Winter 2019 – ST 314 Data Analysis 07

Topics:

- One and Two Proportion Z procedures.
- Describing Visual Displays
- Lessons Covered: 34 - 38
- Textbook Chapter (Optional) : 7,8,9

Grading:

- Points are listed next to each question and should total 25 points overall.
- Grading will be based on the content of the data analysis as well as the overall appearance of the document.
- Late assignments will not be graded.

Deadlines:

- Final Submission: **Monday, February 25th**. All submissions must be PDF files.

Instructions:

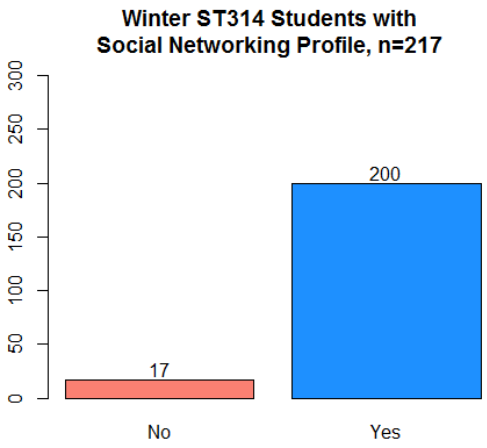
- Clearly label and **type answers** to the questions on the proceeding pages, ***without*** question prompts, in Word, Google Docs, or other word processing software.
- Insert **diagrams or plots as a picture** in an appropriate location.
- Math Formulas need to be typed with Math Type, LaTeX, or clearly using key board symbols such as +, -. *, /, sqrt() and ^
- Submit assignment to the Canvas link as a PDF. Verify the correct document has been uploaded. If not, resubmit. You can submit up to three times.

Allowances:

- You may use any resources listed or posted on the Canvas page for the course.
- You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
- Outside resources need to be cited. Websites such as Chegg, CourseHero, Koofers, etc. are discouraged, but if used need to be cited and used within the boundaries of academic honesty.

Part I. (11 points)

Imagine the engineering department at OSU is interested in understanding the social habits of their students. Specifically, they would like to estimate the proportion of students who have a social networking profile. The provided bar chart was created from Student Information Survey for the Online and Campus class of Winter 2019 term, use this information to answer the following questions.



a. (1 point) From the bar chart, what proportion of students have a social networking profile?
b. (2 point) Make an argument as to why it may be reasonable to use the current student data to represent the population of all current OSU engineering students.
c. (2 point) Make an argument as to why it may not be reasonable to use the current student data to represent the population of all current OSU engineering students.
d. (2 point) Check the sample size conditions for a confidence interval for *p*. Are these met?
e. (2 points) Estimate the proportion of all OSU engineering students with a social networking profile. Use a confidence level of 99%. Show work.
f. (2 points) Interpret your estimate for *p*. Include context, and the point and interval estimates.

Part II. (6 points)

Polling and statistical consulting company Statista claims that 81% of those in the United Stated have some type of social media profile. See report here: https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/

a. (2 point) Based on your confidence interval in part I (d), is it reasonable to assume the actual proportion of all OSU engineering students with a social profile is 0.81?
b. (2 point) Suppose you would like to test the hypotheses:

$$H_0: p = 0.81 \; vs \; H_0: p \neq 0.81$$

Do not perform the test. Instead based on the information from your confidence interval, will you reject or fail to reject the null hypothesis at a significance level of 0.01? Explain. *Hint: See end of lesson 25, week 4.*
c. (2 points) What are some possible characteristics of OSU engineering students that differ from all Americans? Is it possibly these differences may make 0.81 as an unrealistic estimate for p?

Part III. (8 points)
Suppose the 95% confidence interval for a difference in two population proportions, $p_1 - p_2$, is calculated to be between -0.277 to -0.003.

a. (2 points) This interval contains all negative values. What does this tell us about the relationship between p_1 and p_2? Be specific.
b. (2 points) The 95% confidence interval has all negative numbers. For the same data, will the 90% confidence interval also contain all negative values? Explain. What about the 99% confidence interval?
c. (2 points) Calculate the 99% confidence interval for $p_1 - p_2$. Show work.
d. (2 points) What might we conclude when a confidence interval for $p_1 - p_2$ contains both negative and positive values?

NOT GRADED
Part IV. Optional: Sampling Distribution and the Central Limit Theorem Simulations for Proportions!
See the Central Limit Theorem in Action for a Proportion! Statisticians like to use software to simulate scenarios to see how statistical theorems hold up. In this activity, we will randomly generate the population distribution of a categorical variable, then randomly sample from the population using different sample sizes and construct sampling distributions from repeated sampling. If all goes as planned, we should be able to validate the central limit theorem for proportions.

Getting Started:
Go to the following link https://courses.ecampus.oregonstate.edu/statistics/interactives/
Go to simulation 4 listed on the left hand side of the Statistics Interactives.

Q1. For this first part, we will look at the population distribution of a categorical variable whose two outcomes are equally likely. From this population we will take see the affect sample size has on the sampled distribution and ultimately the sampling distribution of sampled proportions.
Instructions for Q1
Choose Binomial Even as the "Distribution Type".
Next, choose a sample size between 2 and 19 under the header "Choose Sample Size".
Choose 1000 for number of simulations under the header "# of Simulations. You may need to wait a few moments for the simulation to run.

What is the shape of the *population* distribution?

Fill out the table according to your simulated values and sample size:

Population Proportion p	**Sampled Proportion $\hat{p}$**	**Sample Size n**

a. According to the central limit theorem, if the n is large enough the distribution of sample proportions will be normal with a mean of p and a standard deviation of $\frac{p(1-p)}{\sqrt{n}}$. Describe the sampling distribution. Does it seem normal? Is there anything strange about the distribution?

b. Recall, one of the conditions for a one proportions z test is $n \times p_0 \geq 10$. If you assume p and p_0 to be equal, then the sample size will need to be at least what value in order for the condition for the hypothesis test to hold?

n must be at least _____

Q2.

a. Now increase the sample size to be between 500 and 1000. Based on the simulated distribution of sampled proportions does the central limit theorem seem to hold? Describe the shape of the distribution of 1000 sample proportions. Is it normally distributed?

b. Fill out the table according to your simulated values and sample size:

Population Proportion p	Sample Size n

c. What are the mean and standard deviation of the simulated sampled proportions? These values should be close to the theoretical values, where $\mu_{\hat{p}} = p \; and \; \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, are they? State your values and give your answers to the questions below.

	Simulated	**Theoretical**
Mean of Sampled Proportions		
Standard Deviation of Sampled Proportions		

Q3. Change the "Distribution Type" to binomial skewed. See how this affects the sampling distribution. Do you need a larger sample size to reach a symmetric distribution? Based on the conditions for a one proportions z test, what is the smallest size n you will need to meet the conditions for inference?