Stat 201: Introduction to Statistics

Standard 31: Confidence Intervals – for Proportion Differences

Telling Which Parameter We’re After

- As statisticians, or data scientists, it’s our job to hear a problem and decide what we’re after
  - We call the parameter of interest the **target parameter**

Parameter	Point Estimate	Key Phrase	Type of Data
$\mu_1 - \mu_2$	$\overline{x_1} - \overline{x_2}$	Mean Difference	Quantitative
$\rho_1 - \rho_2$	$\widehat{p_1} - \widehat{p_2}$	Difference of Proportion, percentage, fraction, rate	Qualitative (Categorical)

Difference of Proportions

- We're often interested in comparing **different** groups of data.
- For confidence intervals, we follow similar steps from our proportions confidence intervals
- For hypothesis testing, we follow similar steps from our hypothesis testing for proportions but a couple of the formulas change

Difference of Proportions

- In the frame of Chapter 7, **sampling distributions**, we need to find the mean and standard deviation of **repeated samples** of proportional differences.

Sampling Distribution for Proportion Difference

- **The sample proportion difference** is the sample proportion of group 1 minus the sample proportion of group 2
  - $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$

Sampling Distribution for Proportion Difference

- **The population mean of all possible sample proportion differences** is the population proportion of group 1 minus the population proportion of group 2
  - $\mu_{\hat{p}_d} = \boldsymbol{\rho_1} - \boldsymbol{\rho_2}$

Sampling Distribution for Proportion Difference

- **The standard error, or the standard deviation of all possible sample proportion differences,** is seen below:

- $\boldsymbol{\sigma_{\hat{p}_d} = \sqrt{\frac{\rho_1(1-\rho_1)}{n_1} + \frac{\rho_2(1-\rho_2)}{n_2}}}$
  - where n_1 & n_2 are the number of people in each group

Difference of Proportions

- In the frame of Chapter 8, **confidence intervals**, we need to find the **point estimate** and **margin of error** so that we can come up with an interval estimate of the **population proportional difference**.

Confidence Intervals for Proportion Difference

- For $\hat{p}_{\mathrm{d}} = \hat{p}_1 - \hat{p}_2$ we have:

$$\mu_{\hat{p}_d} = \rho_1 - \rho_2$$

$$\sigma_{\hat{p}_d} = \sqrt{\frac{\rho_1(1-\rho_1)}{n_1} + \frac{\rho_2(1-\rho_2)}{n_2}}$$

Confidence Intervals: Step Two

- $\widehat{p_1} - \widehat{p_2}$ is our **point-estimate** for the difference of the population proportions
  - Our 'best' guess for the true difference of the population proportions, $\rho_1 - \rho_2$, is our difference of sample proportions, $\widehat{p_1} - \widehat{p_2}$.

Confidence Intervals: Step Two

- We consider $\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$ when we don't know ρ_1 and ρ_2 for the standard error as $\widehat{p_1}$ and $\widehat{p_2}$ can estimate the values of ρ_1 and ρ_2

Confidence Intervals: Step Two

- $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$ is our **margin of error**
  - **As either n increases**, the margin of error decreases causing the width of the confidence interval to narrow
  - **As either n decreases**, the margin of error increases causing the width of the confidence interval to grow wider
  - **If one decreases and the other increases** – we have to plug in the values to see what the overall effect is

Confidence Intervals: Margin of Error

- $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$ is our **margin of error**
  - **As the confidence level decreases**, z decreases causing the margin of error to decrease, causing the width of the confidence interval to narrow
  - **As the confidence level increases**, z increases causing the margin of error to increase, causing the width of the confidence interval to grow wider

Confidence Intervals: Step Two

- $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$ is our **margin of error**
  - $z_{1-\frac{\alpha}{2}}$ is the **confidence coefficient** and is the z value such that $P\left(Z < z_{\left(1-\frac{\alpha}{2}\right)}\right) = 1-\frac{\alpha}{2}$
  - $\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$ is the **estimated std. dev.**

Confidence Intervals – Step Two

- The most common values of Z are listed below
  - **Level of confidence** = (1-∝) * 100%
  - **Error Probability** = ∝ = 1- Level of confidence

Confidence	Error Probability (∝)	$z_{\left(1-\frac{\alpha}{2}\right)}$
.9	.1	1.645
.95	.05	1.96
.99	.01	2.58

  - Our interval will get larger when the margin of error increases
    1) When we increase confidence → increase z → widen interval
    2) When we decrease confidence → decrease z → narrow interval

Confidence Intervals Bounds

$$(\widehat{p_1} - \widehat{p_2}) \pm z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

"We are <u>--%</u> confident that the true difference of population proportions, $\boldsymbol{\rho_1 - \rho_2}$, is between the <u>lower bound</u> and <u>upper bound</u>."

Confidence Intervals for Proportion Difference

- Confidence interval is given by:

$$\hat{p}_{\mathrm{d}} \pm z_{1-\frac{\alpha}{2}}\left(\sigma_{\hat{p}_d}\right)$$

- If the resulting interval, (L,U), has both L and U less than 0 this suggests that the true proportional difference, $p_{\mathrm{d}} = p_1 - p_2$, is negative.
- $p_{\mathrm{d}} = p_1 - p_2 < 0$ indicates that $p_1 < p_2$, that group 2 has the greater proportion.

Confidence Intervals for Proportion Difference

- Confidence interval is given by:

$$\hat{p}_{\mathrm{d}} \pm z_{1-\frac{\alpha}{2}}\left(\sigma_{\hat{p}_d}\right)$$

- If the resulting interval, (L,U), has both L and U greater than 0 this suggests that the true proportional difference, $p_{\mathrm{d}} = p_1 - p_2$, is positive.
- $p_{\mathrm{d}} = p_1 - p_2 > 0$ indicates that $p_1 > p_2$, that group 1 has the greater proportion.

Confidence Intervals for Proportion Difference

- Confidence interval is given by:

$$\hat{p}_{\mathrm{d}} \pm z_{1-\frac{\alpha}{2}}\left(\sigma_{\hat{p}_d}\right)$$

- If the resulting interval, (L,U), contains 0 this suggests that the true proportional difference, $p_{\mathrm{d}} = p_1 - p_2$ can be 0.
- $p_{\mathrm{d}} = p_1 - p_2$ = 0 indicates that $p_1 = p_2$ is plausible and the two groups can have equal proportions.

Confidence Intervals Bounds

- If all the values on the interval are negative then $\rho_1 < \rho_2$
- If all the values on the interval are positive then $\rho_1 > \rho_2$
- If 0 is on the interval then it's possible that $\rho_1 = \rho_2$

Example

- 6,450 transgender and gender non-conforming study participants were asked about whether or not they maintained their family bonds.
- 2773 maintained their family ties of which 887 had attempted suicide
- 3677 experienced rejection from their family of which 1,875 had attempted suicide

Example

- 2773 maintained their family ties of which 887 had attempted suicide

$$\widehat{p_1} = \frac{887}{2773} = .31987$$

- 3677 experienced rejection from their family of which 1,875 had attempted suicide

$$\widehat{p_2} = \frac{1875}{3677} = .50993$$

Example

- **<u>Assumptions:</u>**
  1. Each sample must be obtained through randomization
  2. Samples are independent
  3. If all of the following are true
     - $2773\widehat{p_1} = 887 \geq 15$
     - $2773(1 - \widehat{p_1}) = 1886 \geq 15$
     - $3677\widehat{p_2} = 1875 \geq 15$
     - $3677(1 - \widehat{p_2}) = 1802 \geq 15$

Example

Find a 95% for the true difference of population proportions:

$$(\widehat{p_1} - \widehat{p_2}) \pm z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

$$(.31987 - .50993) \pm (1.959964) \sqrt{\frac{.31987(1-.31987)}{2773} + \frac{.50993(1-.50993)}{3677}}$$

$$(-.2137762 - .1663438)$$

Example

$$(-.2137762 - .1663438)$$

All the values on the interval are negative. This indicates $\rho_1 < \rho_2$ - that the population proportion of transgender and gender non-conforming people who are rejected by their family are more likely to attempt suicide than those that maintained their family ties

Summary!

Sampling Distribution for the Sample Proportion Summary

Shape of sample	Center of sample	Spread of sample
The shape of the distribution is bell shaped if $n * p \geq 15$ $\boldsymbol{and}$ $n * (1 - p) \geq 15$	$\mu_{\hat{p}_d} = p_1 - p_2$	$\sigma_{\hat{p}_d} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Confidence Intervals

Assumptions	1. *Random Sample* 2. At least 5 in each group 3. $n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$
Point Estimate	$\widehat{p_d} = \widehat{p_1} - \widehat{p_2}$
Margin of error	$z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$
Confidence Interval	$\widehat{p_d} \pm z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$

- We are --% confident that the true population proportion difference lays on the confidence interval.