Welcome!

Week 6: Confidence Intervals

TA: Jennifer Lin

PS 210: Introduction to Empirical Methods

2021-10-28

Observational Studies and Large N Methods

- **Observational Studies** are research projects where we don't assign values to the independent variable. We just see what happens
- Examples
    - 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?

Large N Methods

Statistics tests that we use to analyze observational studies

For these methods, we care about the **population** *parameter* and **sample** *statistic*

Parameters

Ppopulation **P**parameter

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

Parameters

Ppopulation **P**parameter

Examples

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

But 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

Sample **S**tatistics

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 (x, a, b)

How do we know how "right" our predictions are?

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?

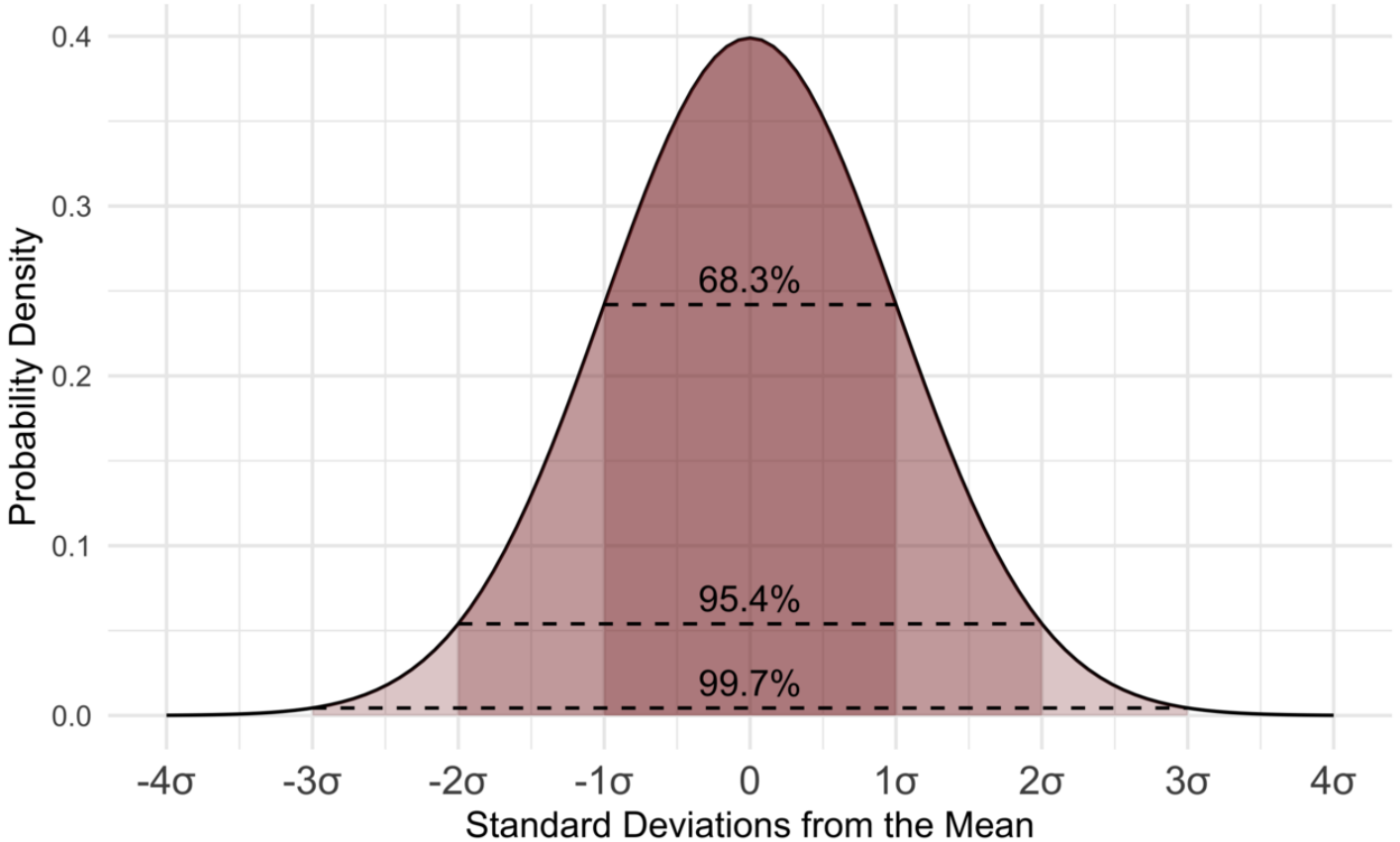
Confidence Intervals

Signals how *sure* we are about our predictions

It shows 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

Capturing the population parameter is our definition of success.


0.4
0.3
0.2
0.1
0.0
Probability Density
68.3%
95.4%
99.7%
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ
Standard Deviations from the Mean

A way to remember...

When calculating confidence intervals, you are casting a net for the true population mean...

The wider the net you cast, the more likely you are to get it.

Anotehr way...

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 the Confidence Interval

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

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

The Confience 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.

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

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

The Mean (or other sample statistics)

We can compute this as we usually compute averages from sample statistics.

This is often denoted as $\overline{x}$ or p, depending on the statistic of interest

The Standard Error

This is based on the sample size and is computed as follows

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

Pulling it Together

To calculate the confidence interval, we add and subtract the product of the z-score and the Standard Error from the mean (or any other statistic)

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

An example

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

Computing the Standard Error

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

Computing the 95% Confidence Interval

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

$$.75 \pm 0.134192$$

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

What happens if I cast a smaller net

... And did a 90% Confidence Interval instead?

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

$$.75 \pm 0.1129678$$

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

... The Interval closes!

And What if I cast a Wider Net?

... And did a 99% Confidence Interval instead?

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

$$.75 \pm 0.1759559$$

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

... The Interval widens!

What if I was extra ambitious...

... and sampled 400 students outside of Norris? And did a 95% confidence Interval

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 So...

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

$$.75 \pm 0.04243525$$

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

... The Interval closes around the mean!

Confidence Intervals for Compairsons

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 Confidence Interval

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

Questions?