Week 6: Confidence Intervals

Jennifer Lin

2021-10-28

Observational Studies and Large N Methods

Sometimes, we want to study things where we cannot randomly assign participants to the independent variable. In that case, we can only conduct **observational studies**. **Observational Studies** are research projects where we don't assign values to the independent variable. We just see what happens.

Some examples of these projects include:

- Does living in a rural area predict conservative political attitudes?
- What are some reasons that people are not getting the COVID-19 vaccine?
- How do people perceive others who are vaccinated or not vaccinated?

To analyze the data in observational studies, we often require large N statistical methods. These are statistics tests that we use to analyze observational studies. For these methods, we care about the **population** *parameter* and **sample** *statistic*.

Population Parameters

Parameters are aspects of the population that we care about. They are often denoted using Greek letters (μ, π, α, β). They capture something that is true for the entire population.

Some examples of this include:

Examples

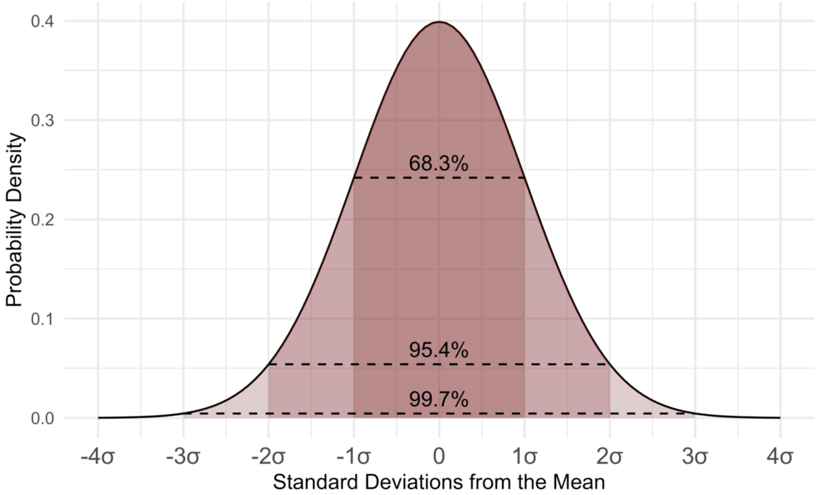
- Total number of registered Republicans and Democrats in the United States
- Total number of people in the US who love vanilla ice cream

However, remember that it is difficult to study the entire population. Think back to the week on survey deign and sampling. We cannot expect to ask everyone in the population. Therefore, we can only expect to **estimate** the population parameters *using statistics*.

Statistics

Statistics are estimates of the population using data that we gather from a subset of the population. We collect the samples using the sampling methods discussed earlier in the quarter (remember: Probability versus nonprobability sampling?). We can calculate sample statistics like mean from the sample and use them to predict the population. Sample statistics are usually denoted by letters (p, a, b).

The question then becomes: How do we know how "right" our predictions are? To answer this question, think back to the first week when we discussed models in this section. Remember that models are inherently missing something. Sampling statistics are models of the population. So how do we know how correct our models are especially since each of the sampling methods can be imperfect? We can use **confidence Intervals**, which signals how *sure* we are about our predictions. They show us a range of values [lower, upper] that is possible for the population parameter to lie given the findings from our sample. The most common confidence interval is the 95%, which shows that we are confident that the true population parameter (usually mean) is within the bounds of the calculated range 95% of the time[1].



Here are some ways to remember the gist of a confidence interval:

- When calculating confidence intervals, you are casting a net for the true population mean. Therefore, the wider the net you cast, the more likely you are to get it.
- When calculating confidence intervals, you are opening an umbrella to keep yourself dry in a rainstorm. The larger the umbrella, the more likely it is that you keep yourself dry.

Calculating Confidence Intervals

To calculate a confidence interval, we need three pieces of information

1. The mean (or parameter of interest)
2. The sample size

[1] Capturing the population parameter is our definition of success.

3. The confidence percentage

The Confidence Percentage

We need to determine how wide of a net to cast to ensure success. Traditionally, this is 95%. But, we can have other percentages as well. From there, we need to get our corresponding z-score[2].

Percent	z-score
90%	1.65
95%	1.96
99%	2.57

The Mean (or any other sample statistic)

For means, we can compute this as we usually compute averages from sample statistics[3]. Essentially, this component is a descriptive statistic that we derive from the sample we have drawn.

The Standard Error

The standard error is the standard deviation of the sampling distribution. This is based on the sample size and is computed as follows:

$$SE = \sqrt{\frac{p \times (1-p)}{N}}$$

Confidence Interval

Pulling this together, the way we calculate a confidence interval is as follows:

$$\overline{x} \pm z(\sqrt{\frac{p \times (1-p)}{N}})$$

An Example of How to Calculate a Confidence Interval

Suppose I polled a sample of Northwestern students to whether or not the approve of the quarter system. I want to predict whether the entire population of students approve the quarter system over the semester system.

To do this, I polled 40 students outside of the Norris University Center. I get the following:

Mean: .75 say they approve of the quarter system

N: 40

I want to compute a 95% Confidence Interval

[2] **z-scores** are standard scores based on a normal distribution with a mean of 0 and a standard deviation of 1 ($N(0,1)$)

[3] For means, this is often denoted as $\overline{x}$

First, I calculate the standard error

$$SE = \sqrt{\frac{.75 \times (1 - .75)}{40}} \approx 0.06846532$$

Then, I plug it into the broader equation

$$.75 \pm 1.96(0.06846532)$$

$$.75 \pm 0.134192$$

And the result is

$$[0.615808, 0.884192] \text{ or } [0.62, 0.88]$$

Bow, suppose I take the same condition but I want to estimate a 90% confidence interval. Here, I can use the same standard error (since no part of that has changed) and rewrite the calculations as follows:

$$.75 \pm 1.65(0.06846532)$$

$$.75 \pm 0.1129678$$

$$[0.6370322, 0.8629678] \text{ or } [0.63, 0.86]$$

As we can see, the interval in the 90% is smaller than the 95% since we are casting a smaller net.

Finally, now I am interested in calculating a 99% confidence interval. I can use the same conditions as I have before and rerun the math. Here is how that looks:

$$.75 \pm 2.57(0.06846532)$$

$$.75 \pm 0.1759559$$

$$[0.5740441, 0.9259559] \text{ or } [0.57, 0.93]$$

Notice that, compared to the 95% interval, this interval is wider and this is because we have casted a bigger net.

Changing up the scenario, suppose I am interested in the same question as before but instead of sampling 40 students, I have sampled 400 students outside of Norris instead[4].

To find the same 95% confidence interval, I can follow the same steps as before, starting with the standard error calculations.

[4] Keep in mind that the 40 student sample gave us a 95% CI of $[0.62, 0.88]$.

$$SE = \sqrt{\frac{.75 \times (1 - .75)}{400}} \approx 0.02165064$$

And then, we can continue the calculations by plugging everything into the main equation.

$$.75 \pm 1.96(0.02165064)$$

$$.75 \pm 0.04243525$$

$$[0.7075647, 0.7924353] \text{ or } [0.70, 0.79]$$

Notice that, with the bigger sample, even on the same confidence level, our confidence interval closes around the mean. This is because a larger sample allows us to get a better sense of the population, making it a more precise estimate (as opposed to accurate since we are still not sure if our sample is actually good at capturing the true mean.)

Expanding to Two Groups for Compairson

Now suppose I want to expand this for comparisons

$$SE = \sqrt{\frac{p_i \times (1 - p_i)}{N_i} + \frac{p_j \times (1 - p_j)}{N_j}}$$

Where i and j are our two groups. The full calculation is as follows:

$$(p_i - p_j) \pm z(\sqrt{\frac{p_i \times (1 - p_i)}{N_i} + \frac{p_j \times (1 - p_j)}{N_j}})$$