Inference for a 1 and 2 Proportions

Nate Wells

Math 141, 4/7/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Perform hypothesis tests for proportions using the theory-based method
- Investigate the theoretical distribution for differences in proportions
- Calculate confidence intervals and conduct hypothesis tests for differences in proportions

Section 1

Single Proportions

The Sampling Distribution for Sample Proportion

- Consider a population variable that takes only two levels, success S and failure F. Let p be the proportion of success in the population.

The Sampling Distribution for Sample Proportion

- Consider a population variable that takes only two levels, success S and failure F. Let p be the proportion of success in the population.
- Suppose we randomly choose a single observation from a population, and denote the result as 1 if observation is S and 0 if it is F.

The Sampling Distribution for Sample Proportion

- Consider a population variable that takes only two levels, success S and failure F. Let p be the proportion of success in the population.
- Suppose we randomly choose a single observation from a population, and denote the result as 1 if observation is S and 0 if it is F.
- The *mean* of this variable is p, and the *standard deviation* is $\sqrt{p(1-p)}$

The Sampling Distribution for Sample Proportion

- Consider a population variable that takes only two levels, success S and failure F. Let p be the proportion of success in the population.
- Suppose we randomly choose a single observation from a population, and denote the result as 1 if observation is S and 0 if it is F.
- The *mean* of this variable is p, and the *standard deviation* is $\sqrt{p(1-p)}$
- If we instead take an SRS of size n from the population, we can view the sample proportion $\hat{p}$ as a sample mean:

The Sampling Distribution for Sample Proportion

- Consider a population variable that takes only two levels, success S and failure F. Let p be the proportion of success in the population.
- Suppose we randomly choose a single observation from a population, and denote the result as 1 if observation is S and 0 if it is F.
- The *mean* of this variable is p, and the *standard deviation* is $\sqrt{p(1-p)}$
- If we instead take an SRS of size n from the population, we can view the sample proportion $\hat{p}$ as a sample mean:
  - We are averaging across each person in the sample the variable that takes the value 1 if the individual is a success and 0 otherwise.

The Sampling Distribution for Sample Proportion

- Consider a population variable that takes only two levels, success S and failure F. Let p be the proportion of success in the population.
- Suppose we randomly choose a single observation from a population, and denote the result as 1 if observation is S and 0 if it is F.
- The *mean* of this variable is p, and the *standard deviation* is $\sqrt{p(1-p)}$
- If we instead take an SRS of size n from the population, we can view the sample proportion $\hat{p}$ as a sample mean:
  - We are averaging across each person in the sample the variable that takes the value 1 if the individual is a success and 0 otherwise.
- By the central limit theorem, if n is large, then $\hat{p}$ is approximately Normal, with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$

Examples

Using data from the gss General Social Survey. . .

- 47.4% identified as female
- 34.8% obtained a college degree
- 98.2% were 21 or older

Examples

Using data from the gss General Social Survey. . .

- 47.4% identified as female
- 34.8% obtained a college degree
- 98.2% were 21 or older

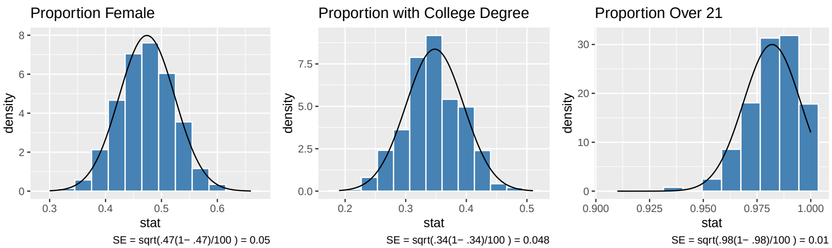
If we draw samples of size 100 from the GSS, the sampling distributions look like. . .

Examples

Using data from the gss General Social Survey...

- 47.4% identified as female
- 34.8% obtained a college degree
- 98.2% were 21 or older

If we draw samples of size 100 from the GSS, the sampling distributions look like...

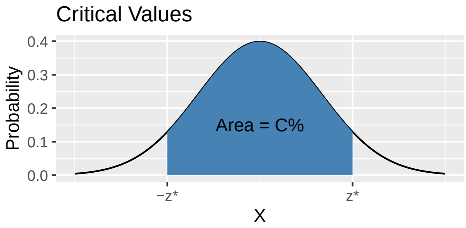


Critical Values

- 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.

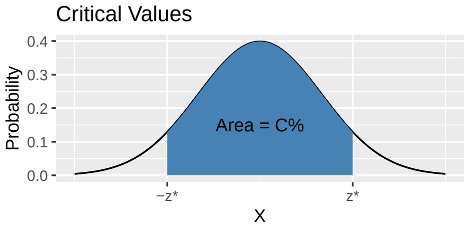
Critical Values

- 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.



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.



- Previously, we saw that for Normal distributions, 95% of observations are within 2 standard deviations of the mean. So the critical value for 95% confidence is

$$z^* = 2$$

Confidence Intervals

When a sample statistic is approximately Normally distribution, the $C\%$ confidence interval is

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

where z^* is the critical value for $C\%$ confidence and SE is the standard error for the statistic.

Confidence Intervals

When a sample statistic is approximately Normally distribution, the $C\%$ confidence interval is

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

where z^* is the critical value for $C\%$ confidence and SE is the standard error for the statistic.

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

Confidence Intervals

When a sample statistic is approximately Normally distribution, the $C\%$ confidence interval is

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

where z^* is the critical value for $C\%$ confidence and SE is the standard error for the statistic.

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

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

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

An Example

On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 47% of 1,012 Americans agreed with this decision. Use the theory-based method at 99% confidence to estimate the true proportion of Americans that agreed with this decision.

An Example

On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 47% of 1,012 Americans agreed with this decision. Use the theory-based method at 99% confidence to estimate the true proportion of Americans that agreed with this decision.

- Our sample statistic is $\hat{p} = 0.47$

```
p_hat<-0.47
p_hat
```

```
## [1] 0.47
```

An Example

On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 47% of 1,012 Americans agreed with this decision. Use the theory-based method at 99% confidence to estimate the true proportion of Americans that agreed with this decision.

- Our sample statistic is $\hat{p} = 0.47$

```
p_hat<-0.47
p_hat
```

```
## [1] 0.47
```

- The critical value z^* for 99% confidence is $z^* = 2.58$

```
z<-qnorm(.995, 0 , 1)
z
```

```
## [1] 2.575829
```

An Example

On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 47% of 1,012 Americans agreed with this decision. Use the theory-based method at 99% confidence to estimate the true proportion of Americans that agreed with this decision.

- Our sample statistic is $\hat{p} = 0.47$

```
p_hat<-0.47
p_hat
```

```
## [1] 0.47
```

- The critical value z^* for 99% confidence is $z^* = 2.58$

```
z<-qnorm(.995, 0 , 1)
z
```

```
## [1] 2.575829
```

- The standard error for $\hat{p}$ is $SE = 0.016$

```
SE<-sqrt(p_hat*(1- p_hat)/1012)
SE
```

```
## [1] 0.01568905
```

An Example

- The theory-based confidence interval is $(0.43, 0.51)$

```
CI_low<-p_hat-z*SE
CI_high<-p_hat+z*SE
```

```
##      CI_low   CI_high
## 1 0.4295877 0.5104123
```

An Example

- The theory-based confidence interval is $(0.43, 0.51)$

```
CI_low<-p_hat-z*SE
CI_high<-p_hat+z*SE
```

```
##      CI_low   CI_high
## 1 0.4295877 0.5104123
```

- How does this compare to the bootstrap method?

An Example

- The theory-based confidence interval is $(0.43, 0.51)$

```
CI_low<-p_hat-z*SE
CI_high<-p_hat+z*SE
```

```
##      CI_low   CI_high
## 1 0.4295877 0.5104123
```

- How does this compare to the bootstrap method?

```
health %>% specify(response = agree, success = "yes") %>%
  generate(reps=10000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = .99, type = "se", point_estimate = p_hat)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.429    0.511
```

z-Scores

- The **z-score** for a statistic X with standard error SE and mean μ under the Null hypothesis is

$$Z = \frac{X - \mu}{SE}$$

z-Scores

- The **z-score** for a statistic X with standard error SE and mean μ under the Null hypothesis is

$$Z = \frac{X - \mu}{SE}$$

  - The z-score for a statistic X measures how far away it is from the mean, in units of standard error

z-Scores

- The **z-score** for a statistic X with standard error SE and mean μ under the Null hypothesis is

$$Z = \frac{X - \mu}{SE}$$

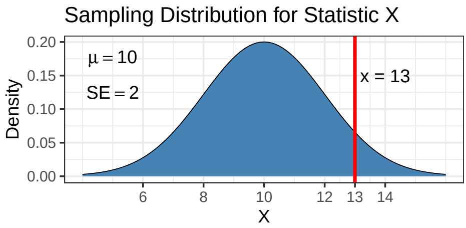
  - The z-score for a statistic X measures how far away it is from the mean, in units of standard error
  - If X is approximately Normal with mean μ and standard deviation σ, then its z-score is approximately **standard** Normal (mean $= 0$, sd $= 1$).

z-Scores

- The **z-score** for a statistic X with standard error SE and mean μ under the Null hypothesis is

$$Z = \frac{X - \mu}{SE}$$

  - The z-score for a statistic X measures how far away it is from the mean, in units of standard error
  - If X is approximately Normal with mean μ and standard deviation σ, then its z-score is approximately **standard** Normal (mean $= 0$, sd $= 1$).

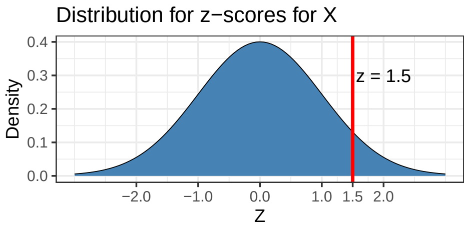


z-Scores

- The **z-score** for a statistic X with standard error SE and mean μ under the Null hypothesis is

$$Z = \frac{X - \mu}{SE}$$

  - The z-score for a statistic X measures how far away it is from the mean, in units of standard error
  - If X is approximately Normal with mean μ and standard deviation σ, then its z-score is approximately **standard** Normal (mean $= 0$, sd $= 1$).



P-Values

- By location-scale invariance, if X is Normal with mean μ and standard error SE and Z is standard Normal, then

$$P(X > x) = P\left(Z > \frac{x-\mu}{\mathrm{SE}}\right)$$

P-Values

- By location-scale invariance, if X is Normal with mean μ and standard error SE and Z is standard Normal, then

$$P(X > x) = P\left(Z > \frac{x-\mu}{\mathrm{SE}}\right)$$

- If we want to compute a P-Value for test statistic x, we can instead compute a P-value for its z-score z:

P-value	$=$	$P(Z > z)$	if H_a is one-sided right
P-value	$=$	$P(Z < z)$	if H_a is one-sided left
P-value	$=$	$2 \cdot P(Z > \lvert z \rvert)$	if H_a is two-sided

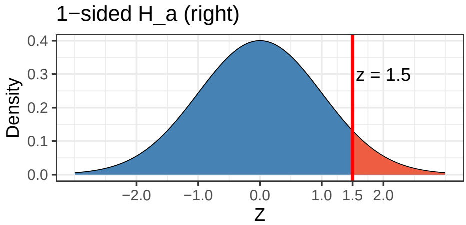
P-Values

- By location-scale invariance, if X is Normal with mean μ and standard error SE and Z is standard Normal, then

$$P(X > x) = P\left(Z > \frac{x-\mu}{\mathrm{SE}}\right)$$

- If we want to compute a P-Value for test statistic x, we can instead compute a P-value for its z-score z:

$$\begin{array}{lcll} \text{P-value} & = & P(Z > z) & \text{if } H_a \text{ is one-sided right} \\ \text{P-value} & = & P(Z < z) & \text{if } H_a \text{ is one-sided left} \\ \text{P-value} & = & 2 \cdot P(Z > |z|) & \text{if } H_a \text{ is two-sided} \end{array}$$



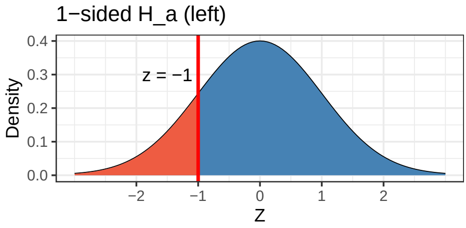
P-Values

- By location-scale invariance, if X is Normal with mean μ and standard error SE and Z is standard Normal, then

$$P(X > x) = P\left(Z > \frac{x-\mu}{\mathrm{SE}}\right)$$

- If we want to compute a P-Value for test statistic x, we can instead compute a P-value for its z-score z:

$$\begin{array}{lcll} \text{P-value} & = & P(Z > z) & \text{if } H_a \text{ is one-sided right} \\ \text{P-value} & = & P(Z < z) & \text{if } H_a \text{ is one-sided left} \\ \text{P-value} & = & 2 \cdot P(Z > |z|) & \text{if } H_a \text{ is two-sided} \end{array}$$



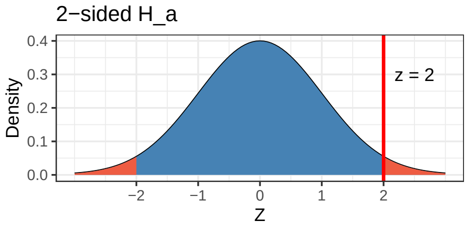
P-Values

- By location-scale invariance, if X is Normal with mean μ and standard error SE and Z is standard Normal, then

$$P(X > x) = P\left(Z > \frac{x-\mu}{\mathrm{SE}}\right)$$

- If we want to compute a P-Value for test statistic x, we can instead compute a P-value for its z-score z:

$$\begin{aligned} \text{P-value} &= P(Z > z) && \text{if } H_a \text{ is one-sided right} \\ \text{P-value} &= P(Z < z) && \text{if } H_a \text{ is one-sided left} \\ \text{P-value} &= 2 \cdot P(Z > |z|) && \text{if } H_a \text{ is two-sided} \end{aligned}$$



Hypothesis Tests

By the central limit theorem, if $H_0 : p = p_0$ is true, then for large n, the standard error for the sample statistic $\hat{p}$ is

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

Hypothesis Tests

By the central limit theorem, if $H_0 : p = p_0$ is true, then for large n, the standard error for the sample statistic $\hat{p}$ is

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

Theorem

To test $H_0 : p = p_0$ against $H_a : p \neq p_0$ (or the one-sided alternative) we use the standardized test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

If n is large enough so that both $n\hat{p}$ and $n(1 - \hat{p})$ are at least 10, then the p-value for the test is computed using the standard Normal distribution.

Rock-Paper-Scissors

In Rock-Paper-Scissors, each player chooses one of 3 symbols (Rock, Paper, Scissors). Are all three options chosen with equal frequency?

Rock-Paper-Scissors

In Rock-Paper-Scissors, each player chooses one of 3 symbols (Rock, Paper, Scissors). Are all three options chosen with equal frequency?

A study observed that for 119 people playing Rock-Paper-Scissors in an official tournament, 66 players selected Rock on their first turn.

Rock-Paper-Scissors

In Rock-Paper-Scissors, each player chooses one of 3 symbols (Rock, Paper, Scissors). Are all three options chosen with equal frequency?

A study observed that for 119 people playing Rock-Paper-Scissors in an official tournament, 66 players selected Rock on their first turn.

- We test $H_0 : p_0 = 1/3$ against $H_a : p_0 \neq 1/3$, where p_0 is the theoretical frequency a player chooses rock on the first turn.

Rock-Paper-Scissors

In Rock-Paper-Scissors, each player chooses one of 3 symbols (Rock, Paper, Scissors). Are all three options chosen with equal frequency?

A study observed that for 119 people playing Rock-Paper-Scissors in an official tournament, 66 players selected Rock on their first turn.

- We test $H_0 : p_0 = 1/3$ against $H_a : p_0 \neq 1/3$, where p_0 is the theoretical frequency a player chooses rock on the first turn.
- The sample statistic is $\hat{p} = 0.55$

```
p_hat<-66/119
p_hat
```

```
## [1] 0.5546218
```

Rock-Paper-Scissors

In Rock-Paper-Scissors, each player chooses one of 3 symbols (Rock, Paper, Scissors). Are all three options chosen with equal frequency?

A study observed that for 119 people playing Rock-Paper-Scissors in an official tournament, 66 players selected Rock on their first turn.

- We test $H_0 : p_0 = 1/3$ against $H_a : p_0 \neq 1/3$, where p_0 is the theoretical frequency a player chooses rock on the first turn.
- The sample statistic is $\hat{p} = 0.55$

```
p_hat<-66/119
p_hat
```

```
## [1] 0.5546218
```

- The standard error is $SE = 0.04$

```
SE<- sqrt((1/3)*(1-(1/3))/119)
SE
```

```
## [1] 0.04321358
```

Rock-Paper-Scissors

In Rock-Paper-Scissors, each player chooses one of 3 symbols (Rock, Paper, Scissors). Are all three options chosen with equal frequency?

A study observed that for 119 people playing Rock-Paper-Scissors in an official tournament, 66 players selected Rock on their first turn.

- We test $H_0 : p_0 = 1/3$ against $H_a : p_0 \neq 1/3$, where p_0 is the theoretical frequency a player chooses rock on the first turn.
- The sample statistic is $\hat{p} = 0.55$

```
p_hat<-66/119
p_hat
```

```
## [1] 0.5546218
```

- The standard error is $SE = 0.04$

```
SE<- sqrt((1/3)*(1-(1/3))/119)
SE
```

```
## [1] 0.04321358
```

- The test statistic is $z = 5.12$

```
z<- (p_hat - 1/3)/ SE
z
```

```
## [1] 5.120809
```

Rock-Paper-Scissors

- The P-Value (probability of observing a sample proportion as extreme as 66/119) is 0.0000003

```
Pval<- 2*pnorm(-z, mean = 0, sd = 1)
Pval
```

```
## [1] 3.04227e-07
```

Rock-Paper-Scissors

- The P-Value (probability of observing a sample proportion as extreme as 66/119) is 0.0000003

```
Pval<- 2*pnorm(-z, mean = 0, sd = 1)
Pval
```

```
## [1] 3.04227e-07
```

- We reject the null hypothesis in favor of the alternative at significance $\alpha = 0.05$.

Rock-Paper-Scissors

- The P-Value (probability of observing a sample proportion as extreme as 66/119) is 0.0000003

```
Pval<- 2*pnorm(-z, mean = 0, sd = 1)
Pval
```

```
## [1] 3.04227e-07
```

- We reject the null hypothesis in favor of the alternative at significance $\alpha = 0.05$.

How does this compare to the simulation based test?

Rock-Paper-Scissors

- The P-Value (probability of observing a sample proportion as extreme as 66/119) is 0.0000003

```
Pval<- 2*pnorm(-z, mean = 0, sd = 1)
Pval
```

```
## [1] 3.04227e-07
```

- We reject the null hypothesis in favor of the alternative at significance $\alpha = 0.05$.

How does this compare to the simulation based test?

```
rps %>% specify(response = choice, success = "rock") %>%
  hypothesize(null = "point", p = 1/3) %>%
  generate(reps = 5000, type = "simulate") %>%
  calculate(stat = "prop") %>%
  get_p_value(obs_stat = p_hat, direction = "both")
```

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

Section 2

Difference in Proportions

Difference in Proportions

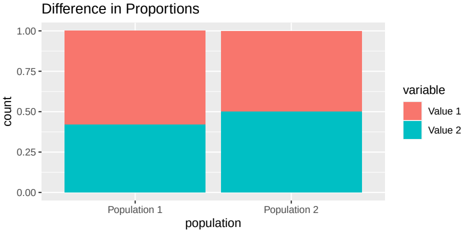
- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.

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.

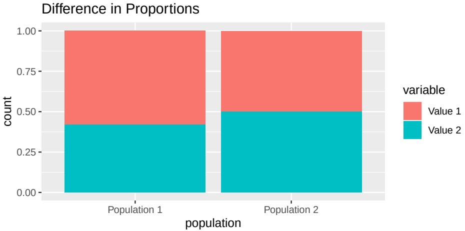
Difference in Proportions

- 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

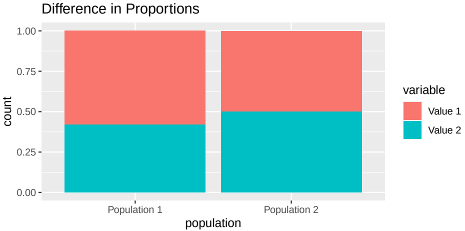
- Suppose we have two populations and wish to compare the proportions p_1 and p_2 of the level of a categorical variable in each population.
- That is, we want to know the value of the difference $p_1 - p_2$ in proportion.



- 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.



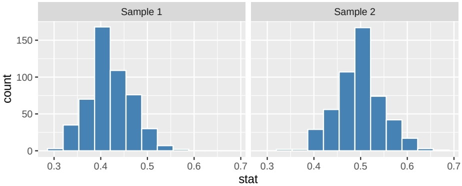
- A reasonable point estimate for $p_1 - p_2$ is the difference in sample proportions $\hat{p}_1 - \hat{p}_2$ for a sample taken from the 1st and 2nd populations.
- As long as we can verify that the statistic $\hat{p}_1 - \hat{p}_2$ has an approximately Normal distribution, we can use the same techniques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

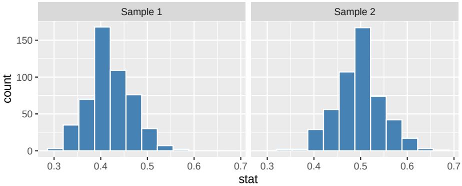
Distribution for $\hat{p}_1 - \hat{p}_2$

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



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

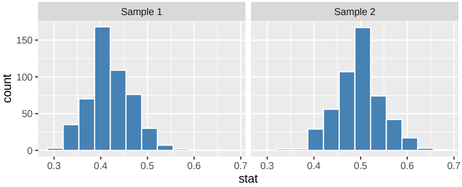
- We know that individually, both $\hat{p}_1$ and $\hat{p}_2$ are approximately normal:



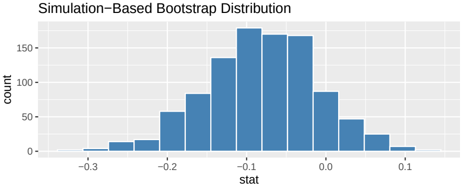
- What about $\hat{p}_1 - \hat{p}_2$?

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

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

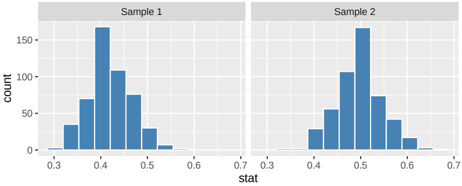


- What about $\hat{p}_1 - \hat{p}_2$?

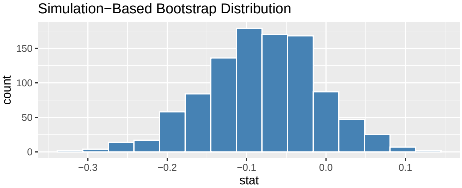


Distribution for $\hat{p}_1 - \hat{p}_2$

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



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



- In general, 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 approximatly normal ($\geq$ 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 approximatly normal ($\geq$ 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 approximatly 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

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

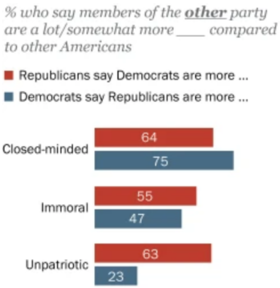
- Importantly, we know the distribution is Normal and we have the standard error
  - We can use `qnorm` to find critical values for confidence intervals and `pnorm` to compute P-values for hypothesis tests

Partisanship

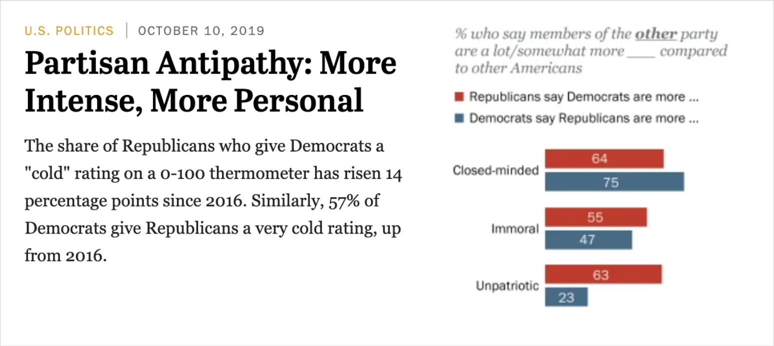
U.S. POLITICS | OCTOBER 10, 2019

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

Let's use the Normal approximation.

Confidence Intervals

Let's use the Normal approximation.

Elsewhere in the study, we find the number of Republicans and Democrats surveyed were 4948 and 4947, respectively.

```
n_r<-4948
n_d<-4947

p_hat_r<-0.64
p_hat_d<-0.75
```

Confidence Intervals

Let's use the Normal approximation.

Elsewhere in the study, we find the number of Republicans and Democrats surveyed were 4948 and 4947, respectively.

```
n_r<-4948
n_d<-4947

p_hat_r<-0.64
p_hat_d<-0.75
```

- Our standard error is therefore 0.009

```
SE<-sqrt(p_hat_r*(1-p_hat_r)/n_r + p_hat_d*(1-p_hat_d)/n_d  )
SE
```

```
## [1] 0.00919054
```

Confidence Intervals

Let's use the Normal approximation.

Elsewhere in the study, we find the number of Republicans and Democrats surveyed were 4948 and 4947, respectively.

```
n_r<-4948
n_d<-4947

p_hat_r<-0.64
p_hat_d<-0.75
```

- Our standard error is therefore 0.009

```
SE<-sqrt(p_hat_r*(1-p_hat_r)/n_r + p_hat_d*(1-p_hat_d)/n_d  )
SE
```

```
## [1] 0.00919054
```

- At a 95% confidence level, the critical value is $z^* = 1.96$

```
z<-qnorm(.975)
z
```

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

Confidence Intervals II

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

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

```
ci_low<-p_hat_r - p_hat_d - z*SE
ci_high<-p_hat_r - p_hat_d + z*SE

c(ci_low, ci_high)
```

```
## [1] -0.12801313 -0.09198687
```

Confidence Intervals II

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

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

```
ci_low<-p_hat_r - p_hat_d - z*SE
ci_high<-p_hat_r - p_hat_d + z*SE

c(ci_low, ci_high)
```

```
## [1] -0.12801313 -0.09198687
```

- Note that both endpoints of the interval are less than 0, suggesting that the true difference in proportions between Republicans and Democrats is negative

Confidence Intervals II

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

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

```
ci_low<-p_hat_r - p_hat_d - z*SE
ci_high<-p_hat_r - p_hat_d + z*SE

c(ci_low, ci_high)
```

```
## [1] -0.12801313 -0.09198687
```

- Note that both endpoints of the interval are less than 0, suggesting that the true difference in proportions between Republicans and Democrats is negative
    - i.e. a greater proportion of Democrats hold the view that Republicans as closed-minded compared to the converse

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.

- We'll use the `pew` data set.

Confidence Interval via infer

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

- We'll use the `pew` data set.

```
pew %>% group_by(party,close_minded) %>%
  summarize(N = n()) %>%
  mutate(prop = N / sum(N))
```

```
## # A tibble: 4 x 4
## # Groups:   party [2]
##   party      close_minded     N  prop
##   <fct>      <fct>        <int> <dbl>
## 1 Democrat   no            1237 0.250
## 2 Democrat   yes           3710 0.750
## 3 Republican no            1781 0.360
## 4 Republican yes           3167 0.640
```

Confidence Interval via infer II

```
boot<-pew %>%
  specify(close_minded ~ party, success = "yes" ) %>%
  generate(reps = 1000, type = "bootstrap" ) %>%
  calculate( "diff in props", order = c("Republican", "Democrat") )
```

Confidence Interval via infer II

```
boot<-pew %>%
  specify(close_minded ~ party, success = "yes" ) %>%
  generate(reps = 1000, type = "bootstrap" ) %>%
  calculate( "diff in props", order = c("Republican", "Democrat") )
```

```
interval <-boot %>% get_confidence_interval(level = .95, type = "se",
              point_estimate = p_hat_r - p_hat_d)
interval
```

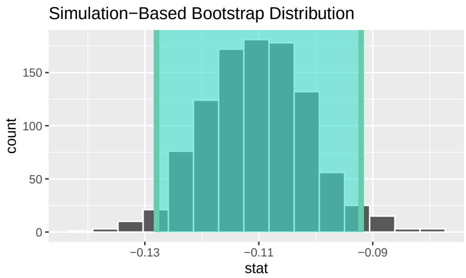
```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.128  -0.0920
```

Confidence Interval via infer II

```
boot<-pew %>%
  specify(close_minded ~ party, success = "yes" ) %>%
  generate(reps = 1000, type = "bootstrap" ) %>%
  calculate( "diff in props", order = c("Republican", "Democrat") )
```

```
interval <-boot %>% get_confidence_interval(level = .95, type = "se",
              point_estimate = p_hat_r - p_hat_d)
interval
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.128  -0.0920
```



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

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

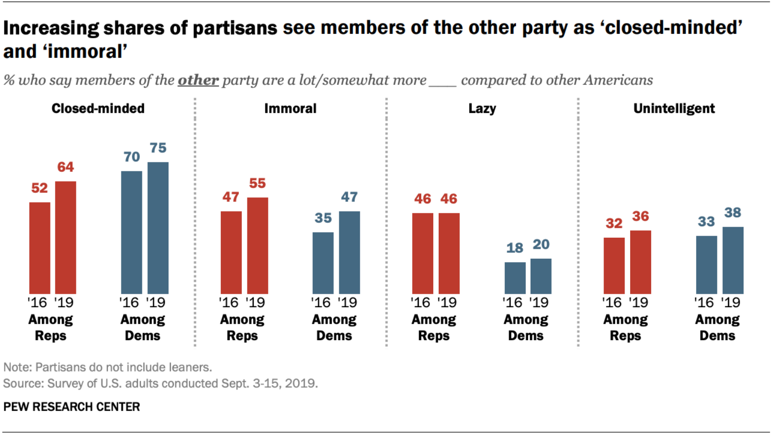
- **If the null hypothesis is true**, collecting a sample of sizes n_1 and n_2 from each population is the same as collecting a single sample of size $n_1 + n_2$.
  - So we may instead consider the pooled proportion $\hat{p}$ given by

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

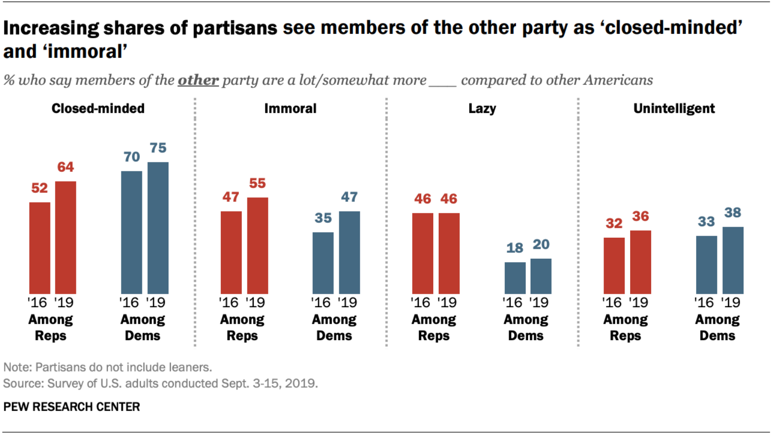
- This gives a standard error for the null distribution of

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

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

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

- Let's use the Normal approximation. In 2016, the number of participants was 4948 and in 2019, the number was 2947. This gives a pooled proportion of $\hat{p} = 0.725$

Hypothesis Tests

We test

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

- Let's use the Normal approximation. In 2016, the number of participants was 4948 and in 2019, the number was 2947. This gives a pooled proportion of $\hat{p} = 0.725$

```
n_16<-4948
n_19<-4947

p_hat_16<-.7
p_hat_19<-.75

p_hat<-(p_hat_16*n_16 + p_hat_19*n_19)/(n_16 + n_19)

p_hat
```

```
## [1] 0.7249975
```

Hypothesis Tests

We test

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

- Let's use the Normal approximation. In 2016, the number of participants was 4948 and in 2019, the number was 2947. This gives a pooled proportion of $\hat{p} = 0.725$

```
n_16<-4948
n_19<-4947

p_hat_16<-.7
p_hat_19<-.75

p_hat<-(p_hat_16*n_16 + p_hat_19*n_19)/(n_16 + n_19)

p_hat
```

```
## [1] 0.7249975
```

- The standard error for the null distribution is 0.009

```
SE <- sqrt( p_hat*(1- p_hat)/n_16 + p_hat*(1- p_hat)/n_19 )
SE
```

```
## [1] 0.008977568
```

Hypothesis Tests II

- Our test statistic is

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

```
z <- (p_hat_16 - p_hat_19)/SE
z
```

```
## [1] -5.569437
```

Hypothesis Tests II

- Our test statistic is

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

```
z <- (p_hat_16 - p_hat_19)/SE
z
```

```
## [1] -5.569437
```

- The P-value for this statistic is 0.00000002

```
P_value<-2*pnorm(z,0 ,1)
P_value
```

```
## [1] 2.555634e-08
```

Hypothesis Tests II

- Our test statistic is

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

```
z <- (p_hat_16 - p_hat_19)/SE
z
```

```
## [1] -5.569437
```

- The P-value for this statistic is 0.00000002

```
P_value<-2*pnorm(z,0 ,1)
P_value
```

```
## [1] 2.555634e-08
```

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

Hypothesis Tests II

- Our test statistic is

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

```
z <- (p_hat_16 - p_hat_19)/SE
z
```

```
## [1] -5.569437
```

- The P-value for this statistic is 0.00000002

```
P_value<-2*pnorm(z,0 ,1)
P_value
```

```
## [1] 2.555634e-08
```

- 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 popualtions truly had the same proportion.

Hypothesis Test via `infer`

Let's now use the `pew2` data

Hypothesis Test via infer

Let's now use the pew2 data

```
pew2 %>% group_by(year,close_minded) %>%
  summarize(N = n()) %>%
  mutate(prop = N / sum(N))
```

```
## # A tibble: 4 x 4
## # Groups:   year [2]
##   year  close_minded     N  prop
##   <fct> <fct>        <int> <dbl>
## 1 2016  no            1484 0.300
## 2 2016  yes           3464 0.700
## 3 2019  no            1237 0.250
## 4 2019  yes           3710 0.750
```

Hypothesis Tests via infer II

```
nulldist<-pew2 %>%
  specify(close_minded ~ year, success = "yes" ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute" ) %>%
  calculate( "diff in props", order = c("2016", "2019") )
```

Hypothesis Tests via infer II

```
nulldist<-pew2 %>%
  specify(close_minded ~ year, success = "yes" ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute" ) %>%
  calculate( "diff in props", order = c("2016", "2019") )
```

```
p_value <-nulldist %>% get_p_value(obs_stat = (p_hat_16 - p_hat_19),
              direction = "both")

p_value
```

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

Hypothesis Tests via infer II

```
nulldist<-pew2 %>%
  specify(close_minded ~ year, success = "yes" ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute" ) %>%
  calculate( "diff in props", order = c("2016", "2019") )
```

```
p_value <-nulldist %>% get_p_value(obs_stat = (p_hat_16 - p_hat_19),
              direction = "both")

p_value
```

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