Randomized Experiments

Teppei Yamamoto

Keio University

Introduction to Causal Inference
Spring 2016

Randomization Solves the Selection Problem

Recall the selection bias formula:

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

How can we eliminate the bias term?

Randomization Solves the Selection Problem

Recall the selection bias formula:

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

How can we eliminate the bias term?

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

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

This implies Bias $= 0$.

Are Experiments Feasible in Social Science?

- Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments

Are Experiments Feasible in Social Science?

- Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments
- *Abbreviated* list of examples (from Green 2008):
    - Program evaluation: development programs, education programs, SAT prep classes, weight loss programs, fundraising, diversity training, deliberative polls, virginity pledging, advertising campaigns, wilderness programs, mental exercise for elderly
    - Public policy evaluation: teacher pay, class size, speed traps, vouchers, alternative sentencing, job training, health insurance subsidies, tax compliance, public housing, jury selection, police interventions
    - Behavioral research: persuasion, mobilization, education, income, interpersonal influence, conscientious health behaviors, media exposure, deliberation, discrimination
    - Research on institutions: rules for authorizing decisions, rules of succession, manner in which an organization is founded, monitoring performance, transparency, corruption, electoral systems, information

Proliferation of Field Experiments in Political Science

How about political science?

Proliferation of Field Experiments in Political Science

How about political science?

- Voter mobilization (Nickerson, Gerber and Green)
- Voting mechanisms (Olken)
- Health Insurance Reform (King et al.)
- Race-based discrimination in labor markets (Bertrand and Mullainathan)
- Corruption (Ferraz and Finnan)
- Information interventions for Elites (Butler)
- Monitoring interventions (Ichino)
- Many more in the pipeline...

Example: Social Pressure Experiment

- Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty.
- Two aspects of this type of utility, intrinsic satisfaction from behaving in accordance with a norm and extrinsic incentives to comply
- Gerber, Green, and Larimer (2008) test intrinsic motives in a large scale field experiment by applying varying degrees of extrinsic pressure on voters using to series of mailings to 180,002 households before the August 2006 primary election in Michigan.
  - Y_i: voted in primary (yes/no)
  - D_i: type of mailing

Example: Social Pressure Experiment

- Civic Duty:
  - Encouraged to vote
- Hawthorne:
  - Encouraged to vote
  - Told that researchers would be checking on whether they voted
- Self:
  - Encouraged to vote
  - Told that whether one votes is a matter of public record
  - Shown whether members of their own household voted in the last two elections
- Neighbors:
  - Like Self but in addition recipients are shown whether the neighbors on the block voted in the last two elections elections

Example: Social Pressure Experiment

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______
9999 BRIAN JOSEPH JACKSON		Voted	______
9991 JENNIFER KAY THOMPSON		Voted	______
9991 BOB R THOMPSON		Voted	

Example: Social Pressure Experiment

	Control (Not Mailed)	Civic Duty (Encouraged to vote)	Hawthorne (Encouraged & Monitored)	Self (Encouraged, Monitored, Shown Own Past Voting)	Neighbors (Encouraged, Monitored, Shown Own & Others' Past Voting)
Percent Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Identification vs. Estimation

Goal of causal inference: Learn about a counterfactual quantity of interest (QoI) using *finite, observed* data.

Identification vs. Estimation

Goal of causal inference: Learn about a counterfactual quantity of interest (QoI) using *finite, observed* data.

Causal inference thus involves 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?

Identification vs. Estimation

Goal of causal inference: Learn about a counterfactual quantity of interest (QoI) using *finite, observed* data.

Causal inference thus involves 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 (Manski):
IDENTIFICATION PRECEDES ESTIMATION

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 $N_0 = N - N_1$

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:

- For now, we assume we have data on the entire population (N = sample size = population size)

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:

- For now, we assume we have data on the entire population (N = sample size = population size)
- 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

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:

- For now, we assume we have data on the entire population (N = sample size = population size)
- 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

Randomization (simple or complete) implies: $\{Y_{1i},\ Y_{0i}\} \perp\!\!\!\perp D_i$

Identification of Average Treatment Effect

Identification assumption (guaranteed by random assignment):

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

Quantity of interest:

$$\tau_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}] = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$$

Identification of Average Treatment Effect

Identification assumption (guaranteed by random assignment):

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

Quantity of interest:

$$\tau_{ATE} \ \equiv\ \mathbb{E}[Y_{1i} - Y_{0i}] \ =\ \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$$

Is τ_{ATE} identified?

Identification of Average Treatment Effect

Identification assumption (guaranteed by random assignment):

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

Quantity of interest:

$$\tau_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}] = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$$

Is τ_{ATE} identified?

$$\begin{aligned}
\mathbb{E}[Y_i | D_i = 1] &= \mathbb{E}[D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i} | D_i = 1] \\
&= \mathbb{E}[Y_{1i} | D_i = 1]
\end{aligned}$$

Identification of Average Treatment Effect

Identification assumption (guaranteed by random assignment):

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

Quantity of interest:

$$\tau_{ATE} \ \equiv\ \mathbb{E}[Y_{1i} - Y_{0i}] \ =\ \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$$

Is τ_{ATE} identified?

$$\begin{aligned} \mathbb{E}[Y_i|D_i = 1] &= \mathbb{E}[D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] \ =\ \mathbb{E}[Y_{1i}] \quad (\because \text{ random assignment}) \end{aligned}$$

Similarly, $\mathbb{E}[Y_i|D_i = 0] \ =\ \mathbb{E}[Y_{0i}]$

Identification of Average Treatment Effect

Identification assumption (guaranteed by random assignment):

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

Quantity of interest:

$$\tau_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}] = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$$

Is τ_{ATE} identified?

$$\begin{aligned}\mathbb{E}[Y_i|D_i = 1] &= \mathbb{E}[D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}] \quad (\because \text{ random assignment})\end{aligned}$$

Similarly, $\mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_{0i}]$
So it follows that

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] = \underbrace{\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]}_{\text{observed difference in means}} \\ &= \frac{1}{N_1}\sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - D_i) Y_i\end{aligned}$$

Identification of ATT and ATC

Imagine a population with 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}		
1	2	1	2	?		
2	0	1	0	?		
3	1	0	?	1		
4	3	0	?	3		

Identification of ATT and ATC

Imagine a population with 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	$\mathbb{E}[Y_{1i}\mid D_i]$	$\mathbb{E}[Y_{0i}\mid D_i]$
1	2	1	2	?		
2	0	1	0	?		
3	1	0	?	1		
4	3	0	?	3		

Identification of ATT and ATC

Imagine a population with 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	$\mathbb{E}[Y_{1i} \mid D_i]$	$\mathbb{E}[Y_{0i} \mid D_i]$
1	2	1	2	?	$\mathbb{E}[Y_{1i} \mid D_i = 1] = 1$	$\mathbb{E}[Y_{0i} \mid D_i = 1] =?$
2	0	1	0	?		
3	1	0	?	1	$\mathbb{E}[Y_{1i} \mid D_i = 0] =?$	$\mathbb{E}[Y_{0i} \mid D_i = 0] = 2$
4	3	0	?	3		

Identification of ATT and ATC

Imagine a population with 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	$\mathbb{E}[Y_{1i}\|D_i]$	$\mathbb{E}[Y_{0i}\|D_i]$
1	2	1	2	?	$\mathbb{E}[Y_{1i}\|D_i=1]=1$	$\mathbb{E}[Y_{0i}\|D_i=1]=$
2	0	1	0	?		$\mathbb{E}[Y_{0i}\|D_i=0]=2$
3	1	0	?	1	$\mathbb{E}[Y_{1i}\|D_i=0]=$	$\mathbb{E}[Y_{0i}\|D_i=0]=2$
4	3	0	?	3	$\mathbb{E}[Y_{1i}\|D_i=1]=1$	

Random assignment ($\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i$) implies:

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

Identification of ATT and ATC

Imagine a population with 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	$\mathbb{E}[Y_{1i}\|D_i]$	$\mathbb{E}[Y_{0i}\|D_i]$
1	2	1	2	?	$\mathbb{E}[Y_{1i}\|D_i = 1] = 1$	$\mathbb{E}[Y_{0i}\|D_i = 1] =$
2	0	1	0	?		$\mathbb{E}[Y_{0i}\|D_i = 0] = 2$
3	1	0	?	1	$\mathbb{E}[Y_{1i}\|D_i = 0] =$	$\mathbb{E}[Y_{0i}\|D_i = 0] = 2$
4	3	0	?	3	$\mathbb{E}[Y_{1i}\|D_i = 1] = 1$	

Random assignment ($\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i$) implies:

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

So we have:

$$\tau_{ATT} = \tau_{ATC} = \tau_{ATE} = 1 - 2 = -1$$

Identification of ATT and ATC

Imagine a population with 4 units:

i	Y_i	D_i	Y_{1i}	Y_{0i}	$\mathbb{E}[Y_{1i} \vert D_i]$	$\mathbb{E}[Y_{0i} \vert D_i]$
1	2	1	2	?	$\mathbb{E}[Y_{1i} \vert D_i = 1] = 1$	$\mathbb{E}[Y_{0i} \vert D_i = 1] =$
2	0	1	0	?		$\mathbb{E}[Y_{0i} \vert D_i = 0] = 2$
3	1	0	?	1	$\mathbb{E}[Y_{1i} \vert D_i = 0] =$	$\mathbb{E}[Y_{0i} \vert D_i = 0] = 2$
4	3	0	?	3	$\mathbb{E}[Y_{1i} \vert D_i = 1] = 1$	

Random assignment ($\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i$) implies:

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

So we have:

$$\tau_{ATT} = \tau_{ATC} = \tau_{ATE} = 1 - 2 = -1$$

Under simple or complete random assignment, observed difference in means identifies ATE, ATT, and ATC.

Variance Due to Random Assignment

Now we know that observed difference in means gives a good point estimate for ATE. How about its estimation uncertainty?

Variance Due to Random Assignment

Now we know that observed difference in means gives a good point estimate for ATE. How about its estimation uncertainty?

Observed difference in means in the population:

$$\begin{aligned}\tilde{\tau} &\equiv \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \\ &= \frac{1}{N_1}\sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - D_i) Y_i\end{aligned}$$

What is the variance (= standard error^2) of $\tilde{\tau}$?

Variance Due to Random Assignment

Now we know that observed difference in means gives a good point estimate for ATE. How about its estimation uncertainty?

Observed difference in means in the population:

$$\begin{aligned}\tilde{\tau} &\equiv \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \\ &= \frac{1}{N_1}\sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - D_i) Y_i\end{aligned}$$

What is the variance (= standard error^2) of $\tilde{\tau}$?

- Here, variance is non-zero due to random assignment to treatment
- Even if we have data on entire population, $\tilde{\tau}$ still has uncertainty due to randomness in D_i

Variance Due to Random Sampling from the Population

- So far, we have assumed for simplicity that our data represent the entire population
- In reality, we have a sample from the population

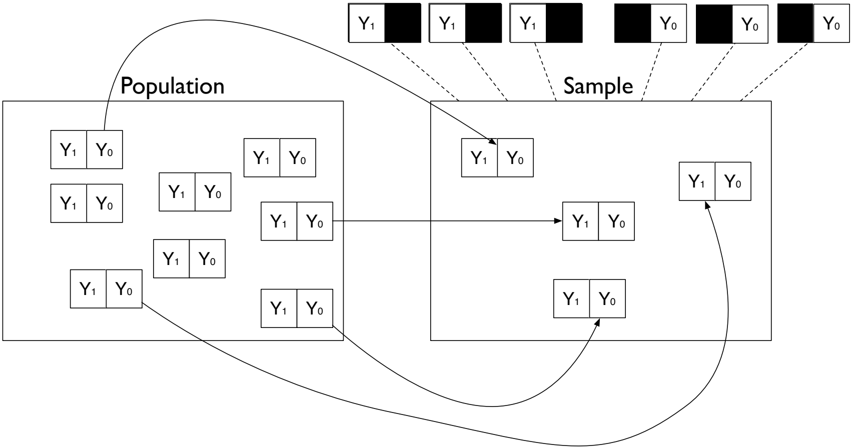
Variance Due to Random Sampling from the Population

- So far, we have assumed for simplicity that our data represent the entire population
- In reality, we have a sample from the population
- Sampling introduces an additional layer of uncertainty in causal inference:
  1. n units are randomly sampled from the population
  2. n_1 units are then randomly assigned to the treatment

Variance Due to Random Sampling from the Population

- So far, we have assumed for simplicity that our data represent the entire population
- In reality, we have a sample from the population
- Sampling introduces an additional layer of uncertainty in causal inference:
  1. n units are randomly sampled from the population
  2. n_1 units are then randomly assigned to the treatment
- How does this affect our inference, in terms of:
  - point estimates – is observed difference in means still unbiased?
  - uncertainty estimates – how can we incorporate both sources of variation?

What's the Estimand?



SATE and PATE

Setup:

- A random sample of units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: $Y_i(0)$, $Y_i(1)$
- Observed outcome: $Y_i = Y_i(D_i)$
- Number of treated/untreated units: $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_1$

SATE and PATE

Setup:

- A random sample of units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: $Y_i(0)$, $Y_i(1)$
- Observed outcome: $Y_i = Y_i(D_i)$
- Number of treated/untreated units: $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_1$

We now have two different ATEs, one for sample and one for population:

- Sample average treatment effect (SATE):

$$SATE \;=\; \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

SATE and PATE

Setup:

- A random sample of units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: $Y_i(0)$, $Y_i(1)$
- Observed outcome: $Y_i = Y_i(D_i)$
- Number of treated/untreated units: $n_1 = \sum_{i=1}^n D_i$ and $n_0 = n - n_1$

We now have two different ATEs, one for sample and one for population:

- Sample average treatment effect (SATE):

$$SATE = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

- Population average treatment effect (PATE):

$$PATE = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \text{(note what } \mathbb{E} \text{ here represents!)}$$

SATE and PATE

Setup:

- A random sample of units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: $Y_i(0)$, $Y_i(1)$
- Observed outcome: $Y_i = Y_i(D_i)$
- Number of treated/untreated units: $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_1$

We now have two different ATEs, one for sample and one for population:

- Sample average treatment effect (SATE):

$$SATE = \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\}$$

- Population average treatment effect (PATE):

$$PATE = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \text{(note what } \mathbb{E} \text{ here represents!)}$$

Random assignment of treatment still implies: $\{Y_i(1),\ Y_i(0)\} \perp\!\!\!\perp D_i$

Estimation of Sample Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i) Y_i$$

Is this estimator unbiased for SATE?

Estimation of Sample Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i) Y_i$$

Is this estimator unbiased for SATE?

Recall: Because our QoI here is *sample* ATE, unbiasedness is defined over repeated random treatment assignment (not sampling).

Estimation of Sample Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

Is this estimator unbiased for SATE?

Recall: Because our QoI here is *sample* ATE, unbiasedness is defined over repeated random treatment assignment (not sampling).

Define $\mathcal{O} \equiv$ the current sample. Then:

$$\mathbb{E}(\hat{\tau} \mid \mathcal{O}) \;=$$

Estimation of Sample Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

Is this estimator unbiased for SATE?

Recall: Because our QoI here is *sample* ATE, unbiasedness is defined over repeated random treatment assignment (not sampling).

Define $\mathcal{O} \equiv$ the current sample. Then:

$$\mathbb{E}(\hat{\tau} \mid \mathcal{O}) \;=\; \frac{1}{n_1}\sum_{i=1}^{n} \underbrace{\mathbb{E}(D_i \mid \mathcal{O})}_{=\;n_1/n} Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} \underbrace{\{1-\mathbb{E}(D_i \mid \mathcal{O})\}}_{=\;n_0/n} Y_i(0)$$

Estimation of Sample Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

Is this estimator unbiased for SATE?

Recall: Because our QoI here is *sample* ATE, unbiasedness is defined over repeated random treatment assignment (not sampling).

Define $\mathcal{O} \equiv$ the current sample. Then:

$$\begin{aligned}\mathbb{E}(\hat{\tau} \mid \mathcal{O}) \;&=\; \frac{1}{n_1}\sum_{i=1}^{n}\underbrace{\mathbb{E}(D_i \mid \mathcal{O})}_{=\;n_1/n} Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\underbrace{\{1-\mathbb{E}(D_i \mid \mathcal{O})\}}_{=\;n_0/n} Y_i(0) \\ &=\; \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \;=\; SATE\end{aligned}$$

$\hat{\tau}$ is therefore unbiased for SATE in a randomized experiment.

Estimation of Population Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i) Y_i$$

Is this estimator unbiased for PATE?

Estimation of Population Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i) Y_i$$

Is this estimator unbiased for PATE?

$$\mathbb{E}(\hat{\tau}) \;=$$

Estimation of Population Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i) Y_i$$

Is this estimator unbiased for PATE?

$$\begin{aligned}
\mathbb{E}(\hat{\tau}) &= \mathbb{E}\left\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\right\} \quad \text{(law of iterated expectations)} \\
&= \mathbb{E}(SATE) \quad \text{(random assignment)} \\
&= PATE \quad \text{(random sampling)}
\end{aligned}$$

Estimation of Population Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

Is this estimator unbiased for PATE?

$$\begin{aligned}
\mathbb{E}(\hat{\tau}) &= \mathbb{E}\left\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\right\} \quad \text{(law of iterated expectations)} \\
&= \mathbb{E}(SATE) \quad \text{(random assignment)} \\
&= PATE \quad \text{(random sampling)} \quad \text{Yes!}
\end{aligned}$$

Estimation of Population Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

Is this estimator unbiased for PATE?

$$\begin{aligned}
\mathbb{E}(\hat{\tau}) &= \mathbb{E}\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\} \quad \text{(law of iterated expectations)} \\
&= \mathbb{E}(SATE) \quad \text{(random assignment)} \\
&= PATE \quad \text{(random sampling)} \quad \text{Yes!}
\end{aligned}$$

- Note that this requires a true random sampling from the population
- Often in social science, obtaining such a sample is impossible
- In such a case, focus on SATE and interpret as such (estimate still *internally valid*, but no longer *externally valid*)

Estimation of Population Average Treatment Effect

Estimator = Observed difference in means in the sample:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

Is this estimator unbiased for PATE?

$$\begin{aligned}
\mathbb{E}(\hat{\tau}) &= \mathbb{E}\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\} \quad \text{(law of iterated expectations)} \\
&= \mathbb{E}(SATE) \quad \text{(random assignment)} \\
&= PATE \quad \text{(random sampling)} \quad \text{Yes!}
\end{aligned}$$

- Note that this requires a true random sampling from the population
- Often in social science, obtaining such a sample is impossible
- In such a case, focus on SATE and interpret as such (estimate still *internally valid*, but no longer *externally valid*)
- Also, because of random assignment, $PATE = PATT$

Variance Estimation for Within-Sample Inference

Again consider the difference-in-means estimator:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i) Y_i$$

What is the standard error of $\hat{\tau}$ *as an estimator of* SATE?

Variance Estimation for Within-Sample Inference

Again consider the difference-in-means estimator:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

What is the standard error of $\hat{\tau}$ *as an estimator of* SATE?

Results:

- Exact variance for SATE is *unidentifiable*

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) \;=\; \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right)$$

$$\text{where} \begin{cases} S_1^2 & = & \text{Sample variance of } Y_i \text{ for the treated (identified)} \\ S_0^2 & = & \text{Sample variance of } Y_i \text{ for the untreated (identified)} \\ S_{01} & = & \text{Sample covariance of } Y_i(1) \text{ and } Y_i(0) \text{ (unidentified)} \end{cases}$$

- Can use the usual formula for conservative inference:

$$\widehat{\mathbb{V}(\hat{\tau} \mid \mathcal{O})} \;\equiv\; \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} \;\geq\; \mathbb{V}(\hat{\tau} \mid \mathcal{O})$$

Details of the SATE Variance Results

When D_i is assigned by complete randomization, we can show:

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

where

$$\begin{aligned} S_1^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad \text{where} \quad \bar{Y}_1 = \frac{1}{N}\sum_{j=1}^{N} Y_{1j} \\ &= \end{aligned}$$

Details of the SATE Variance Results

When D_i is assigned by complete randomization, we can show:

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) = \frac{1}{N}\left(\frac{N_0}{N_1} S_1^2 + \frac{N_1}{N_0} S_0^2 + 2S_{01}\right)$$

where

$$\begin{aligned} S_1^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad \text{where} \quad \bar{Y}_1 = \frac{1}{N}\sum_{j=1}^{N} Y_{1j} \\ &= \text{Sample variance of the potential outcome under treatment} \end{aligned}$$

Details of the SATE Variance Results

When D_i is assigned by complete randomization, we can show:

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

where

$$\begin{aligned}
S_1^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad \text{where} \quad \bar{Y}_1 = \frac{1}{N}\sum_{j=1}^{N} Y_{1j} \\
&= \text{Sample variance of the potential outcome under treatment} \\
S_0^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2 \quad \text{where} \quad \bar{Y}_0 = \frac{1}{N}\sum_{j=1}^{N} Y_{0j} \\
&=
\end{aligned}$$

Details of the SATE Variance Results

When D_i is assigned by complete randomization, we can show:

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \; = \; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

where

$$\begin{aligned}
S_1^2 \quad &= \quad \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad \text{where} \quad \bar{Y}_1 = \frac{1}{N}\sum_{j=1}^{N} Y_{1j} \\
&= \quad \text{Sample variance of the potential outcome under treatment} \\
S_0^2 \quad &= \quad \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2 \quad \text{where} \quad \bar{Y}_0 = \frac{1}{N}\sum_{j=1}^{N} Y_{0j} \\
&= \quad \text{Sample variance of the potential outcome under control} \\
S_{01} \quad &= \quad \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right) \\
&=
\end{aligned}$$

Details of the SATE Variance Results

When D_i is assigned by complete randomization, we can show:

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1} S_1^2 + \frac{N_1}{N_0} S_0^2 + 2 S_{01}\right)$$

where

$$\begin{aligned}
S_1^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad \text{where} \quad \bar{Y}_1 = \frac{1}{N}\sum_{j=1}^{N} Y_{1j} \\
&= \text{Sample variance of the potential outcome under treatment} \\
S_0^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2 \quad \text{where} \quad \bar{Y}_0 = \frac{1}{N}\sum_{j=1}^{N} Y_{0j} \\
&= \text{Sample variance of the potential outcome under control} \\
S_{01} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right) \\
&= \text{Sample covariance of } Y_{1i} \text{ and } Y_{0i}
\end{aligned}$$

Details of the SATE Variance Results

When D_i is assigned by complete randomization, we can show:

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

where

$$\begin{aligned}
S_1^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad \text{where} \quad \bar{Y}_1 = \frac{1}{N}\sum_{j=1}^{N} Y_{1j} \\
&= \text{Sample variance of the potential outcome under treatment} \\
S_0^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2 \quad \text{where} \quad \bar{Y}_0 = \frac{1}{N}\sum_{j=1}^{N} Y_{0j} \\
&= \text{Sample variance of the potential outcome under control} \\
S_{01} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right) \\
&= \text{Sample covariance of } Y_{1i} \text{ and } Y_{0i}
\end{aligned}$$

Note: Don't confuse sampling variance with sample variance!

Unidentifiability of Sampling Variance

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2$$

$$S_{01} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right)$$

Which of these are identified?

Unidentifiability of Sampling Variance

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2$$

$$S_{01} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right)$$

Which of these are identified?

- S_1^2: Identified by the observed variance in the treatment group

$$\tilde{S}_1^2 \ = \ \frac{1}{N_1 - 1}\sum_{i=1}^{N} D_i (Y_i - \tilde{Y}_1)^2 \quad \text{where} \quad \tilde{Y}_1 = \frac{1}{N_1}\sum_{j=1}^{N} D_j Y_j$$

Unidentifiability of Sampling Variance

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2$$

$$S_{01} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right)$$

Which of these are identified?

- S_1^2: Identified by the observed variance in the treatment group

$$\tilde{S}_1^2 \; = \; \frac{1}{N_1 - 1}\sum_{i=1}^{N} D_i (Y_i - \tilde{Y}_1)^2 \quad \text{where} \quad \tilde{Y}_1 = \frac{1}{N_1}\sum_{j=1}^{N} D_j Y_j$$

- S_0^2: Identified by the observed variance in the control group

$$\tilde{S}_0^2 \; = \; \frac{1}{N_0 - 1}\sum_{i=1}^{N}(1 - D_i)(Y_i - \tilde{Y}_0)^2 \quad \text{where} \quad \tilde{Y}_0 = \frac{1}{N_0}\sum_{j=1}^{N}(1 - D_j) Y_j$$

Unidentifiability of Sampling Variance

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2$$

$$S_{01} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right)$$

Which of these are identified?

- S_1^2: Identified by the observed variance in the treatment group

$$\tilde{S}_1^2 \;=\; \frac{1}{N_1 - 1}\sum_{i=1}^{N} D_i (Y_i - \tilde{Y}_1)^2 \quad \text{where} \quad \tilde{Y}_1 = \frac{1}{N_1}\sum_{j=1}^{N} D_j Y_j$$

- S_0^2: Identified by the observed variance in the control group

$$\tilde{S}_0^2 \;=\; \frac{1}{N_0 - 1}\sum_{i=1}^{N}(1 - D_i)(Y_i - \tilde{Y}_0)^2 \quad \text{where} \quad \tilde{Y}_0 = \frac{1}{N_0}\sum_{j=1}^{N}(1 - D_j) Y_j$$

- S_{01}: Unidentified because we never observe Y_{1i} and Y_{0i} together

Unidentifiability of Sampling Variance

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)^2 \quad S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i} - \bar{Y}_0\right)^2$$

$$S_{01} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - \bar{Y}_1\right)\left(Y_{0i} - \bar{Y}_0\right)$$

Which of these are identified?

- S_1^2: Identified by the observed variance in the treatment group

$$\tilde{S}_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N} D_i(Y_i - \tilde{Y}_1)^2 \quad \text{where} \quad \tilde{Y}_1 = \frac{1}{N_1}\sum_{j=1}^{N} D_j Y_j$$

- S_0^2: Identified by the observed variance in the control group

$$\tilde{S}_0^2 = \frac{1}{N_0 - 1}\sum_{i=1}^{N}(1 - D_i)(Y_i - \tilde{Y}_0)^2 \quad \text{where} \quad \tilde{Y}_0 = \frac{1}{N_0}\sum_{j=1}^{N}(1 - D_j) Y_j$$

- S_{01}: Unidentified because we never observe Y_{1i} and Y_{0i} together

Therefore, we can never estimate the sampling variance of $\tilde{\tau}$ without bias

Conservativeness of Usual Variance Estimator

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

Can we make a valid inference about our target estimand, $\mathbb{V}(\tilde{\tau} \mid \mathcal{O})$?

Conservativeness of Usual Variance Estimator

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

Can we make a valid inference about our target estimand, $\mathbb{V}(\tilde{\tau} \mid \mathcal{O})$?

Consider the usual variance formula for difference in means:

$$\mathbb{V}\widehat{(\tilde{\tau} \mid \mathcal{O})} \;\equiv\; \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$$

Conservativeness of Usual Variance Estimator

$$\mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \;=\; \frac{1}{N}\left(\frac{N_0}{N_1}S_1^2 + \frac{N_1}{N_0}S_0^2 + 2S_{01}\right)$$

Can we make a valid inference about our target estimand, $\mathbb{V}(\tilde{\tau} \mid \mathcal{O})$?

Consider the usual variance formula for difference in means:

$$\widehat{\mathbb{V}(\tilde{\tau} \mid \mathcal{O})} \;\equiv\; \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$$

It turns out that we can show:

- $\widehat{\mathbb{V}(\tilde{\tau} \mid \mathcal{O})}$ is identified (as shown on the previous slide)
- $\widehat{\mathbb{V}(\tilde{\tau} \mid \mathcal{O})} \geq \mathbb{V}(\tilde{\tau} \mid \mathcal{O}) \Longrightarrow$ usual formula is always conservative
- $\widehat{\mathbb{V}(\tilde{\tau} \mid \mathcal{O})} = \mathbb{V}(\tilde{\tau} \mid \mathcal{O})$ if and only if $\tau_i = SATE$ for all i (i.e. constant effect)

Therefore, usual formula is still useful, even though it is always too big

Analyzing Classical Experiment Using Regression

Now, for a binary treatment ($D_i \in \{0, 1\}$) we can show...

- Simple regression coefficient is *numerically equal* to difference in means:

$$\hat{\beta}_{OLS} \equiv \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(D_i - \overline{D})}{\sum_{i=1}^{n}(D_i - \overline{D})^2} = \tilde{\tau}$$

- Heteroskedasticity-robust variance (the HC2 variant) is also *numerically equal* to the usual variance formula:

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

This implies...

Analyzing Classical Experiment Using Regression

Now, for a binary treatment ($D_i \in \{0, 1\}$) we can show...

- Simple regression coefficient is *numerically equal* to difference in means:

$$\hat{\beta}_{OLS} \equiv \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(D_i - \overline{D})}{\sum_{i=1}^{n}(D_i - \overline{D})^2} = \tilde{\tau}$$

- Heteroskedasticity-robust variance (the HC2 variant) is also *numerically equal* to the usual variance formula:

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

This implies, in a completely randomized experiment, you can simply:

1. Regress Y_i on D_i and get the coefficient on D_i
2. Calculate the robust standard error (`vcovHC` in the `sandwich` package in R, with `type = "HC2"`)
3. Do t-test, calculate confidence intervals, etc. as usual

Variance Estimation for Population Inference

Now for the same difference-in-means estimator:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i)Y_i$$

What is the standard error of $\hat{\tau}$ *as an estimator of* PATE?

Variance Estimation for Population Inference

Now for the same difference-in-means estimator:

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$$

What is the standard error of $\hat{\tau}$ *as an estimator of* PATE?

Good news: For PATE, the usual formula is unbiased

$$\mathbb{E}\left[\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}\right] \;=\; \mathbb{V}(\hat{\tau})$$

- Intuitively: This formula was always too large for SATE; we overestimated the variability
- But for PATE, because we have additional uncertainty, this becomes an unbiased estimator

Asymptotic Inference for SATE and PATE

What happens in a large sample? (i.e. $n \to \infty$)

Asymptotic Inference for SATE and PATE

What happens in a large sample? (i.e. $n \to \infty$)

What happens to SATE?

Asymptotic Inference for SATE and PATE

What happens in a large sample? (i.e. $n \to \infty$)

What happens to SATE? $SATE \xrightarrow{P} PATE$ because of LLN

Asymptotic Inference for SATE and PATE

What happens in a large sample? (i.e. $n \to \infty$)

What happens to SATE? $SATE \xrightarrow{P} PATE$ because of LLN

What happens to $\hat{\tau}$?

Asymptotic Inference for SATE and PATE

What happens in a large sample? (i.e. $n \to \infty$)

What happens to SATE? $SATE \xrightarrow{P} PATE$ because of LLN

What happens to $\hat{\tau}$?

- Because of LLN, $\hat{\tau}$ is consistent for PATE:

$$\hat{\tau} \xrightarrow{P} PATE$$

Asymptotic Inference for SATE and PATE

What happens in a large sample? (i.e. $n \to \infty$)

What happens to SATE? $SATE \xrightarrow{P} PATE$ because of LLN

What happens to $\hat{\tau}$?

- Because of LLN, $\hat{\tau}$ is consistent for PATE:

$$\hat{\tau} \xrightarrow{P} PATE$$

- And because of CLT, $\hat{\tau}$ is asymptotically normal:

$$\hat{\tau} \overset{a.}{\sim} \mathcal{N}(PATE, \mathbb{V}(\hat{\tau}))$$

In a nutshell, standard tools work out when you have a large sample!

Example: Social Pressure Experiment

R Code

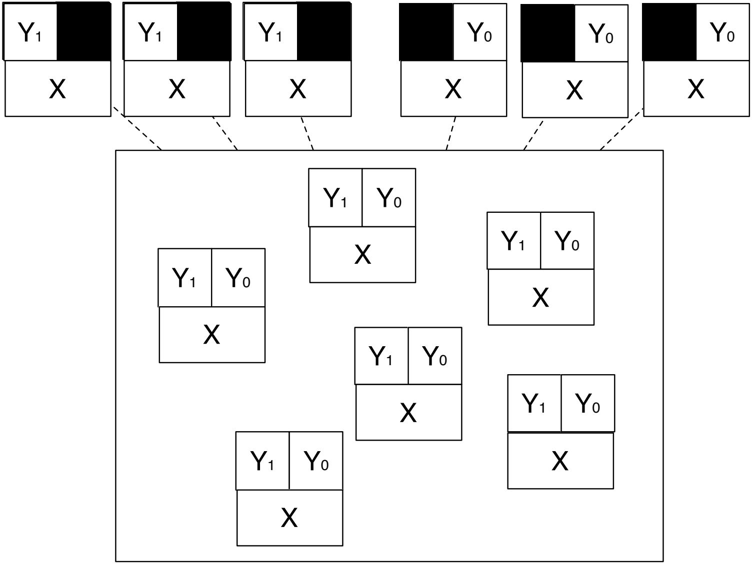
```
> library(foreign)
> d <- read.dta("gerber.dta")
>
> aa <- lm(voted ~ treatment, data = d)
> coef(aa)  # ATE point estimates

        (Intercept)  treatment Hawthorne treatment Civic Duty
              1.297                0.026                0.018
 treatment Neighbors       treatment Self
              0.081                0.049

> library(sandwich)
> sqrt(diag(vcovHC(aa, type="HC2")))  # ATE conservative standard errors

        (Intercept)  treatment Hawthorne treatment Civic Duty
             0.0010               0.0026               0.0026
 treatment Neighbors       treatment Self
             0.0027               0.0026
```

Covariates and Experiments



Covariate Adjustment in Randomized Experiments

- Randomization balances both observed and unobserved pre-treatment covariates between the treated and untreated in large samples
- In small samples, you may get unlucky and suffer from inbalance

Covariate Adjustment in Randomized Experiments

- Randomization balances both observed and unobserved pre-treatment covariates between the treated and untreated in large samples
- In small samples, you may get unlucky and suffer from inbalance
- Common practice: Conduct balance checks with respect to observed pre-treatment covariates
  - Compare means, standard deviations etc. between the treated and untreated; can also regress treatment indicator on covariates
  - Visual inspection of histograms/density plots

What if you found inbalance?

Covariate Adjustment in Randomized Experiments

- Randomization balances both observed and unobserved pre-treatment covariates between the treated and untreated in large samples
- In small samples, you may get unlucky and suffer from inbalance
- Common practice: Conduct balance checks with respect to observed pre-treatment covariates
  - Compare means, standard deviations etc. between the treated and untreated; can also regress treatment indicator on covariates
  - Visual inspection of histograms/density plots

What if you found inbalance?

- Can correct inbalance via regression, matching, weighting, etc.
- Post-randomization adjustment can also improve efficiency

Covariate Adjustment in Randomized Experiments

- Randomization balances both observed and unobserved pre-treatment covariates between the treated and untreated in large samples
- In small samples, you may get unlucky and suffer from inbalance
- Common practice: Conduct balance checks with respect to observed pre-treatment covariates
  - Compare means, standard deviations etc. between the treated and untreated; can also regress treatment indicator on covariates
  - Visual inspection of histograms/density plots

What if you found inbalance?

- Can correct inbalance via regression, matching, weighting, etc.
- Post-randomization adjustment can also improve efficiency
- But it may also produce bias, such as:
  - Bias due to model misspecification
  - Bias due to post-hoc analysis ("p-hacking")
  - Bias due to incorrectly adjusting post-treatment covariates
- Opinions vary: Best to show stable results with or without adjustment

Example: Social Pressure Experiment

With $n \simeq 180,000$, covariates are almost perfectly balanced:

R Code

```
> d <- read.dta("gerber.dta")
> covars <- subset(d, select = c("hh_size", "g2002", "g2000", "p2004",
                                 "p2002", "p2000", "sex", "yob"))
> aggregate(covars, by = list(d$treatment), mean)
      Group.1 hh_size g2002 g2000 p2004 p2002 p2000   sex  yob
1     Control    2.18 0.811 0.843  1.40 0.389 0.252 0.499 1956
2   Hawthorne    2.18 0.813 0.844  1.40 0.394 0.250 0.499 1956
3  Civic Duty    2.19 0.811 0.842  1.40 0.389 0.254 0.500 1956
4   Neighbors    2.19 0.811 0.842  1.41 0.387 0.251 0.500 1956
5        Self    2.18 0.811 0.840  1.40 0.392 0.251 0.500 1956

> aggregate(covars, by = list(d$treatment), sd)
      Group.1 hh_size g2002 g2000 p2004 p2002 p2000 sex  yob
1     Control   0.788 0.392 0.363 0.490 0.488 0.434 0.5 14.4
2   Hawthorne   0.789 0.390 0.362 0.491 0.489 0.433 0.5 14.4
3  Civic Duty   0.802 0.391 0.365 0.490 0.487 0.435 0.5 14.5
4   Neighbors   0.805 0.391 0.365 0.491 0.487 0.434 0.5 14.6
5        Self   0.782 0.391 0.366 0.490 0.488 0.434 0.5 14.4
```

Example: Social Pressure Experiment

Check balance by regressing treatment indicators on covariates:

R Code

```
> d$self <- as.numeric(d$treatment) == 5
> fit <- lm(self ~ hh_size + g2002 + g2000 + p2004 + p2002 + p2000 +
                   sex + yob, data = d)
> summary(fit)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.31e-01   8.22e-02    1.59    0.112
hh_size     -6.97e-04   7.23e-04   -0.96    0.335
g2002        5.30e-04   1.56e-03    0.34    0.734
g2000       -2.91e-03   1.69e-03   -1.72    0.085 .
p2004Yes     5.11e-04   1.10e-03    0.46    0.643
p2002        1.11e-03   1.14e-03    0.97    0.333
p2000       -2.66e-04   1.26e-03   -0.21    0.833
sex          1.66e-04   1.07e-03    0.16    0.877
yob         -8.55e-06   4.19e-05   -0.20    0.838
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
F-statistic: 0.599 on 8 and 344075 DF,  p-value: 0.779
```

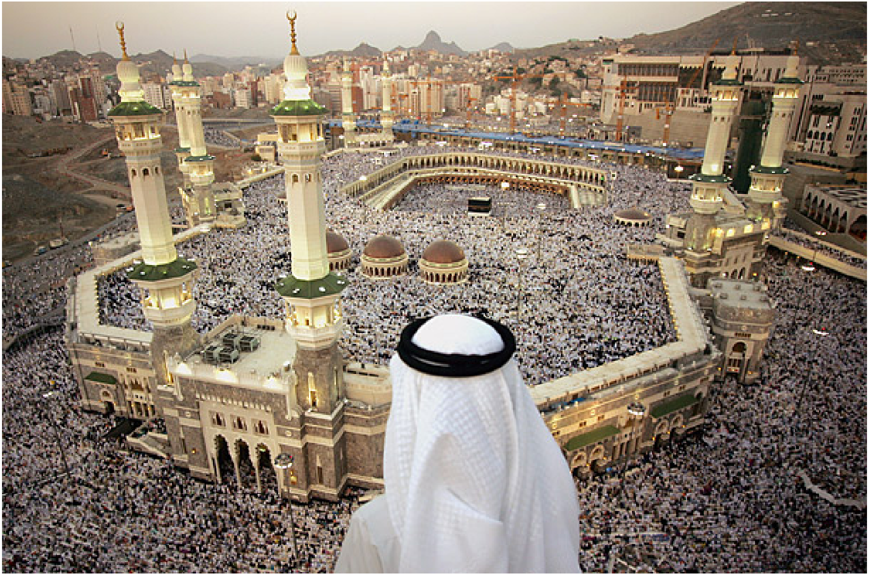
Threats to Internal and External Validity

- Internal validity: can we estimate the treatment effect for our particular sample?
    - Fails when there are differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for
- External validity: can we extrapolate our estimates to other populations?
    - Fails when outside the experimental environment the treatment has a different effect

Most Common Threats to Internal Validity

- Failure of randomization
  - e.g. implementing partners assign their favorites to treatment group; imbalance due to small sample size
- Noncompliance with experimental protocol
  - e.g. failure to treat or "crossover": Some members of the control group receive the treatment and some members of the treatment group go untreated
- Differential attrition
  - e.g. control group subjects are more likely to drop out of a study than treatment group subjects

Example: Klingsmith et al.



Example: Natural Experiment in Pakistan

- Pakistan allocated about 135,000 visas to Saudi Arabia for the Hajj via a randomized lottery.
- Wealthier Pakistanis tend to use private Hajj tour operators rather than the lottery.
- Randomization occurs among individuals grouped into "parties", where parties are stratified by sect, region, and accommodation.
- Compliance with the experiment is imperfect:
  - 99% who win lottery attend the Hajj.
  - 11% who lose lottery still attend the Hajj (via private tours).

Example: Natural Experiment in Pakistan

- Pakistan allocated about 135,000 visas to Saudi Arabia for the Hajj via a randomized lottery.
- Wealthier Pakistanis tend to use private Hajj tour operators rather than the lottery.
- Randomization occurs among individuals grouped into "parties", where parties are stratified by sect, region, and accommodation.
- Compliance with the experiment is imperfect:
  - 99% who win lottery attend the Hajj.
  - 11% who lose lottery still attend the Hajj (via private tours).
- Because randomization is not controlled by researcher, balance checks and qualitative investigation is crucial.

Example: Natural Experiment in Pakistan

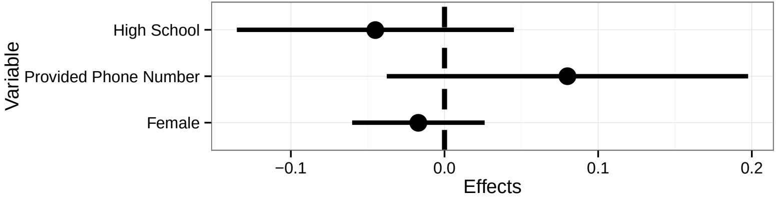
Two pieces of information to bolster the randomization assumption:

- Cursory qualitative information:
  - The lottery selection algorithm was designed and implemented by an independent and reputable third party, and there were no reports of lottery manipulation.

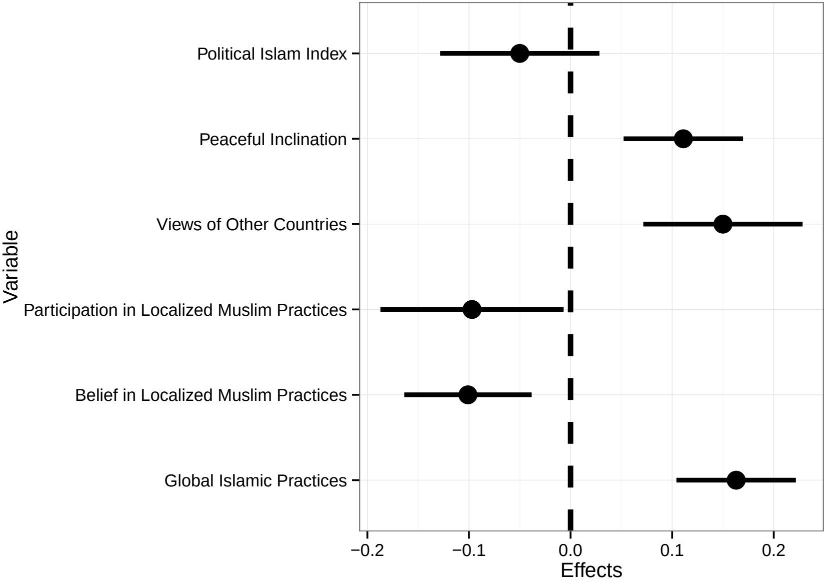
Example: Natural Experiment in Pakistan

Two pieces of information to bolster the randomization assumption:

- Cursory qualitative information:
  - The lottery selection algorithm was designed and implemented by an independent and reputable third party, and there were no reports of lottery manipulation.
- Balance tests:



Example: Natural Experiment in Pakistan



Most Common Threats to External Validity

- Non-representative sample
  - e.g. laboratory experiment using a convenience sample
  - Subjects are randomly sampled, but not from the population of interest
- Non-representative treatment
  - The treatment differs in actual implementations
  - e.g. survey experiment about the effect of media priming on voting
  - Scale effects
  - Actual implementations are not randomized (nor full scale)

Internal vs. External Validity

Which one is more important?

Internal vs. External Validity

Which one is more important?

"One common view is that internal validity comes first. If you do not know the effects of the treatment on the units in your study, you are not well-positioned to infer the effects on units you did not study who live in circumstances you did not study." (Rosenbaum 2010, p. 56)

- Randomization ensures internal validity
- External validity may be partially addressed by comparing the results of several internally valid studies conducted in different circumstances and at different times
- Note that the same external validity issues often apply in observational studies

Cluster Randomization

- So far, we have assumed treatments are assigned at the individual level
- Sometimes random assignment occurs at the cluster level for various reasons:
  - Treatment only makes sense at the group level, but the outcome is measured for individuals
  - Treatment too costly to implement individually
  - SUTVA only plausible if treatment is defined at the group level

(Example: Effect of teaching method on student performance)

Cluster Randomization

- So far, we have assumed treatments are assigned at the individual level
- Sometimes random assignment occurs at the cluster level for various reasons:
  - Treatment only makes sense at the group level, but the outcome is measured for individuals
  - Treatment too costly to implement individually
  - SUTVA only plausible if treatment is defined at the group level

  (Example: Effect of teaching method on student performance)
- Standard errors ignoring cluster randomization are usually too small
- This is due to clustering, i.e., units within the same cluster are typically more similar than units in different clusters

Cluster Randomization

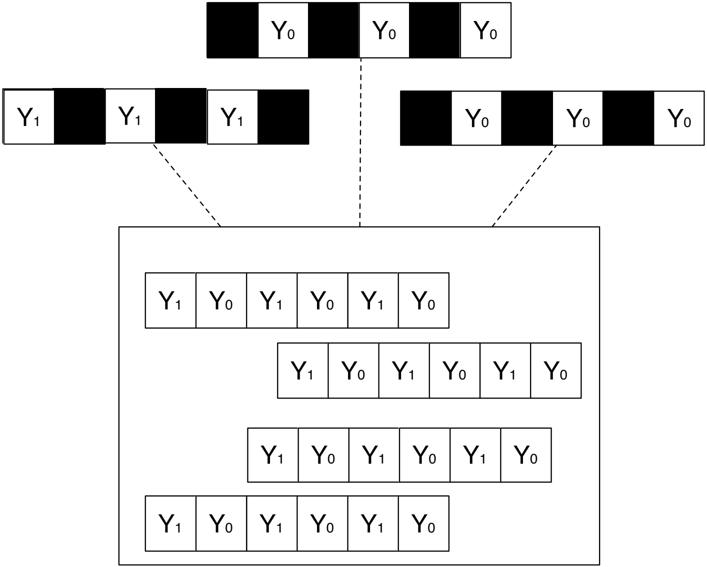
- So far, we have assumed treatments are assigned at the individual level
- Sometimes random assignment occurs at the cluster level for various reasons:
  - Treatment only makes sense at the group level, but the outcome is measured for individuals
  - Treatment too costly to implement individually
  - SUTVA only plausible if treatment is defined at the group level

  (Example: Effect of teaching method on student performance)

- Standard errors ignoring cluster randomization are usually too small
- This is due to clustering, i.e., units within the same cluster are typically more similar than units in different clusters

  "Analyses of group randomized trials that ignore clustering are an exercise in self-deception." (Cornfield 1978)

Randomization at the Group Level



Intracluster Correlation

Recall the Law of Total Variance:

$$\underbrace{\mathbb{V}(Y)}_{\text{total variance}} = \underbrace{\mathbb{E}[\mathbb{V}(Y \mid X)]}_{\text{(mean of) ``within'' variance}} + \underbrace{\mathbb{V}(\mathbb{E}[Y \mid X])}_{\text{``between'' variance}}$$

Intracluster Correlation

Recall the Law of Total Variance:

$$\underbrace{\mathbb{V}(Y)}_{\text{total variance}} = \underbrace{\mathbb{E}[\mathbb{V}(Y \mid X)]}_{\text{(mean of) ``within'' variance}} + \underbrace{\mathbb{V}(\mathbb{E}[Y \mid X])}_{\text{``between'' variance}}$$

This implies the decomposition of heterogeneity in potential outcomes:

$$\underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{\bar{Y}}_d)^2}_{\text{overall variation: } \sigma^2} = \underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{Y}_{dj})^2}_{\text{within-cluster variation: } \sigma^2_W} + \underbrace{\sum_{j=1}^{G} N_j(\bar{Y}_{dj} - \bar{\bar{Y}}_d)^2}_{\text{between-cluster variation: } \sigma^2_B}$$

for $d = 0, 1$, where $\bar{Y}_{dj}$ = mean Y_{dij} in cluster j and $\bar{\bar{Y}}_d$ = global mean Y_{dij}

Intracluster Correlation

Recall the Law of Total Variance:

$$\underbrace{\mathbb{V}(Y)}_{\text{total variance}} = \underbrace{\mathbb{E}[\mathbb{V}(Y \mid X)]}_{\text{(mean of) ``within'' variance}} + \underbrace{\mathbb{V}(\mathbb{E}[Y \mid X])}_{\text{``between'' variance}}$$

This implies the decomposition of heterogeneity in potential outcomes:

$$\underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{\bar{Y}}_d)^2}_{\text{overall variation: } \sigma^2} = \underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{Y}_{dj})^2}_{\text{within-cluster variation: } \sigma_W^2} + \underbrace{\sum_{j=1}^{G} N_j(\bar{Y}_{dj} - \bar{\bar{Y}}_d)^2}_{\text{between-cluster variation: } \sigma_B^2}$$

for $d = 0, 1$, where $\bar{Y}_{dj}$ = mean Y_{dij} in cluster j and $\bar{\bar{Y}}_d$ = global mean Y_{dij}

Then we can define the intracluster correlation: $\rho = \frac{\sigma_B^2}{\sigma^2} = 1 - \frac{\sigma_W^2}{\sigma^2}$

Intracluster Correlation

Recall the Law of Total Variance:

$$\underbrace{\mathbb{V}(Y)}_{\text{total variance}} = \underbrace{\mathbb{E}[\mathbb{V}(Y \mid X)]}_{\text{(mean of) “within” variance}} + \underbrace{\mathbb{V}(\mathbb{E}[Y \mid X])}_{\text{“between” variance}}$$

This implies the decomposition of heterogeneity in potential outcomes:

$$\underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{\bar{Y}}_d)^2}_{\text{overall variation: } \sigma^2} = \underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{Y}_{dj})^2}_{\text{within-cluster variation: } \sigma^2_W} + \underbrace{\sum_{j=1}^{G} N_j(\bar{Y}_{dj} - \bar{\bar{Y}}_d)^2}_{\text{between-cluster variation: } \sigma^2_B}$$

for $d = 0, 1$, where $\bar{Y}_{dj}$ = mean Y_{dij} in cluster j and $\bar{\bar{Y}}_d$ = global mean Y_{dij}

Then we can define the intracluster correlation: $\rho = \frac{\sigma^2_B}{\sigma^2} = 1 - \frac{\sigma^2_W}{\sigma^2}$

Interpretation: When $\rho = \left\{ \begin{array}{c} 1 \\ 0 \end{array} \right\}$, responses are $\left\{ \begin{array}{c} \end{array} \right\}$ within each j

Intracluster Correlation

Recall the Law of Total Variance:

$$\underbrace{\mathbb{V}(Y)}_{\text{total variance}} = \underbrace{\mathbb{E}[\mathbb{V}(Y \mid X)]}_{\text{(mean of) ``within'' variance}} + \underbrace{\mathbb{V}(\mathbb{E}[Y \mid X])}_{\text{``between'' variance}}$$

This implies the decomposition of heterogeneity in potential outcomes:

$$\underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{\bar{Y}}_d)^2}_{\text{overall variation: } \sigma^2} = \underbrace{\sum_{j=1}^{G}\sum_{i=1}^{N_j}(Y_{dij} - \bar{Y}_{dj})^2}_{\text{within-cluster variation: } \sigma_W^2} + \underbrace{\sum_{j=1}^{G} N_j(\bar{Y}_{dj} - \bar{\bar{Y}}_d)^2}_{\text{between-cluster variation: } \sigma_B^2}$$

for $d = 0, 1$, where $\bar{Y}_{dj}$ = mean Y_{dij} in cluster j and $\bar{\bar{Y}}_d$ = global mean Y_{dij}

Then we can define the intracluster correlation: $\rho = \frac{\sigma_B^2}{\sigma^2} = 1 - \frac{\sigma_W^2}{\sigma^2}$

Interpretation: When $\rho = \left\{ \begin{array}{c} 1 \\ 0 \end{array} \right\}$, responses are $\left\{ \begin{array}{c} \text{identical} \\ \text{uncorrelated} \end{array} \right\}$ within each j

Inference in Cluster-Randomized Experiments

- We can show cluster randomization *inflates* the sampling variance (compared to complete randomization) approximately by:

$$\frac{\mathbb{V}(\hat{\tau}_{clustered})}{\mathbb{V}(\hat{\tau}_{complete})} \simeq 1 + (\bar{N} - 1)\rho, \quad \text{where} \quad \bar{N} = \frac{1}{G}\sum_{j=1}^{G} N_j$$

  This is known as the Moulton factor, or the design effect.

Inference in Cluster-Randomized Experiments

- We can show cluster randomization *inflates* the sampling variance (compared to complete randomization) approximately by:

$$\frac{\mathbb{V}(\hat{\tau}_{clustered})}{\mathbb{V}(\hat{\tau}_{complete})} \simeq 1 + (\bar{N} - 1)\rho, \quad \text{where} \quad \bar{N} = \frac{1}{G}\sum_{j=1}^{G} N_j$$

  This is known as the Moulton factor, or the design effect.
  $\longrightarrow$ Calculate the usual SE and multiply by the sqrt of this

Inference in Cluster-Randomized Experiments

- We can show cluster randomization *inflates* the sampling variance (compared to complete randomization) approximately by:

$$\frac{\mathbb{V}(\hat{\tau}_{clustered})}{\mathbb{V}(\hat{\tau}_{complete})} \simeq 1 + (\bar{N} - 1)\rho, \quad \text{where} \quad \bar{N} = \frac{1}{G}\sum_{j=1}^{G} N_j$$

  This is known as the Moulton factor, or the design effect.
  $\longrightarrow$ Calculate the usual SE and multiply by the sqrt of this

- Like in complete randomization, valid inference can be done via regression:
  - Simple regression (= difference in means) estimator is unbiased for τ if clusters are equally sized
  - Possible bias if cluster sizes vary and correlated with potential outcomes
  - Valid (conservative) standard error estimates can be obtained via cluster-robust standard errors

Inference in Cluster-Randomized Experiments

- We can show cluster randomization *inflates* the sampling variance (compared to complete randomization) approximately by:

$$\frac{\mathbb{V}(\hat{\tau}_{clustered})}{\mathbb{V}(\hat{\tau}_{complete})} \simeq 1 + (\bar{N} - 1)\rho, \quad \text{where} \quad \bar{N} = \frac{1}{G}\sum_{j=1}^{G} N_j$$

  This is known as the Moulton factor, or the design effect.
  $\longrightarrow$ Calculate the usual SE and multiply by the sqrt of this

- Like in complete randomization, valid inference can be done via regression:
  - Simple regression (= difference in means) estimator is unbiased for τ if clusters are equally sized
  - Possible bias if cluster sizes vary and correlated with potential outcomes
  - Valid (conservative) standard error estimates can be obtained via cluster-robust standard errors

- When G is small, ρ will be poorly estimated and cluster SEs will be unreliable
- When given choice, increase # of clusters, instead of sample size per cluster

Example: Field Experiment in Benin

Wantchekon (2003)

Example: Field Experiment in Benin

TABLE 1
DESCRIPTION OF THE EXPERIMENTAL DISTRICTS

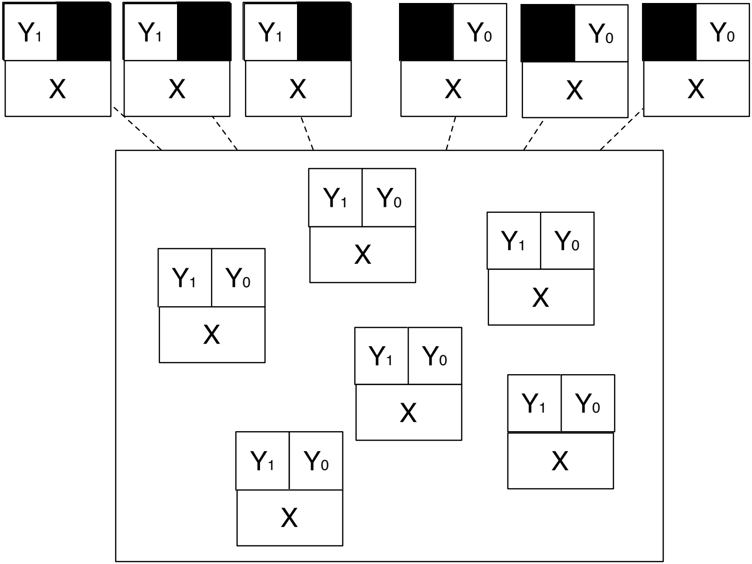
District	*Exp. Candidate*	*Exp. Villages*	*Treatment*	*Ethnicity*
Kandi	Kerekou	Kassakou	clientelism	Bariba (92%)
		Keferi	public policy	Bariba (90%)
Nikki	Kerekou	Ouenou	clientelism	Bariba (89%)
		Kpawolou	public policy	Bariba (88%)
Bembereke	Saka Lafia	Bembereke Est	clientelism	Bariba (86%)
		Wannarou	public policy	Bariba (88%)
Perere	Saka Lafia	Tisserou	clientelism	Bariba (93%)
		Alafiarou	public policy	Bariba (94%)
Abomey-Bohicon	Soglo	Agnangnan	clientelism	Fon (99%)
		Gnidjazoun	public policy	Fon (99%)
Ouidah-Pahou	Soglo	Acadjame	clientelism	Fon (99%)
		Ahozon	public policy	Fon (99%)
Aplahoue	Amoussou	Boloume	clientelism	Adja (99%)
		Avetuime	public policy	Adja (96%)
Dogbo-Toviklin	Amoussou	Dékandji	clientelism	Adja (99%)
		Avedjin	public policy	Adja (99%)
Parakou	Ker./Lafia	Guema	competition	Bariba (80%)
		Thiam	competition	Bariba (82%)
Come	Am./Soglo	Kande	competition	Adja (90%)
		Tokan	competition	Adja (95%)

Example: Field Experiment in Benin

DIFFERENCE IN MEANS BETWEEN TREATMENT AND CONTROL VILLAGES FOR EACH TYPE OF CANDIDATE[a]

Type of Candidate[b]	*Public*	*Clientelist*	*Control*	*Public–Control*	*Clientelist–Control*
Northern	.322 (.032)	.674 (.032)	.565 (.035)	−.243 (.048)***	.109 (.047)**
	208	218	200		
Southern	.840 (.025)	.890 (.021)	.741 (.029)	.099 (.039)***	.149 (.036)***
	219	228	224		
Incumbent	.693 (.032)	.897 (0.21)	.835 (.027)	−.141 (.042)***	.062 (.033)*
	202	214	194		
Opposition	.493 (.033)	.681 (.033)	.509 (.031)	−.015 (.047)	.172 (.045)***
	225	232	230		
Local	.385 (.032)	.603 (.033)	.509 (.033)	−.124 (.046)***	.094 (.047)**
	226	224	230		
National	.816 (.027)	.968 (.012)	.835 (.027)	−.019 (.038)	.133 (.028)***
	201	222	194		

Revisiting Covariates in Experiments



Block Randomization

- We discussed pros/cons for covariate adjustment *after* randomization

Block Randomization

- We discussed pros/cons for covariate adjustment *after* randomization
- A better approach: Adjustment *before* randomization
- Basic idea: If you have data on pre-treatment characteristics X_i, why leave it to "pure chance" to balance them?

Block Randomization

- We discussed pros/cons for covariate adjustment *after* randomization
- A better approach: Adjustment *before* randomization
- Basic idea: If you have data on pre-treatment characteristics X_i, why leave it to "pure chance" to balance them?
- Example: $n = 4$ with two males and two females
  - Complete randomization will place two females in the same treatment group 1/3 of the time
  - If that happens, how can we tell the treatment effect from gender difference?

Block Randomization

- We discussed pros/cons for covariate adjustment *after* randomization
- A better approach: Adjustment *before* randomization
- Basic idea: If you have data on pre-treatment characteristics X_i, why leave it to "pure chance" to balance them?
- Example: $n = 4$ with two males and two females
  - Complete randomization will place two females in the same treatment group $1/3$ of the time
  - If that happens, how can we tell the treatment effect from gender difference?
- Pre-stratify the sample, and then randomize completely within each stratum

Block Randomization

- We discussed pros/cons for covariate adjustment *after* randomization
- A better approach: Adjustment *before* randomization
- Basic idea: If you have data on pre-treatment characteristics X_i, why leave it to "pure chance" to balance them?
- Example: $n = 4$ with two males and two females
  - Complete randomization will place two females in the same treatment group 1/3 of the time
  - If that happens, how can we tell the treatment effect from gender difference?
- Pre-stratify the sample, and then randomize completely within each stratum
- Blocking will perfectly balance X_i
- Randomization will balance the rest in expectation

Block Randomization

- We discussed pros/cons for covariate adjustment *after* randomization
- A better approach: Adjustment *before* randomization
- Basic idea: If you have data on pre-treatment characteristics X_i, why leave it to "pure chance" to balance them?
- Example: $n = 4$ with two males and two females
  - Complete randomization will place two females in the same treatment group 1/3 of the time
  - If that happens, how can we tell the treatment effect from gender difference?
- Pre-stratify the sample, and then randomize completely within each stratum
- Blocking will perfectly balance X_i
- Randomization will balance the rest in expectation

"Block what you can; randomize what you cannot." (Box)

Estimation in Block-Randomized Experiments

Setup:

- Units: $i = 1, ..., N$; Blocks: $j = 1, ..., M$
- N_j: $\#$ of units in block j
- p_j: Treatment probability in block j ($= \frac{N_{1j}}{N_j}$ for within-block complete randomization)

Estimation in Block-Randomized Experiments

Setup:

- Units: $i = 1, ..., N$; Blocks: $j = 1, ..., M$
- N_j: # of units in block j
- p_j: Treatment probability in block j ($= \frac{N_{1j}}{N_j}$ for within-block complete randomization)

If probability of treatment is identical in each j (i.e. $p_j = p$ for all j), then the pooled difference in means is unbiased for the ATE:

$$\hat{\tau} \equiv \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) Y_i$$

Estimation in Block-Randomized Experiments

Setup:

- Units: $i = 1, \ldots, N$; Blocks: $j = 1, \ldots, M$
- N_j: # of units in block j
- p_j: Treatment probability in block j ($= \frac{N_{1j}}{N_j}$ for within-block complete randomization)

If probability of treatment is identical in each j (i.e. $p_j = p$ for all j), then the pooled difference in means is unbiased for the ATE:

$$\hat{\tau} \equiv \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) Y_i$$

If p_j varies across blocks, *weighted average of block-specific difference-in-means* will be unbiased:

$$\hat{\tau}_B \equiv \sum_{j=1}^{M} \frac{N_j}{N} \hat{\tau}_j \quad \text{where} \quad \hat{\tau}_j = \frac{1}{N_{1j}} \sum_{i=1}^{N_j} D_{ij} Y_{ij} - \frac{1}{N_{0j}} \sum_{i=1}^{N_j} (1 - D_{ij}) Y_{ij}$$

Estimation in Block-Randomized Experiments

Setup:

- Units: $i = 1, ..., N$; Blocks: $j = 1, ..., M$
- N_j: # of units in block j
- p_j: Treatment probability in block j ($= \frac{N_{1j}}{N_j}$ for within-block complete randomization)

If probability of treatment is identical in each j (i.e. $p_j = p$ for all j), then the pooled difference in means is unbiased for the ATE:

$$\hat{\tau} \;\equiv\; \frac{1}{N_1}\sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - D_i)Y_i$$

If p_j varies across blocks, *weighted average of block-specific difference-in-means* will be unbiased:

$$\hat{\tau}_B \;\equiv\; \sum_{j=1}^{M} \frac{N_j}{N}\hat{\tau}_j \quad \text{where} \quad \hat{\tau}_j = \frac{1}{N_{1j}}\sum_{i=1}^{N_j} D_{ij} Y_{ij} - \frac{1}{N_{0j}}\sum_{i=1}^{N_j}(1 - D_{ij})Y_{ij}$$

where $N_{1j} = \#$ of treated units in group j and $N_{0j} = N_j - N_{1j}$

Why would $\hat{\tau}$ be biased?

Estimation in Block-Randomized Experiments

Setup:

- Units: $i = 1, ..., N$; Blocks: $j = 1, ..., M$
- N_j: # of units in block j
- p_j: Treatment probability in block j ($= \frac{N_{1j}}{N_j}$ for within-block complete randomization)

If probability of treatment is identical in each j (i.e. $p_j = p$ for all j), then the pooled difference in means is unbiased for the ATE:

$$\hat{\tau} \;\equiv\; \frac{1}{N_1}\sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - D_i) Y_i$$

If p_j varies across blocks, *weighted average of block-specific difference-in-means* will be unbiased:

$$\hat{\tau}_B \;\equiv\; \sum_{j=1}^{M} \frac{N_j}{N} \hat{\tau}_j \quad \text{where} \quad \hat{\tau}_j = \frac{1}{N_{1j}}\sum_{i=1}^{N_j} D_{ij} Y_{ij} - \frac{1}{N_{0j}}\sum_{i=1}^{N_j}(1 - D_{ij}) Y_{ij}$$

where $N_{1j} =$ # of treated units in group j and $N_{0j} = N_j - N_{1j}$

Why would $\hat{\tau}$ be biased? Because D_{ij} is not independent of blocks when p_j varies.

Inference for Block-Randomized Experiments

Because the randomizations in each block are independent, the sampling variance of the weighted-average estimator is simply:

$$\mathbb{V}(\hat{\tau}_B) = \sum_{j=1}^{M} \left(\frac{N_j}{N}\right)^2 \mathbb{V}(\hat{\tau}_j)$$

and the component variance can be estimated (conservatively) via the Neyman formula for each block:

$$\widehat{\mathbb{V}(\hat{\tau}_j)} \;=\; \frac{S_{1j}^2}{N_{1j}} + \frac{S_{0j}^2}{N_{0j}}$$

Inference for Block-Randomized Experiments

Because the randomizations in each block are independent, the sampling variance of the weighted-average estimator is simply:

$$\mathbb{V}(\hat{\tau}_B) = \sum_{j=1}^{M} \left(\frac{N_j}{N} \right)^2 \mathbb{V}(\hat{\tau}_j)$$

and the component variance can be estimated (conservatively) via the Neyman formula for each block:

$$\widehat{\mathbb{V}(\hat{\tau}_j)} \;=\; \frac{S_{1j}^2}{N_{1j}} + \frac{S_{0j}^2}{N_{0j}}$$

Note: If the within-block randomization occurred at the cluster level, use the cluster-robust variance estimates for the component variances

Regression for Block-Randomized Experiments

Like in the classical experiment, one can use linear regression to obtain unbiased estimates in block-randomized experiments.

Regression for Block-Randomized Experiments

Like in the classical experiment, one can use linear regression to obtain unbiased estimates in block-randomized experiments.

- If $p_j = p$, *OLS with block dummies* (or "fixed effects", B_{ij}) will yield unbiased estimate for ATE:

$$Y_i = \alpha + \tau D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i, \quad \text{where} \quad \mathbb{E}[\hat{\tau}_{OLS}] = \tau$$

  Valid uncertainty estimates can then be obtained via the HC2 robust SE (or clustered SE if randomization was clustered within blocks)

Regression for Block-Randomized Experiments

Like in the classical experiment, one can use linear regression to obtain unbiased estimates in block-randomized experiments.

- If $p_j = p$, *OLS with block dummies* (or "fixed effects", B_{ij}) will yield unbiased estimate for ATE:

$$Y_i = \alpha + \tau D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i, \quad \text{where} \quad \mathbb{E}[\hat{\tau}_{OLS}] = \tau$$

  Valid uncertainty estimates can then be obtained via the HC2 robust SE (or clustered SE if randomization was clustered within blocks)

- If p_j varies by block, use *weighted least squares* instead of OLS, where the weight is inverse probability of treatment/control for each unit:

$$w_{ij} \;=\; \left\{ \begin{array}{lll} 1/p_j & \text{if} & D_i = 1 \\ 1/(1-p_j) & \text{if} & D_i = 0 \end{array} \right\} \text{for } i \text{ in block } j$$

Why Does Blocking (Usually) Help?

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \qquad \text{(complete randomization)}$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i^* \qquad \text{(block randomization)}$$

Assuming homoskedasticity for simplicity, we have:

$$V[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \qquad \text{with } \widehat{\sigma}_\varepsilon^2 =$$

Why Does Blocking (Usually) Help?

$$
\begin{aligned}
Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i && \text{(complete randomization)} \\
Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i^* && \text{(block randomization)}
\end{aligned}
$$

Assuming homoskedasticity for simplicity, we have:

$$
\begin{aligned}
V[\widehat{\tau}_{CR}] &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-2} = \frac{SSR_{\widehat{\varepsilon}}}{N-2} \\
V[\widehat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \text{ with } \widehat{\sigma_{\varepsilon^*}}^2 =
\end{aligned}
$$

Why Does Blocking (Usually) Help?

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \qquad \text{(complete randomization)}$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i^* \qquad \text{(block randomization)}$$

Assuming homoskedasticity for simplicity, we have:

$$V[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-2} = \frac{SSR_{\widehat{\varepsilon}}}{N-2}$$

$$V[\widehat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma_{\varepsilon^*}}^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon^*}_i^2}{N-J-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{N-M-1}$$

where R_j^2 is

Why Does Blocking (Usually) Help?

$$
\begin{aligned}
Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i && \text{(complete randomization)} \\
Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i^* && \text{(block randomization)}
\end{aligned}
$$

Assuming homoskedasticity for simplicity, we have:

$$
\begin{aligned}
V[\widehat{\tau}_{CR}] &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} && \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-2} = \frac{SSR_{\widehat{\varepsilon}}}{N-2} \\
V[\widehat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} && \text{with } \widehat{\sigma_{\varepsilon^*}}^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon^*}_i^2}{N-J-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{N-M-1}
\end{aligned}
$$

where R_j^2 is R^2 from regression of D_i on all B_{ij} variables and a constant.

So $\widehat{V}[\widehat{\tau}_{BR}] < \widehat{V}[\widehat{\tau}_{CR}]$ if

Why Does Blocking (Usually) Help?

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \qquad \text{(complete randomization)}$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{M} \beta_j B_{ij} + \varepsilon_i^* \qquad \text{(block randomization)}$$

Assuming homoskedasticity for simplicity, we have:

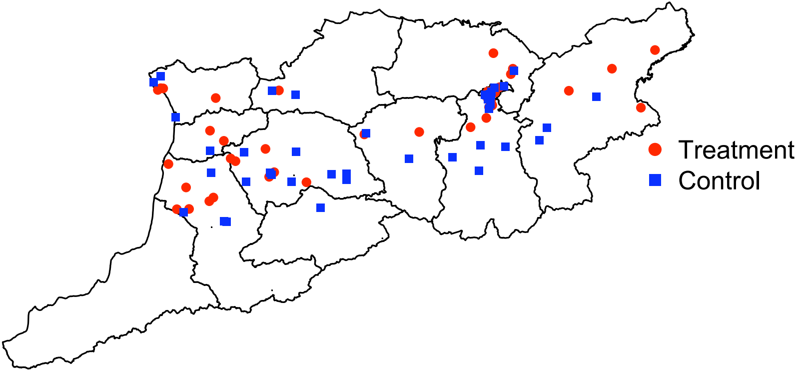
$$V[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-2} = \frac{SSR_{\widehat{\varepsilon}}}{N-2}$$

$$V[\widehat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma_{\varepsilon^*}}^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon^*}_i^2}{N-J-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{N-M-1}$$

where R_j^2 is R^2 from regression of D_i on all B_{ij} variables and a constant.

So $\widehat{V}[\widehat{\tau}_{BR}] < \widehat{V}[\widehat{\tau}_{CR}]$ if $\frac{SSR_{\widehat{\varepsilon^*}}}{N-M-1} < \frac{SSR_{\widehat{\varepsilon}}}{N-2}$

Example: Anti-Vote Fraud Experiment in Georgia



Driscoll and Hidalgo (2013)

Example: Anti-Vote Fraud Experiment in Georgia

Without block fixed effects:

R Code

```
Call:
lm(formula = total.complaints ~ tr.complaints, data = exp.data)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.423e-17  9.657e-02   0.000   1.0000
tr.complaints  2.619e-01  1.366e-01   1.918   0.0586 .
```

With block fixed effects:

R Code

```
Call:
lm(formula = total.complaints ~ tr.complaints + block, data = exp.data)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.3690     0.4089   5.794 3.84e-07 ***
tr.complaints   0.2619     0.1247   2.100 0.040492 *
block10u       -2.5000     0.4949  -5.051 5.54e-06 ***
block11r       -2.5000     0.5715  -4.375 5.73e-05 ***
block11u       -2.0000     0.4949  -4.041 0.000173 ***
block12r       -2.5000     0.5715  -4.375 5.73e-05 ***
```

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{0i}] \quad \text{vs.} \quad H_1 : \mathbb{E}[Y_{1i}] \neq \mathbb{E}[Y_{0i}]$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{0i}] \quad \text{vs.} \quad H_1 : \mathbb{E}[Y_{1i}] \neq \mathbb{E}[Y_{0i}]$$

- Fisher's exact test with small N:

$$H_0 : Y_{1i} = Y_{0i} \quad \text{vs.} \quad H_1 : Y_{1i} \neq Y_{0i}$$

  This H_0 is called the sharp null hypothesis

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{0i}] \quad \text{vs.} \quad H_1 : \mathbb{E}[Y_{1i}] \neq \mathbb{E}[Y_{0i}]$$

- Fisher's exact test with small N:

$$H_0 : Y_{1i} = Y_{0i} \quad \text{vs.} \quad H_1 : Y_{1i} \neq Y_{0i}$$

  This H_0 is called the sharp null hypothesis

- Key idea: Under the sharp null, we "observe" all potential outcomes!

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{0i}] \quad \text{vs.} \quad H_1 : \mathbb{E}[Y_{1i}] \neq \mathbb{E}[Y_{0i}]$$

- Fisher's exact test with small N:

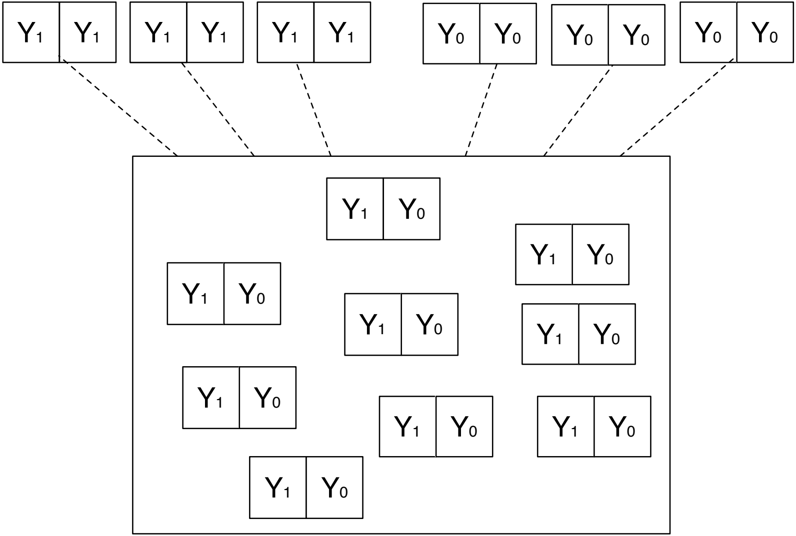
$$H_0 : Y_{1i} = Y_{0i} \quad \text{vs.} \quad H_1 : Y_{1i} \neq Y_{0i}$$

  This H_0 is called the sharp null hypothesis

- Key idea: Under the sharp null, we "observe" all potential outcomes!

- Let Ω be the set of all possible ways to assign treatments
- Fisher's exact test procedure:
  1. Calculate a statistic $\hat{\alpha}$ (e.g. difference in means) from the data
  2. Obtain the null distribution of the statistic by calculating the same statistic $\hat{\alpha}(\omega)$ under the sharp null for every possible $\omega \in \Omega$
  3. Compare $\hat{\alpha}$ to the null distribution and see how "extreme" it is

Potential Outcomes Under the Sharp Null



Example: Lady Tasting Tea

Suppose that we assign 4 individuals out of 8 to the treatment:

									$\hat{\alpha}$
Y_i	12	4	6	10	6	0	1	1	
Actual D_i	1	1	1	1	0	0	0	0	6

Example: Lady Tasting Tea

Suppose that we assign 4 individuals out of 8 to the treatment:

									$\hat{\alpha}$
Y_i	12	4	6	10	6	0	1	1	
Actual D_i	1	1	1	1	0	0	0	0	6
									$\hat{\alpha}(\omega)$
$\omega = 1$	1	1	1	1	0	0	0	0	6
$\omega = 2$	1	1	1	0	1	0	0	0	4

Example: Lady Tasting Tea

Suppose that we assign 4 individuals out of 8 to the treatment:

									$\hat{\alpha}$
Y_i	12	4	6	10	6	0	1	1	
Actual D_i	1	1	1	1	0	0	0	0	6
									$\hat{\alpha}(\omega)$
$\omega = 1$	1	1	1	1	0	0	0	0	6
$\omega = 2$	1	1	1	0	1	0	0	0	4
$\omega = 3$	1	1	1	0	0	1	0	0	1

Example: Lady Tasting Tea

Suppose that we assign 4 individuals out of 8 to the treatment:

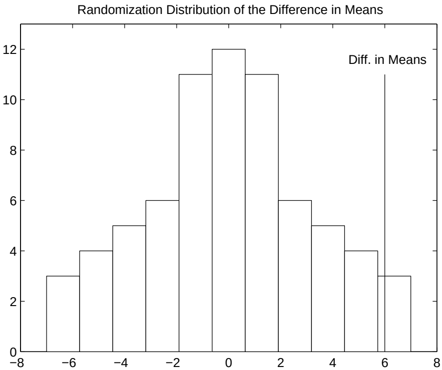
									$\hat{\alpha}$
Y_i	12	4	6	10	6	0	1	1	
Actual D_i	1	1	1	1	0	0	0	0	6
									$\hat{\alpha}(\omega)$
$\omega = 1$	1	1	1	1	0	0	0	0	6
$\omega = 2$	1	1	1	0	1	0	0	0	4
$\omega = 3$	1	1	1	0	0	1	0	0	1
$\omega = 4$	1	1	1	0	0	0	1	0	1.5
			$\vdots$						$\vdots$
$\omega = 70$	0	0	0	0	1	1	1	1	-6

- Calculate the exact p-value such that

$$p \equiv \Pr\left(|\hat{\alpha}(\omega)| \geq |\hat{\alpha}|\right)$$

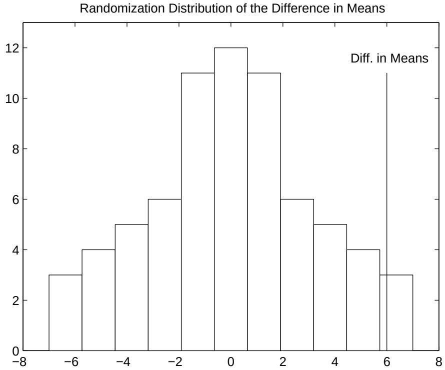
- Reject the null hypothesis if $p \leq 0.05$, for example

Testing in Small Samples: Fisher's Exact Test



$$p = \Pr(|\hat{\alpha}(\omega)| \geq 6) = 0.0857$$

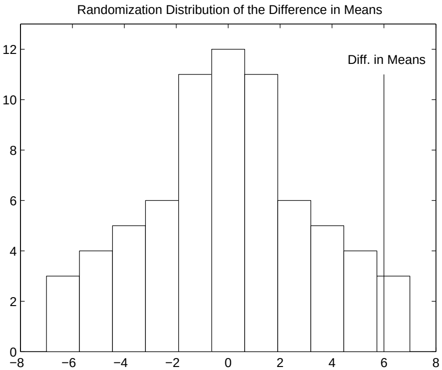
Testing in Small Samples: Fisher's Exact Test



$$p \;=\; \Pr(|\hat{\alpha}(\omega)| \geq 6) \;=\; 0.0857$$

Which assumptions are needed?

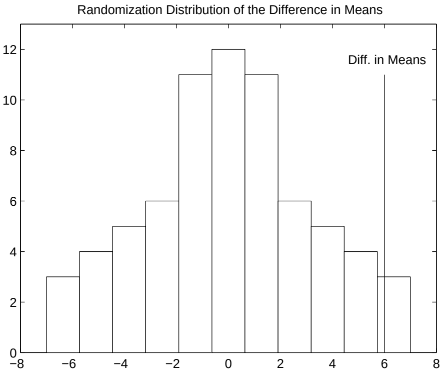
Testing in Small Samples: Fisher's Exact Test



$$p = \Pr(|\hat{\alpha}(\omega)| \geq 6) = 0.0857$$

Which assumptions are needed? None! Randomization as "reasoned basis for causal inference" (Fisher 1935)

Testing in Small Samples: Fisher's Exact Test



$$p \;=\; \Pr(|\hat{\alpha}(\omega)| \geq 6) \;=\; 0.0857$$

Which assumptions are needed? None! Randomization as "reasoned basis for causal inference" (Fisher 1935)

Drawback: The sharp null is often uninteresting (how often is a causal effect exactly zero for every single unit?)

Summary

- Random assignment solves the identification problem for causal inference based on minimal assumptions that researchers can control
- Random assignment balances observed and unobserved confounders, which is why it is considered the gold standard for causal inference
- Regression is a useful tool for analyzing experiments; simple regression with robust SE yields valid estimates
- Covariate adjustment via regression can improve efficiency, and estimates are often robust to alternative model specifications
- Possible tradeoff between internal validity and external validity
- Clustered randomization increases statistical uncertainty, which needs to be incorporated in reported results
- Block randomization can reduce statistical uncertainty; block what you can!
- Fisherian randomization inference focuses on a sharp null to deal with small sample size