STA-1380 – Elementary Statistics
Week #7

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:

- Point Estimators
- CI Notation and Structure

Gathering samples and their data collected can only be interpreted in so many ways. We are taught to assume that the $\mu_{\hat{p}} = p$ and $\mu_{\bar{x}} = \mu$, but what if our samples contained multiple outliers that skewed the results and makes our data remains unusable? To protect against these possibilities, statisticians use a device known as confidence interval to account for the possibilities of error while still using the data procured.

Confidence Intervals seek to provide a range of possible population parameter values. In other words, rather than using a single value, we will use a set area to which inferences about the population can be made. The term confidence comes from the percentages correlating to the correct value being chosen. A 90% confidence interval is formed using a function that produces a correct interval 90% of the time. Not that a single interval is right 90% of the time, but on average will be right 90% of the time.

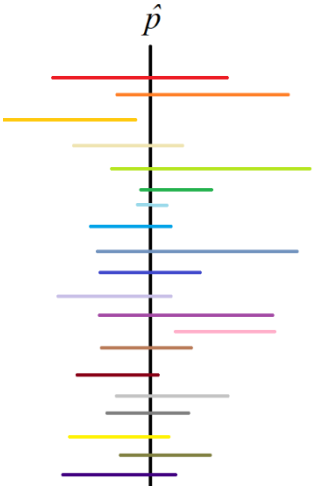
The Confidence Interval DOES NOT mean that there is a 90% chance of the interval being correct or that there is a 90% probability that the interval is correct. One rule of thumb is

that you must always 'tiptoe' around phrases that imply certainty, you cannot imply certainty in a Statistics course.

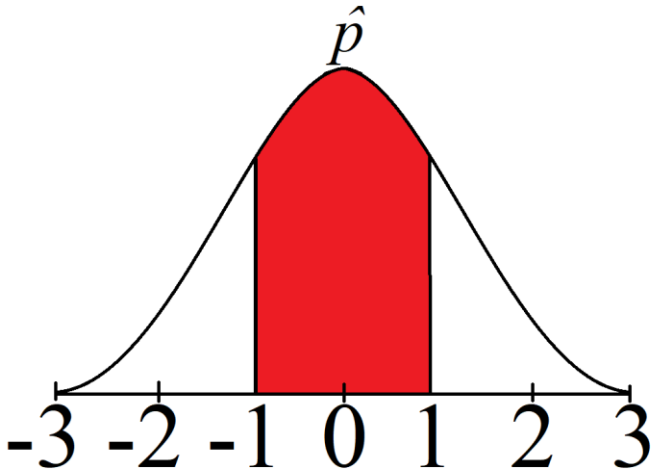
One way to correctly interpret a CI is by declaring your 'confidence' in the tools used, the other is to reference repeated trials. One way is to say that "We are 90% confident that true population parameter of *X* is found within the Interval of (Low, High)." Another is to say that "In repeated trials we would expect that 90% of the intervals contain the true population parameter of *X* and for it to be between Low and High." By doing this, you are not outright saying your choice is correct, but that you have reason to believe it to be, and you suggest others believe it as well.

When working with Confidence Intervals, it is important to remember that you are using data from samples, therefore, you will be using the $\hat{p}$ as both the point estimator and in making the Standard Error. Recall that the Standard Error is Std. Dev. over the square root of the sample size.

Confidence Intervals can be visualized with graphs, lines, and formulas:



*A line Chart demonstrating 90% Confidence Intervals



* A Confidence Interval on a Standardized Normal Distribution

The line charts are often used as a demonstration to show how in repeated samples, the confidence intervals capture the population parameter roughly as much as the confidence value. These are meant to show how varying Confidence intervals still manage to capture the population parameter regardless of width or size and how certain intervals may miss the mean despite being made the exact same way as the others.

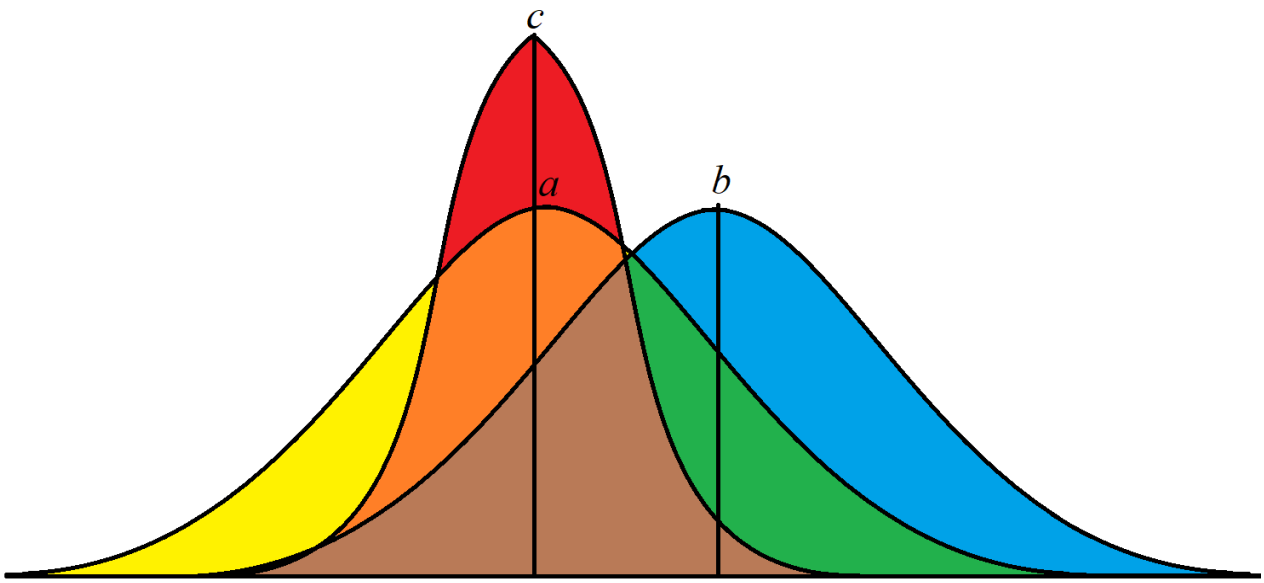
The graphs of a Confidence Interval will resemble that of an 'In-between' Normal Distribution, where the data you are looking for is in between two points, not including the tails of the graph. The area encompassing these graphs will be determined on the size of the interval. The larger the interval, the larger the range of data points. In other words, the more confident you wish to be in your interval, the larger the width of each side must be.

The formulas of a Confidence Interval are the nitty-gritty of this chapter and allow statisticians to form their inferences with real data points. These will be discussed in detail in the Notation and Structure Highlight.

Highlight #1
"Point Estimates"

Definition: A single value gained from a Sample or Sampling Distribution used to estimate the true value of its corresponding population parameter.

Example: $\mu_{\hat{p}}$ and $\mu_{\bar{x}}$ are both Unbiased Point Estimates for population proportions and means.
$\hat{p}$ and $\bar{x}$ are both Point Estimates for the population proportion and means.



*A figure depicting the differences between Bias and Variance between Estimators

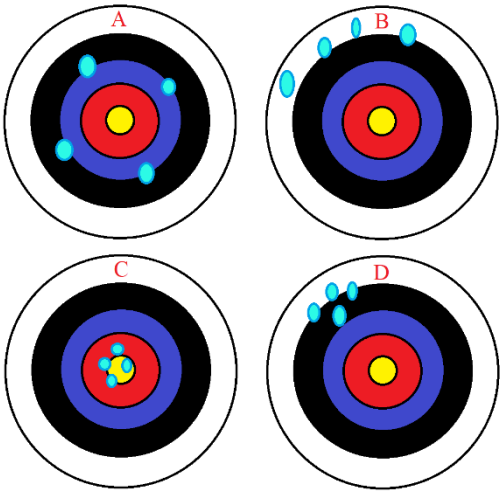
In last week's guide, we went over how the Sampling Distribution's center points, either for a mean or a proportion, was assumed to be equal to the population proportion. This is because the nature of a Sampling Distribution accounts for many trials that slowly but surely normalize the data. Those values are called Unbiased Point Estimators, which is a fancy term for saying that the values do a pretty good job at estimating what the true population values are. But what about regular $\hat{p}$ and $\bar{x}$? How can those two values be used to estimate the parameters of a population?

There are two things to look out for when choosing a point Estimator. Suppose that a scientist wishes to take a sample of Brazilian Tree Frogs to find what proportion of them are female. The Scientist has three potential values they can use as the point estimate for the population parameter decided (proportion of females).

When comparing value *a* with value *b*, the scientist realizes that they share the same spread of the data, but one accurately portrays the true population proportion, and the other is significantly off. The difference between *a* and *b* is that they share the same shape parameter, or that they have an equal variance, but one value, *b*, is a bias estimator. This means that it will not be able to accurately depict the true population because the center of the data is skewed. Between choices *a* and *b*, the Scientist should choose *a*.

When comparing value *a* with value *c*, the scientist realizes that they share the same location, but one has a significantly higher peak with smaller tails compared to the other. The difference between *a* and *c* is that they share the same location parameter, meaning that they are both unbiased estimators of the true population parameter, but one value, *a*, has significantly more variance. This means that it will not be able to accurately depict the true population proportion because more of the data will be spread out farther apart. Between choices *a* and *c*, the scientist should choose *c*.

Another way to think about this process is through a series of Arrow shots:



Target A represents an unbiased estimate with a high variance (Centered, Scattered)
Target B represents a biased estimate with a high variance (Not Centered, Scattered)
Target C represents an unbiased estimate with a low variance (Centered, Clumped)
Target D represents a biased estimate with a low variance (Not Centered, Clumped)

When given the choice between multiple potential point estimates, order them in terms of bias, then variance. Meaning that of the options, choose C first, A second, D third, and B fourth. You will use these Point Estimates as the center points of your confidence intervals.

Highlight #2
"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})

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

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.

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 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. The formula for SE using data from the sample proportions is shown above as the second SE equation. As for a Z* or a $Z_{a/2}$, this is an altered for a Z-score used for a CI.

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.

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

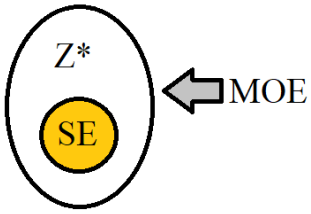
a. Z* = $Z_{a/2}$, and Z-score = Z* + a/2. (That is to say that the Z-score will always have a/2 more area under the curve than its Z* counterpart).

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. SE vs MOE:

a. 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)



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.