Introduction to Inference

Confidence Intervals for Proportions

1

Recap of CI's

- The confidence interval (CI): $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$

for the proportion of a normal population illustrates several important properties that are shared by all confidence intervals. A small margin of error says that we have pinned down the parameter quite precisely. The margin of error is

$$\text{margin of error} = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

We can decrease the margin of error by:

- ✓ Using a smaller **z***. Smaller z* is the same as smaller confidence interval C. We have less confidence but more accuracy.
- ✓ Using a larger **n**. Increasing the sample size **n** reduces the margin of error for any fixed confidence level.

2

Sample size, n

- Sample size **n** for desired **margin of error**.

To determine the sample size n that will yield a confidence interval for a population mean with a **specified margin of error m**,

$$z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} \le m, \text{ then solve for n.}$$

Always use **confidence intervals** when your goal is to estimate a population parameter.

3

Calculation of CI's.

- Calculation of Confidence Intervals is straight forward.

($\hat{p}$) ± (**z***)(standard error of the estimate)

Estimate ± margin of error

Definition: The standard error of a statistic is the estimated standard deviation of the statistic, $\sqrt{\frac{\hat{p}\hat{q}}{n}}$

4

Watch for the following mistakes

- **Don't suggest that the parameter varies**. A statement like "There is a 95% chance that the true proportion is between 42.7% and 51.3%" sounds as though you think the population proportion wanders around and sometimes happens to fall between 42.7% and 51.3%. When you interpret a CI, make it clear that the ***INTERVAL*** varies from sample to sample.
- **Don't assume that other samples will be just like yours**. Statements like "95% of the samples will fall between 42.7% and 51.3%" are just plain wrong! Remember the distribution is based off of the parameter, not the sample statistic!!! The sample goes 1.96 SD to the left and right of 47, but the actual sampling distribution goes 1.96 SD to the left and right of the true center!

5

Watch for the following mistakes

- **Don't be certain about the parameter**. Saying "Between 42.7% and 51.3% of sea fans are infected" asserts that the population cannot be outside the interval. We can't be certain about the parameter without taking a census.
- **Don't forget that we are making a statement about the parameter**. Don't say "I'm 95% confident that $\hat{p}$ is between 42.7% and 51.3%." Of course it is–we just calculated it!!! We already know the what percent of the sample is infected, what we want to know is what percent of the population, *p*, is infected!!!
- **Don't claim to know too much**. You can say, "I'm 95% confident that between 42.7% and 51.3% of all sea fans are infected" only if your sample is an SRS from the entire population. Otherwise this statement is a type of extrapolation.
- **What is the correct way of describing the CI?**

6

Watch for the following mistakes

- ***Make sure that the margin of error is small enough to be useful.*** For example, let's say you are trying to determine the average height of high school students. If you claim that you are 95% confident that the true mean of heights of high school students is between 12 in. and 120 in. (or between 1 ft and 10 ft tall), you are not giving anyone any useful information!
- ***Beware of various types of bias.*** If you send out a survey to 2400 people and only a small proportion of respondents return the questionnaire, it has essentially turned into a voluntary response study – which is of little value for inferring information about the population. It's almost always better to spend resources on increasing the response rate than on surveying a larger group.

Watch for the following mistakes

- ***Assumption violation.*** It is **important** to validate that your assumptions are true by always checking your conditions. In addition, ALWAYS WATCH FOR:
  - Biased samples
  - Violation of independence
  - Skewed Distributions with small sample size

Things to Remember About Inference

- Make sure that the data comes from an Random Sample
- No matter what, a poorly designed study will always produced results that are suspect and invalid.
- Outliers can have a large effect on CI's, so beware.
- If the sample size is small and not normal, the intervals will be tainted. Try to remove outliers if possible.

Critical Values

- **Critical values** are the factors used to calculate the Margin of Error
- **Margin of Errors** express the maximum expected difference between the true population parameter and a sample estimate of that parameter. To be meaningful, the margin of error should be qualified by a probability statement

Examples

- What is the approximate critical value if you want to have a 99.7% Confidence Level?
  - According to the 68-95-99.7 rule, we need to go approximately three standard deviation to obtain the middle 99.7%. So the approximate z^* is 3.
- What is the true critical value if you want to have a 95% Confidence Level?
  - Since the 68-95-99.7 rule gives only approximations, we need to find the standard deviations to obtain the middle 95%. invNorm(p) gives the z-score for the area with an area of p. Therefore, the z* is 1.96 given by invNorm(.025) ≈ 1.96

Examples

- What is the true critical value if you want to have a 99.9% Confidence Level?
  - $z^* \approx 3.291$
- What is the true critical value if you want to have a 80% Confidence Level?
  - $z^* \approx 1.282$
- What is the true critical value if you want to have a 75% Confidence Level?
  - $z^* \approx 1.15$

Example

- Find the proportion of A's of the Precalc test in Mr. B's class using an 80%, 96%, and 99.99% confidence interval:
  - Here are 40 randomly selected test scores:

    98, 81, 78, **93**, 69, 77, 82, **96**, **92**, 81
    73, 86, **96**, **95**, 65, 75, 76, 86, 73, 85
    60, 62, 81, **98**, **91**, 58, 73, 42, 71, 88
    90, 73, 75, 87, 60, 79, **95**, 84, 85, 80