Inference for 2 Proportions

Prof. Wells

STA 209, 4/21/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Calculate confidence intervals for proportions
- Use the formula for standard error to determine necessary sample size
- Investigate the theoretical distribution for differences in proportions
- Calculate confidence intervals and conduct hypothesis tests for differences in proportions

Section 1

Confidence Intervals

Critical Values

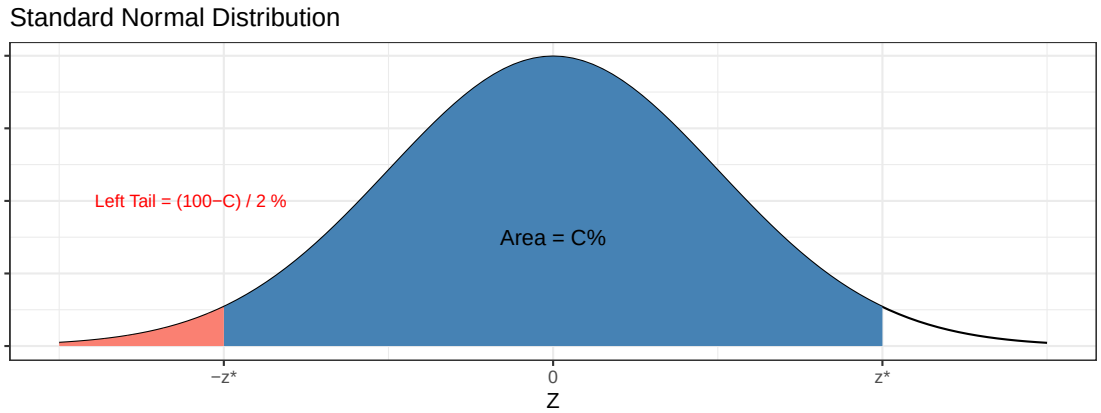
- The **critical value** z^* for a $C\%$ confidence interval is the value so that $C\%$ of area is between $-z^*$ and z^* in the standard Normal distribution

Critical Values

- The **critical value** z^* for a $C\%$ confidence interval is the value so that $C\%$ of area is between $-z^*$ and z^* in the standard Normal distribution
  - That is, the critical value of $C\%$ confidence is the $C + \frac{1-C}{2}$ percentile of the standard Normal distribution

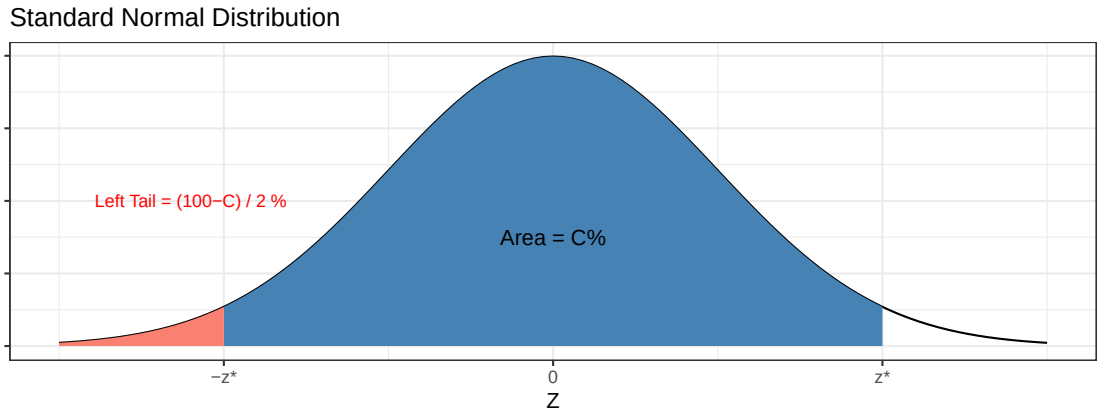
Critical Values

- The **critical value** z^* for a $C\%$ confidence interval is the value so that $C\%$ of area is between $-z^*$ and z^* in the standard Normal distribution
  - That is, the critical value of $C\%$ confidence is the $C + \frac{1-C}{2}$ percentile of the standard Normal distribution



Critical Values

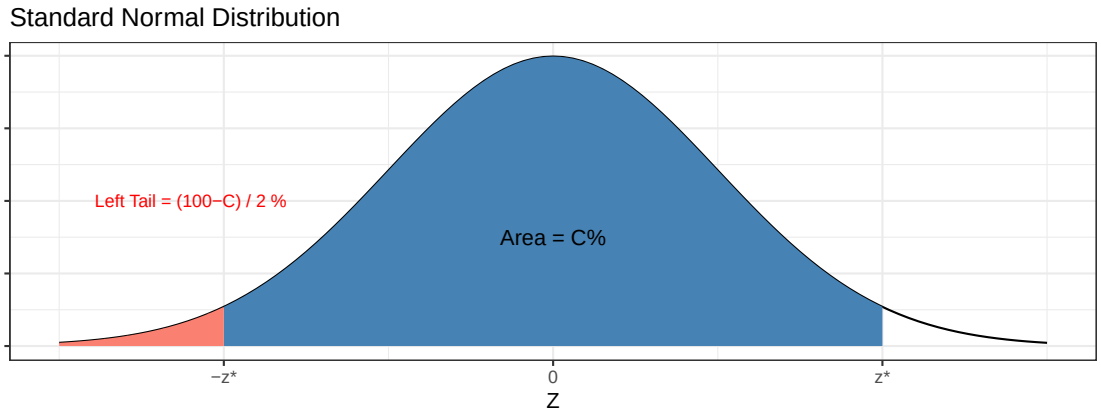
- The **critical value** z^* for a $C\%$ confidence interval is the value so that $C\%$ of area is between $-z^*$ and z^* in the standard Normal distribution
  - That is, the critical value of $C\%$ confidence is the $C + \frac{1-C}{2}$ percentile of the standard Normal distribution



- The critical value for 95% confidence is the $95 + \frac{100-95}{2} = 97.5$ percentile

Critical Values

- The **critical value** z^* for a $C\%$ confidence interval is the value so that $C\%$ of area is between $-z^*$ and z^* in the standard Normal distribution
  - That is, the critical value of $C\%$ confidence is the $C + \frac{1-C}{2}$ percentile of the standard Normal distribution



- The critical value for 95% confidence is the $95 + \frac{100-95}{2} = 97.5$ percentile

```
qnorm(.975, mean = 0, sd = 1) # The 97.5 percentile is the .975 quantile
```

```
## [1] 1.959964
```

Confidence Intervals

If the sample statistic is approximately Normal, the $C\%$ confidence interval is

$$\text{statistic} \pm z^* \cdot SE$$

where z^* is the critical value confidence and SE is the standard error of the statistic

Confidence Intervals

If the sample statistic is approximately Normal, the $C\%$ confidence interval is

$$\text{statistic} \pm z^* \cdot SE$$

where z^* is the critical value confidence and SE is the standard error of the statistic

- The standard error for a sample proportion $\hat{p}$ is $SE = \sqrt{\frac{p(1-p)}{n}}$.

Confidence Intervals

If the sample statistic is approximately Normal, the $C\%$ confidence interval is

$$\text{statistic} \pm z^* \cdot SE$$

where z^* is the critical value confidence and SE is the standard error of the statistic

- The standard error for a sample proportion $\hat{p}$ is $SE = \sqrt{\frac{p(1-p)}{n}}$.
  - But since we don't know p, we estimate it in the SE formula with $\hat{p}$:

$$SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Intervals

If the sample statistic is approximately Normal, the $C\%$ confidence interval is

$$\text{statistic} \pm z^* \cdot SE$$

where z^* is the critical value confidence and SE is the standard error of the statistic

- The standard error for a sample proportion $\hat{p}$ is $SE = \sqrt{\frac{p(1-p)}{n}}$.
  - But since we don't know p, we estimate it in the SE formula with $\hat{p}$:

$$SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Theorem

Suppose an SRS of size n is collected from a population with parameter p. If n is large enough so that both $n\hat{p}$ and $n(1-\hat{p})$ are at least 10, then the confidence interval for p is

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"
  - 48% of respondents indicated Donald Trump, while 41% indicated Joe Biden. Then remaining 11% indicated another preference.

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"
  - 48% of respondents indicated Donald Trump, while 41% indicated Joe Biden. Then remaining 11% indicated another preference.
  - Due to sampling, it's unlikely that **exactly** 48% of Iowans planned to vote for Donald Trump in the 2020 election. But we can create a confidence interval to estimate the true proportion p.

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"
  - 48% of respondents indicated Donald Trump, while 41% indicated Joe Biden. Then remaining 11% indicated another preference.
  - Due to sampling, it's unlikely that **exactly** 48% of Iowans planned to vote for Donald Trump in the 2020 election. But we can create a confidence interval to estimate the true proportion p.
- Using the poll data, $\hat{p} = 0.48$, which means the standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.48(1-0.48)}{814}} = 0.0175$$

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"
  - 48% of respondents indicated Donald Trump, while 41% indicated Joe Biden. Then remaining 11% indicated another preference.
  - Due to sampling, it's unlikely that **exactly** 48% of Iowans planned to vote for Donald Trump in the 2020 election. But we can create a confidence interval to estimate the true proportion p.
- Using the poll data, $\hat{p} = 0.48$, which means the standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.48(1-0.48)}{814}} = 0.0175$$

- Previously, we calculated the critical value z^* for 95% confidence: $z^* = 1.96$

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"
  - 48% of respondents indicated Donald Trump, while 41% indicated Joe Biden. Then remaining 11% indicated another preference.
  - Due to sampling, it's unlikely that **exactly** 48% of Iowans planned to vote for Donald Trump in the 2020 election. But we can create a confidence interval to estimate the true proportion p.
- Using the poll data, $\hat{p} = 0.48$, which means the standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.48(1-0.48)}{814}} = 0.0175$$

- Previously, we calculated the critical value z^* for 95% confidence: $z^* = 1.96$
- Putting this all together, our confidence interval is

$$\hat{p} \pm z^* \cdot SE \quad \Longleftrightarrow \quad 0.48 \pm 1.96 \cdot 0.0175 \quad \Longleftrightarrow \quad (0.4457, 0.5143)$$

Presidential Polling

- An October 2020 poll by the firm Selzer & Co, sponsored by the Des Moines Register, asked 814 likely Iowa voters: "If the general election were held today, for whom would you vote?"
  - 48% of respondents indicated Donald Trump, while 41% indicated Joe Biden. Then remaining 11% indicated another preference.
  - Due to sampling, it's unlikely that **exactly** 48% of Iowans planned to vote for Donald Trump in the 2020 election. But we can create a confidence interval to estimate the true proportion p.
- Using the poll data, $\hat{p} = 0.48$, which means the standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.48(1-0.48)}{814}} = 0.0175$$

- Previously, we calculated the critical value z^* for 95% confidence: $z^* = 1.96$
- Putting this all together, our confidence interval is

$$\hat{p} \pm z^* \cdot SE \quad \Longleftrightarrow \quad 0.48 \pm 1.96 \cdot 0.0175 \quad \Longleftrightarrow \quad (0.4457, 0.5143)$$

- The poll estimated between 44.6% and 51.4% of Iowans intended to vote for Trump, with confidence 95%.

Confidence Intervals in ‘infer1

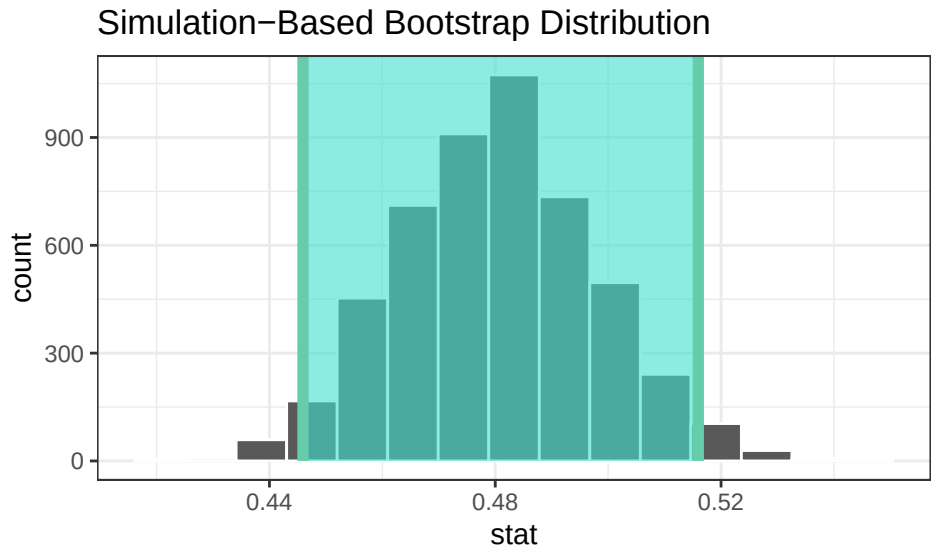
- How does this compare to the bootstrap method?

Confidence Intervals in 'infer1

- How does this compare to the bootstrap method?

```
pres_poll %>% specify(response = vote, success = "Trump") %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.446    0.516
```



Sample Sizes

- One advantage of the theory-based method is it allows us to determine the sample size needed for a desired margin of error.

Sample Sizes

- One advantage of the theory-based method is it allows us to determine the sample size needed for a desired margin of error.

$$\text{MoE} = z^* \cdot SE = z^* \cdot \sqrt{\frac{p(1-p)}{n}}$$

Sample Sizes

- One advantage of the theory-based method is it allows us to determine the sample size needed for a desired margin of error.

$$\text{MoE} = z^* \cdot SE = z^* \cdot \sqrt{\frac{p(1-p)}{n}}$$

- Suppose we want to estimate p to within Margin of Error of 0.01, with 95% confidence. We can solve the Margin of Error equation for n.

$$\text{MoE} = z^* \cdot \sqrt{\frac{p(1-p)}{n}} \quad \Longleftrightarrow \quad n = \left(\frac{z^*}{\text{MoE}}\right)^2 p(1-p)$$

Sample Sizes

- One advantage of the theory-based method is it allows us to determine the sample size needed for a desired margin of error.

$$\text{MoE} = z^* \cdot SE = z^* \cdot \sqrt{\frac{p(1-p)}{n}}$$

- Suppose we want to estimate p to within Margin of Error of 0.01, with 95% confidence. We can solve the Margin of Error equation for n.

$$\text{MoE} = z^* \cdot \sqrt{\frac{p(1-p)}{n}} \quad \Longleftrightarrow \quad n = \left(\frac{z^*}{\text{MoE}}\right)^2 p(1-p)$$

- There is a problem! We don't know p (it's what we are trying to estimate). And we also don't have $\hat{p}$ either (we need to determine a sample size before we gather data)

Sample Sizes

- One advantage of the theory-based method is it allows us to determine the sample size needed for a desired margin of error.

$$\text{MoE} = z^* \cdot SE = z^* \cdot \sqrt{\frac{p(1-p)}{n}}$$

- Suppose we want to estimate p to within Margin of Error of 0.01, with 95% confidence. We can solve the Margin of Error equation for n.

$$\text{MoE} = z^* \cdot \sqrt{\frac{p(1-p)}{n}} \quad \Longleftrightarrow \quad n = \left(\frac{z^*}{\text{MoE}}\right)^2 p(1-p)$$

- There is a problem! We don't know p (it's what we are trying to estimate). And we also don't have $\hat{p}$ either (we need to determine a sample size before we gather data)
  - Instead, we'll use our best guess for p using information available. We can also default to using $p = 0.5$ (corresponding to the most conservative estimate of sample size)

Sample Sizes

- One advantage of the theory-based method is it allows us to determine the sample size needed for a desired margin of error.

$$\text{MoE} = z^* \cdot SE = z^* \cdot \sqrt{\frac{p(1-p)}{n}}$$

- Suppose we want to estimate p to within Margin of Error of 0.01, with 95% confidence. We can solve the Margin of Error equation for n.

$$\text{MoE} = z^* \cdot \sqrt{\frac{p(1-p)}{n}} \quad \Longleftrightarrow \quad n = \left(\frac{z^*}{\text{MoE}}\right)^2 p(1-p)$$

- There is a problem! We don't know p (it's what we are trying to estimate). And we also don't have $\hat{p}$ either (we need to determine a sample size before we gather data)
  - Instead, we'll use our best guess for p using information available. We can also default to using $p = 0.5$ (corresponding to the most conservative estimate of sample size)
- In this case, using $p = 0.5$, the necessary sample size is

$$n = \left(\frac{1.96}{0.01}\right)^2 0.5 \cdot (1 - 0.5) = 9604$$

Section 2

Difference in Proportions

Difference in Proportions

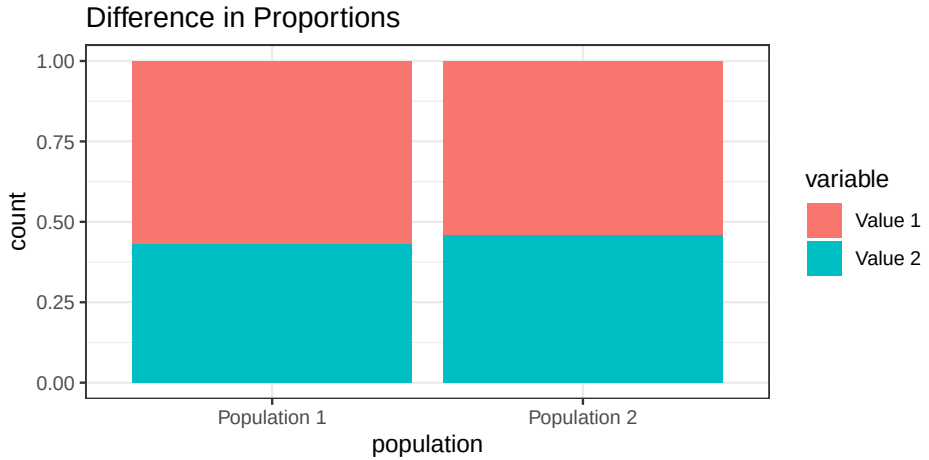
- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.

Difference in Proportions

- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.
- That is, we want to know the value of the difference $p_1 - p_2$ in proportion.

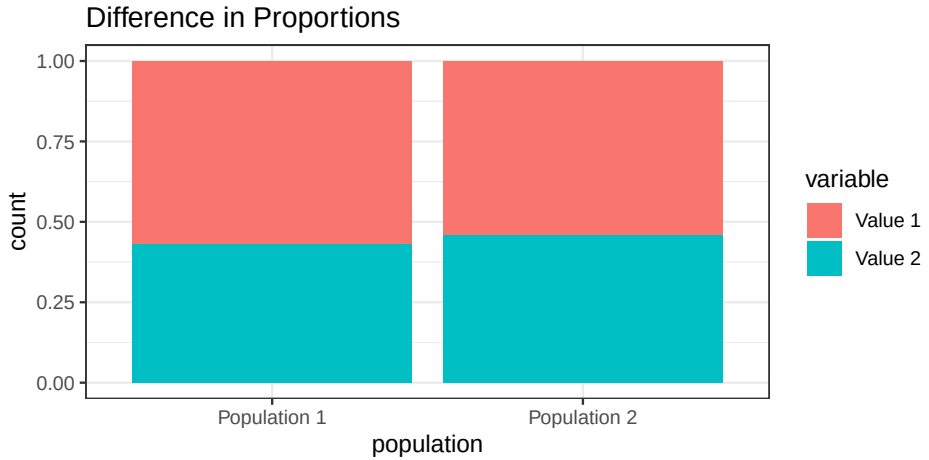
Difference in Proportions

- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.
- That is, we want to know the value of the difference $p_1 - p_2$ in proportion.



Difference in Proportions

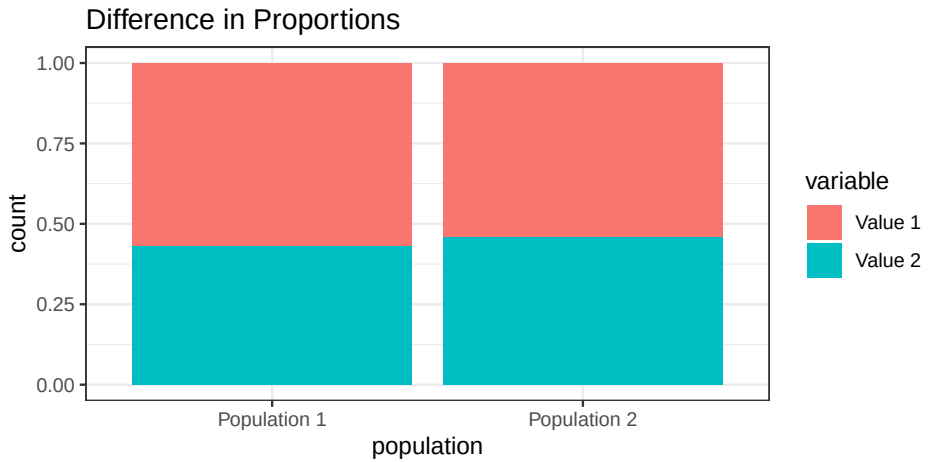
- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.
- That is, we want to know the value of the difference $p_1 - p_2$ in proportion.



- A reasonable point estimate for $p_1 - p_2$ is the difference in sample proportions $\hat{p}_1 - \hat{p}_2$ for a sample taken from the 1st and 2nd populations.

Difference in Proportions

- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.
- That is, we want to know the value of the difference $p_1 - p_2$ in proportion.



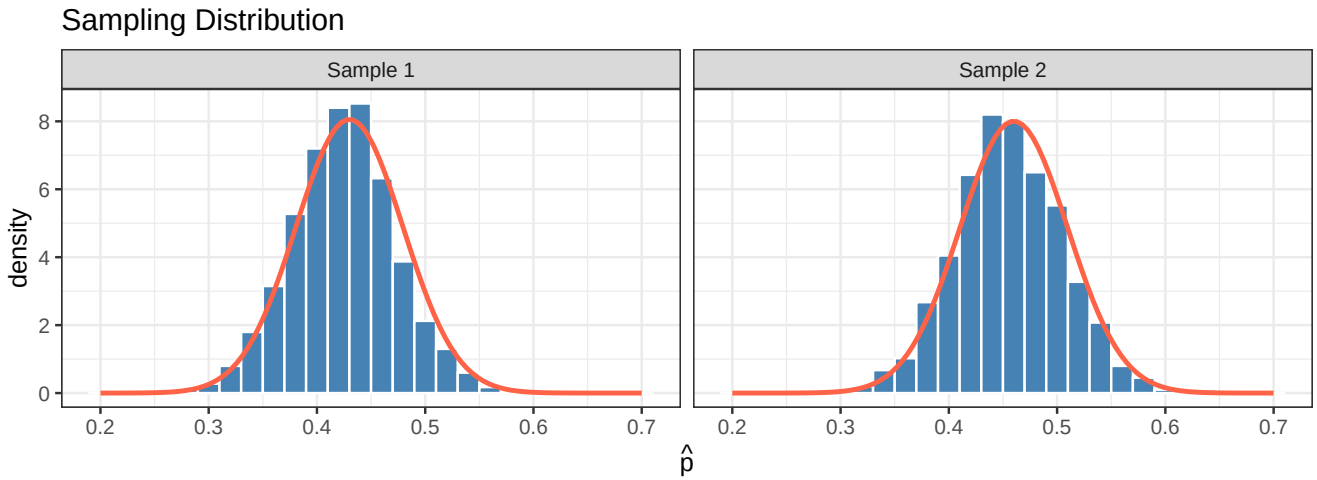
- A reasonable point estimate for $p_1 - p_2$ is the difference in sample proportions $\hat{p}_1 - \hat{p}_2$ for a sample taken from the 1st and 2nd populations.
- As long as we can verify that the statistic $\hat{p}_1 - \hat{p}_2$ has an approximately Normal distribution, we can use the same techniques we used for single sample proportions.

Distribution for $\hat{p}_1 - \hat{p}_2$

- We know that individually, both $\hat{p}_1$ and $\hat{p}_2$ are approximately Normal:

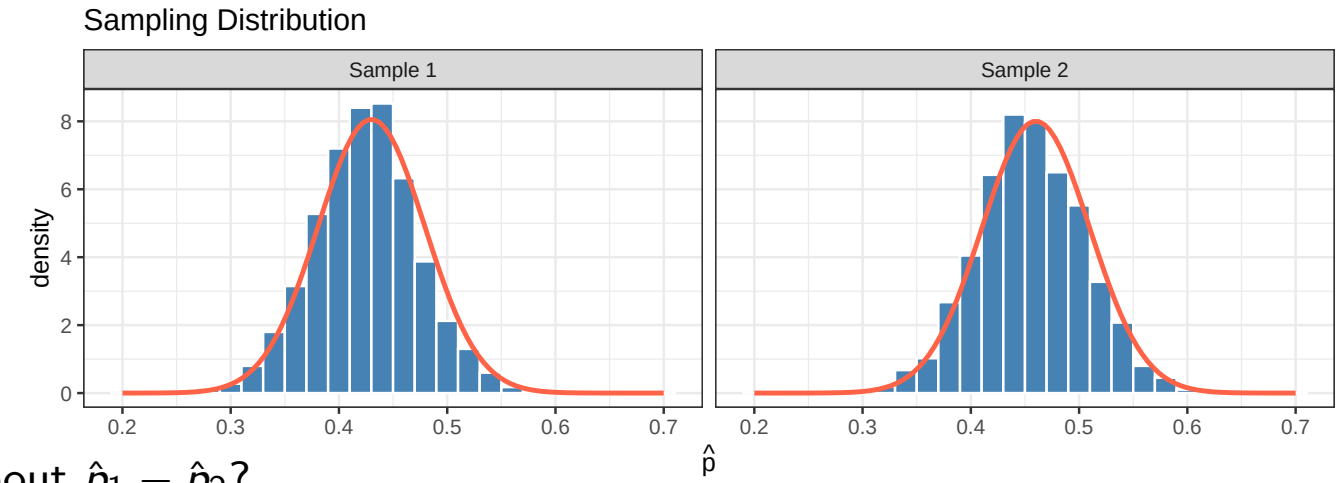
Distribution for $\hat{p}_1 - \hat{p}_2$

- We know that individually, both $\hat{p}_1$ and $\hat{p}_2$ are approximately Normal:



Distribution for $\hat{p}_1 - \hat{p}_2$

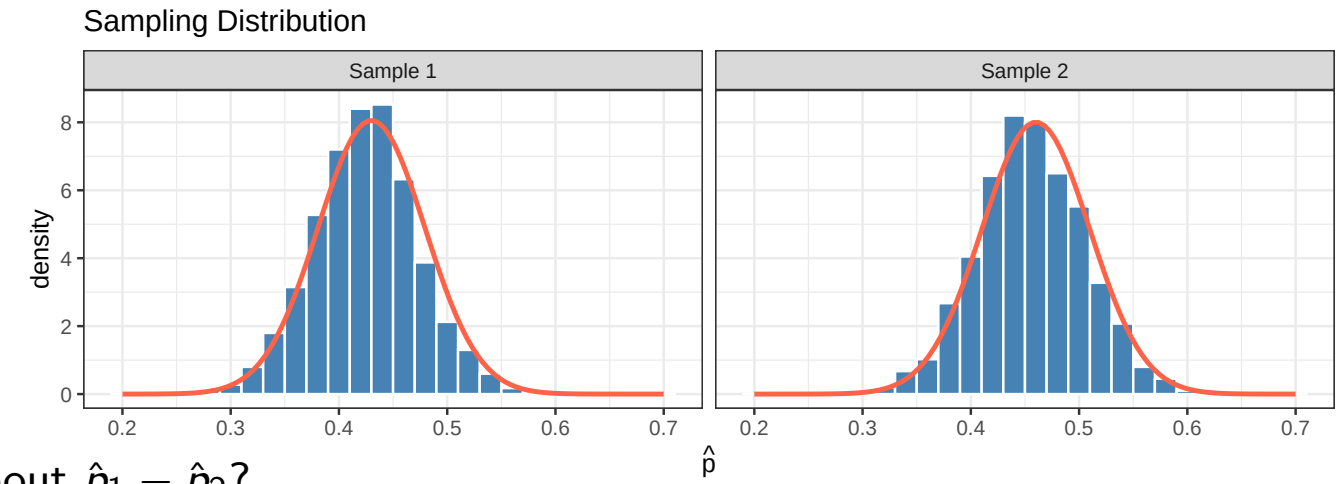
- We know that individually, both $\hat{p}_1$ and $\hat{p}_2$ are approximately Normal:



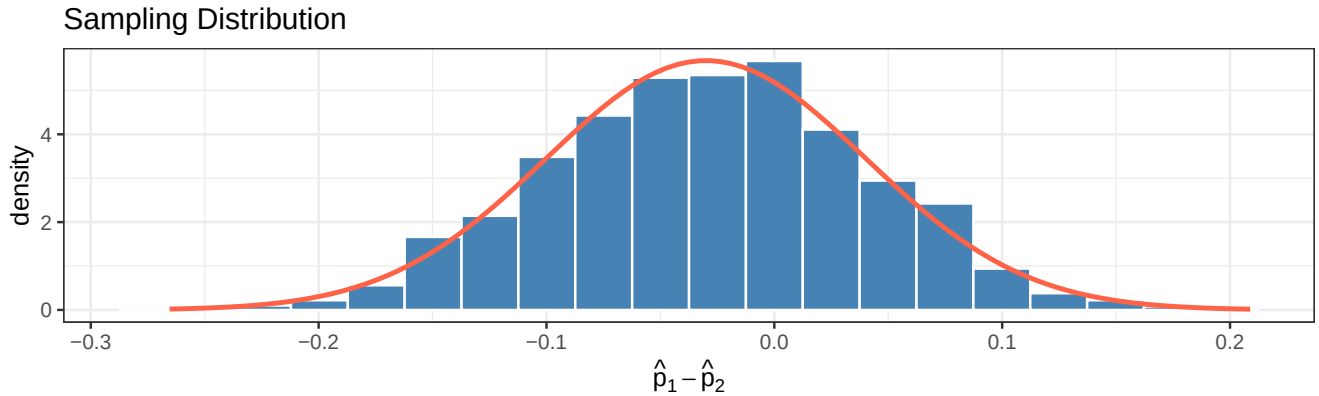
- What about $\hat{p}_1 - \hat{p}_2$?

Distribution for $\hat{p}_1 - \hat{p}_2$

- We know that individually, both $\hat{p}_1$ and $\hat{p}_2$ are approximately Normal:

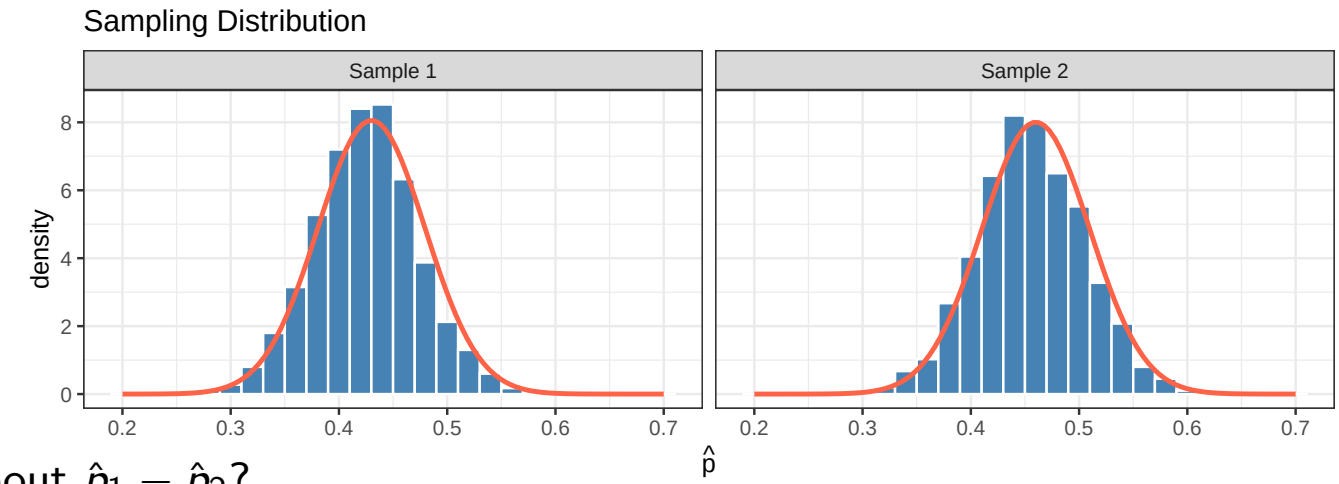


- What about $\hat{p}_1 - \hat{p}_2$?

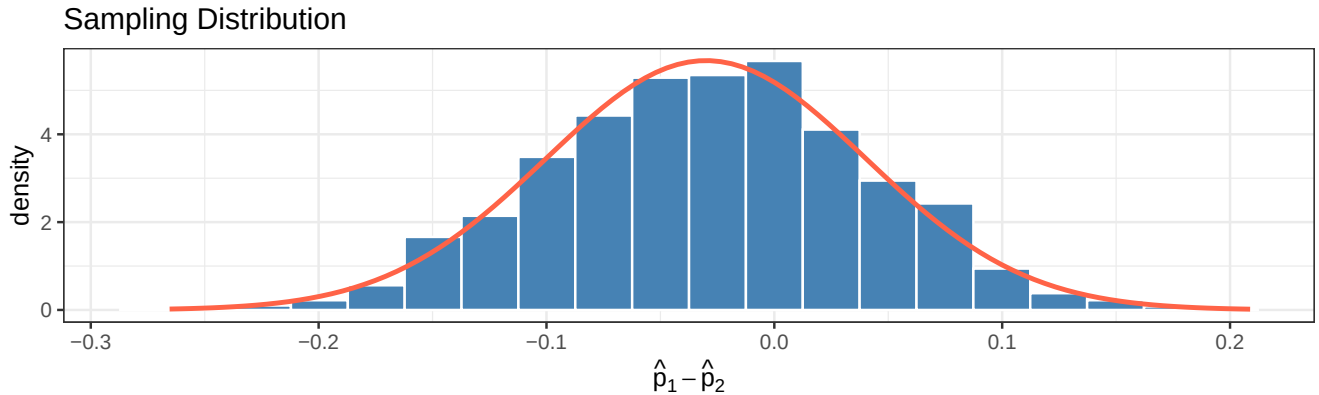


Distribution for $\hat{p}_1 - \hat{p}_2$

- We know that individually, both $\hat{p}_1$ and $\hat{p}_2$ are approximately Normal:



- What about $\hat{p}_1 - \hat{p}_2$?



- The sum or difference of **independent** Normal variables will also be Normal, with variance equal to the sum of individual variances.

Conditions for Theory-based Normal Approximation

Theorem

The difference $\hat{p}_1 - \hat{p}_2$ is approximately Normal when

1. *Each sample proportion is approximately normal (≥ 10 success/failure)*
2. *The two samples are independent of each other*

In this case, the standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Conditions for Theory-based Normal Approximation

Theorem

The difference $\hat{p}_1 - \hat{p}_2$ is approximately Normal when

1. *Each sample proportion is approximately normal (≥ 10 success/failure)*
2. *The two samples are independent of each other*

In this case, the standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Importantly, we know the distribution is Normal and we have the standard error

Conditions for Theory-based Normal Approximation

Theorem

The difference $\hat{p}_1 - \hat{p}_2$ is approximately Normal when

1. *Each sample proportion is approximately normal (≥ 10 success/failure)*
2. *The two samples are independent of each other*

In this case, the standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

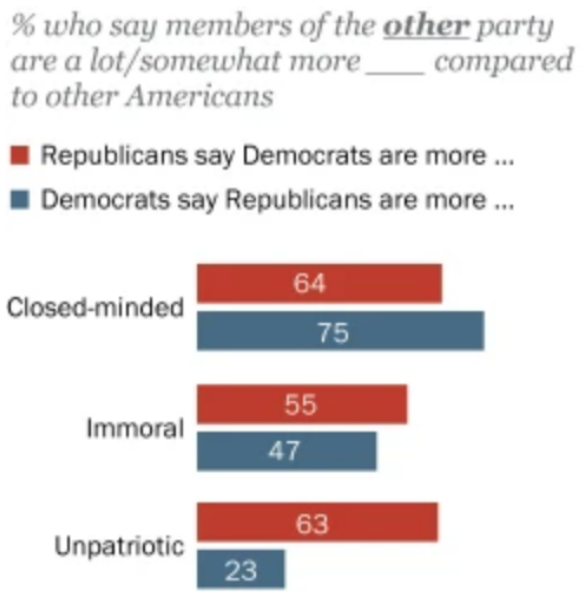
- Importantly, we know the distribution is Normal and we have the standard error
  - We can use `qnorm` to find critical values for confidence intervals and `pnorm` to compute P-values for hypothesis tests

Partisanship

U.S. POLITICS | OCTOBER 10, 2019

Partisan Antipathy: More Intense, More Personal

The share of Republicans who give Democrats a "cold" rating on a 0-100 thermometer has risen 14 percentage points since 2016. Similarly, 57% of Democrats give Republicans a very cold rating, up from 2016.

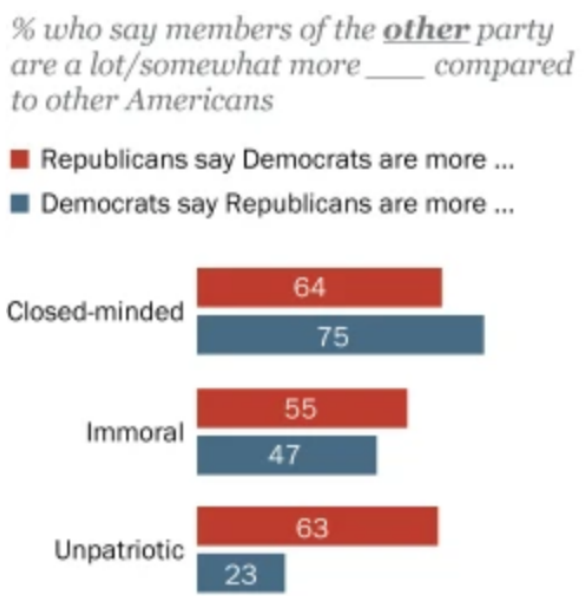


Partisanship

U.S. POLITICS | OCTOBER 10, 2019

Partisan Antipathy: More Intense, More Personal

The share of Republicans who give Democrats a "cold" rating on a 0-100 thermometer has risen 14 percentage points since 2016. Similarly, 57% of Democrats give Republicans a very cold rating, up from 2016.



- Was there really a difference in the proportion of Democrats that view Republicans as close-minded compared to Republicans that view Democrats the same? Or is the difference just due to random sampling?

Confidence Intervals

- Recall that the formula for a confidence interval for $p_r - p_d$ is

$$(\hat{p}_r - \hat{p}_d) \pm z^* \cdot SE$$

Confidence Intervals

- Recall that the formula for a confidence interval for $p_r - p_d$ is

$$(\hat{p}_r - \hat{p}_d) \pm z^* \cdot SE$$

- From the study, we determine sample proportions and sample sizes:

$$\hat{p}_r = 0.64 \quad n_r = 4948 \qquad \hat{p}_d = 0.75 \quad n_d = 4947$$

Confidence Intervals

- Recall that the formula for a confidence interval for $p_r - p_d$ is

$$(\hat{p}_r - \hat{p}_d) \pm z^* \cdot SE$$

- From the study, we determine sample proportions and sample sizes:

$$\hat{p}_r = 0.64 \quad n_r = 4948 \qquad \hat{p}_d = 0.75 \quad n_d = 4947$$

- Our standard error is therefore

$$SE = \sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{n_r} + \frac{\hat{p}_d(1-\hat{p}_d)}{n_d}} = \sqrt{\frac{0.64(1-0.64)}{4948} + \frac{0.75(1-0.75)}{4947}} = 0.009$$

Confidence Intervals

- Recall that the formula for a confidence interval for $p_r - p_d$ is

$$(\hat{p}_r - \hat{p}_d) \pm z^* \cdot SE$$

- From the study, we determine sample proportions and sample sizes:

$$\hat{p}_r = 0.64 \quad n_r = 4948 \qquad \hat{p}_d = 0.75 \quad n_d = 4947$$

- Our standard error is therefore

$$SE = \sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{n_r} + \frac{\hat{p}_d(1-\hat{p}_d)}{n_d}} = \sqrt{\frac{0.64(1-0.64)}{4948} + \frac{0.75(1-0.75)}{4947}} = 0.009$$

- Using `qnorm` in R, the critical value z^* for 95% confidence is

```
qnorm(.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

Confidence Intervals

- Recall that the formula for a confidence interval for $p_r - p_d$ is

$$(\hat{p}_r - \hat{p}_d) \pm z^* \cdot SE$$

- From the study, we determine sample proportions and sample sizes:

$$\hat{p}_r = 0.64 \quad n_r = 4948 \qquad \hat{p}_d = 0.75 \quad n_d = 4947$$

- Our standard error is therefore

$$SE = \sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{n_r} + \frac{\hat{p}_d(1-\hat{p}_d)}{n_d}} = \sqrt{\frac{0.64(1-0.64)}{4948} + \frac{0.75(1-0.75)}{4947}} = 0.009$$

- Using `qnorm` in R, the critical value z^* for 95% confidence is

```
qnorm(.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

- Assembling these pieces, the confidence interval for $p_r - p_d$ is

$$(0.64 - 0.75) \pm 1.96 \cdot 0.009 \qquad \Longleftrightarrow \qquad (-0.128, -0.092)$$

Confidence Intervals

- Recall that the formula for a confidence interval for $p_r - p_d$ is

$$(\hat{p}_r - \hat{p}_d) \pm z^* \cdot SE$$

- From the study, we determine sample proportions and sample sizes:

$$\hat{p}_r = 0.64 \quad n_r = 4948 \qquad \hat{p}_d = 0.75 \quad n_d = 4947$$

- Our standard error is therefore

$$SE = \sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{n_r} + \frac{\hat{p}_d(1-\hat{p}_d)}{n_d}} = \sqrt{\frac{0.64(1-0.64)}{4948} + \frac{0.75(1-0.75)}{4947}} = 0.009$$

- Using `qnorm` in R, the critical value z^* for 95% confidence is

```
qnorm(.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

- Assembling these pieces, the confidence interval for $p_r - p_d$ is

$$(0.64 - 0.75) \pm 1.96 \cdot 0.009 \qquad \Longleftrightarrow \qquad (-0.128, -0.092)$$

- It is plausible that true difference in proportion is between -9.2% and -12.8%

Confidence Interval via `infer`

- Alternatively, we can use `infer` to compute confidence intervals.

Confidence Interval via infer

- Alternatively, we can use infer to compute confidence intervals.

```
pew %>%
  specify(response = close_minded, explanatory = party, success = "yes" ) %>%
  generate(reps = 5000, type = "bootstrap" ) %>%
  calculate( "diff in props", order = c("Republican", "Democrat") ) %>%
  get_ci(level = .95, type = "percentile")
```

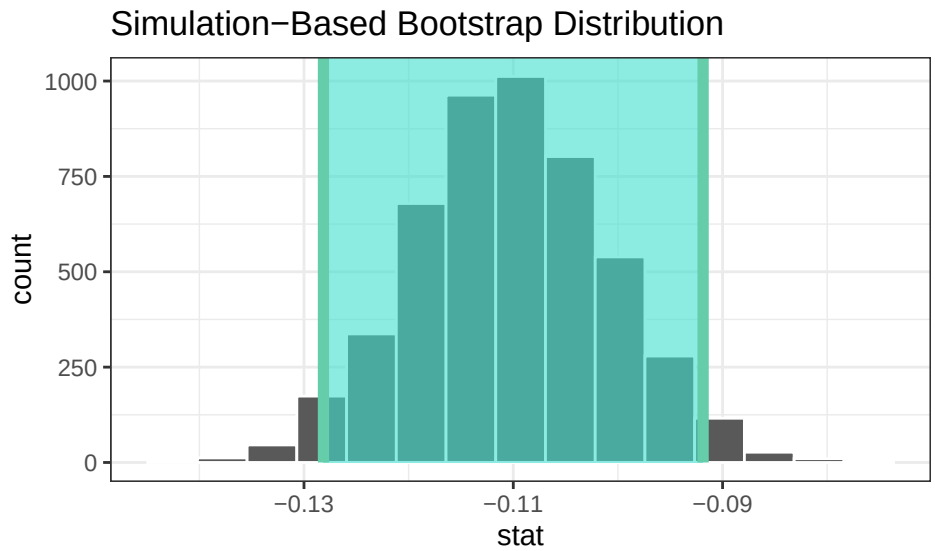
```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.128  -0.0919
```

Confidence Interval via `infer`

- Alternatively, we can use `infer` to compute confidence intervals.

```
pew %>%
  specify(response = close_minded, explanatory = party, success = "yes" ) %>%
  generate(reps = 5000, type = "bootstrap" ) %>%
  calculate( "diff in props", order = c("Republican", "Democrat") ) %>%
  get_ci(level = .95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.128  -0.0919
```



Pooled sample for Hypothesis Tests

- Suppose we are interested in testing the following hypotheses

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \neq p_2$$

Pooled sample for Hypothesis Tests

- Suppose we are interested in testing the following hypotheses

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \neq p_2$$

- **If the null hypothesis is true**, collecting a sample of sizes n_1 and n_2 from each population is the same as collecting a single sample of size $n_1 + n_2$.

Pooled sample for Hypothesis Tests

- Suppose we are interested in testing the following hypotheses

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \neq p_2$$

- **If the null hypothesis is true**, collecting a sample of sizes n_1 and n_2 from each population is the same as collecting a single sample of size $n_1 + n_2$.
  - So we may instead consider the pooled proportion $\hat{p}$ given by

$$\hat{p} = \frac{\text{overall successes}}{\text{overall sample size}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Pooled sample for Hypothesis Tests

- Suppose we are interested in testing the following hypotheses

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \neq p_2$$

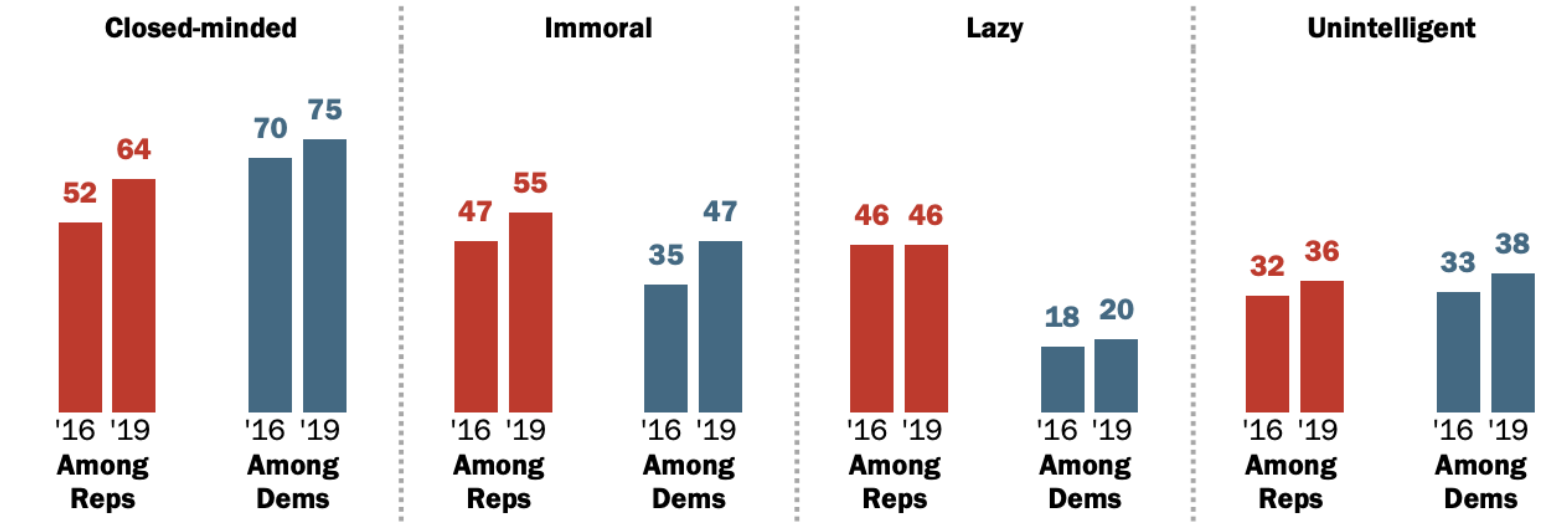
- **If the null hypothesis is true**, collecting a sample of sizes n_1 and n_2 from each population is the same as collecting a single sample of size $n_1 + n_2$.
  - So we may instead consider the pooled proportion $\hat{p}$ given by

$$\hat{p} = \frac{\text{overall successes}}{\text{overall sample size}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

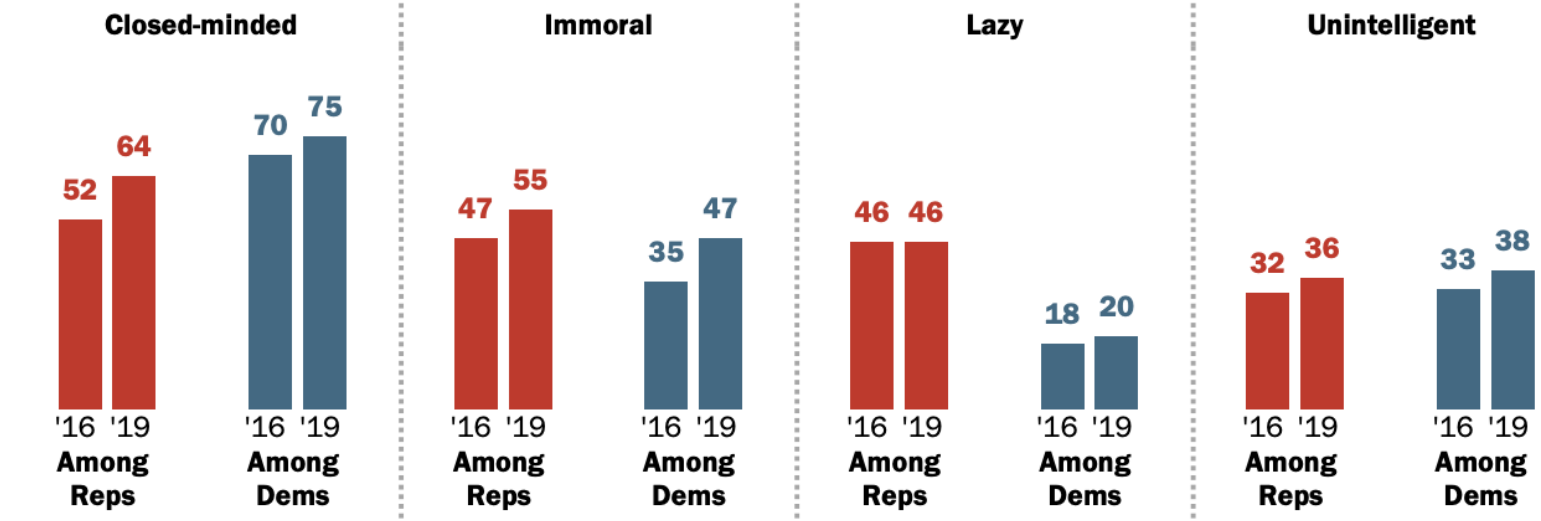
- This gives a standard error for the null distribution of

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Partisanship over Time



Partisanship over Time



- Was there really a change in the proportion of Democrats that view Republicans as close-minded between 2016 and 2019?

Hypothesis Tests

- We test

$$H_0 : p_{16} = p_{19} \qquad H_a : p_{16} \neq p_{19}$$

Hypothesis Tests

- We test

$$H_0 : p_{16} = p_{19} \qquad H_a : p_{16} \neq p_{19}$$

- In the study, we find

$$\hat{p}_{16} = 0.7 \quad n_{16} = 4948 \qquad \hat{p}_{19} = 0.75 \quad n_{19} = 4947$$

which gives a pooled proportion of

$$\hat{p} = \frac{n_{16}\hat{p}_{16} + n_{19}\hat{p}_{19}}{n_{16} + n_{19}} = 0.725$$

Hypothesis Tests

- We test

$$H_0 : p_{16} = p_{19} \qquad H_a : p_{16} \neq p_{19}$$

- In the study, we find

$$\hat{p}_{16} = 0.7 \quad n_{16} = 4948 \qquad \hat{p}_{19} = 0.75 \quad n_{19} = 4947$$

which gives a pooled proportion of

$$\hat{p} = \frac{n_{16}\hat{p}_{16} + n_{19}\hat{p}_{19}}{n_{16} + n_{19}} = 0.725$$

- The standard error for the null distribution is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{16}} + \frac{\hat{p}(1-\hat{p})}{n_{19}}} = \sqrt{\frac{0.725(1-0.725)}{4948} + \frac{0.725(1-0.725)}{4947}} = 0.009$$

Hypothesis Tests

- We test

$$H_0 : p_{16} = p_{19} \qquad H_a : p_{16} \neq p_{19}$$

- In the study, we find

$$\hat{p}_{16} = 0.7 \quad n_{16} = 4948 \qquad \hat{p}_{19} = 0.75 \quad n_{19} = 4947$$

which gives a pooled proportion of

$$\hat{p} = \frac{n_{16}\hat{p}_{16} + n_{19}\hat{p}_{19}}{n_{16} + n_{19}} = 0.725$$

- The standard error for the null distribution is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{16}} + \frac{\hat{p}(1-\hat{p})}{n_{19}}} = \sqrt{\frac{0.725(1-0.725)}{4948} + \frac{0.725(1-0.725)}{4947}} = 0.009$$

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

Hypothesis Tests

- We test

$$H_0 : p_{16} = p_{19} \qquad H_a : p_{16} \neq p_{19}$$

- In the study, we find

$$\hat{p}_{16} = 0.7 \quad n_{16} = 4948 \qquad \hat{p}_{19} = 0.75 \quad n_{19} = 4947$$

which gives a pooled proportion of

$$\hat{p} = \frac{n_{16}\hat{p}_{16} + n_{19}\hat{p}_{19}}{n_{16} + n_{19}} = 0.725$$

- The standard error for the null distribution is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{16}} + \frac{\hat{p}(1-\hat{p})}{n_{19}}} = \sqrt{\frac{0.725(1-0.725)}{4948} + \frac{0.725(1-0.725)}{4947}} = 0.009$$

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

- Without computing a p-value, does this seem to be statistically significant?

P-Value

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

P-Value

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

- By the CLT, z-scores are approximately standard Normal, so we compute p-values using `pnorm`.
  - Since H_a was two-sided, and $z < 0$, we compute the area in the left tail, and double.

P-Value

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

- By the CLT, z-scores are approximately standard Normal, so we compute p-values using `pnorm`.
  - Since H_a was two-sided, and $z < 0$, we compute the area in the left tail, and double.

```
2*pnorm(-5.569, mean = 0 ,sd = 1)
```

```
## [1] 0.00000002562
```

P-Value

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

- By the CLT, z-scores are approximately standard Normal, so we compute p-values using `pnorm`.
  - Since H_a was two-sided, and $z < 0$, we compute the area in the left tail, and double.

```
2*pnorm(-5.569, mean = 0 ,sd = 1)
```

```
## [1] 0.00000002562
```

- The test is significant at $\alpha = 0.01$ and we reject the null hypothesis.

P-Value

- Our test statistic is

$$z = \frac{\hat{p}_{16} - \hat{p}_{19}}{SE} = \frac{0.7 - 0.75}{0.009} = -5.57$$

- By the CLT, z-scores are approximately standard Normal, so we compute p-values using `pnorm`.
  - Since H_a was two-sided, and $z < 0$, we compute the area in the left tail, and double.

```
2*pnorm(-5.569, mean = 0 ,sd = 1)
```

```
## [1] 0.00000002562
```

- The test is significant at $\alpha = 0.01$ and we reject the null hypothesis.
  - It is unlikely that the observed difference in proportions is due to chance, if the populations truly had the same proportion.

Hypothesis Test via `infer`

- Repeating our analysis, this time using `infer`

Hypothesis Test via infer

- Repeating our analysis, this time using infer

```
pew2 %>% specify(response = close_minded, explanatory = year, success = "yes" ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute" ) %>%
  calculate( "diff in props", order = c("2016", "2019") ) %>%
  get_p_value(obs_stat = (0.7 - 0.75), direction = "both")
```

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

Hypothesis Test via infer

- Repeating our analysis, this time using infer

```
pew2 %>% specify(response = close_minded, explanatory = year, success = "yes" ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute" ) %>%
  calculate( "diff in props", order = c("2016", "2019") ) %>%
  get_p_value(obs_stat = (0.7 - 0.75), direction = "both")
```

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

- Why did the infer method report a p-value of 0?

Hypothesis Test via infer

- Repeating our analysis, this time using infer

```
pew2 %>% specify(response = close_minded, explanatory = year, success = "yes" ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute" ) %>%
  calculate( "diff in props", order = c("2016", "2019") ) %>%
  get_p_value(obs_stat = (0.7 - 0.75), direction = "both")
```

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

- Why did the infer method report a p-value of 0?