Math 140 Three-Part Categorical & Quantitative Data Project
Project Instructions / Fall 2018 / Teachout

You will be doing a real world data analysis project during the semester. It is broken up into three parts, collecting the data, population estimates with confidence intervals, and a relationship hypothesis test. Each part is graded separately. See homework schedule for due dates.

Part 1: Collecting Data, create Excel Spread Sheet, Data Collection Essay

Collecting your Data

Collect data yourself. This cannot be data found on the web. Your data needs to have one categorical question with only two possible answers and one quantitative question for each person or object. You should have at least 30 for each of your categorical variables. (For example, if your categorical question was "Do you smoke, yes or no?", then you should have 30 or more people that smoke and 30 or more people that do not smoke.) Note: This does not mean that you have to have 30 exactly. The actual data will probably be 30 smokers and 73 non-smokers. More data is better!!

Essay

You will write an essay on how you collected the data. What was your population of interest? Was the data random? If it was not random, what method did you use? List all possible sources of bias you can think of in your data collecting. How well you think this data will represent the population of interest?

Create Excel Spread Sheet

Put the categorical data into one column of an excel spreadsheet. Put the corresponding quantitative data the next column next to the categorical data. Be careful to not mess up the order of the values. Now separate the quantitative data into groups. The category makes up the two groups. For example smokers and non-smokers. If the quantitative variable was age, I would put the ages of the smokers in one column and the ages of the non-smokers in another column. When you have only two groups (like smokers and nonsmokers), you should have a total of four columns of data in your excel spreadsheet (raw categorical data, raw quantitative data, two separated quantitative data sets).

Note: You will need to turn in your data collecting essay as well as the excel spread sheet.

Part 1 Grading Rubric:

- **Collecting Data (50% of grade)**
- **Bias, Sampling Technique, Population Essay (25% of grade)**
- **Separating Quantitative Data by Group (5% of grade)**
- **Excel Spread Sheet with four columns (original categorical data, original quantitative data, separated quantitative data for each group). (20% of grade)**

Part 2: Use Statcato or StatCrunch to create Confidence Intervals for the categorical and quantitative data that you collected.

Categorical Data Confidence Intervals Directions:

Check the assumptions for the proportion for each of the categorical variables and tell how well it will apply to the population.

Use Statcato or Statcrunch to create a confidence interval for each categorical variable. Convert the proportions into percentages. For each confidence interval, include the Statcato or StatCrunch printout with the sample proportions and confidence interval. For each variable, write a sentence to explain the confidence interval.

Also use Statcato or Statcrunch to create a two-population proportion confidence interval to estimate the difference between the population proportions (percentages). For each two-population proportion confidence interval, write a sentence to explain the confidence interval. Include the Statcato or StatCrunch printout with the sample percent difference and the two population confidence interval. Specifically analyze whether or not there was a significant difference between the percentages and explain why.

Quantitative Data Confidence Intervals:

Check the assumptions for the mean average of each group's quantitative variable and tell how well it will apply to the population. You will need to include a histogram for each group's quantitative variable.

Use Statcato or Statcrunch to create a confidence interval for each group's population mean average of the quantitative variable. For each confidence interval, include the Statcato or StatCrunch printout with the sample mean and confidence interval. For each group's mean, write a sentence to explain the confidence interval.

Also use Statcato or Statcrunch to create a two-population mean average confidence interval to estimate the difference between the population means. You should compare the means from your two groups. Write a sentence to explain the two-population mean confidence interval. Include the Statcato or StatCrunch printout with the sample mean difference and the two population confidence interval. Specifically analyze whether or not there was a significant difference between the mean averages and explain why.

Part 2 Grading Rubric:

- Assumptions check, Statcato or StatCrunch confidence interval proportion printout, sentence explaining population proportion for group 1. **(12% of grade)**

- Assumptions check, Statcato or StatCrunch confidence interval proportion printout, sentence explaining population proportion for group 2. **(12% of grade)**

- Histogram, assumptions check, Statcato or StatCrunch confidence interval mean average printout, sentence explaining population mean for group 1. **(16% of grade)**

- Histogram, assumptions check, Statcato or StatCrunch confidence interval mean average printout, sentence explaining population mean for group 2. **(16% of grade)**

- Assumptions check, Statcato or StatCrunch confidence interval printout for two-population proportion difference between groups 1 and 2, sentence explaining the confidence interval, explain whether or not there was a significant difference between the percentages and why. **(20% of grade)**

- Histograms, assumptions check, Statcato or StatCrunch confidence interval printout for two-population mean average difference between groups 1 and 2, sentence explaining the confidence interval, explain whether or not there was a significant difference between the mean averages and why. **(24% of grade)**

<u>Part 3</u>: Categorical / Quantitative Relationship Hypothesis Test (Two Population Mean T-test)

In part 1 of the project, you collected some categorical and quantitative data and separated your quantitative data by groups. We want to determine if the categorical variables has a significant relationship with the quantitative variable.

Use Statcato or Statcrunch to perform the two-population mean hypothesis test. Your hypothesis test should have the null and alternative hypothesis, the T test statistic, the P-value, whether or not you reject the null hypothesis and the standard conclusion. You should also have sentences explaining the assumptions, test statistic meaning, the P-value meaning, and the conclusion. Include the Statcato or Statcrunch printout. Are the mean averages significantly different? Explain how you know. Could the sample data happened by random chance? Explain how you know. Was there a relationship between the categorical and quantitative variables? Explain why.

Part 3 Grading Rubric: (9% of grade for each of following)

- **Statcato or StatCrunch two-population mean hypothesis test printout**
- **Two histograms checking shape (one for each group)**
- **Check and explaining all the assumptions correctly**
- **Correct Ho and Ha (including relationships)**
- **Test Statistic with Sentence Explaining it**
- **P-value with Sentence Explaining it**
- **Reject Ho or Fail to reject Ho and explain why.**
- **Standard Conclusion sentence addressing evidence, claim, and relationship**
- **Are means significantly different? Sentences explaining.**
- **Could the data have happened by random chance? Sentences explaining.**
- **Is there a relationship between the categorical data and the quantitative data? Sentences explaining.**

Sample null and alternative hypothesis

Ho: $\mu 1 = \mu 2$ (There is no relationship between the categorical and quantitative variables)

Ha: $\mu 1 \neq \mu 2$ (There is a significant relationship between the categorical and quantitative variables)

Assumptions for 1 population proportion (%)

- *Data set is random or representative of the population*
- *At least 10 success*
- *At least 10 failures*
- *Population size at least 10 times larger than the sample size.*

Assumptions for 2 population proportion (%)

- *Both data sets random or representative of the population*
- *Both data sets have at least 10 success*
- *Both data sets have at least 10 failures*
- *Independent Groups*
- *Population sizes at least 10 times larger than the sample sizes.*

Assumptions for 1 population mean average

- *Data set is random or representative of the population*
- *Sample size of 30 or bell shaped (Include histograms)*
- *Population size at least 10 times larger than the sample size.*

Assumptions for 2 population mean average

- *Both data sets random or representative of the population*
- *Both data sets have a sample size of 30 or bell shaped (Include histograms)*
- *Matched Pair or Independent Groups? Remember matched pair is a one-to-one pairing (not just something in common)*
- *Population sizes at least 10 times larger than the sample sizes.*