STA-1380 – Elementary Statistics
Week #8

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 Continued"

Key Points:

- Manipulating Margin of Error
- Quantile Plots
- Student T Distributions
- CI Notation and Structure for Means

In previous weeks we have used the Normal Distribution as the center point of all our data from μ and σ being the location and shape parameters to Confidence Intervals isolating Z-scores to figure out where a population's true proportion could be to varying degrees of confidence. However, for the Confidence Intervals of means, this will not often be the case. Confidence Intervals for Means will use a T- Distribution when the Std. Dev of the population is not known.

In reality, the true population parameters of μ and σ will not be known. Despite this fact, statistical inference can still be done using another distribution known as the Student-T Distribution. This Distribution functions the same as a Normal Distribution in the sense that it has a bell curve, with the exception that its peak is higher and its tails are wider. This Distribution is built to account for statistical errors and anomalies while sampling to be able to mimic the true population proportion.

Highlight #1
"Manipulating Margin of Error"

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

Formula: MOE = (Z*) x (SE) SE = $\sqrt{\frac{\hat{p} \bullet \hat{q}}{n}}$

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*.

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

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:

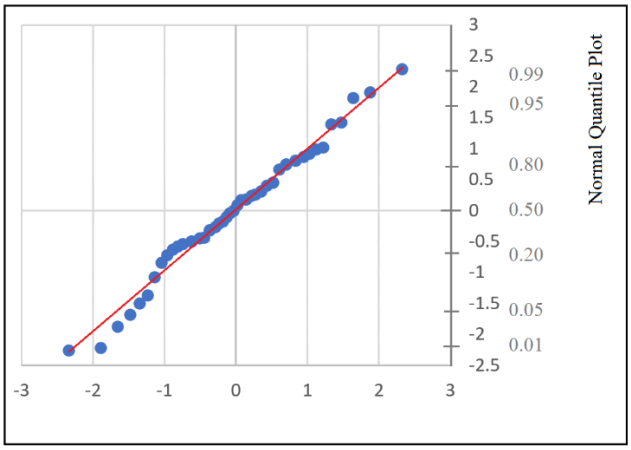
$$n = \frac{(1.96)^2 \text{ x } (.5*.5)}{(.03)^2}$$

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'.

Highlight #2
"Quantile 'QQ' Plots"

Definition: A scatter plot designed to test the normality of a data set by the straightness of the data and the minimal spread between points.

Example: Graphs



*A figure depicted Frog Weights with an $\bar{x}$ of 7.31 and an s of .667

Without knowing the distribution of the population, statistics have no way of knowing if the sample conducted is large enough to conform to the normality standards of the central limit theorem. One such way of testing if the data collected is normally distributed is by using a Normal Quantile Plot.

The QQ-plot collects the raw data and through a series of calculations, turns the data points into Z-scores by using the sample means and standard deviations. The graph then places these against what a perfectly normal distribution's Z-scores should be. If the population was perfectly normal, the data would appear to mimic a y=x curve.

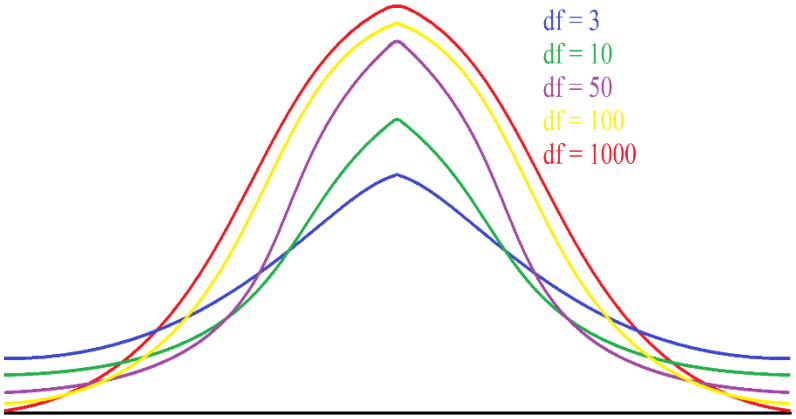
One rule of thumb if you ever question a sample's normality is to place a pencil over the data. If this pencil covers most of the graph, it means the data has a linear trend and is normally distributed. If the data points begin to curve and wildly stray from the y=x graph, your data is not normally distributed a CI cannot be taken. You will not need to know how to calculate these graphs. You will only need to interpret them, understand their purpose, and infer if they prove a sample can be used for a Student T-Distribution.

For Graphs that depict a non-normal distribution, the way the data curves can tell you how the population is Implied to curve. A Concave up curve (U shape) means that the data is skewed right (The majority of the data is on the left side, but data points on the far right pull the mean). A Concave down curve (Hill shape) means that the data is skewed left. (The majority of the data is to the right of the mean.

Highlight #3
"Student T- Distribution"

Definition: A single bell shaped symmetrically curved continuous distribution with a location parameter of $\bar{x}$ and a shape parameter of s designed with a specific degree of freedom in mind used when the population Std. Dev. is unknown.

Example: PDF Graph, T-Score Tables



*A Graph depicted 5 T-distributions

Table of Student T-Distribution (One-tailed)

df	0.1	0.05	0.025	0.01	0.005
1	3.077683537	6.313751515	12.70620474	31.82051595	63.65674116
2	1.885618083	2.91998558	4.30265273	6.964556734	9.924843201
3	1.637744354	2.353363435	3.182446305	4.540702859	5.84090931
4	1.533206274	2.131846786	2.776445105	3.746947388	4.604094871
5	1.475884049	2.015048373	2.570581836	3.364929999	4.032142984
6	1.439755747	1.943180281	2.446911851	3.142668403	3.707428021
7	1.414923928	1.894578605	2.364624252	2.997951567	3.499483297
8	1.39681531	1.859548038	2.306004135	2.896459448	3.355387331
9	1.383028738	1.833112933	2.262157163	2.821437925	3.249835542
10	1.372183641	1.812461123	2.228138852	2.763769458	3.169272673
11	1.363430318	1.795884819	2.20098516	2.718079184	3.105806516
12	1.356217334	1.782287556	2.17881283	2.680997993	3.054539589
13	1.350171289	1.770933396	2.160368656	2.650308838	3.012275839
14	1.345030374	1.761310136	2.144786688	2.624494068	2.976842734
15	1.340605608	1.753050356	2.131449546	2.602480295	2.946712883
16	1.336757167	1.745883676	2.119905299	2.583487185	2.920781622
17	1.33337939	1.739606726	2.109815578	2.566933984	2.89823052
18	1.330390944	1.734063607	2.10092204	2.55237963	2.878440473
19	1.327728209	1.729132812	2.093024054	2.539483191	2.860934606
20	1.325340707	1.724718243	2.085963447	2.527977003	2.84533971
30	1.310415025	1.697260887	2.042272456	2.457261542	2.749995654
40	1.303077053	1.683851013	2.02107539	2.423256779	2.704459267
50	1.298713694	1.675905025	2.008559112	2.403271917	2.677793271
60	1.295821094	1.670648865	2.000297822	2.390119473	2.660283029
70	1.293762898	1.666914479	1.994437112	2.380807482	2.647904624
80	1.292223583	1.664124579	1.990063421	2.373868273	2.638690596
90	1.291028899	1.661961084	1.986674541	2.368497476	2.631565166
100	1.290074761	1.660234326	1.983971519	2.364217366	2.625890521
1000	1.282398721	1.646378817	1.962339081	2.330082675	2.580754698

* A Table depicting the T-scores with a set probability and degree of freedom

The Student T-distribution is a tool used widely throughout statistics for when a population parameter is not known. This distribution is often used in real world studies due to the nature of how difficult censuses are hard to take, and the nature of its design. There are 3 parts that make up the T-Distribution: $\bar{x}$, s, and df. Degrees of Freedom, df, is a statistical term for how much leeway is given to each graph with df = $(n - 1)$.

As seen in the PDF graphs above, the lower the degree of freedom, the taller the tail and the lower the peak. As the degrees of freedom increase, the distribution slowly shapes until it resembles that of a Normal Distribution with thin tails and a taller peak. This is also shown in the T-table. The value of a $T_{.05}$ is 1.646, which is almost exactly the 1.645 that a Normal Distribution has.

A T-Distribution functions the same as a Normal Distribution, but because it is primarily used for Confidence Intervals and Hypothesis Testing (a concept we will cover in the next chapter), the T-tables only depict the right tail or 'one-tail' values. Treat a T-table as if you would a Z-alpha table, with the addition task of matching a df to the value.

Highlight #4
"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 ± T*(SE)

Formulas: σ Known: $CI = \bar{x} \pm Z_{a/2} \frac{\sigma}{\sqrt{n}}$ σ Unknown: $CI = \bar{x} \pm t_{a/2,df} \frac{s}{\sqrt{n}}$

The CI for a population mean is structured the same way that a CI for a population proportion is in that there is the Point Estimate (PE) and the Margin of Error (MOE). The margin of Error is calculated by multiplying the Reliability Factor (either the Z* or $t_{a/2}$,df) to the Standard Error. The Reliability Factor is a term used to describe the specific value associated with a sample size and Confidence Level (CL). For 95% confidence, the Z* is 1.96 if the std. dev. is known and varies depending on sample size if std. dev is unknown.

DF = $(n - 1)$, with numbers from 1-30 usually shown on the chart, then every 10 until 100, then it jumps up again to 1000. This is because as the sample size increases, the difference between each new T-Distribution's Reliability Factor becomes less and less until it is almost negligible. If a sample size is in between numbers on the T-Table, round down a lower df, this is because the higher number will make the sample appear more normally distributed.

In certain questions the population std. dev. might be known. In this instance, use a Sampling Distribution to calculate the MOE using a Z* score from the Normal-Table or JMP output and the std. dev. precalculated for the sample size. However, STA-1380 will often focus on T-Distributions during these chapters for you to become familiar with them. In this instance, use the second formula addressed above for when the std. dev. is unknown or not stated.

JMP, TI-80 series graphing calculators, and other software capable of Statistical evaluation will often calculate the sample statistics ($\bar{x}$, s) so that you do not have to do it by hand. However, certain outputs become difficult to read or can be confusing at times, practice recognizing what each data point means on an output so that when the exams come around, you will not be confused on which results to use in your explanations.

Like a CI for proportions, a CI for means can tell you about the true population's tendencies. For example, if at a certain confidence level, a specific threshold is not within the interval, that means that it is not plausible for that threshold to be accurate to the population. For example: If the previous μ was stated to be 45 and the interval you collect at 95% confidence is (47.34, 49.78), then you have evidence to support that the previous mean is no longer accurate.

Check Your Learning

1. An author wishes to see if there is an optimal number of pages a book can be before a reader loses interest while still having a captivating story. To do this, they look to find the true mean number of pages from the 'New York Times Best Sellers' list. They collect a random sample from a NYT catalogue and gain the following results:

766 706 607 609 224 251 317 636 482 532 685 518 464 685 329 374
436 512 537 564 483 352 552 493 488 327 401 539 387 498 391 425

a. Using any software of your choice, compute a QQ-plot to determine the normality of the data.

b. If it appears normal, create a 95% Confidence Interval and explain the results in the context of the problem.

2. In a 20x0 census, it was found that during the summer, the State of Texas had an average temperature of 94 degrees with a standard deviation of .4 degrees. Due to the heat waves you experienced during Welcome Week, you have reason to doubt that the true mean temperature may no longer be the case. Going back to the summer before your freshman year, you collect your data.

a. Compute a 90% CI for the population mean with the given results: 12 random days from the summer are collected, with a mean of 99 degrees with a standard deviation of 1.5.

b. Say before you even take a sample, you only wanted your MOE to be ± .15 degrees, what is the smallest sample size you can take to reach this desired CI length?

Things Students Struggle With

1. How to read a Stats Output:

a. With $\bar{x}$, s, df, Confidence Intervals for the means and for the Std. Dev., the outputs can be confusing if you are not familiar with the notation or properly instructed. In certain cases, an output may be all you are given, and no data points are ever shown. For times like those, it is vital to understand how an output is formatted.

For JMP, Confidence Intervals will generate values for the mean and std dev. Do

not use the CI associated with the Std Dev. Only use the CI that is on the same line as the mean.

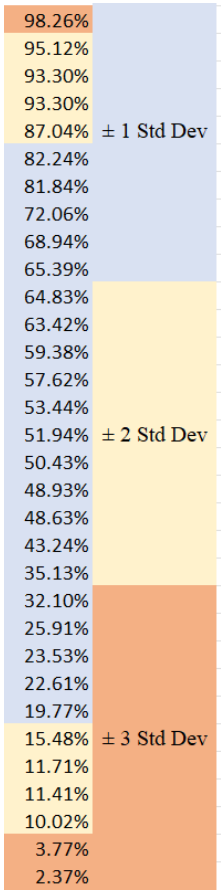
2. What is considered Normally Distributed?:

a. If given a visual histogram (bar graph), look to see if it is roughly symmetrical and bell-shaped, this will be enough for the STA-1380 course.

If given the percentiles (QQ-plot results), look to see if the percentiles align with that of the empirical rule. The empirical rule, also known as the "68-95-99.7 rule," is the portions of data that should fall between the first 3 standard deviations.

While looking at the percentiles or the graph, check to see if over half the data points are within the first std dev, (16% through 84%), and that almost all the data points are within the first 2 std dev, (2.5% through 97.5%), In smaller samples, you should see very little in either end of the tails, this will prove the 3rd part of the rule.

Here is a visual representation of the Empirical Rule on percentiles

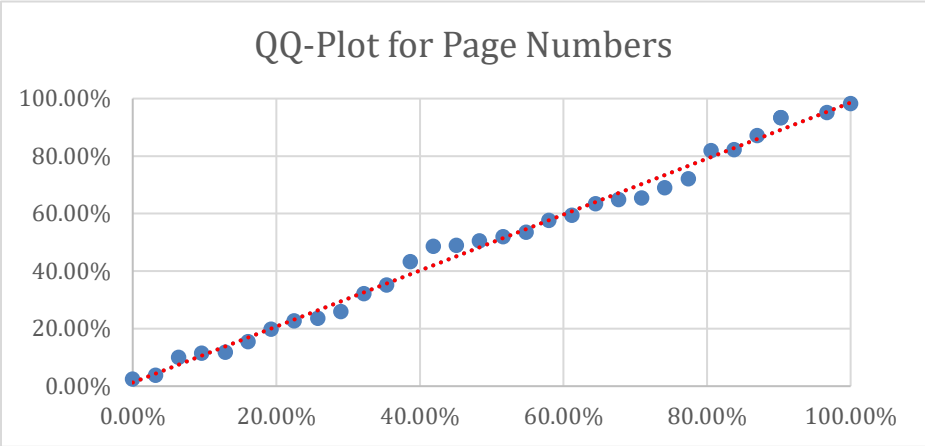


Concluding Comments

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.



b. Because the QQ-plot showed the sample was normally distributed, the CI computed is such: CI = $\bar{x} \pm (t_{a/2} * (s / (n^{\wedge} {}^{1}/_{2})))$ CI = 486.56 ± (2.042)*(132.44/(5.656854)) = We are 95% confident that the true mean of the page numbers for New York Times bestselling books is between 438.81 and 534.31 pages.

2. a. CI = $\bar{x} \pm (Z_{.1} * (\sigma / (n^{\wedge} {}^{1}/_{2})))$ CI = 99 ± (1. * (1.645 / (3.464))) = (98.810, 99.190).

b. Use the known population std dev. The necessary sample size is 28 summer days for a MOE of .15 degrees.