STA-1380 – Elementary Statistics
Week #9

Hello! Welcome to the additional online **Weekly Resources** for the course of **STA-1380**. Following a traditional calendar semester, these will be some of the topics your professors will go over. If you do not see material your section is going over for the week, please look at the other resources listed for this course. In addition to these resources, there might be **Group Tutoring** for this course, please see our website for more details. These sessions will go over these materials in more detail as well as any questions about the material.

Any additional help or services can be found through the **Baylor Tutoring Website.** Visit to schedule a free 30-minute private tutoring session, drop-in times for your course, the Baylor Tutoring YouTube channel, or any additional tutoring resources.

Contacts: Sid Rich M-Th 9am-8pm (Fall and Spring class days)   Office Phone: 254-710-4135

Topic of the Week:
"Confidence Intervals"

Key Points:

- SE, MOE, and Length
- Estimating Sample Size
- CI Notation

As addressed in last week's guides, the Confidence Interval is used to give a small range of values the population proportion is estimated to be. During this week, we will cover how to calculate a Confidence Interval and what the components are for each part of the notation. Be mindful of previous chapters and what concepts you covered. STA - 1380 is a course designed to build off of itself and will use ideas such as the sampling distribution from chapter four throughout the rest of this chapter.

For this week's guides, we will be focusing on estimating the true population proportion and how to calculate the sample size necessary for a specific Margin of Error. For Z* values, these will be the 'Special Z-scores' denoted on the Table below the T-distribution in the back of your book. For more information on Point Estimates or a deeper understand of what a Confidence Interval is, see last week's guides for clarification and conceptual frameworks.
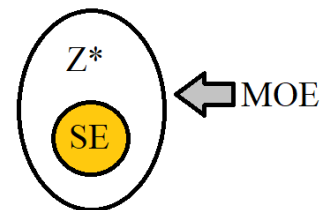
Highlight #1
"Standard Error, Margin of Error and Length"

Formulas: $SE = \dfrac{\sigma}{\sqrt{n}}$ or $\sqrt{\dfrac{\hat{p} \cdot \hat{q}}{n}}$ $MOE = (Z^*_{a/2})(\frac{\sigma_{\hat{p}}}{\sqrt{n}})$

The Standard Error, or 'SE', is the variable applied to the MOE that accounts for the sample size and standard deviations of either the true population, or the sample if the population Std. Dev. is not known. This formula should be very familiar! In fact it's the exact same formula we used to find the standard deviation for a sampling distribution. In the last guide we talked about how Confidence Intervals use Sampling Distribution as their form of spread. The formula for SE using data from the sample proportions is shown above as the second SE equation.

The term Margin of error is denoted as 'E' in your textbooks as 'error' for short. The Margin of error is the numerical value that is either added or subtracted from the PE to create the upper and lower bounds of the Confidence Interval. This is created by multiplying the SE with a particularly chosen Z*. The Length is x2 the MOE. It is the entire size of the Confidence Interval and captures the area both to the left and right of the Point Estimate.

With so many acronyms in STA-1380, it's hard to keep track of which 'Error' to use and were. A helpful memory clue is that Margin of Error is a bigger acronym than Standard Error, therefore, it holds more value. Another way to think of it is like an Egg. The MOE is the entire egg, while the SE is just the yolk. (Taking this analogy one step further, the difference is that the MOE also has the Z* incorporated into it as the white of the egg)

Highlight #1
"Confidence Interval Notation and Structure"

Definition: The formula used to compute a confidence interval and the Z*-scores required to do so.

Notation: CI = MVUE ± MOE or CI = PE ± Z*(SE) or CI = PE ± Z* (σ_{PE})

There are many ways to write the Confidence Interval formula. One way to read it aloud is by saying, "A 'Confidence Interval' equals the 'Minimum Variance Unbiased Estimator' plus or minus the 'Margin of Error'." The way your course will teach it is, "A 'Confidence Interval' equals the 'Point Estimate' plus or minus the 'Z-star Score' times the 'Standard Error'."

The MVUE is a large acronym used in higher statistic courses to declare that it is the point estimate that resembles the true population parameter the best. In many cases MVUE will equal the PE, these terms will be interchangeable for the STA-1380 course. For proportions this is often denoted as $\hat{p}$, because with just one sample, it is the only point estimator available.

Recall that because Z-scores read from the lower tail to a certain point, and that Confidence intervals measure the area between two points in the middle of the curve, disregarding the tails. Because of this discontinuity, Statisticians opt to using an adjusted Z-score often referred to as 'Z-Star' or 'Z-Alpha over 2' If you recall from the previous guides, Z-alpha scores typically read from the right tail going leftward until a certain probability under the curve is met. The '$Z_{a/2}$' works the same way as its name's sake, only with ½ of the probability denoted by *a*.

Every Confidence Interval starts with its Confidence Level. This is the percentage to which the Interval can be trusted to give an accurate estimate on the true population parameter. Alpha, or *a* is the anti-probability, or the area under the curve we do not want our CI to account for. For a 95% CI, we would want 5% of the Area under the curve and thus, the probability, to be ignored. Because the CI is centered over the mean on a Normal Distribution and '0' over the Standardized Curve, we wish to ignore equal parts of the curve on either side of the middle, both tails to be precise.

Because Alpha is 5% in this example, we wish to ignore 2.5% on either tail. This is where $Z_{a/2}$ comes into play. $Z_{a/2}$ just so happens to be the exact Z-score needed to find the upper bound of a 95% confidence interval because the area to the right of $Z_{5/2}$ is equal to the exact amount we wish to avoid. Because the CI's formula already accounts for an addition and subtraction, we do not need to convert $Z_{a/2}$ to a Z-score to find the lower bounds. For Example, if you wished to find the upper and lower Z-scores for a 90% CI, you would look for the $Z_{a/2}$ where *a* is .10, and the $Z_{.05}$. On a CDF table, this would be a Z-score with 95% of the probability before it and 5% after it or a Z-score of 1.645.

Therefore, to compute a Confidence Interval, you calculate the Z* needed using the alpha from the problem, calculate the Standard Error using the standard deviation of the Sampling Distribution, multiply the SE by the Z*, then finally take the Point Estimate from your sample and add and subtract the Margin of Error to compute the upper and lower bounds.

Highlight #3
"Estimating Sample Size"

Definition: A process used to isolate individual aspects of the MOE formula to target certain sample sizes, confidence levels, and CI length.

Formula:
$$n = \frac{Z^{*2}(\hat{p} \bullet \hat{q})}{MOE^2}$$

When it comes down to the Margin of Error, the sample size of *n* is the main deciding factor for many reasons. Whether it be the cost of the study that limits how many people can be sampled, or that the niche variable that has a small *N* by nature, *n* is often the part of the study that research wishes to identify prior conducting the study.

Due to the nature of the formula, it is possible to find the smallest sample size necessary to have a certain margin of error regardless of the sample results. To solve so for any specific variable within MOE, manipulate the equation to solve for that aspect. The most common manipulation is to solve for *n*.

This formula is capable of finding the smallest *n* can be to get a target MOE length. For example, if a researcher wishes to find what sample size must be taken to get a target MOE of ± 3% and 95% confidence, then the researcher can calculate what *n* must be. If the variance of the population is known, (P-hat * Q-hat), then the researcher can use it. If that value is not known, then the researcher should plug in .5 for $\hat{p}$. This is because assuming the proportion is 50% results in the largest potential value the variance could be, which means that there is no room for potentially sampling too few individuals.

For the Scientist's question, the formula would look like this:

$$n = \frac{(1.96)^2 \text{ x } (.5*.5)}{(.03)^2}$$

Solving for *n* would equal 1068, since you cannot have a percentage of a person, you always round up. Note that Z* is squared and so is the MOE. The variance is not $\hat{p}$ squared, but $(\hat{p} * (1- \hat{p}))$, only for the 'worst-case' scenario will 'P-hat' equal 'Q-hat'.

Check Your Learning

1. Martha wants to see what proportion of students at Baylor are their parents' only children. She conducts a sample spanning all grades and majors to find that of the 1000 students sampled, 337 students were raised as the 'only-child' and 663 were found to have siblings.

   a. Construct a 95% Confidence Interval for p.

   b. Interpret what this interval means in terms of the context of the problem.

2. A Professor is grading the exams of "Stats 101," an online test done to assess the statistical knowledge of their students before the course begins. The Professor has hundreds of exams to assess to determine what proportion of their class has a minimum score of 75% on the test, this is the minimum score considered to be passing. If over 80% of the class passes, the professor can begin the course on Chapter 2, saving over 5 days of class time. If the population proportion is under 80%, they will have to start at chapter 1. Rather than hand-grade every test, they collect a random sample of 50 tests and compute the sample proportion of passing grades to be 88%.

   a. Using a 90% Confidence Interval, will the professor have to start on Chapter 1 or Chapter 2?

   b. Using a 95% Confidence Interval, will the professor have to start on Chapter 1 or Chapter 2? What does this say about an increase in CI level compared to the result?

Things Students Struggle With

1. Z* vs $Z_{a/2}$ vs Z-scores

   a. Z* = $Z_{a/2}$, and Z-score = Z* + a/2. (That is to say that the Z-score will always have a/2 more area under the curve than its Z* counterpart).

   b. On a CDF Table, A Z-score of 1.645 will have 95% of the area to the left of it, a Z-score of 1.960 will have 97.5% of the area to the right of it, and a Z-score of 2.576 will have 99.5% of the area to the left of it. Therefore it is clear to see the relationship between Z*, $Z_{a/2}$, and Z.

   c. Some important Z* scores to know: 90% = 1.645, 95% = 1.960, and 99% = 2.576

2. 'P-hat' versus 'P'

   a. When using Confidence Intervals, it is important to remember that we are basing these off of the samples. As such, the standard error will use the 'P-Hat' or sample proportion for its estimates. In later chapters we will use the population proportion when confirming whether or not the point estimate provides useful data. But for now, use 'P-hat' in your CI equations.

Concluding Comments

That's it for this week! Please reach out if you have any questions and don't forget to visit the Tutoring Center website for further information at https://www.baylor.edu/tutoring

---

Answers to CYL

1. **a.** CI = (.3077, .3663)
   b. "We are 95% confident that the true proportion of 'only-child' students at Baylor is between .3077 and .3663 of the population."

2. **a.** CI = (.8044, .9556); the professor can start on Chapter 2 since .80 is not within the Interval and the entire interval is above the decision line.
   b. CI = (.7899, .9701); the professor must start on Chapter 1 since .80 is within the Interval and therefore, it is plausible that the true population proportion of their students is below the decision line.
   This demonstrates that as Confidence Levels Increase, so does the Width of the Interval.