

- Letters like $\vec{\mu}$ denote **estimands** (unobserved quantities of interest)
- A hat $\hat{\mu}$ denotes **estimators** (rules to calculate a quantity)

Latin practical

- Letters like $\vec{X}$ denote **actual variables** in our data
- A bar $\bar{X}$ denotes an **estimate** calculated from our data

$$\text{Data} \rightarrow \text{Estimate} \rightarrow \text{Estimator} -$$

What can we do?

We want to know

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

But we can only calculate

$$\hat{\tau} = E[Y(1)|D = 1] - E[Y(0)|D$$

What can we do?

We want to know

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

But we can only calculate

$$\underbrace{\hat{\tau} = E[Y(1)|D = 1] - E[Y}_{\text{Difference in (observed) means between treat}}$$

What would make these two equivalent?

What can we do?

We want to know

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

But we can only calculate

$$\underbrace{\hat{\tau} = E[Y(1)|D = 1] - E[Y}_{\text{Difference in (observed) means between treat}}$$

What would make these two equivalent?

What can we do?

We want to know

$$\underbrace{\tau = E[Y(1)] - E[Y(0)]}_{\text{Difference in average potential outcomes}}$$

But we can only calculate

$$\underbrace{\hat{\tau} = E[Y(1)|D = 1] - E[Y}_{\text{Difference in (observed) means between treat}}$$

We want **treatment assignment** to be ~~**ignorable**~~

Random assignment

- If we can claim that units are assigned to conditions **independently** from potential outcomes D
- Then we can claim that $\hat{\tau}$ is a *valid* approximation of τ
- So that the *difference in means* is an **unbiased estimator** of the *ATE*
- Random assignment guarantees this **in expectation**

Implications

If random assignment works:

- No reverse causation
- No selection bias
- No omitted variable bias

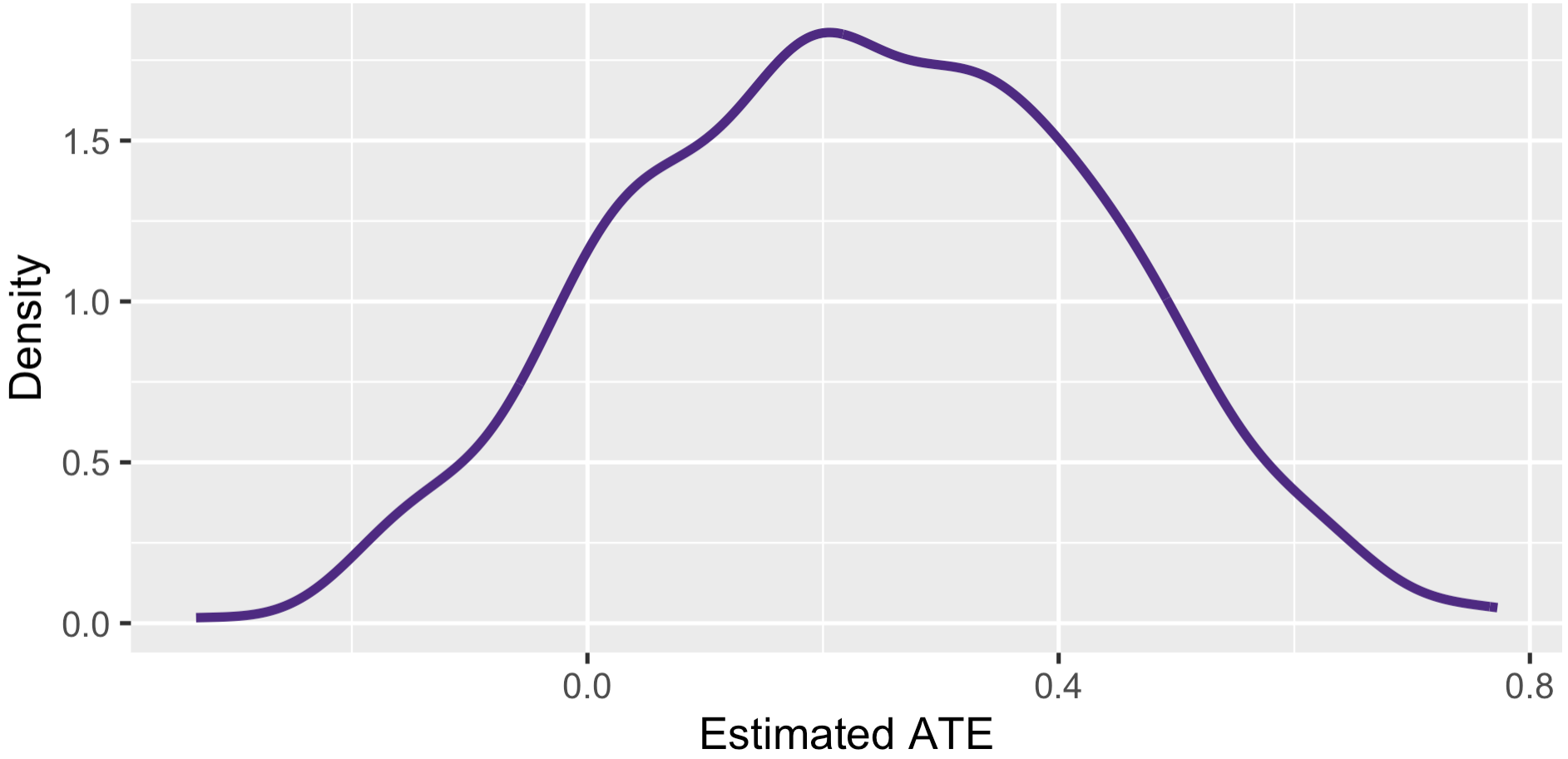
But you can always get unlucky!

Example

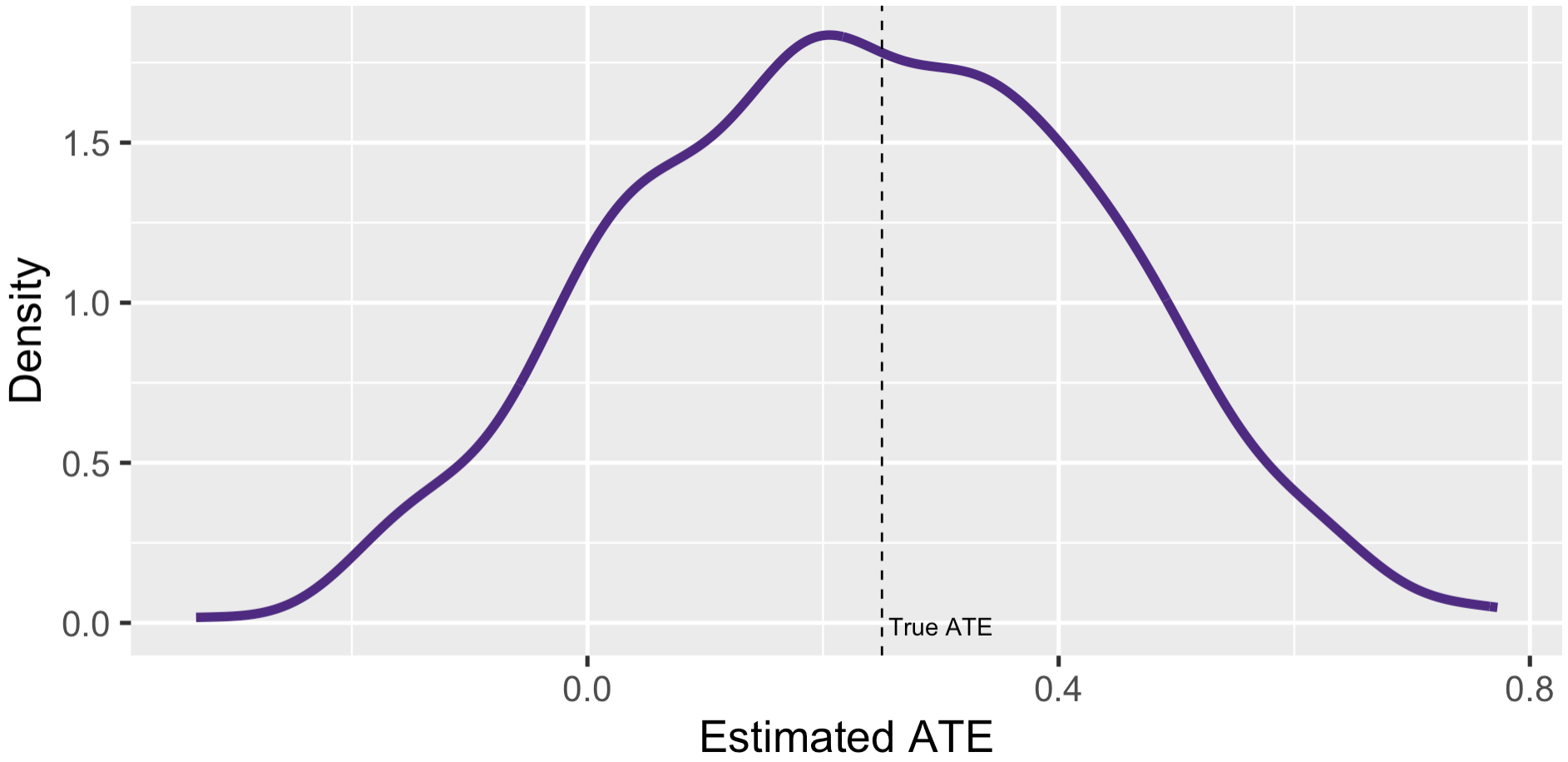
	ID	Y0	Y1	D	Y
1	001	0.484520880	0.73452088	1	0.73452088
2	002	0.770743000	1.02074300	0	0.77074300
3	003	0.186039041	0.43603904	1	0.43603904
4	004	-0.086703465	0.16329654	0	-0.08670346
5	005	0.681404050	0.93140405	0	0.68140405
6	006	1.970049082	2.22004908	0	1.97004908
7	007	0.034679490	0.28467949	0	0.03467949
8	008	-0.150141526	0.09985847	0	-0.15014153
9	009	1.687850716	1.93785072	1	1.93785072
10	010	0.784055885	1.03405589	1	1.03405589
11	011	-0.637762442	-0.38776244	0	-0.63776244
12	012	-0.277068057	-0.02706806	1	-0.02706806
13	013	0.823002236	1.07300224	1	1.07300224
14	014	-2.189373502	-1.93937350	1	-1.93937350
15	015	-0.222094995	0.02790501	1	0.02790501

- **True ATE:** 0.25
- **Estimate:** 0.246 (p = 0.191)

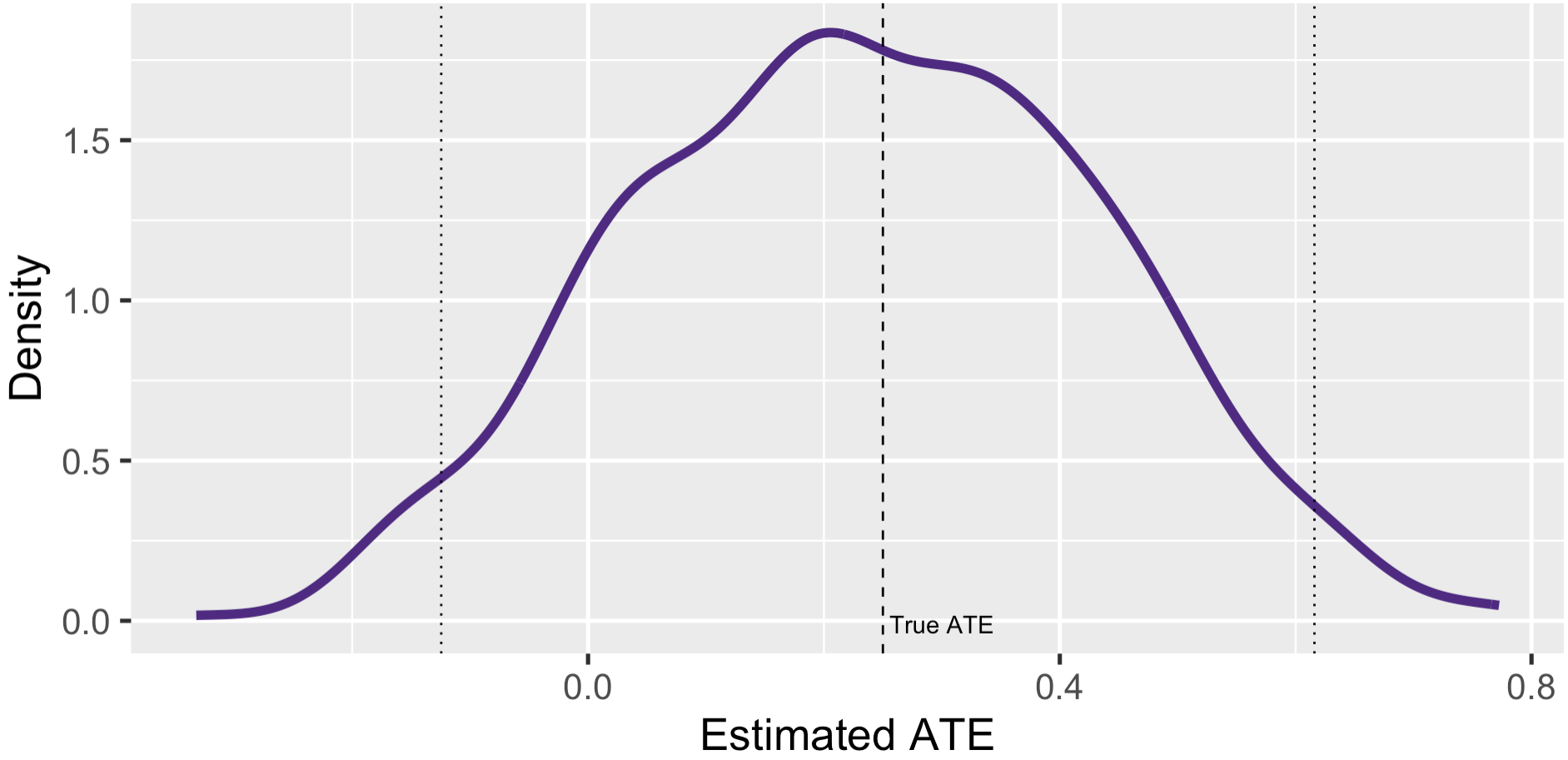
Redo experiment 1,000 times



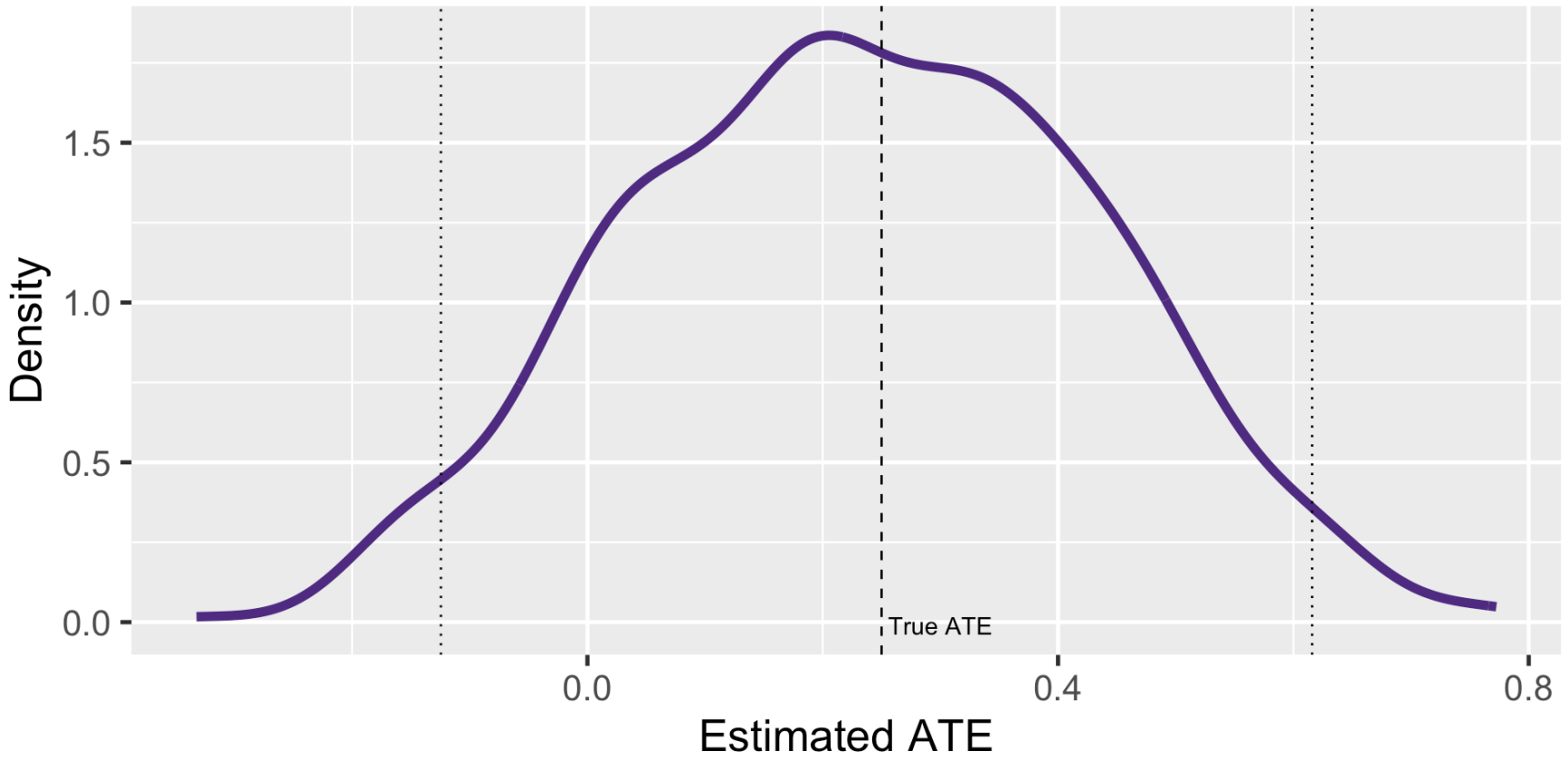
Redo experiment 1,000 times



Redo experiment 1,000 times



Redo experiment 1,000 times



95% confidence interval: [-0.12 , 0.62]

Types of experiment

- Survey experiments
- Field experiments
- Laboratory experiments

Depends on *how treatments are delivered*

Examples

Tomz and Weeks (2013): “Public Opinion and the Democratic Peace”

- Surveys in the UK () and US ()
- April-May 2010 $n = 762$ $n = 1273$
- **Outcome:** Support for military strike
- 2x2x2 survey experiment

Tomz, Michael R. and Jessica L.P. Weeks. 2013. “Public Opinion and the Democratic Peace.” *American Political Science Review* 107 (4): 849-865

Vignette design

UK

- **Political regime:** Democracy/not a democracy
- **Military alliances:** Ally/not an ally
- **Military power:** As strong/half as strong

US

- **Political regime:** Democracy/not a democracy
- **Military alliances:** Ally/not an ally
- **Trade:** High level/not high level

Results for democracy

TABLE 1. The Effect of Democracy on Willingness to Strike

	United Kingdom (between)	United States (between)	United States (within)
Not a democracy	34.2	53.3	50.0
Democracy	20.9	41.9	38.5
Effect of democracy	−13.3	−11.4	−11.5
95% C.I.	(−19.6 to −6.9)	(−17.0 to −5.9)	(−14.7 to −8.3)

Kalla et al (2018): Are You My Mentor?

- Correspondence experiment with legislators in the US
- Also known as *audit* experiments $N = 8189$
- Send email about fake student seeking advice to become politician
- Cue gender with student's name

Kalla, Joshua, Frances Rosenbluth, and Dawn Langan Teele. 2018. "Are You My Mentor? A Field Experiment on Gender, Ethnicity, and Political Self-Starters." *Journal of Politics* 80 (1): 337-341

Sample email

From: [Treatment: Student Sex]
To: [Legislator's email]
Subject: Help on a class project?

Dear [LEGISLATOR],

My name is [MALE/FEMALE] and I am a college sophomore. I'm interviewing politicians for a class project to learn about how they entered their field and what advice they might have for students interested in politics. As someone who really cares about my community, one day I hope to be a politician. What advice would you give to me?

Sincerely, [MALE/FEMALE]

Figure 1. Treatment wording

Findings

Outcome	Male Sender	Female Sender	p-value
Received reply	0.25	0.27	0.15
Meaningful response	0.11	0.13	0.47
Praised	0.05	0.06	0.17
Offer to help	0.03	0.05	0.09
Warned against running	0.01	0.02	0.14
Substantive advice	0.07	0.08	0.33
Word count (logged)	1.00	1.10	0.06
Character count	145.00	170.00	0.04

Why not much difference by gender?

Adapted from Table 1

Schnall et al (2008): “Disgust as Embodied Moral Judgment”

- Students at University of Virginia (n = 43, 18 male)
- Offered to participate in study for course credit
- **Outcome:** Moral judgment questions (several scenarios)
- **Treatment:** Extraneous disgust (dirty room)
- **Control:** No disgust (clean room)

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2008. “Disgust as Embodied Moral Judgment.” *Personality and Social Psychology Bulletin* 34 (8): 1023-1153

Conditions

> On the desk there was a transparent plastic cup with the dried up remnants of a smoothie and a pen that was chewed up. Next to the desk was a trash can overflowing with garbage including greasy pizza boxes and dirty-looking tissues. For the no-disgust condition, the same desk was used, but it was covered up with a clean white tablecloth. A new chair was provided, and none of the disgusting objects were present. A new and unchewed pen was provided for filling out the questionnaires.

Findings

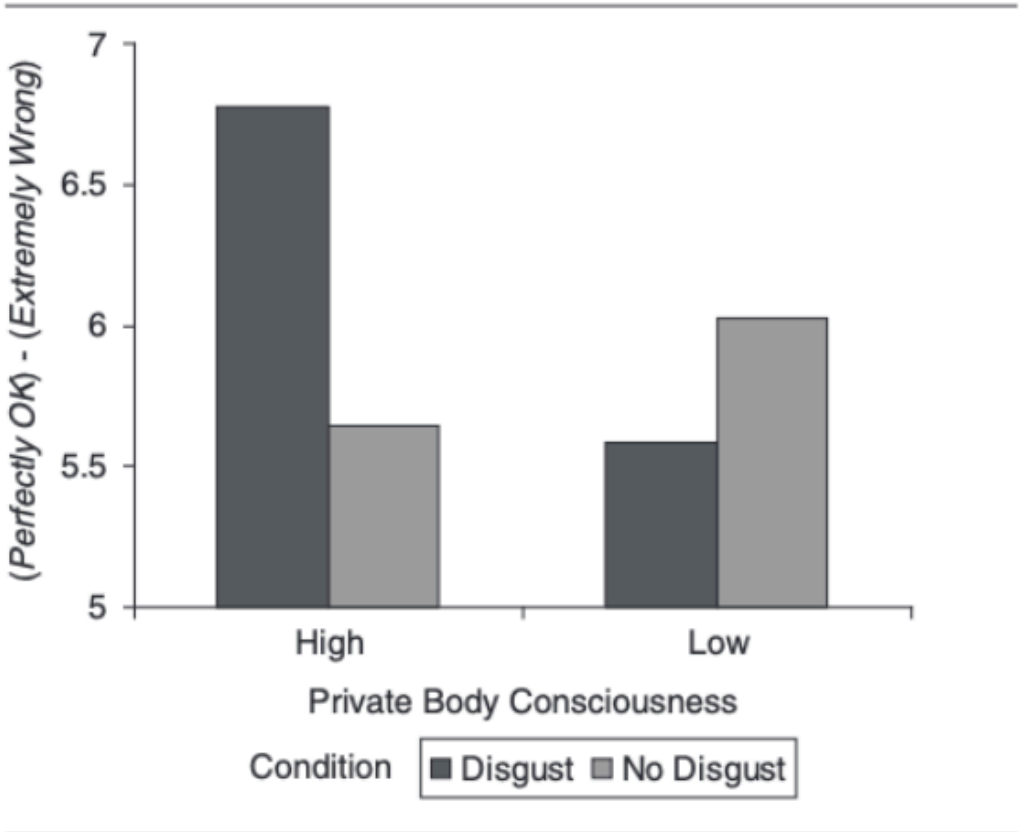


Figure 1 Judgments of wrongness of moral actions as a function of Private Body Consciousness and condition (Experiment 2).

Private body consciousness is a person's general attention to internal physical states

Summary

Last week:

- Random **sampling** enables inference from sample to population because of *asymptotic properties* (central limit theorem)

This week:

- Random **assignment** enables causal inference in experiments because of *finite-sample properties* (weak law of large numbers)

Can combine both, but often do not need to

Experiments

POLI SCI 210

Introduction to Empirical Methods in Political Science

Last time

- Experiments as the **gold standard** for causal inference
- Thanks to random assignment, we can rule out:

1. Reverse causation
2. Selection bias
3. Omitted variable bias

What exactly can we learn from experiments?

Learning from experiments

How do you prove that a policy intervention works?

We want to make statements about *causation*

- **Example:** *Universal income improves political participation*

To back up those statements, we need to rule out **confounding factors**

- *Those who enroll on universal income programs are already more inclined to participate*

What kind of critique is this?

Ruling out confounders

We learned that **random assignment** allows us to rule out potential confounders

We can claim that treatment assignment is **ignorable** or **independent** from other factors

Challenge: This is only true *in expectation*

A small experiment

ID	Female	Y(0)	Y(1)
1	0	0	0
2	0	0	1
3	1	0	1
4	1	1	1

- are the **potential outcomes** under control `(0)` and treatment `(1)`, respectively $Y(*)$
- means person's life improves, means life stays the same $Y(*) = 1$ $Y(*) = 0$

A small experiment

ID	Female	Y(0)	Y(1)
1	0	0	0
2	0	0	1
3	1	0	1
4	1	1	1

- We have:
  - One person for which the policy would do nothing
  - Two people for which the policy improves life
  - One person who improves their life either way

Assign treatment at random

ID	Female	Y(0)	Y(1)	Z
1		0	0	0
2		0	1	0
3		1	1	1
4		1	1	1

- We happened to randomly assign the policy to the two women
- We only observe the potential outcomes that corresponds to the treatment status

Revealing outcomes

ID	Female	Y(0)	Y(1)	Z	Y obs
1	0	0	0	0	0
2	0	0	1	0	0
3	1	0	1	1	1
4	1	1	1	1	1

- The **true** treatment effect is

$$ATE = E[Y(1)] - E[Y(0)] = 3/4$$

- Which we **cannot observe in the real world**

Revealing outcomes

ID	Female	Y(0)	Y(1)	Z	Y obs
1	0	0	0	0	0
2	0	0	1	0	0
3	1	0	1	1	1
4	1	1	1	1	1

- We can **approximate** the ATE with $\widehat{ATE} = 2/2 - 0/2 = 1$
- We are off the mark! What happens if we redo the experiment?

Redoing the experiment

ID	Female	Y(0)	Y(1)	Z	Y obs
1	0	0	0	1	0
2	0	0	1	0	0
3	1	0	1	1	1
4	1	1	1	0	1

- We still have
$$ATE = 1/2$$
- But now
$$\widehat{ATE} = 1/2 - 1/2 = 0$$
- Off the mark in the opposite direction

Why does this happen?

				Experiment 1		Experiment 2	
ID	Female	Y(0)	Y(1)	Z	Y obs	Z	Y obs
1	0	0	0	0	0	1	0
2	0	0	1	0	0	0	0
3	1	0	1	1	1	1	1
4	1	1	1	1	1	0	1

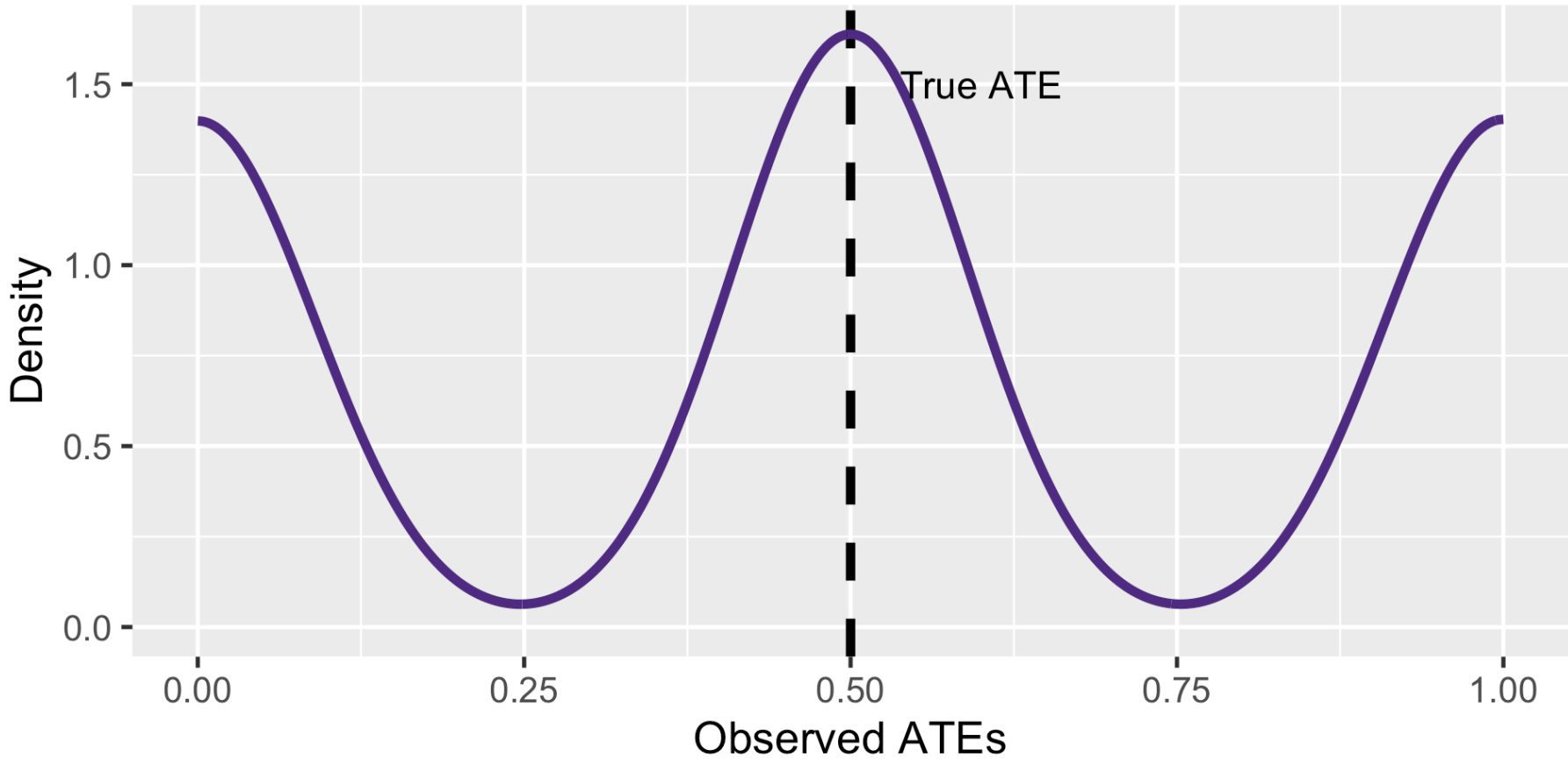
- Perhaps men and women react to treatment differently
- We want to rule out results depending on whether we assign treatments to men or women

Why does this happen?

- **Experiment 1:** 2/2 women in treatment and 0/2 in control `(imbalanced)`
- **Experiment 2:** 1/2 woman in treatment and 1/2 in control `(balanced)`

- Does that mean that experiment 2 is free from **random confounding**?

Redo 1,000 experiments



Half of the time we are spot on, half of the time we are wrong in either direction

What does this mean?

- Experiments only rule out the role of potential confounders **IN EXPECTATION**
- We can sustain this claim in two ways

1. **CLT:** With a sufficiently large sample `(But how large is large enough?)`
2. **WLLN:** By repeating the same experiment multiple times `(Nobody does this)`

In practice

- We only know statitical properties in our simulations
- Need a lot of domain expertise to attribute ATE to policy
- This involves **explaining why it works**
- First step toward knowing whether it would **work somewhere else**

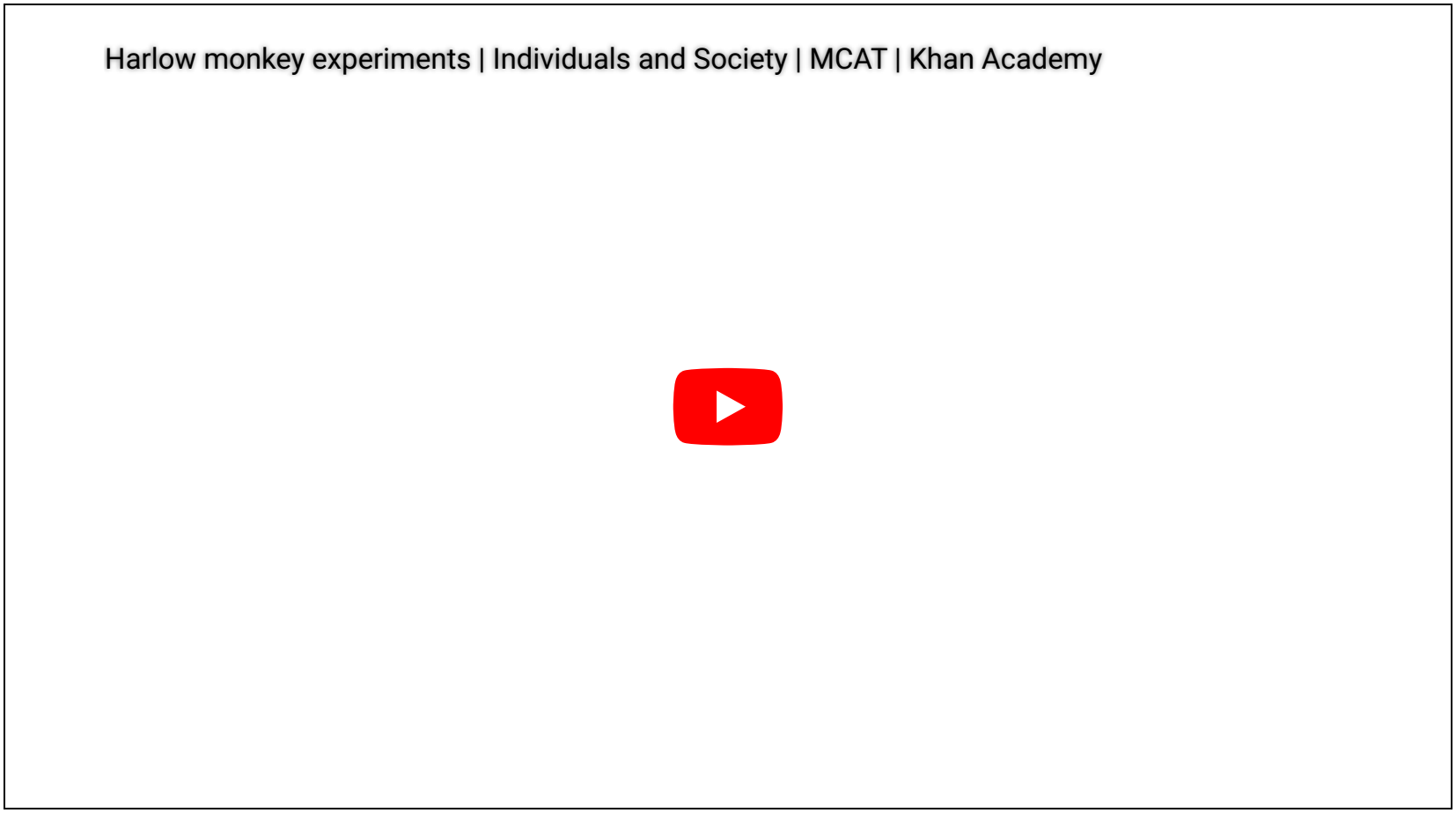
Generalization and extrapolation

- **Critique:** Experiments invest in *internal validity* at the expense of *external validity*

- **Internal validity:** We can `(probabilistically)` attribute effect to policy intervention
- **External validity:** Whether effect *extrapolates* or *generalizes*

- **Extrapolation:** Whether it works *elsewhere*
- **Generalization:** Whether it works *everywhere*

Thinking about external validity



https://youtu.be/9wmvZH5lX_U?si=pSNn03CKvYLH4SNm

Questions

1. That are these monkeys representative of?
2. To what other contexts would this apply?
3. Are the treatments realistic?
4. Is this behavior common/expected?

External validity concerns

Type	Concern

See Shadish et al (2002) and Egami and Hartmant (2023) for details

External validity concerns

Type	Concern
Samples	*Does this apply to a different population?*

See Shadish et al (2002) and Egami and Hartmant (2023) for details

External validity concerns

Type	Concern
Samples	*Does this apply to a different population?*
Contexts	*Does this apply in a different setting?*

See Shadish et al (2002) and Egami and Hartmant (2023) for details

External validity concerns

Type	Concern
Samples	*Does this apply to a different population?*
Contexts	*Does this apply in a different setting?*
Treatments	*Do they resemble real-world phenomena?*

See Shadish et al (2002) and Egami and Hartmant (2023) for details

External validity concerns

Type	Concern
Samples	*Does this apply to a different population?*
Contexts	*Does this apply in a different setting?*
Treatments	*Do they resemble real-world phenomena?*
Outcomes	*Do they reflect actual behaviors?*

See Shadish et al (2002) and Egami and Hartmant (2023) for details

Support factors

- **Example:** A house burns down because the television was left on
- Not all houses with TVs left on burn down, but sometimes they do, perhaps because the wiring was poor
- A **support factor** is one part of the **causal pie**
- **Causal pie:** A set of causes that are jointly but not separately sufficient for a contribution to an effect `(INUS causation)`
- **Analogy:** TUP only works if we have good schools

These ideas come from qualitative methods!

Scales and drills

- **Scaling up:** Whether we can apply intervention to broader area
  - Small scale interventions can become **unfeasible** or **cost-prohibitive** in a larger scale
  - Some policies only work at a small scale!

Scales and drills

- **Drilling down:** Can we apply the results of an intervention to individual units?
  - Just because it works **on average**, it does not mean that everyone will benefit from it
  - May waste money on people for whom the policy does not work
  - This can be **unethical**

Summary

- Experiments help in establishing *cause and effect*
- But do not explain how/why/where
- Need more knowledge to draw definitive conclusions