Confidence Intervals for Proportions

Grinnell College

Review

We have seen how to make CI's for pop. means or the diff. in pop. means

Confidence intervals will always have the following form:

$$\text{statistic} \pm C \times \text{SE}$$

There is a quick-guide available on the course page that will be a good reference for you as we continue.

Outline

Today we will learn how to work with proportions.

- CI for a single population proportion (p)
- CI for a difference in population proportions (p_1 - p_2).

These slides will not be extensive. Once you know how to make CI's for means like you have seen, proportions are not much different.

Means vs. Proportions

Let's think back to what type of variables we are working with.

Quantitative Variables

- use means (and standard deviation) to describe the population
- one group → one mean (μ)
- two groups → diff. in means ($\mu_1 - \mu_2$)

Categorical Variables

- use the proportion of a category to describe the population
- one group → one proportion (p)
- two groups → diff. in proportions (p_1 - p_2)

Proportions – Examples

Let's say we want to find out if a coin is "fair." We could flip the coin a whole bunch of times

- outcomes are Heads or Tails → categorical
- record the 'proportion of heads'
- should be close to 0.5 if coin is fair

In an election year in the U.S., political campaigns (and political groups in general) are interested in finding out how people are going to vote

- who someone will vote for → categorical
- we want to estimate 'proportion who will vote for a candidate'

CIs for Proportions

We will now go through and explain the logic related to how CI's for proportions was developed.

Relationship between Means and Proportions

There is an interesting relationship between means and proportions

For example, consider taking a fair coin and flipping it 10 times. How many heads would you expect to see?

Relationship between Means and Proportions

Say we flip the coin 10 times and get the following sample $\mathcal{S}$.

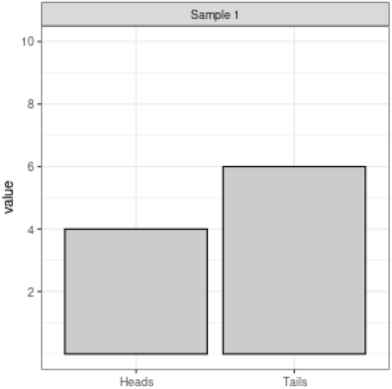
$$\mathcal{S} = \{H, H, T, T, H, T, H, T, T, T\}$$
$$X = \{1, 1, 0, 0, 1, 0, 1, 0, 0, 0\}$$

We can find the *proportion* of heads from our sample $\mathcal{S}$ by simply taking the total number of heads and dividing by the total number of flips, giving

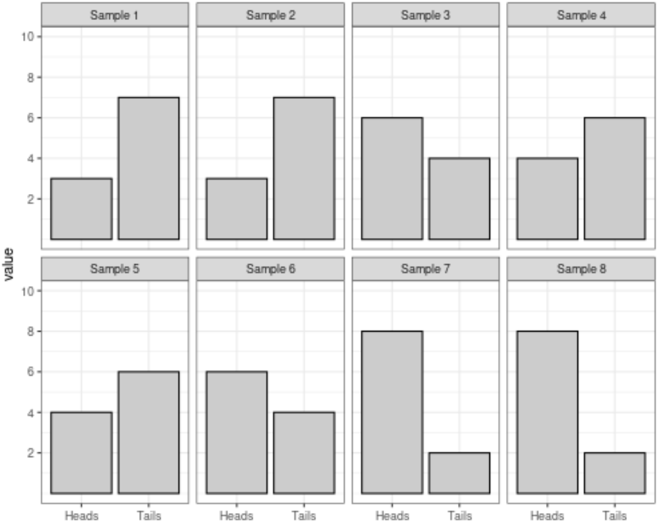
$$\hat{p} = \frac{4}{10}$$

However, if we consider X, which defines H as 1 and T as 0, we can also find the sample mean:

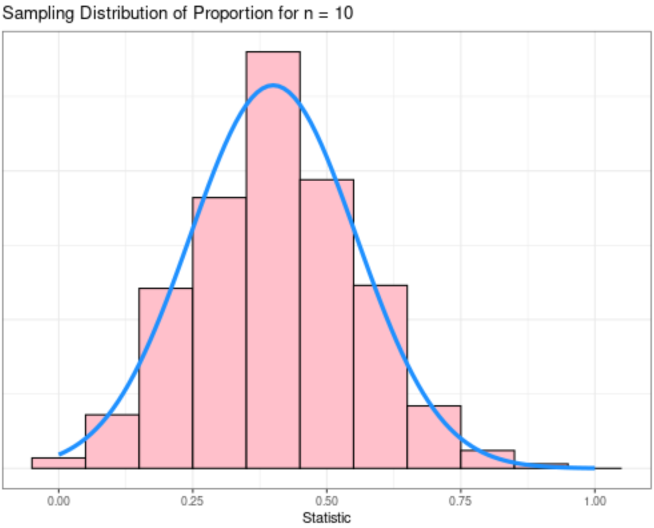
$$\begin{aligned}\overline{x} &= \frac{1}{10}\sum_{i=1}^{n} x_i \\ &= 0.4\end{aligned}$$



Repeated Samples



Sampling Distr. for the Proportion of Heads



- What does this look like?

Central Limit Theorem

For a sample with one proportion, the sampling distribution of our proportion statistic, $\hat{p}$ is approximately

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

There a few rules of thumb relating to the size and the proportion:

1. $n \times p \geq 10$
2. $n \times (1-p) \geq 10$

In particular, it is often difficult to estimate proportions precisely that are near the boundaries (0 and 1)

Back to CI's

So... the sampling dist. for a population mean looks Normal!

- we used this result to make CIs for means
- we are going to use it for CIs for proportions too

Confidence intervals will always have the following form:

$$\text{statistic} \pm C \times \text{SE}$$

- C will be determined by the confidence just like we've seen
- 95% confidence → C=1.96
- we will **always** use the normal distribution for proportions

CI for a population proportion

$$\text{statistic} \pm C \times \text{SE}$$

Want to estimate the population proportion p. Use what we have

- statistic $= \widehat{p}$

We can get the SE from the CLT result for proportions

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

- we don't know the value for p → SE $= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$

CI for a population proportion

Our final formula for estimating p (a population proportion):

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

- $\widehat{p}$ is the sample proportion
- n is our sample size
- z^{*} is the appropriate value from the normal distribution that gives us the Confidence % that we want
  - 95% Confidence $\rightarrow z^{*} = 1.96$
  - 80% Confidence $\rightarrow z^{*} = 1.282$
  - 90% Confidence $\rightarrow z^{*} = 1.645$
  - 99% Confidence $\rightarrow z^{*} = 2.576$

CI for proportion – Conditions

The conditions required for the CI to work well for a pop. proportion:

- Random sample
- $n \times p \geq 10$ (Success Condition)
- $n \times (1 - p) \geq 10$ (Failure Condition)

Success and Failure conditions get their name from counting the # of successes and failures respectively. When checking conditions:

- say that each condition is "met" or "not met"
- i.e.: random sample (met / not met)
- i.e.: Success condition (met / not met)
- i.e.: Failure condition (met / not met)

Example

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months. With your group:

1. Use a normal approximation to construct a 95% confidence interval estimate for the true proportions of babies born at 25 weeks gestation that are expected to survive
2. An article on Wikipedia suggests that 70% of babies born at a gestation period of 25 weeks survive. Is this claim consistent with the Johns Hopkins study?

Example

1. We find that

$$\hat{p} = \frac{31}{39} = 0.795$$

$$SE = \sqrt{\frac{0.795(1 - 0.795)}{39}} = 0.065$$

   From here, we found our 95% CI:

$$0.795 \pm 1.96 \times 0.065 = (0.668, 0.922)$$

2. As 0.7 is contained within our constructed 95% CI, it is consistent with the results of the study by Johns Hopkins

Difference in Proportions

Just like how we sometimes wanted to find a difference in means... we can do the same thing for proportions

- need to be careful that we are not confusing ourselves

When do we want to estimate a proportion vs. difference in proportions?

- we need to have 2 groups defined by a 2nd variable
- this is not the same as having multiple categories for the categorical variable we are working with

Difference in Proportions

Example: Coin flips

- We can estimate proportion of heads or proportion of tails
- if we know one, we know the other by default
- Heads/Tails are different categories of the variable of interest
- we do not need to estimate a difference in proportions

Example: New medical treatment for headaches

- We can compare survival rates between treatment & control groups
- 2nd variable: whether someone received the treatment (Yes/No)
- two groups → need to estimate a *difference* in means

CI for difference in population proportions

$$\text{statistic} \pm C \times \text{SE}$$

Want to estimate the difference in population proportions $(p_1 - p_2)$

- statistic $= \widehat{p}_1 - \widehat{p}_2$
- $\text{SE} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

CI for difference in population proportions

Our final formula for estimating $p_1 - p_2$ (diff. in pop. proportions):

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

- $\widehat{p}_1$, $\widehat{p}_2$ are the sample proportions
- n_1, n_2 are the sample sizes
- z^* is the appropriate value from the normal distribution that gives us the Confidence % that we want
  - 95% Confidence $\rightarrow z^* = 1.96$
  - 80% Confidence $\rightarrow z^* = 1.282$
  - 90% Confidence $\rightarrow z^* = 1.645$
  - 99% Confidence $\rightarrow z^* = 2.576$

CI for diff. in proportions – Conditions

The conditions required for the CI to work well for a difference in proportions are:

- Groups are independent of each other (not influencing each other)
  - satisfied if we have a random sample
- $n_1 \times p_1 \geq 10$
- $n_1 \times (1 - p_1) \geq 10$
- $n_2 \times p_2 \geq 10$
- $n_2 \times (1 - p_2) \geq 10$