Lecture 2: Experimental Methods

Randomization Inference

Raymond Duch

May 7, 2020

Director CESS Nuffield/Santiago

Lecture 2 Road Map

- How we can think about statistical uncertainty
- Hypothesis testing
- Confidence intervals

Statistical uncertainty

Observed Outcomes Local Budget

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30

Potential Outcomes Local Budget

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1	10	15	5
Village 2	15	15	0
Village 3	20	30	10
Village 4	20	15	-5
Village 5	10	20	10
Village 6	15	15	0
Village 7	15	30	15
Average	15	30	15

2 ways of thinking about statistical uncertainty

- Sampling-based inference (Neyman):
  - Experimental subjects are a random draw from some "super-population"
  - Realized ATE has a sampling distribution with reference to that superpopulation
  - Different (random) experimental samples $\rightarrow$ different ATE's from draw to draw
  - Uncertainty arises from random sampling of subjects: How are ATE's distributed in the population?
  - Sampling distribution under the H0 is typically $X \sim \mathcal{N}(\mu, \sigma^2)$

Neyman's plan for inference

1. Define the estimand
2. Find unbiased estimator of the estimand
3. Calculate true sampling variance of the estimator
4. Find unbiased estimator of true sampling variance of estimator
5. Assume approximate normality to obtain p-value and confidence interval
6. Where $H0 : E[Y_i(1)] - E[Y_i(0)] = 0$

2 ways of thinking about statistical uncertainty

- Randomization-based inference (Fisher):
  - Treatment assignments are a random draw from the set of all possible assignment combinations $\rightarrow$ finite sample
  - Realized ATE has a distribution over those possible random assignments
  - Different ways of assigning subjects to treatment $\rightarrow$ different ATE's from allocation to allocation
  - Uncertainty arises from random assignment and missing potential outcomes

This has implications for other methods: use the ones that are *directly justified* by randomization $\rightarrow$ design instead of analysis for covariate adjustment; diff-in-group means estimator; reduce reliance on auxiliary modelling assumptions

The goal of randomization inference is to derive a *sampling distribution* of estimated ATEs. In our application, generated when two of the seven villages listed in Table 2 are assigned to treatment

	Estimated ATE	Frequency with which an estimate occurs
	-1	2
	0	2
	0.5	1
	1	2
	1.5	2
	2.5	1
	6.5	1
	7.5	3
	8.5	3
	9	1
	9.5	1
	10	1
	16	1
Average	5	
Total		21

We can calculate the variation of these estimates:

Sum of squared deviations

$$
\begin{aligned}
&= (-1-5)^2 + (-1-5)^2 + (0-5)^2 + (0-5)^2 + (0.5-5)^2 + (1-5)^2 + (1-5)^2 \\
&+ (1.5-5)^2 + (1.5-5)^2 + (2.5-5)^2 + (6.5-5)^2 + (7.5-5)^2 + (7.5-5)^2 \\
&+ (7.5-5)^2 + (8.5-5)^2 + (8.5-5)^2 + (8.5-5)^2 + (9-5)^2 + (9.5-5)^2 \\
&+ (10-5)^2 + (16-5)^2 = 445
\end{aligned}
$$

$$
\textit{Square root of the average squared deviation} = \sqrt{\frac{1}{21}(445)} = 4.60
$$

Neyman variance estimator

Neyman quantifies the variance of our difference-in-means estimator with the Neyman variance estimator. Formally,

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{N-1}\left\{\frac{m\text{Var}(Y_i(0))}{N-m} + \frac{(N-m)\text{Var}(Y_i(1))}{m} + 2\text{Cov}(Y_i(0), Y_i(1)\right\}}$$

In our application,

$$SE(\widehat{ATE}) = \sqrt{\frac{1}{6}\left\{\frac{(2)(14.29)}{5} + \frac{(5)(42.86)}{2} + (2)(7.14)\right\}} = 4.60$$

You can see that the covariance of the two potential outcomes is fundamentally unobservable, so we assume constant treatment effects, and the sample analog reduces to

$$\widehat{SE} = \sqrt{\frac{\widehat{Var}(Y_i(0))}{N-m} + \frac{\widehat{Var}(Y_1(1))}{m}}$$

Formal Randomization Inference

- Now, randomization inference is different. We only ever observe one particular realization of the randomized treatment assignment
- Yet, given m, N and a binary treatment, there is a set of all possible randomization realizations such that $\Omega = \frac{N!}{m!(N-m)!}$
- For the Abadie and Cattaneo (2018) example, we have $\Omega = \frac{8!}{4!(8-4)!} = 70$, and we are interested in the distribution of $\hat{\tau}(\omega)$, i.e. for each possible realization of the randomized assignment $\omega \in \Omega$, as in the following table

Potential Outcomes Local Budget

2 of 21 possible worlds:

World 1:

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1		15	
Village 2		15	
Village 3	20		
Village 4	20		
Village 5	10		
Village 6	15		
Village 7	15		
Average	16	15	-1

Potential Outcomes Local Budget

World 2:

	Budget share if village head is male	Budget share if village head is female	Treatment Effect
Village 1	10		
Village 2	15		
Village 3	20		
Village 4	20		
Village 5	10		
Village 6		15	
Village 7		30	
Average	15	22.5	7.5

Abadie and Cattaneo 2018

Table 1 Randomization distribution of a difference in means

Panel A: Sample and sample statistic									
Y_i	12	4	6	10	6	0	1	1	
W_i	1	1	1	1	0	0	0	0	$\widehat{\tau} = 6$
Panel B: Randomization distribution									$\widehat{\tau}(\omega)$
$\omega = 1$	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
$...\omega = 70$	0	0	0	0	1	1	1	1	−6

Hypothesis testing

Hypothesis Testing

- We can test certain conjectures that provide us a complete schedule of potential outcomes
- One such conjecture is the *sharp null hypothesis* that the treatment effect is zero for all observations
- Therefore, $H0 : Y_i(1) = Y_i(0)$, i.e. potential outcomes are identical
- Simulated randomizations provide an exact sampling distribution of the estimated average treatment effect under the sharp null hypothesis
- Then, we are interested in $Pr(\hat{\tau}(\omega) \geq \hat{\tau})$ or $Pr(|\hat{\tau}(\omega)| \geq |\hat{\tau}|)$

Abadie and Cattaneo 2018

- The randomization distribution is *exact* since it is computed without error (all missing potential outcomes are imputed)
- We then derive

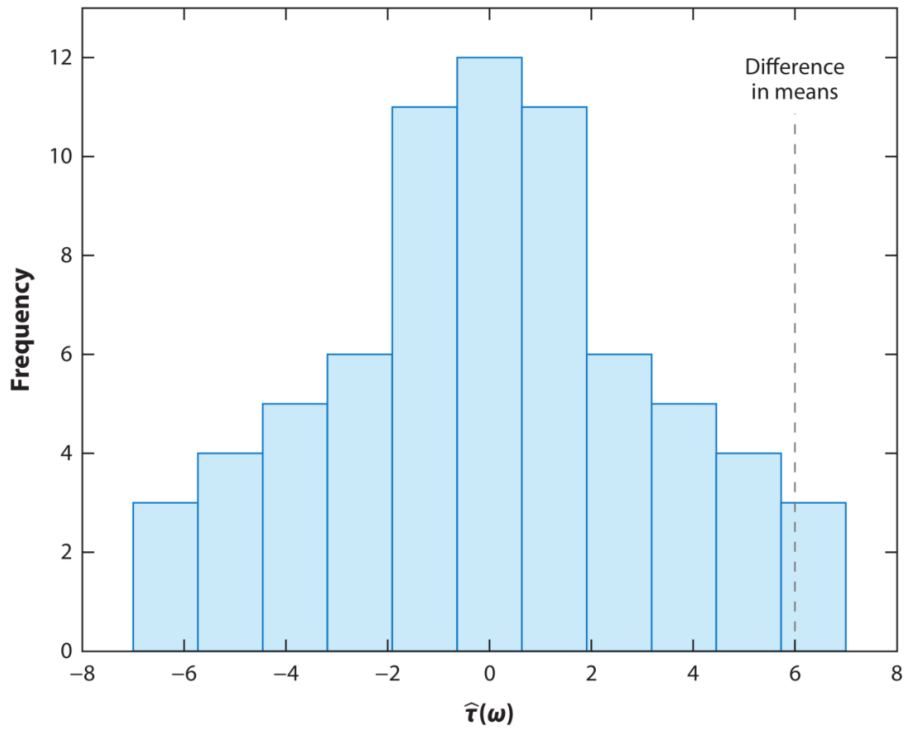


Figure 2

Randomization distribution of the difference in means. The vertical line represents the sample value of $\hat{\tau}$.

Observed Outcome Local Budget

Where does this *exact* randomization distribution come from in practice?

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	?	30

Example: Randomization

- From this table generate an estimate of the ATE of 6.5
- How likely are we to obtain estimate as large as or larger than 6.5 if the true effect were zero for all observations?
- The probability, or p-value, of interest in this case addresses a *one-tailed hypothesis*, namely that female village council heads increase budget allocations to water sanitation

Observed Outcome Local Budget

	Budget share if village head is male	Budget share if village head is female
Village 1	?	15
Village 2	15	(15)
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	?
Village 7	(30)	30
Average	19	15

Observed Outcome Local Budget

	Budget share if village head is male	Budget share if village head is female
Village 1	(15)	15
Village 2	15	?
Village 3	20	?
Village 4	20	?
Village 5	10	?
Village 6	15	(15)
Village 7	(30)	30
Average	16	22.5

Example: Randomization

- Based on the observed outcome in the table, we may calculate the 21 possible estimates of the ATE that could have been generated if the null hypothesis were true: {-7.5, -7.5, -7.5, -4.0, -4.0, -4.0, -4.0, -4.0, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, 3.0, 3.0, 6.5, 6.5, 6.5, 10.0, 10.0}
- How likely are we to obtain an estimate as large as or larger than 6.5 if the true effect were zero for all observations?

Example: 1-tailed test

- The probability, or p-value, of interest in this case addresses a *one-tailed hypothesis*, namely that female village council heads increase budget allocations to water sanitation
- Five of the estimates are as large as 6.5. Therefore, when evaluating the one-tailed hypothesis that female village heads *increase* water sanitation budgets, we would conclude that the probability of obtaining an estimate as large as 6.5 if the null hypothesis were true is $5/21 = 24\%$

Example: 2-tailed test

- If we sought to evaluate the *two-tailed hypothesis* - whether female village council heads either increase or decrease the budget allocation for water sanitation
- We would calculate the p-value of obtaining a number that is greater than or equal to 6.5 or less than or equal to -6.5. A two-tailed hypothesis test would count all instances in which the estimates are at least as great as 6.5 *in absolute value*. Eight of the estimates qualify, so the two-tailed p-value is $8/21 = 38\%$

Lady Testing Tea

- Ronald Fisher, *The Design of Experiments*
- Randomized Tea Experiment:
  - 8 identical cups prepared
  - 4 cups randomly prepared with milk first
  - 4 cups randomly prepared with milk after
- Lady correctly classified all cups!

Lady Tasting Tea

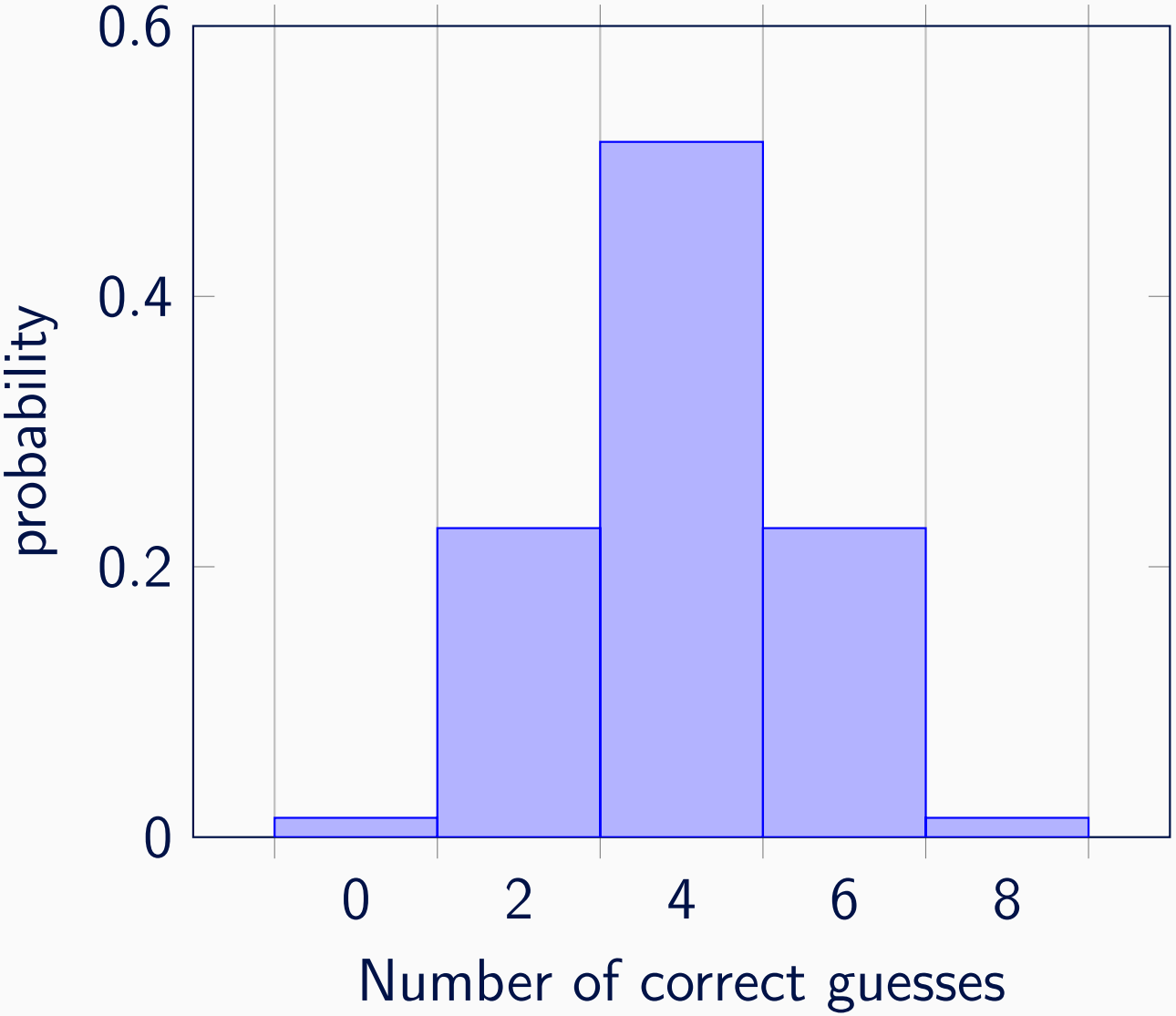
cups	lady's guess	actual order	scenarios			...
1	M	M	T	T	T	
2	T	T	T	T	M	
3	T	T	T	T	M	
4	M	M	T	M	M	
5	M	M	M	M	T	
6	T	T	M	M	T	
7	T	T	M	T	M	
8	M	M	M	M	T	
number of correct guesses		8	4	6	2	...

Lady Tasting Tea

```
> ## truth: enumerate the number of assignment combinations

> true <- c(choose(4, 0) * choose(4, 4), choose(4, 1) *
    choose(4, 3), choose(4, 2) * choose(4, 2), choose(4, 3)
    * choose(4, 1), choose(4, 4) * choose(4, 0))

> ## compute probability: divide it by the total number of
    events

> true <- true/sum(true)

> ## number of correctly classified cups as labels

> names(true) <- c(0,2,4,6,8)

> true
        0         2         4         6         8
0.01428571 0.22857143 0.51428571 0.22857143 0.01428571
```

To get 6 hits, she must label as milk-first three of the four true milk-first cups, and must mislabel as milk-first one of the four tea-first cups

Lady Tasting Tea



Lady Tasting Tea: Simulate

```
> ## Simulations

> sims <- 1000

> guess <- c("M", "T", "T", "M", "M", "T", "T", "M") # lady's guess

> correct <- rep(NA, sims) # place holder for number of correct guesses

> for (i in 1:sims) {
+   cups <- sample(c(rep("T", 4), rep("M", 4)), replace = FALSE)
+   correct[i] <- sum(guess == cups) # number of correct guesses
}

> ## comparison
> prop.table(table(correct)) - true
correct
           0            2            4            6            8
 0.001714286  0.004428571 -0.015285714  0.007428571  0.001714286
```

Fisher Exact Test: Definitions

The total probability of observing data as extreme or more extreme if the null hypothesis is true.

Tests the probability of getting a table that is as strong due to the chance of sampling. The word strong is defined as the proportion of the cases that are diagonal with the most cases.

Fisher Exact Test

```
> ## rows: actual assignments
> ## columns: reported guesses

> ## all correct
> x <- matrix(c(4, 0, 0, 4), byrow = TRUE, ncol = 2, nrow = 2)

> ## six correct
> y <- matrix(c(3, 1, 1, 3), byrow = TRUE, ncol = 2, nrow = 2)

> rownames(x) <- colnames(x) <- rownames(y) <- colnames(y) <- c("M",    T")

>␣x
 ␣␣M␣T
M␣4␣0
T␣0␣4
>␣y
 ␣␣M␣T
M␣3␣1
T␣1␣3
```

Cooperman 2017

Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation in *Political Analysis* 25:277-288.

- How do you estimate the variance of the effect of rainfall on political outcomes?
- 1 year $=$ 1 possible randomized assignment (historical weather patterns from 73 years)
- Compares ATE to sampling distribution of estimates under the exact Null of no effect

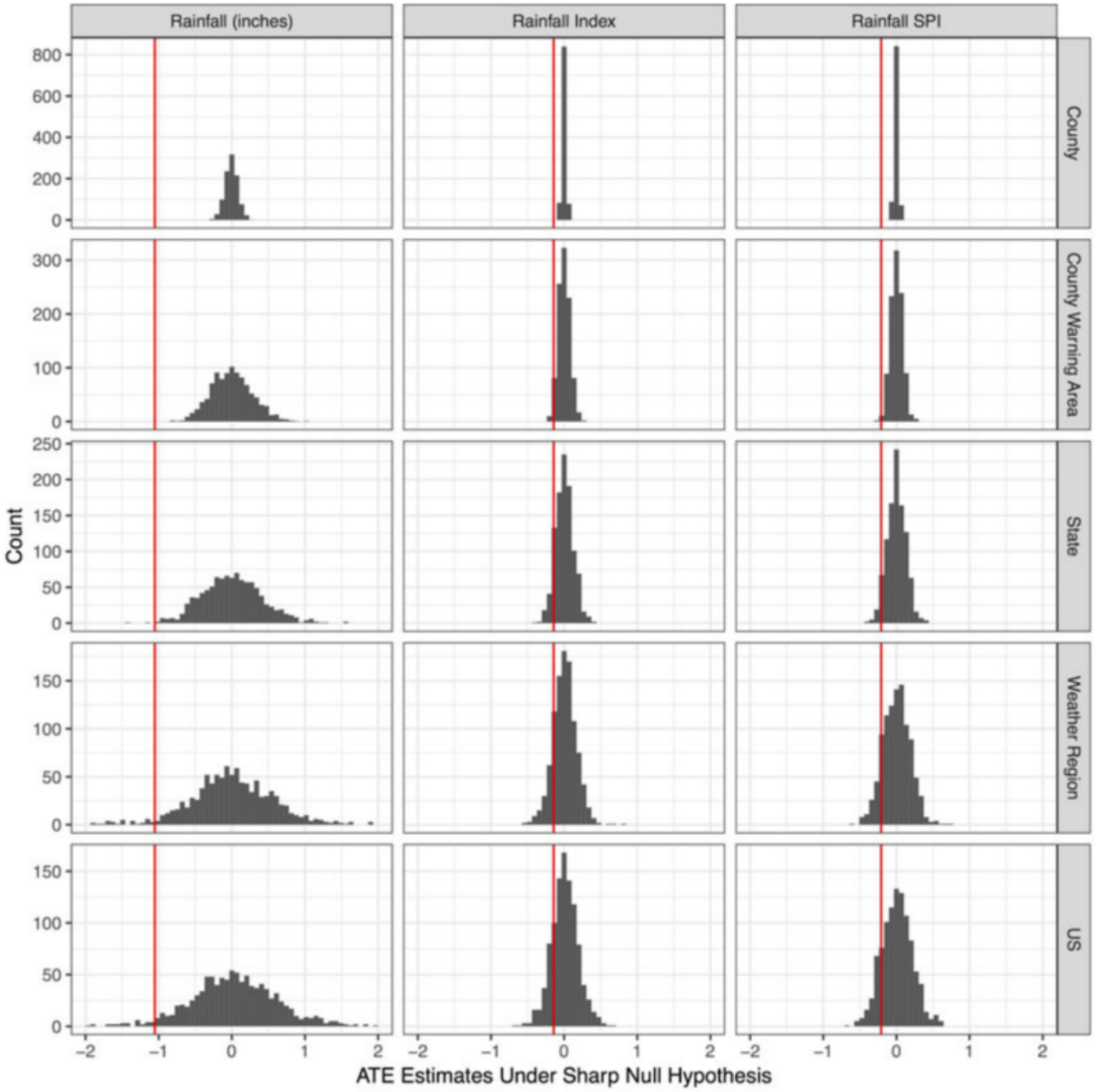


Figure 4. Sampling Distribution of ATE Estimates for Rainfall by Measurement and Cluster Unit. *Note:* Figure represents sampling distributions of estimated ATEs under the sharp null hypothesis of no effect. Potential randomizations are random draws from county-level rainfall data from 1940 to 2012 on election or potential election days. Differences in measurement are reflected in each column. The first column uses rainfall (in), the second column uses a z-score index, and the third column uses a Standardized Precipitation Index. The dotted lines represent the estimated ATEs of −1.052 for rainfall (in), −0.143 for rainfall index, and −0.208 for rainfall SPI from Table 2. Different assumptions about the cluster unit are reflected in each row. County assignment

Commments on RI p-values

- One obtains arbitrarily precise p-values without relying on distributional assumptions
- The same method can be used for a wide variety of applications and test statistics (e.g., the difference-in-means estimator, regression, difference-in-variance, etc.)
- It forces the researcher to take a moment to think carefully about what the null hypothesis is and how it should be tested

Confidence Intervals

Confidence Intervals

- We cannot estimate the dispersion of the estimates without making simplifying assumptions
- A simple approach is to assume that the treatment effect for every subject is equal to the estimated ATE (this is different from the sharp $H0 : ATE = 0$!)
- For subjects in the control condition, missing $Y_i(1)$ values are imputed by adding the estimated ATE to the observed values of $Y_i(0)$
- For subjects in the treatment condition, missing $Y_i(0)$ values are imputed by subtracting the estimated ATE from the observed values of $Y_i(1)$
- Complete schedule of potential outcomes, which we may then use to simulate all possible random allocations

Effect of winning visa lottery on attitudes toward people from other countries

- We cannot estimate the dispersion of the estimates without making simplifying assumptions
- Winners and losers were asked to rate the Saudi, Indonesian, Turkish, African, European, and Chinese people on a five-point scale ranging from very negative (-2) to very positive (+2)
- Adding the responses to all six items creates an index ranging from -12 to +12
- Average in the treatment group is 2.34
- 1.87 in the control group

Pakistani Muslims Lottery

Ratings of people from other countries	Control(%)	Treatment (%)
-12	0	0.2
-9	0.22	0
-8	0	0.2
-6	0.45	0.2
-5	0	0.2
-4	0.45	0.59
-3	0	0.2
-2	1.12	0.98
-1	1.56	2.75
0	27.23	18.63
1	18.3	13.14
2	24.33	25.29
3	8.48	10.98
4	5.8	9.61
5	3.35	3,92
6	3.79	7.25
7	2.23	2.55
8	0.89	1.37
9	0.22	0.78
10	0.45	0
11	0.67	0.2
12	0.45	0.98
TOTAL	100	100
N	(448)	(510)

Estimate our 95% interval

- We add 0.47 to the observed outcomes in the control group in order to approximate their values
- We subtract 0.47 for the treatment group's observed outcomes in order to approximate their values
- Simulating 100,000 random allocations using this schedule of potential outcomes and sorting the estimated ATEs in ascending order
- We find that the 2,500th estimate is 0.16 and the 97,501st estimate is 0.79, so the 95% interval is [0.16, 0.79]

Example from Matthew Blackwell

- Suppose we are targeting 6 people for donations to Harvard.
- As an encouragement, we send 3 of them a mailer with inspirational stories of learning from our graduate students.
- Afterwards, we observe them giving between \$0 and \$5.
- Simple example to show the steps of RI in a concrete case.

Randomization Distribution

	Mailer	Contrib		
Unit	**D**	**Y**	**Y(0)**	**Y(1)**
Donald	1	3	(3)	3
Carly	1	5	(5)	5
Ben	1	0	(0)	0
Ted	0	4	4	(4)
Marco	0	0	0	(0)
Scott	0	1	1	(1)

T(diff)=|8/3 - 5/3| = 1

Randomization Distribution

	Mailer	Contrib		
Unit	D	**Y**	**Y(0)**	**Y(1)**
Donald	1	3	(3)	3
Carly	1	5	(5)	5
Ben	0	0	(0)	0
Ted	1	4	4	(4)
Marco	0	0	0	(0)
Scott	0	1	1	(1)

T(diff)=|12/3 - 1/3| = 3.67

T(diff)=|8/3 - 5/3| = 1

T(diff)=|9/3 - 4/3| = 1.67

Randomization Distribution

D1	D2	D3	D4	D5	D6	| Diff in means |
1	1	1	0	0	0	1.00
1	1	0	1	0	0	3.67
1	1	0	0	1	0	1.00
1	1	0	0	0	1	1.67
1	0	1	1	0	0	0.33
1	0	1	0	1	0	2.33
1	0	1	0	0	1	1.67
1	0	0	1	1	0	0.33
1	0	0	1	0	1	1.00
1	0	0	0	1	1	1.67
0	1	1	1	0	0	1.67
0	1	1	0	1	0	1.00
0	1	1	0	0	1	0.33
0	1	0	1	1	0	1.68