Key to Part 1 of Test 4

1. ***Write at least a paragraph, in your own words, to explain what a sampling distribution is and how it relates to making statistical inferences. Be sure to use the term "sampling variability" in your answer. You may use an example to illustrate your answer or you can just answer in general terms. Your response should be college level writing.***

(Answers will vary but should mention or describe inferential statistics and include the words "sampling variability".)

Inferential statistical analysis involves taking some information (a "statistic") from a sample and trying to see what that statistic says about the same value in a population (a "parameter"). The problem is sampling variability; i.e., that the values in samples (statistics) will likely be different from the population paramter and from one another, so we can't rely on the results of a sample to tell us exactly what is true of the population.

A sampling distribution is a probability distribution that accounts for all of that sampling variability by (theoretically) taking all possible samples of a fixed size from a population and seeing how the corresponding sample statistics are distributed. As long as certain conditions are met, the sampling distributions for certain statistics (proportions, means, etc.) will have predictable patterns, curves that can be graphed that allow us to assess probabilities.

2. ***If you had census data from a population would you use a hypothesis test to learn something about that population? Explain.***

(Answers will vary.) We would NOT use a hypothesis test because hypothesis tests are used to see what a sample tells us about a population. If we had census data, then we would know everything about the population already, so conducting a hypothesis test would be unnecessary (and wrong!).

3. ***Explain why we use the t-distribution instead of the z-distribution (the Normal distribution) for making inferences about population means.***

In order to make an inference about a population mean, we have to know about the sampling distribution of the mean. As long as certain conditions are met, the Central Limit Theorem tells us the sampling distribution of the sample means is normally distributed (z-distribution) with standard error based on sigma, the population standard deviation. But sigma is rarely known so the best we can do is use s, the sample standard deviation, which introduces more variability since s will change from sample to sample. The t-distribution accounts for that variability by adjusting the curve according to sample size (degrees of freedom).

4. ***Suppose you poll 30 students in the Math Lab and find that their average GPA is 3.26, with a standard deviation of 0.81.***

   a. ***By hand, construct a 95% confidence interval for the mean GPA of all students who use the Math Lab. Interpret the confidence interval in the context of the problem.***

Parking Lot: $n = 30$ $\bar{x} = 3.26$ $s = 0.81$ $SE = \frac{0.81}{\sqrt{30}} = 0.148$ df = 30 -1 = 29 t* = 2.045 for 95% confidence level	Note: Conditions will be checked in part (d) as part of the hypothesis test. CI: point estimate ± margin of error $= \bar{x} \pm t^{*}SE$ $= 3.26 \pm 2.045(0.148)$ $= 3.26 \pm 0.303$ $= (3.26 - 0.303,\ 3.26 + 0.303)$ $= (2.96, 3.56)$ We are 95% confident that the true (population) **mean** GPA for students who use the Math Lab is between 2.96 and 3.56. (In reality, since the conditions of random sample and independence weren't met, as well as there being the very real possibility of positive bias due to self-reporting, this result is very suspect!)

b. ***Based on the confidence interval, could you conclude that there are no individual students with a 4.0 GPA who use the Math Lab? Explain, based on what the confidence does or does not tell us about individuals in a population.***

No, we can't conclude anything about any individual from the values in a confidence interval. Confidence intervals tell us only about averages; they provide no information about individuals!

c. ***Suppose the average GPA of all Cuesta College students is 2.73. Could we conclude, based on just the confidence interval, that students who use the Math Lab have, on average, significantly higher GPA's than general Cuesta students? Explain how you can tell.***

Examining the confidence interval, we see that all the values in the confidence interval (which are all possible mean values for GPA for all students who use the Math Lab) are above 2.73, the average GPA of all Cuesta students. Because of this we can conclude that the average GPA of Math Lab users is significantly higher than that of general Cuesta students. If 2.73 had been captured by the CI, then we would have concluded there was not a significant difference.

d. ***Test (using a 1-Sample t-Test) whether students in the Math Lab have, on average, significantly higher GPA's than general Cuesta students. Include all 4 steps of the hypothesis test. Write the hypothesis with symbols and with words. You may use Minitab for the Compute step, but draw a well-labeled curve that illustrates the sampling distribution, the sample mean, the t-value, and the P-value.***

Step 1: Hypothesize $\mu =$ mean GPA of students who use the Math Lab

$H_0 : \mu = 2.73$ The mean GPA of students who use the Math Lab is the same as the mean for all Cuesta students.

$H_a : \mu > 2.73$ The mean GPA of students who use the Math Lab is greater than the mean for all Cuesta students.

Step 2: Prepare. We'll use a .05 level of significance and the 1-Sample t-Test for means.

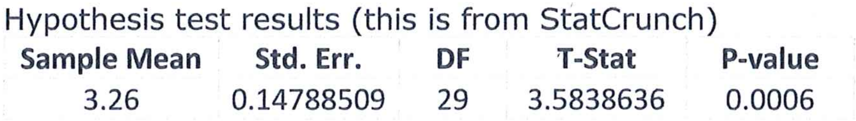
Check conditions/make assumptions.

(1) **Sample is random and independent?** Neither condition is likely to be met. Sample was probably not random, depending on how the poll was conducted. (Did every student who uses the Math Lab have an equally likely chance of being selected? Unlikely.). The sample observations are probably not independent and likely are positively biased. (If students were asked (polled) in a group, they may have inflated their GPA depending on the other people sitting around them.)

(2) **Sample size/Population considerations**. The sample is large ($n = 30 \geq 25$), so no need to check that the underlying population is normal.

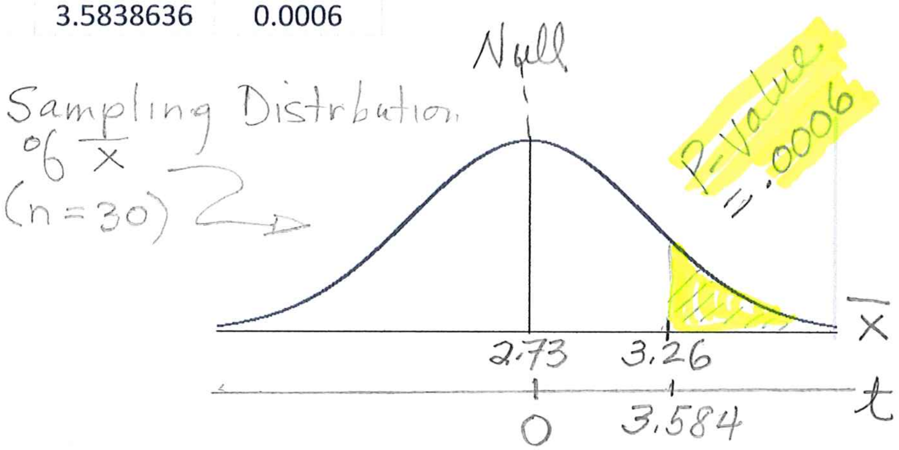
(3) **Large population?** Is the population of all users of Math Lab $\geq$ 10(30) = 300 students? We'll have to asume this (and it isn't unreasonable).

Step 3: Compute: One sample t hypothesis test:



Hypothesis test results (this is from StatCrunch)

Sample Mean	Std. Err.	DF	T-Stat	P-value
3.26	0.14788509	29	3.5838636	0.0006



Step 4: Interpret: Since the P-value is much smaller than the level of significance (.0006 < .05), we will reject the null and accpt the alternative hypothesis.

We have strong evidence that the mean GPA of students who use the Math Lab is significantly higher than the mean GPA of the general Cuesta student population.

e. Can we conclude that using the Math Lab <u>causes</u> students to have higher GPA's? Explain why or why not.

No, we can NOT conclude causation because this was just an observational study, not a controlled experiment with random assignment into groups. There are many potential confounders, including the possibility that students who use the Math Lab may be generally more willing to do what's needed to be successful in their classes (hence have higher GPA's because of that), or who may have more time to be on campus to study, etc. To conclude cause-and-effect, we would have to randomly assign a group to use the Math Lab and another to not use the Math Lab, then compare the mean GPA's of the two groups after a period of time (say, a semester).