# Inference and Randomization

EC 607, Set 10

Edward Rubin

27 May 2020

# Prologue

# Schedule

## Last time

An analytical solution to cluter-robust inference

## Today

Inference using (re)randomization †

## Upcoming

The end is near. As is the final.

† These notes follow notes by Kosuke Imai, *Field Experiments* by Gerber and Green, and *Causal Inference for Statistics, Social, and Biomedical Sciences* by Imbens and Rubin.

# Inference and (re)randomization

# Inference and (re)randomization

## Inference recap

Our inference techniques have focused on (asymptotic) **analytical methods**.

1. Choose (or derive) an estimator

2. Derived the estimator's (asymptotic) distribution[†]

3. Construct confidence intervals or hypothesis tests

† And, consequently, standard errors.

## Resampling

**Resampling methods** offers a different, more computationally intense (less asymptoically intense) approach.

# Inference and (re)randomization

## Resampling

**Resampling methods** offers a different, more computationally intense (less asymptoically intense) approach.

A **resampling method** involves repeatedly drawing samples (*resampling*) from a dataset and refitting the model of interest on each sample. We can learn about the behavior of the model through its performance across the many iterations.[†]

† This approach is very similar to our Monte Carlo simulations, except that we will sample *with replacement* from a single dataset.

## Resampling

**Resampling methods** offers a different, more computationally intense (less asymptoically intense) approach.

A **resampling method** involves repeatedly drawing samples (*resampling*) from a dataset and refitting the model of interest on each sample. We can learn about the behavior of the model through its performance across the many iterations.[†]

*Common implementations:* Bootstrap (and jackknife), cross validation, permutation tests/randomization inference

† This approach is very similar to our Monte Carlo simulations, except that we will sample *with replacement* from a single dataset.

# The bootstrap

# The bootstrap

## Basics

**Bootstrapping** resamples, *with replacement*, from the original dataset.

# The bootstrap

## Basics

Bootstrapping resamples, *with replacement*, from the original dataset.

- In each sample, we apply our estimator.
- Then, we consider the distribution/properties of these estimates.

This resampling helps us better understand the uncertainty associated with our estimator (within the current data setting).

## More formally

Let's formalize the bootstrap a bit.

- $Z$ denotes our original dataset (*e.g.*, $Z = [\mathbf{Y} \mid \mathbf{X}]$ in our standard setup).

## More formally

Let's formalize the bootstrap a bit.

- $Z$ denotes our original dataset (*e.g.*, $Z = [\mathbf{Y} \mid \mathbf{X}]$ in our standard setup).
- $\hat{\alpha}(Z)$ refers to the estimate for $\alpha$ derived from our dataset $Z$.

## More formally

Let's formalize the bootstrap a bit.

- $Z$ denotes our original dataset (*e.g.*, $Z = [\mathbf{Y} \mid \mathbf{X}]$ in our standard setup).
- $\hat{\alpha}(Z)$ refers to the estimate for $\alpha$ derived from our dataset $Z$.
- We draw $B$ bootstrap samples $b \in \{1, \ldots, B\}$.

# The bootstrap

## More formally

Let's formalize the bootstrap a bit.

- $Z$ denotes our original dataset (*e.g.*, $Z = [\mathbf{Y} \mid \mathbf{X}]$ in our standard setup).
- $\hat{\alpha}(Z)$ refers to the estimate for $\alpha$ derived from our dataset $Z$.
- We draw $B$ bootstrap samples $\boldsymbol{b} \in \{1,\, \ldots,\, B\}$.
- $Z^{\star 1}$ represents our first bootstrap sample ($\boldsymbol{b} = 1$).

# The bootstrap

## More formally

Let's formalize the bootstrap a bit.

- $Z$ denotes our original dataset (*e.g.*, $Z = [\mathbf{Y} \mid \mathbf{X}]$ in our standard setup).
- $\hat{\alpha}(Z)$ refers to the estimate for $\alpha$ derived from our dataset $Z$.
- We draw $B$ bootstrap samples $b \in \{1, \ldots, B\}$.
- $Z^{\star 1}$ represents our first bootstrap sample ($b = 1$).
- $\hat{\alpha}^{\star 1} = \hat{\alpha}(Z^{\star 1})$ is our estimator evaluated on the first bootstrap sample.

## More formally

Let's formalize the bootstrap a bit.

- $Z$ denotes our original dataset (*e.g.*, $Z = [\mathbf{Y} \mid \mathbf{X}]$ in our standard setup).
- $\hat{\alpha}(Z)$ refers to the estimate for $\alpha$ derived from our dataset $Z$.
- We draw $B$ bootstrap samples $b \in \{1, \ldots, B\}$.
- $Z^{\star 1}$ represents our first bootstrap sample ($b = 1$).
- $\hat{\alpha}^{\star 1} = \hat{\alpha}(Z^{\star 1})$ is our estimator evaluated on the first bootstrap sample.

The bootstrapped standard error of $\hat{\alpha}$ is the standard deviation of the $\hat{\alpha}^{\star b}$

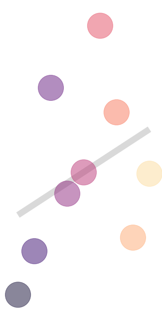
$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left( \hat{\alpha}^{\star b} - \frac{1}{B} \sum_{\ell=1}^{B} \hat{\alpha}^{\star \ell} \right)^2}$$

# The bootstrap

## More graphically

$Z$


$$\hat{\beta} = 0.653$$

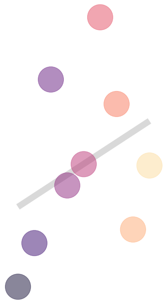


## More graphically

$Z$

| | | |
|---|---|---|
| 7 | 8 | 9 |
| 4 | 5 | 6 |
| 1 | 2 | 3 |

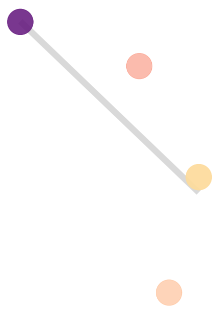
$$\hat{\beta} = 0.653$$



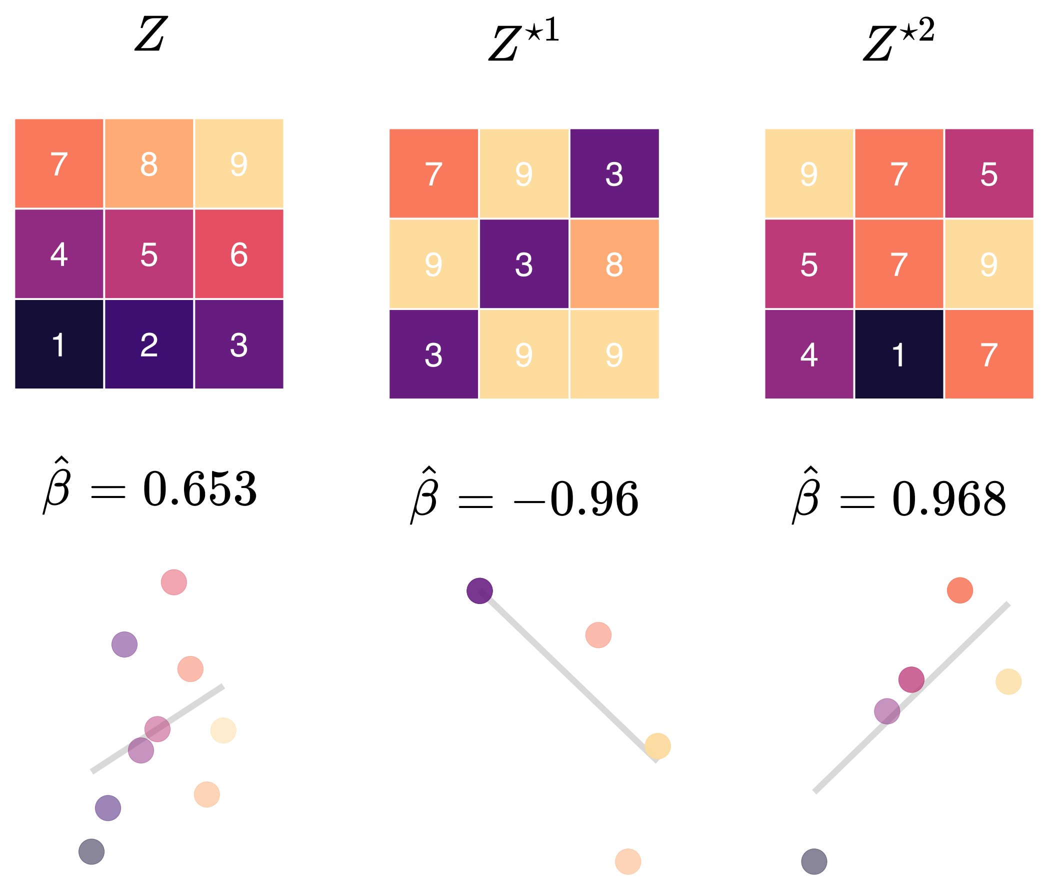
$Z^{\star 1}$

| | | |
|---|---|---|
| 7 | 9 | 3 |
| 9 | 3 | 8 |
| 3 | 9 | 9 |

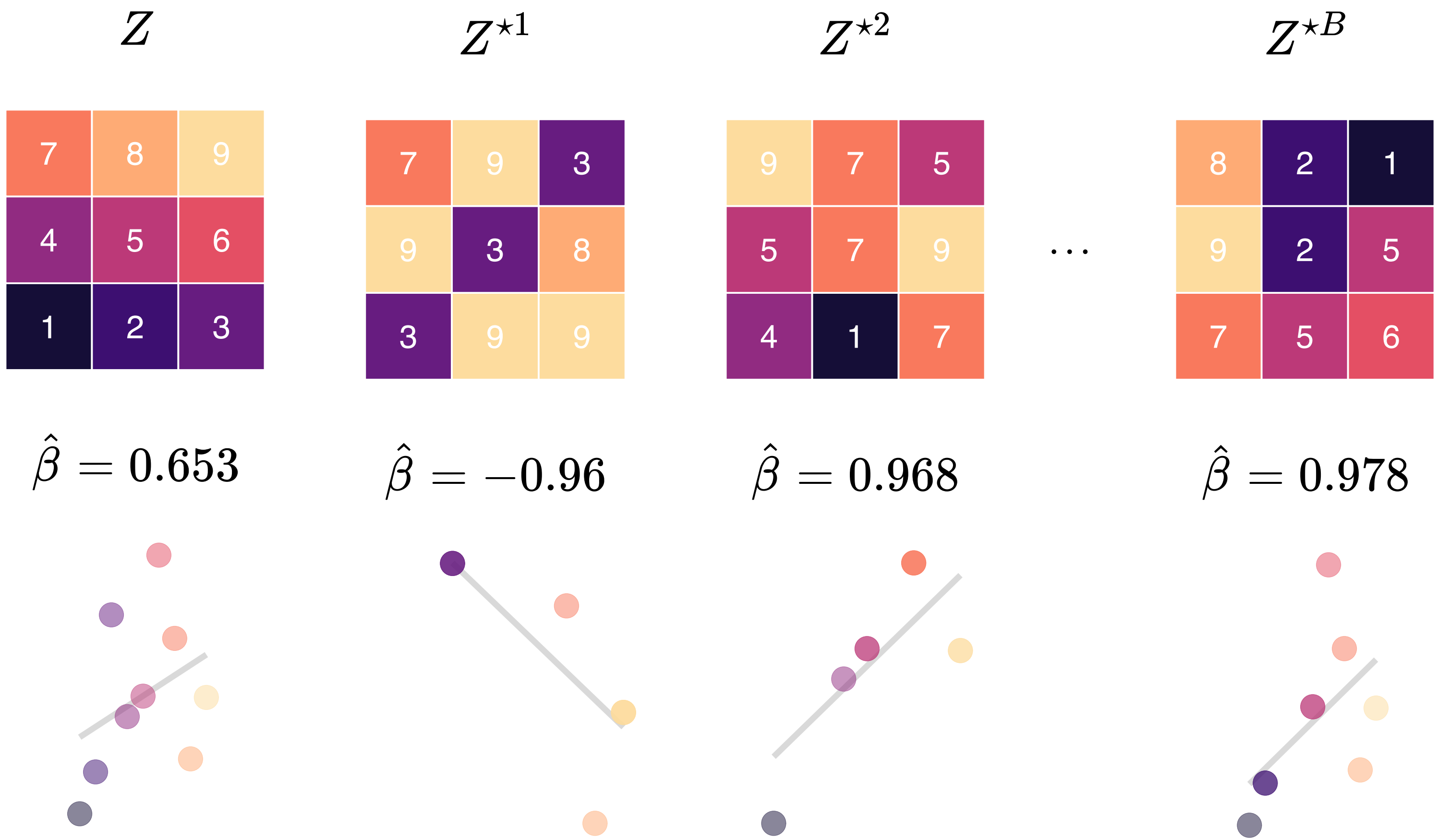
$$\hat{\beta} = -0.96$$



## More graphically



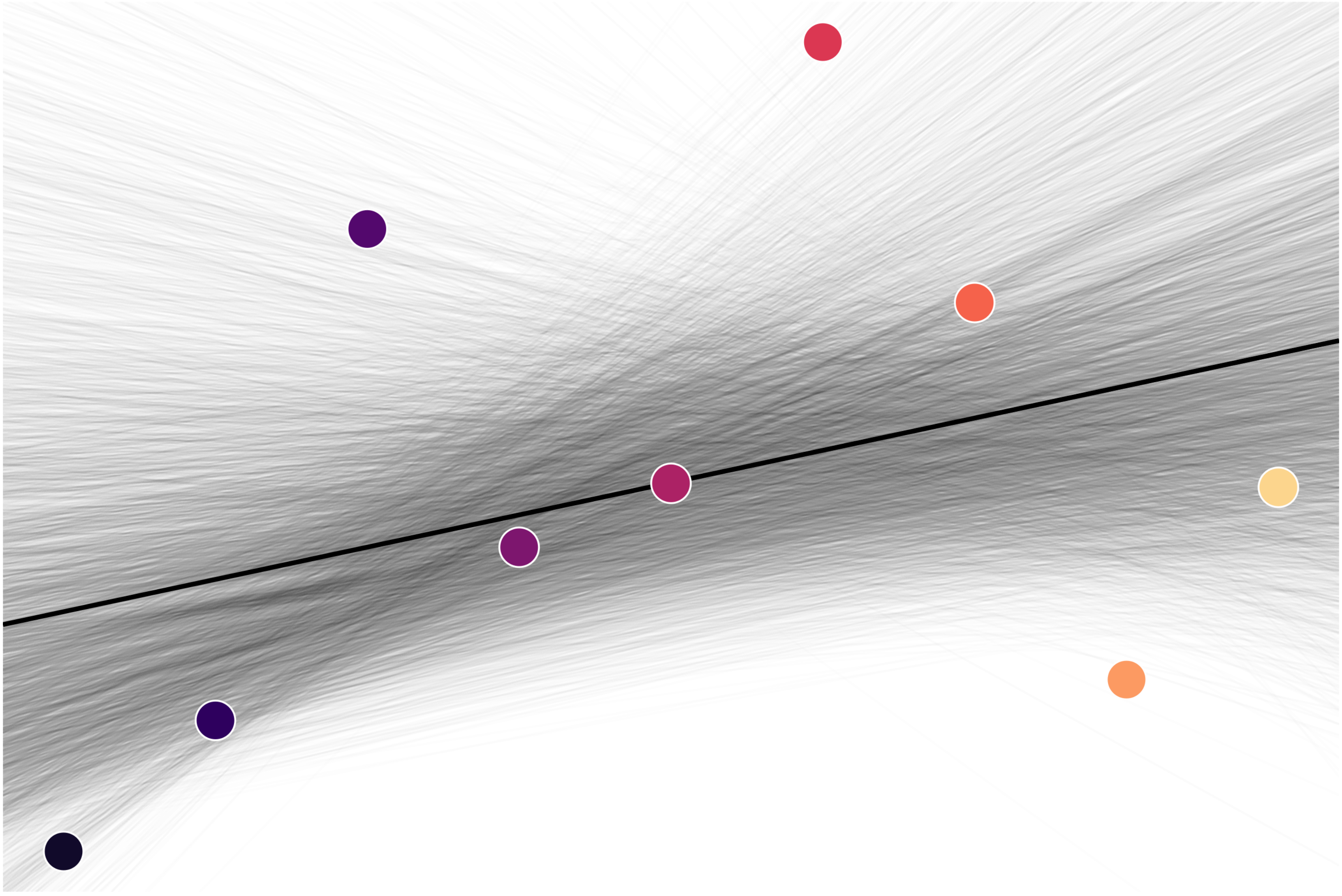
## More graphically



# The bootstrap

Running this bootstrap 10,000 times

```r
plan(multiprocess, workers = 10)
# Set a seed
set.seed(123)
# Run the simulation 1e4 times
boot_df ← future_map_dfr(
  # Repeat sample size 100 for 1e4 times
  rep(n, 1e4),
  # Our function
  function(n) {
    # Estimates via bootstrap
    est ← lm(y ~ x, data = z[sample(1:n, n, replace = T), ])
    # Return a tibble
    data.frame(int = est$coefficients[1], coef = est$coefficients[2])
  },
  # Let furrr know we want to set a seed
  .options = future_options(seed = T)
)
```



# The bootstrap

## Comparison

In this 10,000-sample bootstrap, we calculate a standard error for $\hat{\beta}_1$ of approximately 0.786.

## Comparison

In this 10,000-sample bootstrap, we calculate a standard error for $\hat{\beta}_1$ of approximately 0.786.

If we go the old-fashioned OLS route $\left(s^2 \left(\mathrm{X}'\mathrm{X}\right)^2\right)$, we estimate 0.673.

Not bad.

# Permutation tests

# Permutation tests

## Motivation

Consider the null hypothesis of *no average treatment effect*, *i.e.*,

$$\mathrm{H_0}\colon \overline{\mathbf{Y}}_0 = \overline{\mathbf{Y}}_1 \quad \left(\implies \bar{\tau} = 0\right)$$

# Permutation tests

## Motivation

Consider the null hypothesis of *no average treatment effect*, *i.e.*,

$$\text{H}_0\colon \overline{\mathbf{Y}_0} = \overline{\mathbf{Y}_1} \quad \left(\implies \bar{\tau} = 0\right)$$

We've discussed how randomization avoids the pitfalls of selection bias.

# Permutation tests

## Motivation

Consider the null hypothesis of *no average treatment effect*, *i.e.*,

$$\text{H}_0\text{: } \overline{\mathbf{Y}}_0 = \overline{\mathbf{Y}}_1 \quad \left(\implies \bar{\tau} = 0\right)$$

We've discussed how randomization avoids the pitfalls of selection bias.

Randomization can also clarify inference—helping quantify uncertainty.

# Permutation tests

## Motivation

Consider the null hypothesis of *no average treatment effect*, *i.e.*,

$$\text{H}_0\text{: } \overline{\mathbf{Y}}_0 = \overline{\mathbf{Y}}_1 \quad \left(\implies \bar{\tau} = 0\right)$$

We've discussed how randomization avoids the pitfalls of selection bias.

Randomization can also clarify inference—helping quantify uncertainty.

**Q** How?

# Permutation tests

## Motivation

Consider the null hypothesis of *no average treatment effect*, *i.e.*,

$$\text{H}_0\text{: } \overline{\mathbf{Y}}_0 = \overline{\mathbf{Y}}_1 \quad \left(\implies \bar{\tau} = 0\right)$$

We've discussed how randomization avoids the pitfalls of selection bias.

Randomization can also clarify inference—helping quantify uncertainty.

**Q** How?

**A** We know exactly how the randomness happened (we assigned it), so we don't need parametric assumptions to derive a distribution under $\text{H}_0$!

# Permutation tests

## Motivation

Consider the null hypothesis of *no average treatment effect*, *i.e.*,

$$\text{H}_\text{o}\text{: } \overline{\mathbf{Y}}_0 = \overline{\mathbf{Y}}_1 \quad \left( \implies \overline{\tau} = 0 \right)$$

We've discussed how randomization avoids the pitfalls of selection bias.

Randomization can also clarify inference—helping quantify uncertainty.

Q How?

A We know exactly how the randomness happened (we assigned it), so we don't need parametric assumptions to derive a distribution under $\text{H}_\text{o}$!
We use the **experimental design**, rather than a probability model.

# Permutation tests

## Tea drinkers

*Classic example* Sir R. A. Fisher had a colleague who claimed to be able to tell whether the tea was poured into milk *or* milk was poured into the tea.[†]

† Don't worry, Fisher is known for more than this one experiment.

# Permutation tests

## Tea drinkers

*Classic example* Sir R. A. Fisher had a colleague who claimed to be able to tell whether the tea was poured into milk *or* milk was poured into the tea.†

Being the friend he was, Fisher designed an experiment to determine whether his colleague was telling the truth.

† Don't worry, Fisher is known for more than this one experiment.

# Permutation tests

## Tea drinkers

*Classic example* Sir R. A. Fisher had a colleague who claimed to be able to tell whether the tea was poured into milk *or* milk was poured into the tea.†

Being the friend he was, Fisher designed an experiment to determine whether his colleague was telling the truth.

Fisher randomized the order of 8 cups of tea:

- 4 cups with **m**ilk added first
- 4 cups with **t**ea added first

† Don't worry, Fisher is known for more than this one experiment.

# Permutation tests

## Tea drinkers

*Classic example* Sir R. A. Fisher had a colleague who claimed to be able to tell whether the tea was poured into milk *or* milk was poured into the tea.†

Being the friend he was, Fisher designed an experiment to determine whether his colleague was telling the truth.

Fisher randomized the order of 8 cups of tea:

- 4 cups with **m**ilk added first
- 4 cups with **t**ea added first

Vindication! His colleague got all 8 correct.

† Don't worry, Fisher is known for more than this one experiment.

## Tea drinkers

*Classic example* Sir R. A. Fisher had a colleague who claimed to be able to tell whether the tea was poured into milk *or* milk was poured into the tea.†

Being the friend he was, Fisher designed an experiment to determine whether his colleague was telling the truth.

Fisher randomized the order of 8 cups of tea:

- 4 cups with **m**ilk added first
- 4 cups with **t**ea added first

Vindication! His colleague got all 8 correct.

**Q** With random guessing, how likely is correctly guessing all 8 cups?

† Don't worry, Fisher is known for more than this one experiment.

## Tea drinkers 2

Q With random guessing, how likely is correctly guessing all 8 cups?

# Permutation tests

## Tea drinkers 2

Q With random guessing, how likely is correctly guessing all 8 cups?

This question reflects our understanding of a ***p*-value**.

If Fisher's colleague had no ability and simply guessed ($H_0$), what is the probability she would have guessed all 8 cups correctly?

## Tea drinkers 2

Q With random guessing, how likely is correctly guessing all 8 cups?

This question reflects our understanding of a ***p*-value**.

If Fisher's colleague had no ability and simply guessed ($H_0$), what is the probability she would have guessed all 8 cups correctly?

Fisher's $H_0$: the answers were unrelated to the cups' actual contents.

Under this hypothesis, we can re-randomize the cups and see how many times her answer was perfectly correct.

# Permutation tests

## Tea drinkers 2

Q With random guessing, how likely is correctly guessing all 8 cups?

This question reflects our understanding of a ***p*-value**.

If Fisher's colleague had no ability and simply guessed ($H_0$), what is the probability she would have guessed all 8 cups correctly?

Fisher's $H_0$: the answers were unrelated to the cups' actual contents.

Under this hypothesis, we can re-randomize the cups and see how many times her answer was perfectly correct.

This is the idea behind permutation testing and randomization inference.

# Permutation tests

## Tea drinkers with a vengeance

| *Cup* |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |

# Permutation tests

## Tea drinkers with a vengeance

| *Cup* | *Guess* |
|---|---|
| 1 | **m** |
| 2 | **t** |
| 3 | **t** |
| 4 | **m** |
| 5 | **m** |
| 6 | **t** |
| 7 | **t** |
| 8 | **m** |

# Permutation tests

## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* |
|---|---|---|
| 1 | **m** | **m** |
| 2 | **t** | **t** |
| 3 | **t** | **t** |
| 4 | **m** | **m** |
| 5 | **m** | **m** |
| 6 | **t** | **t** |
| 7 | **t** | **t** |
| 8 | **m** | **m** |
| | | 8/8 |

# Permutation tests

## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* | $P_1$ |
|---|---|---|---|
| 1 | **m** | **m** | **m** |
| 2 | **t** | **t** | ~~**m**~~ |
| 3 | **t** | **t** | ~~**m**~~ |
| 4 | **m** | **m** | **m** |
| 5 | **m** | **m** | ~~**t**~~ |
| 6 | **t** | **t** | **t** |
| 7 | **t** | **t** | **t** |
| 8 | **m** | **m** | ~~**t**~~ |
| | | 8/8 | 4/8 |

# Permutation tests

## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* | $P_1$ | $P_2$ |
|---|---|---|---|---|
| 1 | **m** | **m** | **m** | **m** |
| 2 | **t** | **t** | ~~**m**~~ | ~~**m**~~ |
| 3 | **t** | **t** | ~~**m**~~ | ~~**m**~~ |
| 4 | **m** | **m** | **m** | ~~**t**~~ |
| 5 | **m** | **m** | ~~**t**~~ | **m** |
| 6 | **t** | **t** | **t** | **t** |
| 7 | **t** | **t** | **t** | **t** |
| 8 | **m** | **m** | ~~**t**~~ | ~~**t**~~ |
| | | 8/8 | 4/8 | 4/8 |

# Permutation tests

## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* | $P_1$ | $P_2$ | $P_3$ |
|---|---|---|---|---|---|
| 1 | m | m | m | m | m |
| 2 | t | t | ~~m~~ | ~~m~~ | ~~m~~ |
| 3 | t | t | ~~m~~ | ~~m~~ | ~~m~~ |
| 4 | m | m | m | ~~t~~ | ~~t~~ |
| 5 | m | m | ~~t~~ | m | ~~t~~ |
| 6 | t | t | t | t | ~~m~~ |
| 7 | t | t | t | t | t |
| 8 | m | m | ~~t~~ | ~~t~~ | ~~t~~ |
| | | 8/8 | 4/8 | 4/8 | 2/8 |

# Permutation tests

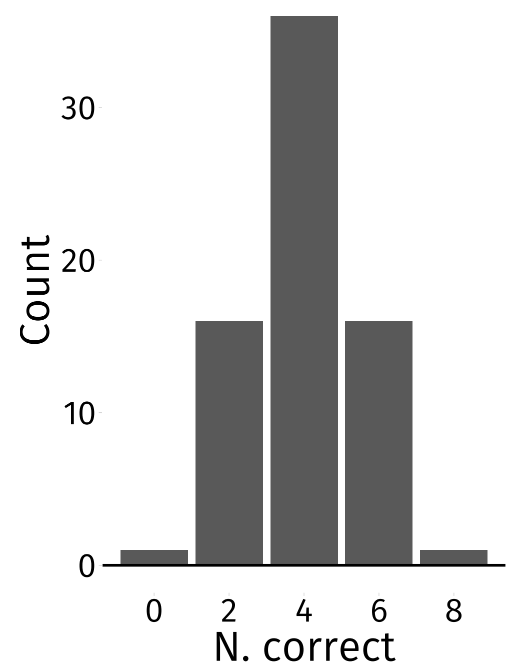
## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* | $P_1$ | $P_2$ | $P_3$ | $\cdots$ | $P_{70}$ |
|---|---|---|---|---|---|---|---|
| 1 | m | m | m | m | m | | ~~t~~ |
| 2 | t | t | ~~m~~ | ~~m~~ | ~~m~~ | | t |
| 3 | t | t | ~~m~~ | ~~m~~ | ~~m~~ | | t |
| 4 | m | m | m | ~~t~~ | ~~t~~ | | ~~t~~ |
| 5 | m | m | ~~t~~ | m | ~~t~~ | | m |
| 6 | t | t | t | t | ~~m~~ | | ~~m~~ |
| 7 | t | t | t | t | t | | ~~m~~ |
| 8 | m | m | ~~t~~ | ~~t~~ | ~~t~~ | | m |
| | | 8/8 | 4/8 | 4/8 | 2/8 | | 4/8 |

# Permutation tests

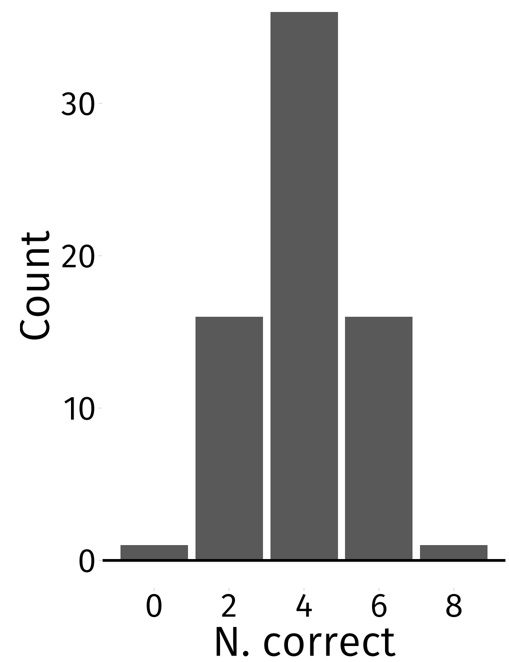
## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* | $P_1$ | $P_2$ | $P_3$ | $\cdots$ | $P_{70}$ |
|---|---|---|---|---|---|---|---|
| 1 | m | m | m | m | m | | ~~t~~ |
| 2 | t | t | ~~m~~ | ~~m~~ | ~~m~~ | | t |
| 3 | t | t | ~~m~~ | ~~m~~ | ~~m~~ | | t |
| 4 | m | m | m | ~~t~~ | ~~t~~ | | ~~t~~ |
| 5 | m | m | ~~t~~ | m | ~~t~~ | | m |
| 6 | t | t | t | t | ~~m~~ | | ~~m~~ |
| 7 | t | t | t | t | t | | ~~m~~ |
| 8 | m | m | ~~t~~ | ~~t~~ | ~~t~~ | | m |
| | | 8/8 | 4/8 | 4/8 | 2/8 | | 4/8 |



## Tea drinkers with a vengeance

| *Cup* | *Guess* | *Truth* | $P_1$ | $P_2$ | $P_3$ | $\cdots$ | $P_{70}$ |
|---|---|---|---|---|---|---|---|
| 1 | m | m | m | m | m | | ~~t~~ |
| 2 | t | t | ~~m~~ | ~~m~~ | ~~m~~ | | t |
| 3 | t | t | ~~m~~ | ~~m~~ | ~~m~~ | | t |
| 4 | m | m | m | ~~t~~ | ~~t~~ | | ~~t~~ |
| 5 | m | m | ~~t~~ | m | ~~t~~ | | m |
| 6 | t | t | t | t | ~~m~~ | | ~~m~~ |
| 7 | t | t | t | t | t | | ~~m~~ |
| 8 | m | m | ~~t~~ | ~~t~~ | ~~t~~ | | m |
| | | 8/8 | 4/8 | 4/8 | 2/8 | | 4/8 |




So our permutation-test-based *p*-value is $1/70 \approx 0.0143$. $\Longrightarrow$ Reject $H_0$.

## Generalization

The procedure for permutation-based hypothesis testing[†] is the same as our "standard" asymptotic-based hypothesis testing.

† Also called *Fisher's exact test*, as you get exact $p$-values.

# Permutation tests

## Generalization

The procedure for permutation-based hypothesis testing[†] is the same as our "standard" asymptotic-based hypothesis testing.

1. **Define hypotheses**, $H_o$ and $H_a$.
2. Choose our **rejection threshold** $\alpha$ (tolerated type-I error rate).
3. Choose a **test statistic** that is a function of our sample.
4. Derive/calculate the **test statistic's distribution *under $H_o$***.
5. **Compute the *p*-value** by comparing test stat. to its $H_o$ distribution.
6. **Conclusions**—reject or fail to reject $H_o$.

† Also called *Fisher's exact test*, as you get exact *p*-values.

# Permutation tests

## Generalization

The procedure for permutation-based hypothesis testing[†] is the same as our "standard" asymptotic-based hypothesis testing.

1. **Define hypotheses**, $H_o$ and $H_a$.
2. Choose our **rejection threshold** $\alpha$ (tolerated type-I error rate).
3. Choose a **test statistic** that is a function of our sample.
4. Derive/calculate the **test statistic's distribution *under $H_o$***.
5. **Compute the *p*-value** by comparing test stat. to its $H_o$ distribution.
6. **Conclusions**—reject or fail to reject $H_o$.

*The difference:* Permutation tests use the randomization's mechanism to construct the test-statistic's exact distribution under $H_o$.

† Also called *Fisher's exact test*, as you get exact *p*-values.

# Permutation tests

## More generally

Fisher focused on testing a **sharp null hypothesis**—no effect *for anyone*, *i.e.*,

$$\text{H}_\text{o}\text{: } \mathbf{Y}_{1i} - \mathbf{Y}_{0i} = 0 \ \ \forall i \quad (\implies \tau_i = 0 \ \ \forall i)$$

against an alternative hypothesis that someone has a non-zero effect

$$\text{H}_\text{a}\text{: } \ \mathbf{Y}_{1i} - \mathbf{Y}_{0i} \neq 0 \text{ for some } i \ (\implies \exists i \ \ \text{s.t.} \ \ \tau_i \neq 0)$$

# Permutation tests

## More generally

Fisher focused on testing a **sharp null hypothesis**—no effect *for anyone*, *i.e.*,

$$\text{H}_\text{o}\text{: } \mathbf{Y}_{1i} - \mathbf{Y}_{0i} = 0 \;\; \forall i \quad (\implies \tau_i = 0 \;\; \forall i)$$

against an alternative hypothesis that someone has a non-zero effect

$$\text{H}_\text{a}\text{: } \;\; \mathbf{Y}_{1i} - \mathbf{Y}_{0i} \neq 0 \text{ for some } i \; (\implies \exists i \;\; \mathbf{s.t.} \;\; \tau_i \neq 0)$$

A **sharp null hypothesis** is specified *for all individuals*, *e.g.*,

$$\text{H}_\text{o}\text{: } \mathbf{Y}_{1i} - \mathbf{Y}_{0i} = C \;\; \forall i$$

which differs from the ATE-based nulls that we normally consider, *e.g.*,

$$\text{H}_\text{o}\text{: } E[\mathbf{Y}_{1i} - \mathbf{Y}_{0i}] = C.$$

# Permutation tests

## Key insight

Our estimate (or test statistic) is a function of

1. individuals' responses $(\mathbf{Y}_i)$
2. individuals' treatment assignments $(\mathbf{D}_i)$

# Permutation tests

## Key insight

Our estimate (or test statistic) is a function of

1. individuals' responses $(\mathbf{Y}_i)$
2. individuals' treatment assignments $(\mathbf{D}_i)$

**Under the sharp null** $\text{H}_0$: $\tau_i = 0$ $(\forall i)$

- $\mathbf{Y}_{0i} = \mathbf{Y}_{1i} = \mathbf{Y}_i\ \ \forall i$ (*i.e.*, changing $\mathbf{D}_i$ will not affect observed $\mathbf{Y}_i$)

# Permutation tests

## Key insight

Our estimate (or test statistic) is a function of

1. individuals' responses $(\mathbf{Y}_i)$
2. individuals' treatment assignments $(\mathbf{D}_i)$

**Under the sharp null** $\text{H}_0$: $\tau_i = \mathbf{0}$ $(\forall \boldsymbol{i})$

- $\mathbf{Y}_{0i} = \mathbf{Y}_{1i} = \mathbf{Y}_i \;\; \forall \boldsymbol{i}$ (*i.e.*, changing $\mathbf{D}_i$ will not affect observed $\mathbf{Y}_i$)

- Permutations of $\mathbf{D}$ construct the *exact* null distribution (unchanged $\mathbf{Y}$).

# Permutation tests

## Key insight

Our estimate (or test statistic) is a function of

1. individuals' responses $(\mathbf{Y}_i)$
2. individuals' treatment assignments $(\mathbf{D}_i)$

**Under the sharp null** $H_0$: $\tau_i = \mathbf{0}$ $(\forall \boldsymbol{i})$

- $\mathbf{Y}_{0i} = \mathbf{Y}_{1i} = \mathbf{Y}_i \;\; \forall \boldsymbol{i}$ (*i.e.*, changing $\mathbf{D}_i$ will not affect observed $\mathbf{Y}_i$)

- Permutations of $\mathbf{D}$ construct the *exact* null distribution (unchanged $\mathbf{Y}$).

The number of possible permutations can get big

# Permutation tests

## Key insight

Our estimate (or test statistic) is a function of

1. individuals' responses $(\mathbf{Y}_i)$
2. individuals' treatment assignments $(\mathbf{D}_i)$

**Under the sharp null** $\text{H}_0$: $\tau_i = \mathbf{0}\ (\forall \boldsymbol{i})$

- $\mathbf{Y}_{0i} = \mathbf{Y}_{1i} = \mathbf{Y}_i\ \ \forall \boldsymbol{i}$ (*i.e.*, changing $\mathbf{D}_i$ will not affect observed $\mathbf{Y}_i$)

- Permutations of $\mathbf{D}$ construct the *exact* null distribution (unchanged $\mathbf{Y}$).

The number of possible permutations can get big—*e.g.*, 500 treated and 500 control has $2.7 \times 10^{299}$ options.

## Key insight

Our estimate (or test statistic) is a function of

1. individuals' responses $(\mathbf{Y}_i)$
2. individuals' treatment assignments $(\mathbf{D}_i)$

**Under the sharp null** $\mathrm{H_0}$: $\tau_i = 0$ $(\forall i)$

- $\mathbf{Y}_{0i} = \mathbf{Y}_{1i} = \mathbf{Y}_i \;\; \forall i$ (*i.e.*, changing $\mathbf{D}_i$ will not affect observed $\mathbf{Y}_i$)

- Permutations of $\mathbf{D}$ construct the *exact* null distribution (unchanged $\mathbf{Y}$).

The number of possible permutations can get big—*e.g.*, 500 treated and 500 control has $2.7 \times 10^{299}$ options. Approximate the distribution by sampling.

# Permutation tests

## Different inference

In his 2019 paper Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results, Alwyn Young 'updates' inference from 53 experimental papers by using randomization-based inference.

> In the average paper, randomization tests of the significance of individual treatment effects find 13% to 22% fewer significant results than are found using authors' methods.

Young (2019)

# Permutation tests

## Different inference?

It's certainly possible authors and methods can be wrong.

However, permutation-based inference itself may generate differences relative to the more standard, derived, asymptotics-based estimators.

*Why?*

# Permutation tests

## Different inference?

It's certainly possible authors and methods can be wrong.

However, permutation-based inference itself may generate differences relative to the more standard, derived, asymptotics-based estimators.

*Why?*

1. We are testing **different null hypotheses** (sharp *vs.* non-sharp).
2. The two estimators have **different asymptotic properties**.[†]

† Thanks go to Alberto Abadie for this point.

# Permutation tests

## On average

The sharp null was central to Fisher's interpretation.

Neyman *et al.* (1935) extended[†] this idea of permutation-based tests to the average treatment effect (testing $H_o$: $E[\mathbf{Y}_{1i}] - E[\mathbf{Y}_{0i}] = 0$).

Neyman and others also added standard errors and confidence intervals.

† Fisher, paraphrased: 🤬

†† *Permutation tests* and *Randomization inference* are not the most strictly defined terms.

# Permutation tests

## On average

The sharp null was central to Fisher's interpretation.

Neyman *et al.* (1935) extended[†] this idea of permutation-based tests to the average treatment effect (testing $\text{H}_0$: $E[\mathbf{Y}_{1i}] - E[\mathbf{Y}_{0i}] = \mathbf{0}$).

Neyman and others also added standard errors and confidence intervals.

These extensions have come to be known as **randomization inference**.[††]

† Fisher, paraphrased: 🤬

†† *Permutation tests* and *Randomization inference* are not the most strictly defined terms.

# Randomization inference

# Randomization inference

## Setup

In order to **generalize our null hypothesis to the average treatment effect**,

$$\mathrm{H_0}\colon \overline{\tau} = 0 \implies E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] = 0$$

we have to give up something.

1. If we want an exact null distribution, then we must **assume a uniform treatment effect**.

# Randomization inference

## Setup

In order to **generalize our null hypothesis to the average treatment effect**,

$$\text{H}_0\colon \overline{\tau} = 0 \implies E[\text{Y}_{1i} - \text{Y}_{0i}] = 0$$

we have to give up something.

1. If we want an exact null distribution, then we must **assume a uniform treatment effect**. (Assuming our way back to a sharp null.)

# Randomization inference

## Setup

In order to **generalize our null hypothesis to the average treatment effect**,

$$\text{H}_0\text{: } \overline{\tau} = 0 \implies E[\text{Y}_{1i} - \text{Y}_{0i}] = 0$$

we have to give up something.

1. If we want an exact null distribution, then we must **assume a uniform treatment effect**. (Assuming our way back to a sharp null.)

2. If we want to avoid assuming $\tau_i = \overline{\tau} \;\; \forall i$, then we have to **accept a non-exact null distribution**. (We don't observe $\text{Y}_{0i}$ for $\text{D}_i = 1$.)

# Randomization inference

## Setup

In order to **generalize our null hypothesis to the average treatment effect**,

$$\text{H}_0\text{: } \overline{\tau} = 0 \implies E[\text{Y}_{1i} - \text{Y}_{0i}] = 0$$

we have to give up something.

1. If we want an exact null distribution, then we must **assume a uniform treatment effect**. (Assuming our way back to a sharp null.)

2. If we want to avoid assuming $\tau_i = \overline{\tau} \;\; \forall i$, then we have to **accept a non-exact null distribution**. (We don't observe $\text{Y}_{0i}$ for $\text{D}_i = 1$.)

If we don't like either option, then we need to go back to deriving asymptotic properties via probability modeling assumptions.

# Randomization inference

## Implementation

Once we decide which simplification we're willing to accept, we proceed similarly to permutation tests:

- shuffle $\mathbf{D}$ in a way that mimics treatment assignment
- collect test statistics from each iteration

## Implementation

Once we decide which simplification we're willing to accept, we proceed similarly to permutation tests:

- shuffle $\mathbf{D}$ in a way that mimics treatment assignment
- collect test statistics from each iteration

*Note* Monte Carlo simulations, bootstrap, permutation tests, and randomization all apply very similar processes.

# Randomization inference

## (Which) Test statistics

We still need to choose a test statistic on which we base the $p$-value.

- The **actual estimate**—difference in means or coefficient
- **Transformed estimates**
- **Quantiles**, *e.g.*, the median
- ***t* statistic**
- **Rank** statistics

# Randomization inference

## (Which) Test statistics

We still need to choose a test statistic on which we base the *p*-value.

- The **actual estimate**—difference in means or coefficient
- **Transformed estimates**
- **Quantiles**, *e.g.*, the median
- ***t* statistic**
- **Rank** statistics

We can also extend this idea to **confidence intervals**.

*E.g.*, Use the point estimates associated with the 2.5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles to construct a 95% confidence interval.

# Randomization inference

## Example

Back to the LaLonde NSW dataset. We previously estimated

- the NSW increased real earnings by $\hat{\beta}_1 \approx$ \$886.30
- (het.-robust) standard error of \$488.20
- *t* statistic $t_{\text{stat}} \approx$ 1.82 with *p*-value $\approx$ 0.0699

# Randomization inference

## Example

Back to the LaLonde NSW dataset. We previously estimated

- the NSW increased real earnings by $\hat{\beta}_1 \approx$ \$886.30
- (het.-robust) standard error of \$488.20
- *t* statistic $t_{\text{stat}} \approx$ 1.82 with *p*-value $\approx$ 0.0699

Let's re-randomize treatment 10,000 times. In each **iteration** $r$, calculate

1. $\hat{\beta}_1^r$, the **point estimate** (the regression coefficient)
2. $t_{\text{stat}}^r$, the ***t* statistic**

# Randomization inference

## Example

Back to the LaLonde NSW dataset. We previously estimated

- the NSW increased real earnings by $\hat{\beta}_1 \approx$ \$886.30
- (het.-robust) standard error of \$488.20
- $t$ statistic $t_{\text{stat}} \approx$ 1.82 with $p$-value $\approx$ 0.0699

Let's re-randomize treatment 10,000 times. In each **iteration** $r$, calculate

1. $\hat{\beta}_1^r$, the **point estimate** (the regression coefficient)
2. $t_{\text{stat}}^r$, the ***t* statistic**

Then calculate the implied $p$-values using the location of $\hat{\beta}_1$ and $t_{\text{stat}}$ in the distributions of $\hat{\beta}_1^r$ and $t_{\text{stat}}^r$, respectively.†

† Very similar exercise for confience intervals.

# Randomization inference

## Example: Re-randomization

The main decision is how to generate treatment.

**Q** Should we permute $\mathbf{D}$ or draw $\mathbf{D}_i$ for each individual?†

† The difference is in whether we hold the number of treated individuals constant.

## Example: Re-randomization

The main decision is how to generate treatment.

**Q** Should we permute $\mathbf{D}$ or draw $\mathbf{D}_i$ for each individual?[†]

**A** How was the original randomization conducted?

† The difference is in whether we hold the number of treated individuals constant.

# Randomization inference

## Example: Re-randomization

The main decision is how to generate treatment.

**Q** Should we permute $\mathbf{D}$ or draw $\mathbf{D}_i$ for each individual?[†]

**A** How was the original randomization conducted?

We'll assume the NSW started with a set number of treatments to disperse.

† The difference is in whether we hold the number of treated individuals constant.

First, we'll write a function that performs one iteration.

```r
# Arguments: 'i' (iteration), 'n_t' (# of trt)
fun_randomization ← function(i) {
  # Sample the treatment vector. NOTE: Sampling WITHOUT replacement
  t_i ← sample(nsw_df$treat, size = nrow(nsw_df), replace = F)
  # Regression using our re-randomized treatment
  est_i ← lm_robust(re78 ~ t_i, data = nsw_df) %>% tidy()
  # Return tibble with iteration, point estimate, and test statistic
  tibble(i, est = est_i[2,"estimate"], t_stat = est_i[2,"statistic"])
}
```

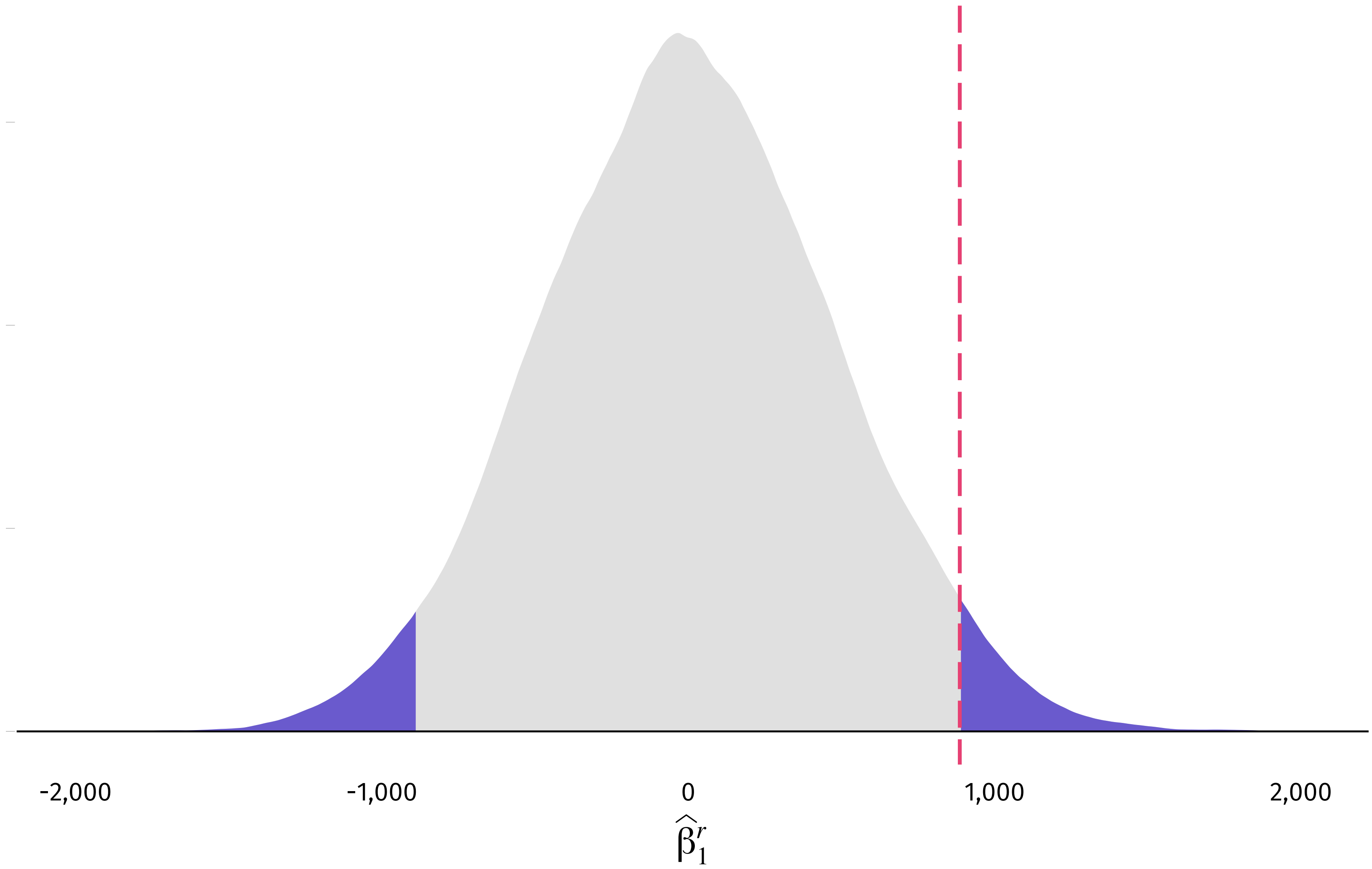
First, we'll write a function that performs one iteration.

```r
# Arguments: 'i' (iteration), 'n_t' (# of trt)
fun_randomization <- function(i) {
  # Sample the treatment vector. NOTE: Sampling WITHOUT replacement
  t_i <- sample(nsw_df$treat, size = nrow(nsw_df), replace = F)
  # Regression using our re-randomized treatment
  est_i <- lm_robust(re78 ~ t_i, data = nsw_df) %>% tidy()
  # Return tibble with iteration, point estimate, and test statistic
  tibble(i, est = est_i[2,"estimate"], t_stat = est_i[2,"statistic"])
}
```

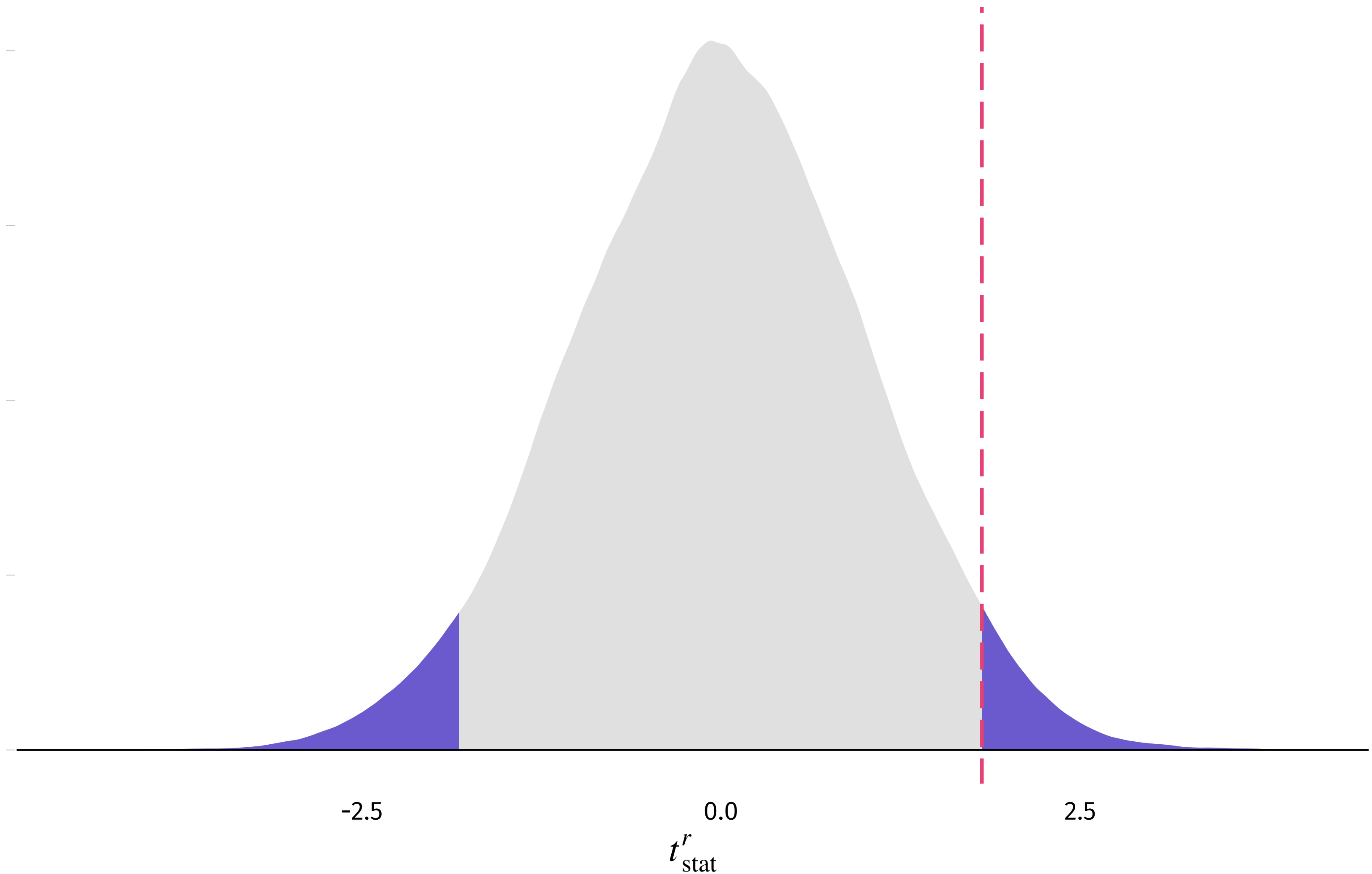
And now run the re-randomization function 10,000 times.

```r
# Set up parallelization and seed
plan(multiprocess, workers = 4); set.seed(1234)
# Run the simulation 1e4 times
random_df <- future_map_dfr(
  1:1e4,
  fun_randomization,
  .options = future_options(seed = T)
)
```

*Result 1* Share $|\hat{\beta}_1^r| > \hat{\beta}_1 = 0.0615$. (Original $p$-value = 0.0699)



*Result 2* Share $\left|t^r_{\text{stat}}\right| > t_{\text{stat}} = 0.0703$. (Original $p$-value $= 0.0699$)



## Confidence intervals

To construct confidence intervals, we **invert** randomization-based **hypothesis tests**, imposing a range of null hypotheses.

# Randomization inference

## Confidence intervals

To construct confidence intervals, we **invert** randomization-based **hypothesis tests**, imposing a range of null hypotheses.

*E.g.,* To construct a 95% C.I. for $\hat{\tau}$

1. Impose the null hypothesis $H_0$: $\tau = \tau_o$ for many values of $\tau_o$.
2. Find all values of $\tau_o$ that do not reject $\hat{\tau}$ at the 5% level.

# Randomization inference

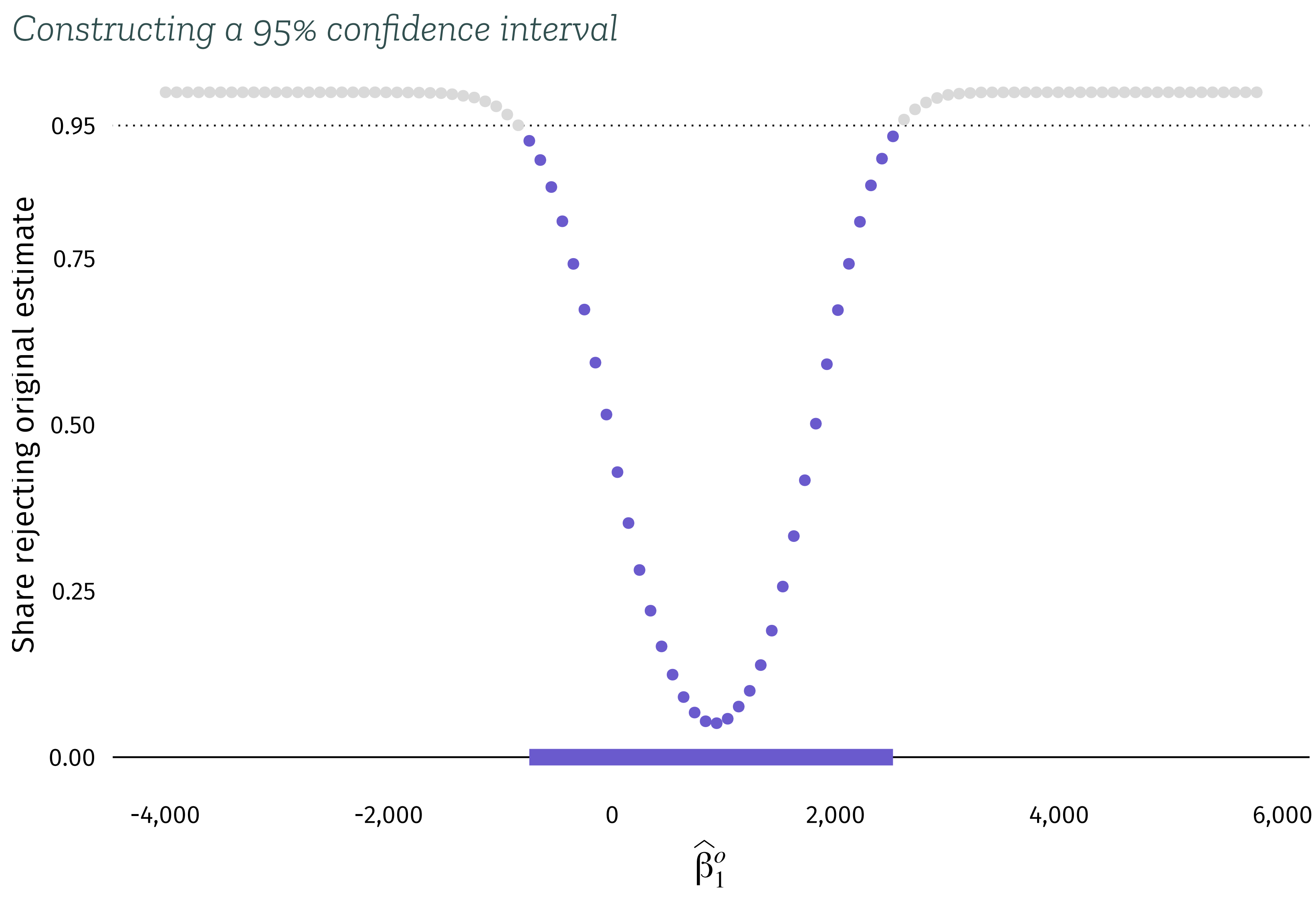
## Confidence intervals

To construct confidence intervals, we **invert** randomization-based **hypothesis tests**, imposing a range of null hypotheses.

*E.g.,* To construct a 95% C.I. for $\hat{\tau}$

1. Impose the null hypothesis $\text{H}_0$: $\tau = \tau_o$ for many values of $\tau_o$.
2. Find all values of $\tau_o$ that do not reject $\hat{\tau}$ at the 5% level.

*Note* We must to be able to clearly impose the null in our "model".


Constructing a 95% confidence interval
Share rejecting original estimate
0.95
0.75
0.50
0.25
0.00
-4,000
-2,000
0
2,000
4,000
6,000
$\widehat{\beta}_1^o$

Athey and Imbens (2016) **on regression and randomization inference**:[†]

> Although these methods [regression] remain the most popular way of analyzing data from randomized experiments, we suggest caution in using them.
>
> … In particular there is a disconnect between the way the conventional assumptions in regression analyses are formulated and the implications of randomization. As a result it is easy for the researcher using regression methods to go beyond analyses that are justified by randomization, and end up with analyses that rely on a difficult-to-assess mix of randomization assumptions, modeling assumptions, and large sample approximation.

† Specifically in the context of experiments, though the concerns should remain in other contexts.

Athey and Imbens (2016) **on regression and randomization inference**:[†]

> Ultimately we recommend that researchers wishing to use regression or other model-based methods rather than the randomization-based methods we prefer, do so with care. For example, using only indicator variables based on partitioning the covariate space, rather than using multi-valued variables as covariates in the regression function preserves many of the finite sample properties that simple comparisons of means have, and leads to regression estimates with clear interpretations. In addition, in many cases the potential gains from regression adjustment can also be captured by careful ex ante design, that is, through stratified randomized experiments to be discussed in the next section, without the potential costs associated with ex post regression adjustment.

† Specifically in the context of experiments, though the concerns should remain in other contexts.

# Randomization and clustering

## The plot thickens

Permutation tests and randomization inference both work because we know[†] the process through which treatment was randomly assigned.

† Or claim to understand.

# Randomization and clustering

## The plot thickens

Permutation tests and randomization inference both work because we know[†] the process through which treatment was randomly assigned.

If treatment is correlated within groups, then our bootstraps, permutations, and re-randomizations need to reflect this dependence.

† Or claim to understand.

# Further reading

## Papers

Bootstrap-Based Improvements for Inference with Clustered Errors
Cameron, Gelbach, and Miller (2008)

The Econometrics of Randomized Experiments Athey and Imbens (2016)

Randomization Inference With Natural Experiments
Ho and Imai (2012)
*Also:* Notes by Kosuke Imai

# Further reading

## Books: Resampling methods and the bootstrap

*An Introduction to Statistical Learning*
James, Witten, Hastie, and Tibshirani

*Elements of Statistical Learning*
Hastie, Tibshirani, and Friedman

## Books: Permutation tests and randomization inference

*Causal Inference for Statistics, Social, and Biomedical Sciences*
Imbens and Rubin

*Field Experiments*
Gerber and Green

# Table of contents