Hypothesis testing

Seung-Ho An, University of Arizona

Agenda

The lady tasting tea

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea popured before the milk. You stop by a local tea shop and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea popured before the milk. You stop by a local tea shop and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

- You're skeptical that she can tell the difference, so you devise a test:

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea popured before the milk. You stop by a local tea shop and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

- You're skeptical that she can tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea popured before the milk. You stop by a local tea shop and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

- You're skeptical that she can tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups to friend in a **random** order

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea popured before the milk. You stop by a local tea shop and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

- You're skeptical that she can tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups to friend in a **random** order
  - Ask friend to pick which 4 of the 8 were milk-first

Friend picks out all 4 milk-first cups correctly!

```
library(TPDdata)
tea
```

```
## # A tibble: 8 x 2
##   truth      guess
##   <chr>      <chr>
## 1 tea-first  tea-first
## 2 milk-first milk-first
## 3 milk-first milk-first
## 4 tea-first  tea-first
## 5 tea-first  tea-first
## 6 milk-first milk-first
## 7 tea-first  tea-first
## 8 milk-first milk-first
```

Could she have been guessing at random? What would guessing look like?

```
set.seed(02138)
one_guess <- tea |>
  mutate(random_guess = sample(guess))
one_guess
```

```
## # A tibble: 8 x 3
##   truth      guess      random_guess
##   <chr>      <chr>      <chr>
## 1 tea-first  tea-first  milk-first
## 2 milk-first milk-first tea-first
## 3 milk-first milk-first tea-first
## 4 tea-first  tea-first  milk-first
## 5 tea-first  tea-first  tea-first
## 6 milk-first milk-first milk-first
## 7 tea-first  tea-first  tea-first
## 8 milk-first milk-first milk-first
```

4 correct in this random guess!

Another guess

```
another_guess <- tea |>
  mutate(random_guess = sample(guess))
another_guess
```

```
## # A tibble: 8 x 3
##   truth      guess      random_guess
##   <chr>      <chr>      <chr>
## 1 tea-first  tea-first  tea-first
## 2 milk-first milk-first tea-first
## 3 milk-first milk-first milk-first
## 4 tea-first  tea-first  tea-first
## 5 tea-first  tea-first  milk-first
## 6 milk-first milk-first milk-first
## 7 tea-first  tea-first  tea-first
## 8 milk-first milk-first milk-first
```

6 correct in this random guess!

We could enumerate all possible guesses. “Guessing” would mean choosing one of these at random:

```
##   Cup 1 Cup 2 Cup 3 Cup 4 Cup 5 Cup 6 Cup 7 Cup 8
## 1  milk  milk  milk  milk   tea   tea   tea   tea
## 2  milk  milk  milk   tea  milk   tea   tea   tea
## 3  milk  milk   tea  milk  milk   tea   tea   tea
## 4  milk   tea  milk  milk  milk   tea   tea   tea
## 5   tea  milk  milk  milk  milk   tea   tea   tea
## 6  milk  milk  milk   tea   tea  milk   tea   tea
```

[snip]

```
##    Cup 1 Cup 2 Cup 3 Cup 4 Cup 5 Cup 6 Cup 7 Cup 8
## 65   tea   tea   tea  milk  milk   tea  milk  milk
## 66  milk   tea   tea   tea   tea  milk  milk  milk
## 67   tea  milk   tea   tea   tea  milk  milk  milk
## 68   tea   tea  milk   tea   tea  milk  milk  milk
## 69   tea   tea   tea  milk   tea  milk  milk  milk
## 70   tea   tea   tea   tea  milk  milk  milk  milk
```

- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly?**

- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly?**
    - Only one way to choose all 4 correct cups

- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly?**
  - Only one way to choose all 4 correct cups
  - But 70 ways of choosing 4 cups among 8

- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly?**
  - Only one way to choose all 4 correct cups
  - But 70 ways of choosing 4 cups among 8
  - Choosing at random: picking each of these 70 with equal probability

- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly?**
    - Only one way to choose all 4 correct cups
    - But 70 ways of choosing 4 cups among 8
    - Choosing at random: picking each of these 70 with equal probability
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%

- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly?**
  - Only one way to choose all 4 correct cups
  - But 70 ways of choosing 4 cups among 8
  - Choosing at random: picking each of these 70 with equal probability
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%
- $\rightarrow$ the guessing hypothesis might be implausible
  - Impossible? No, because of random chance

Hypothesis tests

- Statistical hypothesis testing is a **thought experiment**

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?

Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**
    - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**

- Statistical hypothesis testing is a **thought experiment**
    - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
  - An analyst claims that 20% of Boston households are in poverty

Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
  - An analyst claims that 20% of Boston households are in poverty
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line

Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
  - An analyst claims that 20% of Boston households are in poverty
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line
  - Should you conclude that the analyst is wrong?

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
  - An analyst claims that 20% of Boston households are in poverty
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line
  - Should you conclude that the analyst is wrong?
- Example 2:

- Statistical hypothesis testing is a **thought experiment**
    - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
    - An analyst claims that 20% of Boston households are in poverty
    - You take a sample of 900 households and find that 23% of the sample is under the poverty line
    - Should you conclude that the analyst is wrong?
- Example 2:
    - Trump won 47.5% of the vote in the 2020 election

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
  - An analyst claims that 20% of Boston households are in poverty
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line
  - Should you conclude that the analyst is wrong?
- Example 2:
  - Trump won 47.5% of the vote in the 2020 election
  - Last YouGov poll of 1,363 likely voters said 44% planned to vote for Trump

Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**
  - Could our results just be due to randome chance?
- What would the world look like **if we knew the truth?**
- Example 1:
  - An analyst claims that 20% of Boston households are in poverty
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line
  - Should you conclude that the analyst is wrong?
- Example 2:
  - Trump won 47.5% of the vote in the 2020 election
  - Last YouGov poll of 1,363 likely voters said 44% planned to vote for Trump
  - Could the difference between the poll and the outcome be just due to random chance?

- **Null hypothesis**: Some statement about the population parameters

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - “Devil’s advocate” position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters

- **Null hypothesis**: Some statement about the population parameters
    - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
    - Usually that an observed difference is due to chance
    - Ex: poll drawn from the same population as all voters
    - Denoted H_0

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0
  - It is the opposite of the null hypothesis

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0
  - It is the opposite of the null hypothesis
  - An observed difference is real, not just due to chance

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0
  - It is the opposite of the null hypothesis
  - An observed difference is real, not just due to chance
  - Ex: polling for Trump is systematically wrong

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0
  - It is the opposite of the null hypothesis
  - An observed difference is real, not just due to chance
  - Ex: polling for Trump is systematically wrong
  - Denoted H_1 or H_a

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong
  - Usually that an observed difference is due to chance
  - Ex: poll drawn from the same population as all voters
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0
  - It is the opposite of the null hypothesis
  - An observed difference is real, not just due to chance
  - Ex: polling for Trump is systematically wrong
  - Denoted H_1 or H_a
- **Probabilistic** proof by contradiction: try to "disprove" the null

- Are we polling the same population as the actual voters?

- Are we polling the same population as the actual voters?
    - If so, how likely are we to see polling error this big by chance?

Hypothesis testing example

- Are we polling the same population as the actual voters?
    - If so, how likely are we to see polling error this big by chance?
- What is the parameter we want to learn about?

- Are we polling the same population as the actual voters?
  - If so, how likely are we to see polling error this big by chance?
- What is the parameter we want to learn about?
  - True population mean of the surveys, p

- Are we polling the same population as the actual voters?
  - If so, how likely are we to see polling error this big by chance?
- What is the parameter we want to learn about?
  - True population mean of the surveys, p
  - Null hypothesis: H_0: $p = 0.475$ (surveys drawing from same population)

- Are we polling the same population as the actual voters?
    - If so, how likely are we to see polling error this big by chance?
- What is the parameter we want to learn about?
    - True population mean of the surveys, p
    - Null hypothesis: H_0: $p = 0.475$ (surveys drawing from same population)
    - Alternative hypothesis: H_1: $p \neq 0.475$

- Are we polling the same population as the actual voters?
  - If so, how likely are we to see polling error this big by chance?
- What is the parameter we want to learn about?
  - True population mean of the surveys, p
  - Null hypothesis: H_0: $p = 0.475$ (surveys drawing from same population)
  - Alternative hypothesis: H_1: $p \neq 0.475$
- Data: poll has $\overline{X} = 0.44$ with $n = 1363$

- If the null were true, what should the distribution of the data be?

- If the null were true, what should the distribution of the data be?
    - X_i is 1 if respondent i will vote for Trump

Statistical thought experiment

- If the null were true, what should the distribution of the data be?
  - X_i is 1 if respondent i will vote for Trump
  - Under null, X_i is a coin flip with probability $p = 0.475$ of landing on "Trump"

- If the null were true, what should the distribution of the data be?
  - X_i is 1 if respondent i will vote for Trump
  - Under null, X_i is a coin flip with probability $p = 0.475$ of landing on “Trump”
  - $\Sigma_{i=1}^{2} X_i$ is the number in sample that will vote for Trump

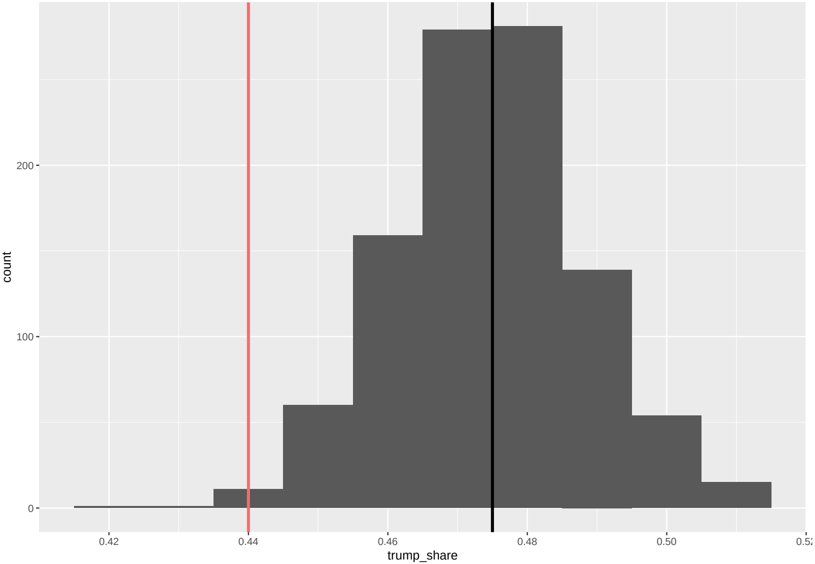
- If the null were true, what should the distribution of the data be?
  - X_i is 1 if respondent i will vote for Trump
  - Under null, X_i is a coin flip with probability $p = 0.475$ of landing on "Trump"
  - $\Sigma_{i=1}^{2} X_i$ is the number in sample that will vote for Trump
- We can simulate sums of coin flips using a function called `rbinom()`

- If the null were true, what should the distribution of the data be?
  - X_i is 1 if respondent i will vote for Trump
  - Under null, X_i is a coin flip with probability $p = 0.475$ of landing on "Trump"
  - $\Sigma_{i=1}^{2} X_i$ is the number in sample that will vote for Trump
- We can simulate sums of coin flips using a function called `rbinom()`
- Compare the distribution of proportions under the null to the observed proportion

- If the null were true, what should the distribution of the data be?
  - X_i is 1 if respondent i will vote for Trump
  - Under null, X_i is a coin flip with probability $p = 0.475$ of landing on "Trump"
  - $\Sigma_{i=1}^{2} X_i$ is the number in sample that will vote for Trump
- We can simulate sums of coin flips using a function called `rbinom()`
- Compare the distribution of proportions under the null to the observed proportion

```
null_dist1 <- tibble(
  trump_share = rbinom(n = 1000, size 1363, prob = 0.475) / 1363
)
ggplot(null_dist1, aes(x = trump_share)) +
  geom_histogram(binwidth=0.01) +
  geom_vline(xintercept = 0.44, color= "indianred1", size = 1.25) +
  geom_vline(xintercept = 0.475, size = 1.25)
```

Simulations of the reference distribution



p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?

p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
    - Smaller p-value $\rightsquigarrow$ stronger evidence against the null

p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
  - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
  - **NOT** the probability that the null is true

p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
    - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
    - **NOT** the probability that the null is true
- p-values are usually **two-sided**:

p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
  - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
  - **NOT** the probability that the null is true
- p-values are usually **two-sided**:
  - Observed error of 0.44 - 0.475 = -0.035 under the null

p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
  - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
  - **NOT** the probability that the null is true
- p-values are usually **two-sided**:
  - Observed error of 0.44 - 0.475 = -0.035 under the null
  - p-value is probability of sample proportions being less that 0.44 **plus**

p-value

The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
  - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
  - **NOT** the probability that the null is true
- p-values are usually **two-sided**:
  - Observed error of 0.44 - 0.475 = -0.035 under the null
  - p-value is probability of sample proportions being less that 0.44 **plus**
  - Probability of sample proportions being greater than 0.475+0.035=0.51

p-value

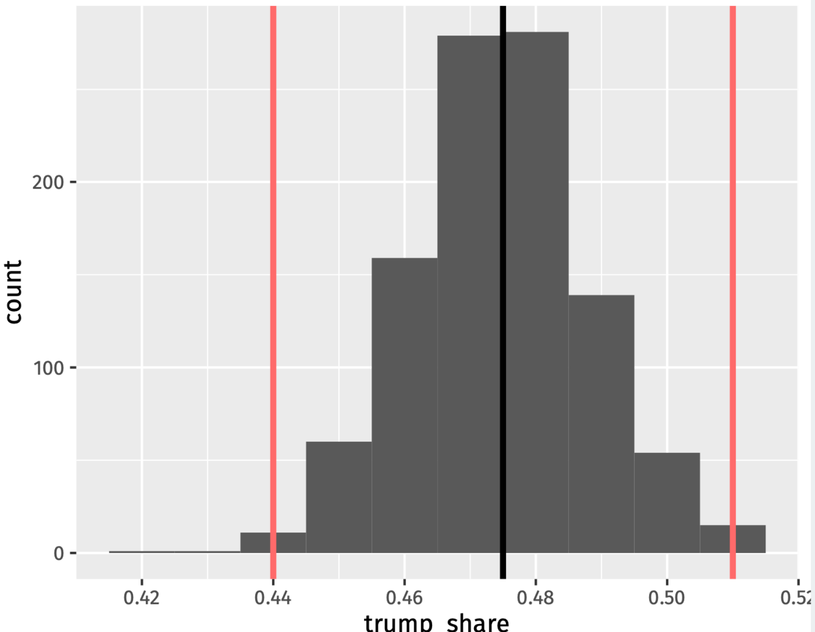
The **p-value** is the probability of observing data as or more extreme as our data under the null

- If the null is true, how often would we expect polling errors this big?
  - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
  - **NOT** the probability that the null is true
- p-values are usually **two-sided**:
  - Observed error of 0.44 - 0.475 = -0.035 under the null
  - p-value is probability of sample proportions being less that 0.44 **plus**
  - Probability of sample proportions being greater than 0.475+0.035=0.51

```
mean(null_dist1$trump_share < 0.44) + mean(null_dist1$trump_share > 0.51)
```

```
## [1] 0.01
```

Two-sided p-value



- Sometimes our hypothesis is directional

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction

One-sided tests

- Sometimes our hypothesis is directional
    - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters

One-sided tests

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative:** polls underestimate Trump support

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative:** polls underestimate Trump support
  - $H_1 : p < 0.475$

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative:** polls underestimate Trump support
  - $H_1 : p < 0.475$
- Makes the p-value one-sided:

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative:** polls underestimate Trump support
  - $H_1 : p < 0.475$
- Makes the p-value one-sided:
  - What's the probability of a random sample underestimating Trump support by as much as we see in the sample?

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative:** polls underestimate Trump support
  - $H_1 : p < 0.475$
- Makes the p-value one-sided:
  - What's the probability of a random sample underestimating Trump support by as much as we see in the sample?
  - Always smaller than a two-sided p-value

One-sided tests

- Sometimes our hypothesis is directional
  - We only consider evidence against the null from one direction
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative:** polls underestimate Trump support
  - $H_1 : p < 0.475$
- Makes the p-value one-sided:
  - What's the probability of a random sample underestimating Trump support by as much as we see in the sample?
  - Always smaller than a two-sided p-value

```
mean(null_dist1$trump_share < 0.44)
```

```
## [1] 0.005
```

- Tests usually end with a decision to reject the null or not

Rejecting the null

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
  - **Test level** αL the threshold for a test

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
  - **Test level** αL the threshold for a test
  - Decision rule: ""reject the null if the p-value is below α"

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
  - **Test level** αL the threshold for a test
  - Decision rule: ""reject the null if the p-value is below α"
  - Otherwise ""fail to reject", not "accept the null"

Rejecting the null

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
  - **Test level** αL the threshold for a test
  - Decision rule: ""reject the null if the p-value is below α"
  - Otherwise ""fail to reject", not "accept the null"
- Common (arbitrary) thresholds:

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
  - **Test level** αL the threshold for a test
  - Decision rule: ""reject the null if the p-value is below α"
  - Otherwise ""fail to reject", not "accept the null"
- Common (arbitrary) thresholds:
  - $p \geq 0.1$ "not statistically significant"

Rejecting the null

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
  - **Test level** αL the threshold for a test
  - Decision rule: ""reject the null if the p-value is below α"
  - Otherwise ""fail to reject", not "accept the null"
- Common (arbitrary) thresholds:
  - $p \geq 0.1$ "not statistically significant"
  - $p < 0.05$ "statistically significant"

- Tests usually end with a decision to reject the null or not
- Choose a threshold below which you'll reject the null
    - **Test level** αL the threshold for a test
    - Decision rule: ""reject the null if the p-value is below α"
    - Otherwise ""fail to reject", not "accept the null"
- Common (arbitrary) thresholds:
    - $p \geq 0.1$ "not statistically significant"
    - $p < 0.05$ "statistically significant"
    - $p < 0.01$ "statistically significant"

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true
  - $\rightsquigarrow$ 5% of the time we will reject the null when it is actually true

Testing errors

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true
  - $\rightsquigarrow$ 5% of the time we will reject the null when it is actually true
- Test errors:

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true
  - $\rightsquigarrow$ 5% of the time we will reject the null when it is actually true
- Test errors:

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

Which types of error is worse?

Testing errors

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true
  - $\rightsquigarrow$ 5% of the time we will reject the null when it is actually true
- Test errors:

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

Which types of error is worse?

- Type I error because it's the worst
  - "Convicting" an innocent null hypothesis

Testing errors

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true
  - $\rightsquigarrow$ 5% of the time we will reject the null when it is actually true
- Test errors:

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

Which types of error is worse?

- Type I error because it's the worst
  - "Convicting" an innocent null hypothesis
- Type II error less serious

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true
    - $\rightsquigarrow$ 5% of the time we will reject the null when it is actually true
- Test errors:

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

Which types of error is worse?

- Type I error because it's the worst
    - "Convicting" an innocent null hypothesis
- Type II error less serious
    - Missed out on an awesome finding

Hypothesis testing using infer

```
library(infer)
gss
```

```
## # A tibble: 500 x 11
##     year   age sex    college   partyid hompop hours income class finrela we
##    <dbl> <dbl> <fct>  <fct>     <fct>    <dbl> <dbl> <ord>  <fct> <fct>    <
##  1  2014    36 male   degree    ind          3    50 $2500~ midd~ below ~  0
##  2  1994    34 female no degree rep          4    31 $2000~ work~ below ~  1
##  3  1998    24 male   degree    ind          1    40 $2500~ work~ below ~  0
##  4  1996    42 male   no degree ind          4    40 $2500~ work~ above ~  1
##  5  1994    31 male   degree    rep          2    40 $2500~ midd~ above ~  1
##  6  1996    32 female no degree rep          4    53 $2500~ midd~ average  1
##  7  1990    48 female no degree dem          2    32 $2500~ work~ below ~  1
##  8  2016    36 female degree    ind          1    20 $2500~ midd~ above ~  0
##  9  2000    30 female degree    rep          5    40 $2500~ midd~ average  1
## 10  1998    33 female no degree dem          2    40 $1500~ work~ far be~  0
## # i 490 more rows
```

What is the average hours worked?

`dplyr` way:

```
gss |>
  summarize(mean(hours))
```

```
## # A tibble: 1 x 1
##   `mean(hours)`
##           <dbl>
## 1          41.4
```

`infer` way:

```
observed_mean <- gss |>
  specify(response = hours) |>
  calculate(stat="mean")
observed_mean
```

```
## Response: hours (numeric)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  41.4
```

Could we get a mean this different from 40 hours if that was the true population average of hours worked?

Could we get a mean this different from 40 hours if that was the true population average of hours worked?

Null and alternative:

$$H_0 : \mu_{hours} = 40$$

$$H_1 : \mu_{hours} \neq 40$$

Could we get a mean this different from 40 hours if that was the true population average of hours worked?

Null and alternative:

$$H_0 : \mu_{hours} = 40$$

$$H_1 : \mu_{hours} \neq 40$$

How do we perform this test using infer? The **bootstrap!**

Specifying the hypotheses

```
gss |>
  specify(response = hours) |>
  hypothesize(null = "point", mu = 40)
```

```
## Response: hours (numeric)
## Null Hypothesis: point
## # A tibble: 500 x 1
##    hours
##    <dbl>
##  1    50
##  2    31
##  3    40
##  4    40
##  5    40
##  6    53
##  7    32
##  8    20
##  9    40
## 10    40
## # i 490 more rows
```

Generating the null distribution

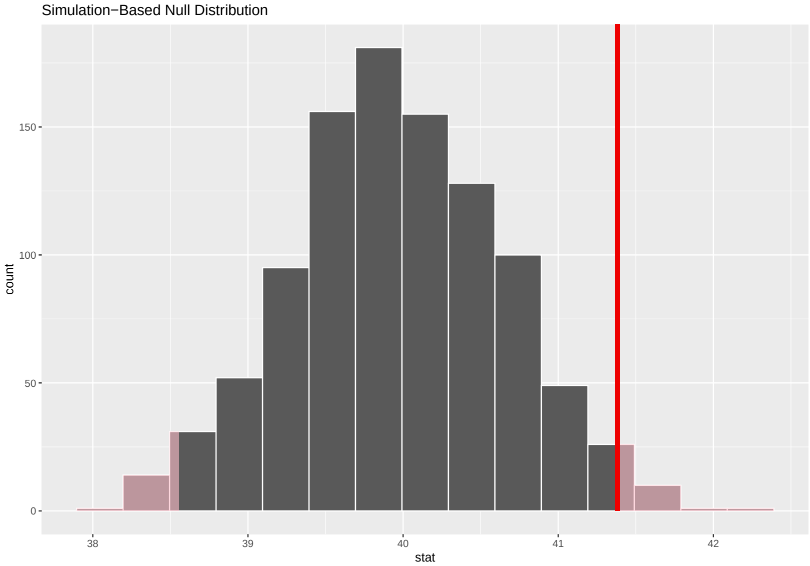
We can use the bootstrap to determine how much variation there will be around 50 in the null distribution

```r
null_dist2 <- gss |>
  specify(response = hours) |>
  hypothesize(null = "point", mu = 40) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean")
null_dist2
```

```
## Response: hours (numeric)
## Null Hypothesis: point
## # A tibble: 1,000 x 2
##    replicate  stat
##        <int> <dbl>
##  1         1  40.3
##  2         2  39.8
##  3         3  40.0
##  4         4  39.2
##  5         5  40.3
##  6         6  40.2
##  7         7  40.4
##  8         8  39.5
##  9         9  39.8
```

We can visualize our bootstrapped null distribution and the p-value as a shaded region:

```
null_dist2 |>
  visualize() +
  shade_p_value(observed_mean,
                direction = "two-sided")
```


Simulation−Based Null Distribution
count
150
100
50
0
38
39
40
41
42
stat

Two-sample tests

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
    - Control: no mailer

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
  - Control: no mailer
  - Civic duty: mailer saying voting is your civic duty

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
  - Control: no mailer
  - Civic duty: mailer saying voting is your civic duty
  - Hawthorne: a “we’re watching you” message

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
    - Control: no mailer
    - Civic duty: mailer saying voting is your civic duty
    - Hawthorne: a “we’re watching you” message
    - Neighbors: naming-and-shaming social pressure mailer

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
  - Control: no mailer
  - Civic duty: mailer saying voting is your civic duty
  - Hawthorne: a "we're watching you" message
  - Neighbors: naming-and-shaming social pressure mailer
- Outcome: whether household members voted or not

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
  - Control: no mailer
  - Civic duty: mailer saying voting is your civic duty
  - Hawthorne: a "we're watching you" message
  - Neighbors: naming-and-shaming social pressure mailer
- Outcome: whether household members voted or not
- We'll focus on Neighbors vs. Control

Social pressure experiment

- Experimental study where each household for 2006 MI primary was randomly assigned to one of 4 conditions:
    - Control: no mailer
    - Civic duty: mailer saying voting is your civic duty
    - Hawthorne: a "we're watching you" message
    - Neighbors: naming-and-shaming social pressure mailer
- Outcome: whether household members voted or not
- We'll focus on Neighbors vs. Control
- Randomized implies samples are **independent**

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	______

Social pressure data

```
social <-read.csv("data/social.csv")
social <- as_tibble(social)
```

- Parameter: **population ATE** $\mu_T - \mu_C$

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment

Two-sample hypotheses

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment
  - μ_C: Turnout rate in the population if everyone received control

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment
  - μ_C: Turnout rate in the population if everyone received control
- Goal: learn about the population difference in means

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment
  - μ_C: Turnout rate in the population if everyone received control
- Goal: learn about the population difference in means
- Usual null hypothesis: no difference in population means (ATE=0)

- Parameter: **population ATE** $\mu_T - \mu_C$
    - μ_T: Turnout rate in the population if everyone received treatment
    - μ_C: Turnout rate in the population if everyone received control
- Goal: learn about the population difference in means
- Usual null hypothesis: no difference in population means (ATE=0)
    - Null: $H_0 : \mu_T - \mu_C = 0$

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment
  - μ_C: Turnout rate in the population if everyone received control
- Goal: learn about the population difference in means
- Usual null hypothesis: no difference in population means (ATE=0)
  - Null: $H_0 : \mu_T - \mu_C = 0$
  - Two-sided alternative: $H_1 : \mu_T - \mu_C \neq 0$

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment
  - μ_C: Turnout rate in the population if everyone received control
- Goal: learn about the population difference in means
- Usual null hypothesis: no difference in population means (ATE=0)
  - Null: $H_0 : \mu_T - \mu_C = 0$
  - Two-sided alternative: $H_1 : \mu_T - \mu_C \neq 0$
- In words: are the differences in sample means just due to chance?

How do we generate draws of the difference in means under the null?

$H_0 : \mu_T - \mu_C = 0$

How do we generate draws of the difference in means under the null?

$H_0 : \mu_T - \mu_C = 0$

If the voting distribution is the same in the treatment and control groups, we could randomly swap who is labelled as treated and who is labelled as control and it shouldn't matter

How do we generate draws of the difference in means under the null?

$H_0 : \mu_T - \mu_C = 0$

If the voting distribution is the same in the treatment and control groups, we could randomly swap who is labelled as treated and who is labelled as control and it shouldn't matter

Permutation test: generate the null distribution by permuting the group labels and see the resulting distribution of differences in proportions

```r
social <- social |>
  filter(messages %in% c("Neighbors", "Control"))

social |>
  mutate(messages_permute = sample(messages)) |>
  select(primary2006, messages, messages_permute)
```

```
## # A tibble: 229,444 x 3
##    primary2006 messages messages_permute
##          <int> <chr>    <chr>
##  1           0 Control  Control
##  2           1 Control  Control
##  3           1 Control  Neighbors
##  4           0 Control  Control
##  5           0 Control  Control
##  6           1 Control  Control
##  7           0 Control  Control
##  8           1 Control  Control
##  9           1 Control  Neighbors
## 10           1 Control  Control
## # i 229,434 more rows
```

Two-sample permutation tests with `infer`

Calculating the difference in proportion

`infer` functions with binary outcomes work best with factor variables:

```
social <- social |>
  mutate(turnout = if_else(primary2006==1, "Voted", "Didn't Vote"))

est_ate <- social |>
  specify(turnout ~ messages, success = "Voted") |>
  calculate(stat = "diff in props", order = c("Neighbors", "Control"))
est_ate
```

```
## Response: turnout (factor)
## Explanatory: messages (factor)
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 0.0813
```

Specifying the relationship of interest

`infer` functions with binary outcomes work best with factor variables:

```
social |>
  specify(turnout ~ messages, success = "Voted")
```

```
## Response: turnout (factor)
## Explanatory: messages (factor)
## # A tibble: 229,444 x 2
##    turnout     messages
##    <fct>       <fct>
##  1 Didn't Vote Control
##  2 Voted       Control
##  3 Voted       Control
##  4 Didn't Vote Control
##  5 Didn't Vote Control
##  6 Voted       Control
##  7 Didn't Vote Control
##  8 Voted       Control
##  9 Voted       Control
## 10 Voted       Control
## # i 229,434 more rows
```

The null for these two-sample tests is called "independence" for the `infer` package because the assumption is that the two variables are statistically independent

```
social |>
  specify(turnout ~ messages, success = "Voted") |>
  hypothesize(null = "independence")
```

```
## Response: turnout (factor)
## Explanatory: messages (factor)
## Null Hypothesis: independence
## # A tibble: 229,444 x 2
##    turnout     messages
##    <fct>       <fct>
##  1 Didn't Vote Control
##  2 Voted       Control
##  3 Voted       Control
##  4 Didn't Vote Control
##  5 Didn't Vote Control
##  6 Voted       Control
##  7 Didn't Vote Control
##  8 Voted       Control
##  9 Voted       Control
## 10 Voted       Control
```

Generating the permutations

We can tell `infer` to do our permutation test by using the argument `type = "permute"` to `generate()`:

```
## Response: turnout (factor)
## Explanatory: messages (factor)
## Null Hypothesis: independence
## # A tibble: 229,444,000 x 3
## # Groups:   replicate [1,000]
##    turnout     messages replicate
##    <fct>       <fct>        <int>
##  1 Voted       Control          1
##  2 Voted       Control          1
##  3 Didn't Vote Control          1
##  4 Voted       Control          1
##  5 Voted       Control          1
##  6 Didn't Vote Control          1
##  7 Didn't Vote Control          1
##  8 Didn't Vote Control          1
##  9 Didn't Vote Control          1
## 10 Didn't Vote Control          1
## # i 229,443,990 more rows
```

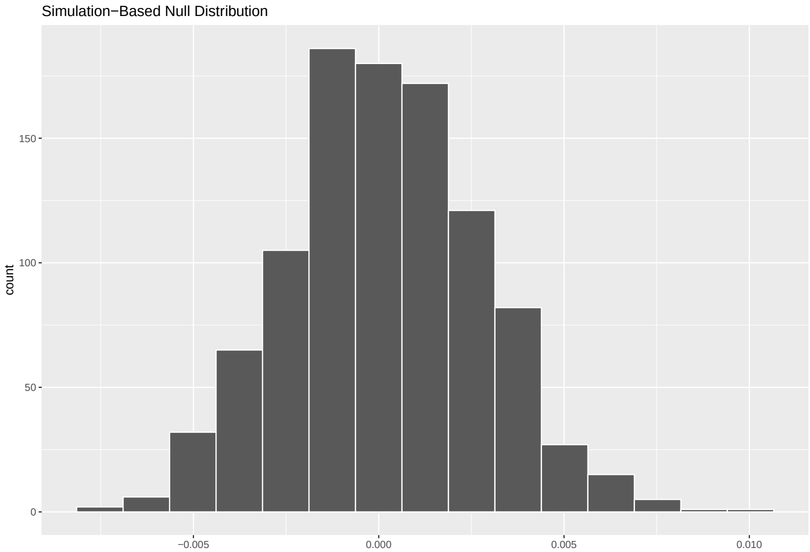
Calculating the diff in proportions in each sample

```
null_dist3 <- social |>
  specify(turnout ~ messages, success = "Voted") |>
  hypothesize(null="independence") |>
  generate(reps = 1000, type="permute") |>
  calculate(stat="diff in props", order = c("Neighbors", "C
```

```r
null_dist3
```

```
## Response: turnout (factor)
## Explanatory: messages (factor)
## Null Hypothesis: independence
## # A tibble: 1,000 x 2
##    replicate        stat
##        <int>       <dbl>
##  1         1  0.00173
##  2         2  0.00235
##  3         3  0.00845
##  4         4 -0.00211
##  5         5 -0.000000932
##  6         6 -0.00295
##  7         7 -0.00267
##  8         8  0.00104
##  9         9 -0.000975
## 10        10  0.00421
## # i 990 more rows
```

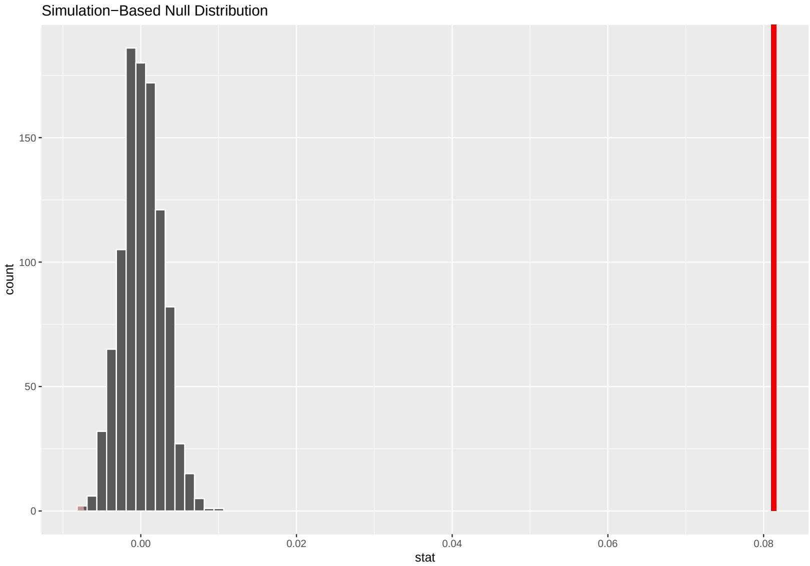
```
null_dist3 |>
  visualize()
```



```r
ate_pval <- null_dist3 |>
  get_p_value(obs_stat = est_ate, direction = "both")
ate_pval
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Visualizing p-values



Confidence intervals vs. hypothesis testing

- Hypothesis testing is probably the dominant interpretive tool in applied social science

Confidence intervals vs. hypothesis testing

- Hypothesis testing is probably the dominant interpretive tool in applied social science
- Nontheless, confidence intervals are generally much more informative

Confidence intervals vs. hypothesis testing

- Hypothesis testing is probably the dominant interpretive tool in applied social science
- Nontheless, confidence intervals are generally much more informative
- Hypothesis tests only allow you to confirm (or fail to confirm) that there's a non-zero effect; but this effect could be so tiny that we don't care

Confidence intervals vs. hypothesis testing

- Hypothesis testing is probably the dominant interpretive tool in applied social science
- Nontheless, confidence intervals are generally much more informative
- Hypothesis tests only allow you to confirm (or fail to confirm) that there's a non-zero effect; but this effect could be so tiny that we don't care
- In hypothesis testing, failing to confirm the null tells us virtually nothing with a confidence interval, it's possible to conclude that there is-at most-a tiny effect (if the confidence interval contains only a small range of values close to zero)

Class exercise

Let's design a research!

Think about your dependent and independent variables, first

1. dependent variable:
2. independent variable:

What would be the unit of your analyses?

What are your null and alternative hypotheses?

Which research methods would you employ to test your hypotheses?