Inference and Randomization

EC 425/525, Set 11

Edward Rubin
03 June 2019

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.

† Parts of these notes closely follow notes from Kosuke Imai.

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.

Inference and (re)randomization

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.

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.[†]

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).

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).

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 α derived from our dataset Z.

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 α 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 α derived from our dataset Z.
- We draw B bootstrap samples $b \in \{1, \ldots, B\}$.
- $Z^{\star 1}$ represents our first bootstrap sample ($b = 1$).

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 α 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 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 α 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 $\hat{\alpha}^{\star b}$

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

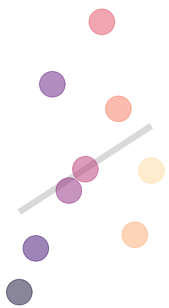
The bootstrap

More graphically

Z

7	8	9
4	5	6
1	2	3

$\hat{\beta} = 0.653$



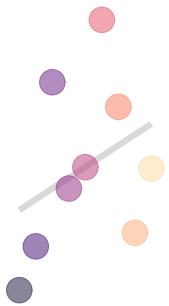
The bootstrap

More graphically

Z

7	8	9
4	5	6
1	2	3

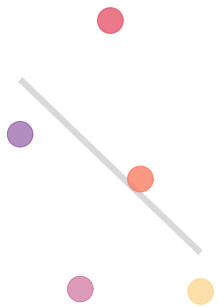
$\hat{\beta} = 0.653$



$Z^{\star 1}$

6	7	5
7	6	9
3	9	9

$\hat{\beta} = -0.961$



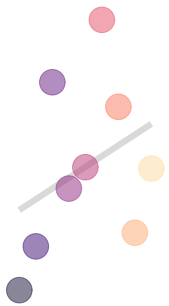
The bootstrap

More graphically

Z

7	8	9
4	5	6
1	2	3

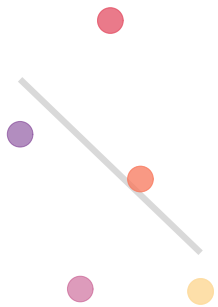
$\hat{\beta} = 0.653$



$Z^{\star 1}$

6	7	5
7	6	9
3	9	9

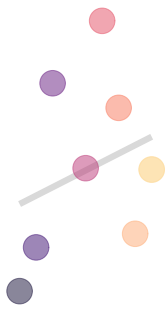
$\hat{\beta} = -0.961$



$Z^{\star 2}$

8	1	5
9	9	7
6	3	2

$\hat{\beta} = 0.51$



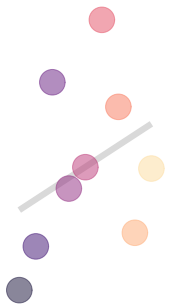
The bootstrap

More graphically

Z

7	8	9
4	5	6
1	2	3

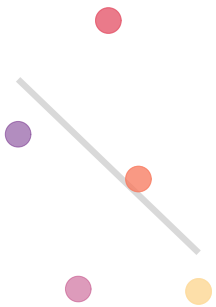
$\hat{\beta} = 0.653$



$Z^{\star 1}$

6	7	5
7	6	9
3	9	9

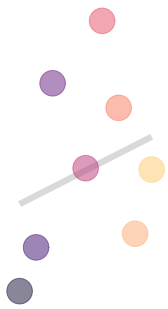
$\hat{\beta} = -0.961$



$Z^{\star 2}$

8	1	5
9	9	7
6	3	2

$\hat{\beta} = 0.51$

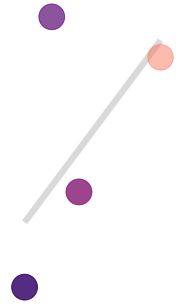


$\cdots$

$Z^{\star B}$

4	4	2
3	2	4
7	2	3

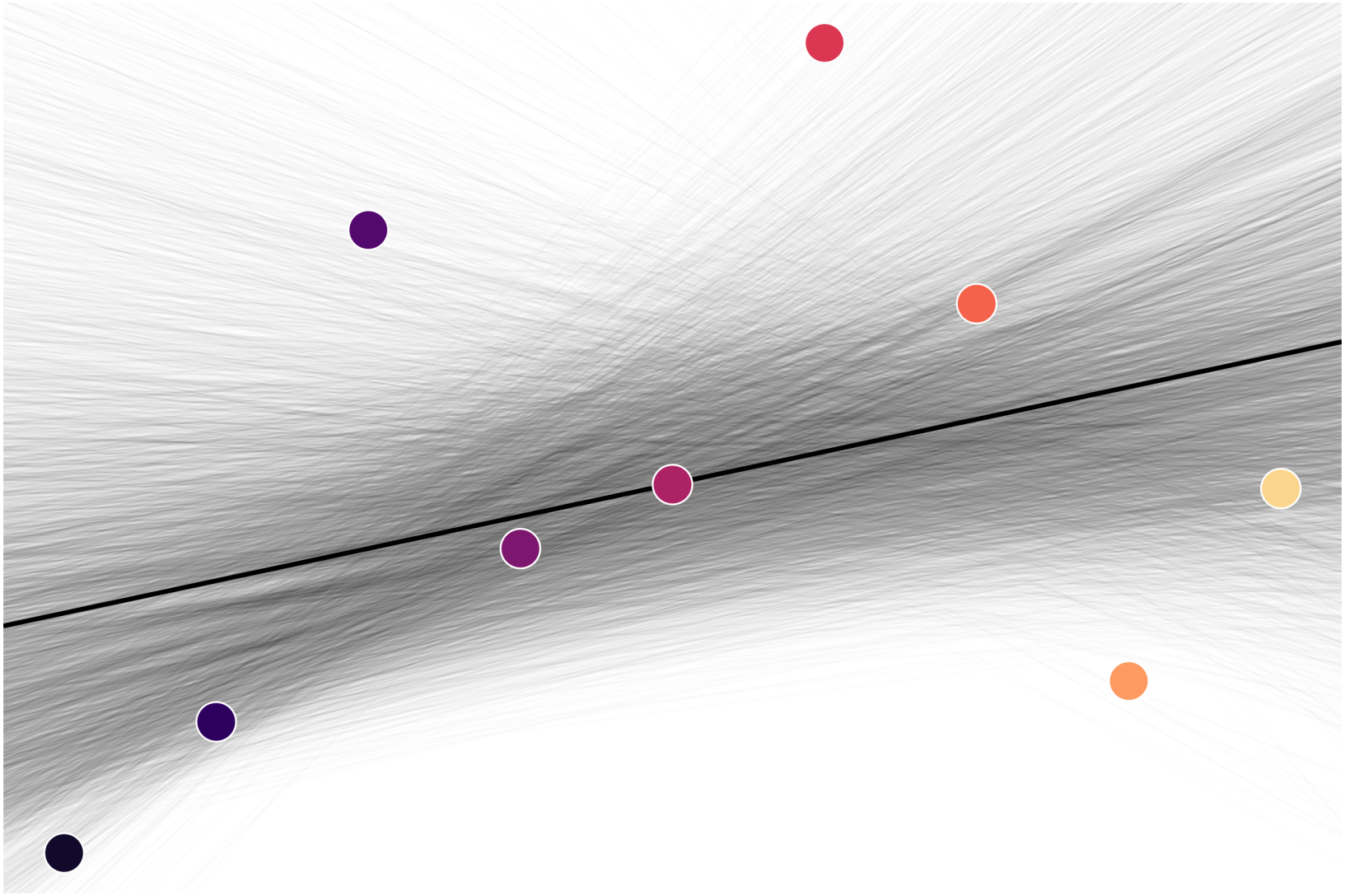
$\hat{\beta} = 1.338$



The bootstrap

Running this bootstrap 10,000 times

```
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 0.777.

The bootstrap

Comparison

In this 10,000-sample bootstrap, we calculate a standard error for $\hat{\beta}_1$ of approximately 0.777.

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

Not bad.

Permutation tests

Permutation tests

Motivation

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

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

Permutation tests

Motivation

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

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

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.*,

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

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.*,

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

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.*,

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

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 H_0!

Permutation tests

Motivation

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

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

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 H_0!
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.

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

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

Permutation tests

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?

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.

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

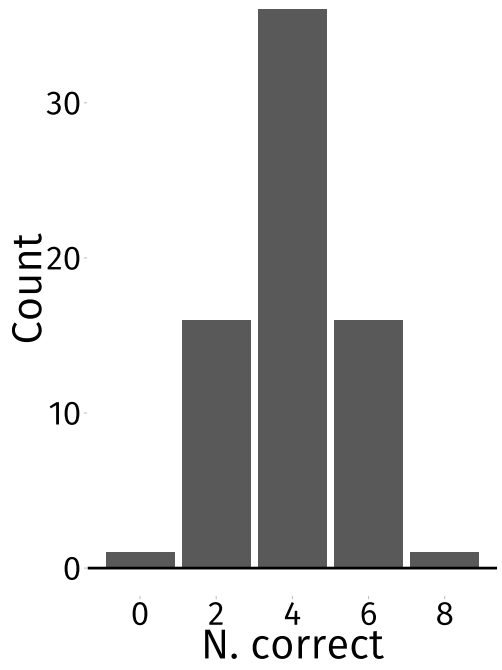
Tea drinkers with a vengeance

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

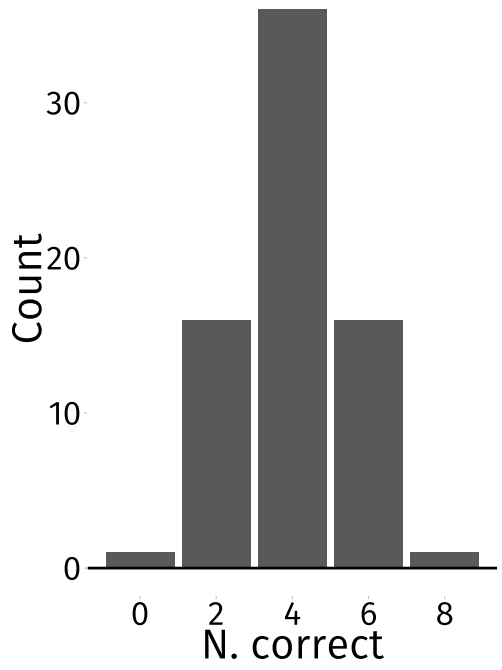
Cup	*Guess*	*Truth*	P_1	P_2	P_3	$\cdots$	P_{70}
1	m	m	m	m	m		t
2	t	t	~~m~~	~~m~~	~~m~~		t
3	t	t	~~m~~	~~m~~	~~m~~		t
4	m	m	m	~~t~~	~~t~~		~~t~~
5	m	m	~~t~~	m	~~t~~		m
6	t	t	t	t	~~m~~		~~m~~
7	t	t	t	t	t		~~m~~
8	m	m	~~t~~	~~t~~	~~t~~		m
		8/8	4/8	4/8	2/8		4/8



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



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

Permutation tests

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** α (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** α (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

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

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

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_0: $\boldsymbol{E}[\mathbf{Y}_{1i}] - \boldsymbol{E}[\mathbf{Y}_{0i}] = 0$).

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 H_0: $\boldsymbol{E}[\mathbf{Y}_{1i}] - \boldsymbol{E}[\mathbf{Y}_{0i}] = 0$).

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

Key insight

Our estimate (or test statistic) is a function of

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

Test statistics

(Which?)

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 and clustering

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.

Table of contents