2.05 Confidence Intervals

Recap: Data Science Workflow

1. Define the problem.
2. Obtain the data.
3. Explore the data.
4. Model the data.
5. Evaluate the model.
6. Answer the problem.

Statistical Inference Focus Areas

Today, we are going to discuss the process of statistical inference.

- That is, how do we get from our statistics (measures of samples) to our parameters (measures of populations)?

In frequentist statistical inference, there are two main ways to generalize from a sample to a population:

- Confidence Intervals
- Hypothesis Tests

Statistical Inference Steps

1. We identify our population.

2. We gather a random sample of data from the population.

3. We calculate some statistic(s) based on our sample.

4. Using statistics, we conduct inference on the parameters.

5. We use our understanding of parameters to make conclusions about population.

Populations

Most data science problems have to do with studying populations in some form or another.

- Examples:
  - All undergraduates currently at Ohio State.
  - All microwaves constructed at my factory this year.
  - All hurricanes to enter the Gulf of Mexico.
  - All people who will vote in the 2020 election.
  - All states (and their average standardized test scores).

Populations

- If we're interested in learning about populations, why don't we just measure the population directly?

- What might we do instead?

Statistical Inference – Example #1

I want to see who will win the California U.S. Senate election in 2020. I call 1,000 registered voters and ask who they will support.

- Population: All eligible voters in California
- Sample: 1,000 registered voters in California
- Statistic: Sample Proportion of votes
- Parameter: Population Proportion of votes

Statistical Inference – Example #2

I developed a new drug (“New Drug”) that I believe reduces the diastolic blood pressure of adults over 50. I lead a clinical trial of 100 patients, where I compare my drug to the standard drug (“Old Drug”).

- Population: All adults aged above 50
- Sample: 100 patients
- Statistic: Sample Mean of blood pressure
- Parameter: Population Mean of blood pressure

What is a Confidence Interval?

- A confidence interval displays the probability that a parameter will fall between a pair of values around the mean
- Confidence intervals measure the degree of certainty in a sample
- They are most often constructed using confidence levels of 95% or 99%.
- Confidence intervals are established using hypothesis tests such as a t-test

Confidence Interval Example

95% confidence interval: [0.48, 0.54]

Estimate = 0.51 ± 0.03 (at 95% confidence)

Point estimate

Margin of error

Common Misconception about Confidence Intervals

- The biggest misconception regarding confidence intervals is that they represent the percentage of data from a given sample that falls between the upper and lower bounds.

- In other words, it would be incorrect to assume that a 99% confidence interval means that 99% of the data in a random sample falls between these bounds.
  - On the contrary, what it actually means is that one can be 99% certain that the range will contain the population mean.