Confidence Intervals for Proportions

Ryan Miller

Introduction

- A $P\%$ **confidence interval** is an interval *estimate of a population parameter* that is constructed using a procedure with a long-run $P\%$ success rate

Introduction

- A $P\%$ **confidence interval** is an interval *estimate of a population parameter* that is constructed using a procedure with a long-run $P\%$ success rate
- Today, we will see two different methods used to construct these intervals for categorical data (proportions and differences in proportions)

1) Using a normal distribution as suggested by Central Limit Theorem
2) Using the exact binomial distribution

Central Limit Theorem (One Proportion)

Central Limit Theorem describes the distribution of sample averages. For estimating a single population proportion p using a sample estimate $\hat{p}$, CLT suggests:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Central Limit Theorem (One Proportion)

Central Limit Theorem describes the distribution of sample averages. For estimating a single population proportion p using a sample estimate $\hat{p}$, CLT suggests:

$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

Notice three things:

1) The sample estimate is *unbiased* for p
2) The variability in estimates we could have observed can be described by the *standard error*, $SE = \sqrt{\frac{p(1-p)}{n}}$
3) This standard error, along with the normal curve, can used to find percentiles containing the $P\%$ of sample estimates

Central Limit Theorem (One Proportion)

Taken together, these suggest confidence intervals of the form:

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

- Where z^* indicates the percentile of the standard normal distribution such that the middle $P\%$ of distribution is between $(-z^*, +z^*)$
  - For example, $z^* = 1.96$ for 95% confidence intervals, because the middle 95% of the standard normal curve lies between -1.96 and $+1.96$

Central Limit Theorem (One Proportion)

Taken together, these suggest confidence intervals of the form:

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

- Where z^* indicates the percentile of the standard normal distribution such that the middle $P\%$ of distribution is between $(-z^*, +z^*)$
  - For example, $z^* = 1.96$ for 95% confidence intervals, because the middle 95% of the standard normal curve lies between -1.96 and $+1.96$
- We call $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ the *standard error* (SE) because it describes the variability (sampling error) of $\hat{p}$ as an estimate
  - We use $\hat{p}$ in place of the unknown population parameter p because it is our *best estimate*

Example

- Consider a random sample of $n = 100$ claims from the `tsa` dataset

```
set.seed(123)
sample_id <- sample(1:nrow(tsa), size = 100)
tsa_sample <- tsa[sample_id,]
sum(tsa_sample$Status == "Denied")
```

```
## [1] 46
```

- In this sample, 46 of 100 claims were denied
  - How would you find a 99% confidence interval estimate of the proportion of *all claims* that are denied?

Example (solution)

- Using CLT results for a single proportion, we simply need to plug-in the proper values into $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Clearly, $\hat{p} = 46/100 = .46$ and $n = 100$
  - All that remains is to find z^* for the middle 99% of the standard normal curve

Example (solution)

- Using CLT results for a single proportion, we simply need to plug-in the proper values into $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Clearly, $\hat{p} = 46/100 = .46$ and $n = 100$
  - All that remains is to find z^* for the middle 99% of the standard normal curve

```r
qnorm(.995, mean = 0, sd = 1, lower.tail = TRUE)
```

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

- Taken together, we arrive at the interval estimate: $.46 \pm 2.58\sqrt{\frac{.46 * .54}{100}} = (0.33, 0.59)$
- The population proportion was $p = 0.417$, so this interval was successful!

Another Perspective

- Recognize that our previous confidence interval was based upon Central Limit Theorem, a result that only holds for large sample sizes
  - Let's now consider a random sample of $n = 10$ claims

```
set.seed(123)
sample_id <- sample(1:nrow(tsa), size = 10)
tsa_sample <- tsa[sample_id,]
sum(tsa_sample$Status == "Denied")
```

```
## [1] 3
```

- The 3 of 10 denials in this sample lead to the 99% CI: $.3 \pm 2.58\sqrt{\frac{.3*.7}{10}} = (-0.07, 0.67)$
- Do you notice any problems with this interval?

- We don't want an interval that suggests negative proportions are plausible!

Exact Binomial Intervals

- We don't want an interval that suggests negative proportions are plausible!
- One way to avoid this issue by focusing on the *number of successes*, $\sum_i x_i$, rather than the *proportion of successes*, $\hat{p} = \sum_i x_i / n$
  - The sample proportion (of successes) follows a normal distribution for *large n*
  - The sample sum of successes follows a binomial distribution for *any n*

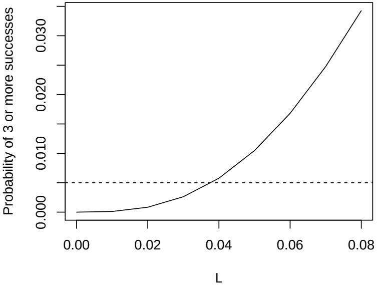
Exact Binomial Intervals

To construct a 99% interval estimate, we need to find two numbers (L, U) that provide a 99% long-run chance of containing p. A simple way to do this is trial and error, that is:

1) Check a bunch of values for L and U
2) Anything where the binomial probability of seeing a result *at least as extreme* as $\sum_i x_i = 3$ is less than 0.5% (half of the excluded 1%) gets *excluded* from the interval

Exact Binomial Intervals

- This process can be more easily understood visually
  - The left panel shows the search for the interval's lower endpoint (a proportion where $P(\sum_i X_i \geq 3) = 0.005$)
  - The right panel shows the search for the interval's upper endpoint (a proportion where $P(\sum_i X_i \leq 3) = 0.005$)



Exact Binomial Intervals

While it's difficult to find these endpoints by hand, it's quite easy for a program like R:

```
output <- binom.test(3,10, conf.level = .99)
output$conf.int[1:2]
```

```
## [1] 0.03700722 0.73511399
```

Exact Binomial vs. CLT Approximation

- Obviously the exact binomial interval worked better for sample of size $n = 10$ (it didn't suggest negative proportions were plausible)
  - But in general, how should we decide between these two approaches?

Exact Binomial vs. CLT Approximation

- Obviously the exact binomial interval worked better for sample of size $n = 10$ (it didn't suggest negative proportions were plausible)
  - But in general, how should we decide between these two approaches?
- If you have access to `R`, there's really no reason not to use the exact approach (after all, it's exact)
- However, the normal approximation (CLT) method will produce a nearly identical result when the following criteria are met:
  - $n\hat{p} \geq 10$
  - $n(1 - \hat{p}) \geq 10$

Conditional Proportions

Let's now consider two different *conditional proportions*:

1) The proportion of denied claims at checkpoints, denoted $p_{\text{de}|\text{chk}}$
2) The proportion of denied claims at baggage checks, denoted $p_{\text{de}|\text{bag}}$

Conditional Proportions

```r
set.seed(123)
sample_id <- sample(1:nrow(tsa), size = 100)
tsa_sample <- tsa[sample_id,]

my_table <- table(tsa_sample$Claim_Site,
                  tsa_sample$Status)
addmargins(my_table)
```

```
##
##                   Approved Denied Settled Sum
##   Checked Baggage       21     40      23  84
##   Checkpoint             9      6       1  16
##   Sum                   30     46      24 100
```

- In *this sample* of $n = 100$, $\hat{p}_{\text{de}|\text{chk}} = 6/16 = 0.375$ and $\hat{p}_{\text{de}|\text{bag}} = 40/84 = 0.476$
- In the *population*, is it possible that claims at checkpoints and baggage checks are equally likely to be denied?

Comparing Proportions

- We can estimate the population proportions using the sample data, let's use 90% confidence intervals

```r
## Checkpoint claims
binom.test(6,16, conf.level = .90)$conf.int[1:2]
```

```
## [1] 0.1777659 0.6089884
```

```r
## Baggage Claims
binom.test(40,84, conf.level = .90)$conf.int[1:2]
```

```
## [1] 0.3823842 0.5712900
```

- Notice the substantial overlap between these intervals, does that mean that claims at baggage checks are equally likely to be denied?

Comparing Proportions

- We can estimate the population proportions using the sample data, let's use 90% confidence intervals

```
## Checkpoint claims
binom.test(6,16, conf.level = .90)$conf.int[1:2]
```

```
## [1] 0.1777659 0.6089884
```

```
## Baggage Claims
binom.test(40,84, conf.level = .90)$conf.int[1:2]
```

```
## [1] 0.3823842 0.5712900
```

- Notice the substantial overlap between these intervals, does that mean that claims at baggage checks are equally likely to be denied?
  - Not necessarily, while confidence intervals report a *range of plausible values*, not all of these values are equally plausible

Comparing Proportions

- In this scenario, it's more efficient to look at the *difference in proportions*
- We observed, $\hat{p}_{\text{de}|\text{chk}} - \hat{p}_{\text{de}|\text{bag}} = 0.375 - 0.476 = -0.101$
  - Is it plausible that the population difference, $p_{\text{de}|\text{chk}} - p_{\text{de}|\text{bag}}$, is zero?

Comparing Proportions

- In this scenario, it's more efficient to look at the *difference in proportions*
- We observed, $\hat{p}_{\text{de}|\text{chk}} - \hat{p}_{\text{de}|\text{bag}} = 0.375 - 0.476 = -0.101$
  - Is it plausible that the population difference, $p_{\text{de}|\text{chk}} - p_{\text{de}|\text{bag}}$, is zero?
- To answer this question, we'll need to apply some probability theory to our previous central limit result

Linear Combinations of Random Variables

- We know that $\hat{p}_{\mathsf{de}|\mathsf{chk}}$ and $\hat{p}_{\mathsf{de}|\mathsf{bag}}$ each have sampling distributions that are normal (CLT)

Linear Combinations of Random Variables

- We know that $\hat{p}_{\text{de}|\text{chk}}$ and $\hat{p}_{\text{de}|\text{bag}}$ each have sampling distributions that are normal (CLT)
- Consider two *independent* random variables, X and Y, that follow $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$ distributions (respectively)
  - A linear combination of these variables, $aX + bY$, will follow a normal distribution with mean $a\mu_X + b\mu_Y$ and standard deviation $\sqrt{a^2\sigma_X + b^2\sigma_Y}$

Linear Combinations of Random Variables

- We know that $\hat{p}_{\text{de}|\text{chk}}$ and $\hat{p}_{\text{de}|\text{bag}}$ each have sampling distributions that are normal (CLT)
- Consider two *independent* random variables, X and Y, that follow $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$ distributions (respectively)
  - A linear combination of these variables, $aX + bY$, will follow a normal distribution with mean $a\mu_X + b\mu_Y$ and standard deviation $\sqrt{a^2\sigma_X + b^2\sigma_Y}$
- Therefore, letting $p_1 = p_{\text{de}|\text{chk}}$ and $p_2 = p_{\text{de}|\text{bag}}$:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

- Note that n_1 is the size of the first group, and n_2 is the size of the second group

Approximate CI for a Difference in Proportions

This distributional result suggests the following formula for $P\%$ confidence intervals:

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Approximate CI for a Difference in Proportions

This distributional result suggests the following formula for $P\%$ confidence intervals:

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- This formula relies upon *two* successful normal approximations, leading to a rather long set of criteria for these intervals to be valid
  - p_1 and p_2 are independent
  - $n_1 \hat{p}_1 \geq 10$
  - $n_1(1 - \hat{p}_1) \geq 10$
  - $n_2 \hat{p}_2 \geq 10$
  - $n_2(1 - \hat{p}_2) \geq 10$

Example

- For our example involving denied claims at checkpoints and baggage checks, we observed $\hat{p}_1 = \hat{p}_{\mathsf{de|bag}} 6/16 = 0.375$ and $\hat{p}_2 = \hat{p}_{\mathsf{de|bag}} = 40/84 = 0.476$

Example

- For our example involving denied claims at checkpoints and baggage checks, we observed $\hat{p}_1 = \hat{p}_{\text{de}|\text{bag}} 6/16 = 0.375$ and $\hat{p}_2 = \hat{p}_{\text{de}|\text{bag}} = 40/84 = 0.476$
- p_1 and p_2 are clearly independent, since no single claim occurs at both a checkpoint and baggage check
  - However, notice $n_1 \hat{p}_1 = 16 * .375 \geq 10$, so we should be cautious using this result

Example

- For our example involving denied claims at checkpoints and baggage checks, we observed $\hat{p}_1 = \hat{p}_{\text{de}|\text{bag}} 6/16 = 0.375$ and $\hat{p}_2 = \hat{p}_{\text{de}|\text{bag}} = 40/84 = 0.476$
- p_1 and p_2 are clearly independent, since no single claim occurs at both a checkpoint and baggage check
  - However, notice $n_1 \hat{p}_1 = 16 * .375 \geq 10$, so we should be cautious using this result
- Nevertheless, let's use the previously stated formula to compute a 90% confidence interval estimate for $p_1 - p_2$:

$$.375 - 0.465 \pm 1.65 \sqrt{\frac{.375(1-.375)}{16} + \frac{.476(1-.476)}{84}} = (-0.31, 0.13)$$

- So, it is plausible that there is no difference in these proportions (in the population!)

Exact Intervals for Differences in Proportions?

- While it is possible to find an exact confidence interval for a difference in proportions, it isn't very often that statisticians do so
- Instead, two proportions are often compared using *odds ratios* (leading many methods for constructing confidence intervals tend to be focused on odds ratios)
  - We will discuss odds ratios later in the semester

Summary

- In this lecture we discussed two methods for constructing $P\%$ confidence intervals for a single proportion
  - A normal approximation approach based upon the Central Limit Theorem
  - An exact approach that uses the binomial distribution

Summary

- In this lecture we discussed two methods for constructing $P\%$ confidence intervals for a single proportion
  - A normal approximation approach based upon the Central Limit Theorem
  - An exact approach that uses the binomial distribution
- We also learned a CLT-based approach that can be used for differences in proportions
  - We will not learn an exact approach for this scenario until later on when we discuss odds ratios