Math 140 Four-Part Categorical & Quantitative Data Project
Project Instructions / Spring 2018 / Teachout

You will be doing a real world data analysis project during the semester. It is broken up into four parts, collecting the data, analyzing 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

Grading Rubric:

- **Collecting Data 50%**
- **Bias, Sampling Technique, Population Essay 30%**
- **Separating Quantitative Data by Group 10%**
- **Excel Spread Sheet of original categorical and quantitative data as well as separated quantitative data for each group. 10%**

Collect data yourself. This cannot be data found on the web. The data needs to be collected from at least 50 people or objects and have one categorical question and one quantitative question for each person or object. You should have at least 10 successes for each of your categorical variables. So you may need to collect more than 50.

You will write an essay on how you collected the data, the method used, your population of interest, possible sources of bias, and how well this data will represent the population of interest. Include sample size requirements
(Quantitative data: at least 30 numbers, Categorical Data: at least 10 success for each variable).

Put the categorical data into one column of an excel spreadsheet. Put the quantitative data into another column of the same excel spreadsheet. Be careful to not mess up the order of the values. Now take each categorical variable and type the quantitative values for that variable into one column. This will create separate quantitative data sets for each group in your categorical data. Include these columns on the same excel spread sheet as your original data.

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

Part 2: Use Statcato to create graphs and statistics and analyze the categorical data and compare the quantitative data for each group.

Grading Rubric:

- **Analyze Categorical Data with Pie Chart and Bar Chart: 25%**
- **Which categorical variable had the highest count and percentage? 5%**
- **Analyze Quantitative Data for first group with a data analysis paragraph. Include the Statcato statistics printout, Histogram, Boxplot, Dotplot, Shape, Center, Average, Spread, Typical Values, and Unusual Values: 30%**
- **Analyze Quantitative Data for second group with a data analysis paragraph. Include the Statcato statistics printout, Histogram, Boxplot, Dotplot, Shape, Center, Average, Spread, Typical Values, and Unusual Values: 30%**
- *(If needed)* **Analyze Quantitative Data for third, fourth or fifth group with a data analysis paragraph. Include the Statcato statistics printout, Histogram, Boxplot, Dotplot, Shape, Center, Average, Spread, Typical Values, and Unusual Values (Adjusted grading for more groups)**
- **Which group had the highest average? (5%)**
- **Which group had the most typical spread? (5%)**

Categorical data: Create a bar chart and pie chart and find all the sample percentages for each categorical variable. Compare the percentages to explore any key features or surprising results.

For the quantitative data for each group, create a histogram, dotpot and boxplot, give the shape, the best center (average), two numbers that typical values fall in between, outlier cutoff points, and a complete list of all unusually high and unusually low values in the data set. Which group has the highest average? Which group has the most typical spread? Write a data analysis paragraph for each group.

You will need to turn in your data analysis essays with sentences and analysis as well as the graphs and statistics printout from Statcato.

Part 3: Use Statcato to create Confidence Intervals for the categorical and quantitative data that you collected.

Categorical Data Confidence Intervals:

Check the assumptions 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. Include the sample percent, Standard Error and the Margin of Error for each categorical variable. For each variable, write a sentence to explain the Standard Error, a sentence to explain the Margin of Error, and a sentence to explain the confidence interval.

Also use Statcato or Statcrunch to create many two-population proportion confidence interval to estimate the difference between the population percentages. You should compare every two categorical variables. For each two-population confidence interval, write a sentence to explain the Standard Error, a sentence to explain the Margin of Error, and a sentence to explain the confidence interval. Specifically analyze whether or not there was a significant difference and explain why.

Quantitative Data Confidence Intervals:

For each group, check the assumptions for the quantitative variable and tell how well it will apply to the population.

For each group's quantitative data, create a confidence interval to estimate the population average.

If you used the median average, use bootstrapping and lock5stat.com to create a bootstrap distribution for the median and find the confidence interval. Include the sample median, Standard Error, Margin of Error, and the confidence interval. Write a sentence to explain the Standard Error, a sentence to explain the Margin of Error and a sentence to explain the confidence interval. Include a picture of the bootstrap distribution.

If you used the mean average, use bootstrapping and lock5stat.com to create a bootstrap distribution for the mean and find the confidence interval. Include the sample mean, Standard Error, Margin of Error and the confidence interval. Write a sentence to explain the Standard Error, a sentence to explain the Margin of Error and a sentence to explain the confidence interval. Include a picture of the bootstrap distribution.

Also use Statcato or Statcrunch to create many two-population confidence interval to estimate the difference between the population means. You should compare the means for every two groups. For each two-population mean confidence interval, write a sentence to explain the Standard Error, a sentence to explain the Margin of Error, and a sentence to explain the confidence interval. Specifically analyze whether or not there was a significant difference and explain why.

You will need to turn in the Statcato printouts and Lock5stat.com bootstrap printouts with your confidence intervals, standard errors, margin of errors, and sample values and an essay with your sentences explaining assumptions, standard errors, margin of errors and confidence intervals.

Part 4: Categorical / Quantitative Relationship Hypothesis Test (ANOVA or Two Population T)

In part 1 of the project, you used your categorical data to separated your quantitative data by groups. Put these columns into Statcato to perform a hypothesis test. We want to determine if the categorical variables have a significant relationship with the quantitative variable.

Categorical Data with only two responses:

If your original categorical question had only two responses (like yes and no), then you will have two columns of quantitative data (one for the yes group and one for the no group). Use this data and Statcato to perform a two population mean (two-tailed) hypothesis test to see if there is a relationship between the categorical variables and the quantitative variable you collected at the beginning of the class. Check your assumptions for the two population T-test. Was this a matched pair test or independent groups? This includes graphs to check shape. Your hypothesis test should have the null and alternative hypothesis, the T test statistic, the critical value, the P-value, whether or not you reject the null hypothesis and the standard conclusion. You should also have sentences explaining the assumptions, null and alternative hypothesis, test statistic meaning, critical value meaning, and the P-value meaning.

Sample null and alternative hypothesis
Ho: $\mu1 = \mu2$ (There is no relationship between the categorical and quantitative variables)
Ha: $\mu1 \neq \mu2$ (There is a significant relationship between the categorical and quantitative variables)

Categorical Data with three or more responses:

If your original categorical question had three or more responses (like liberal, conservative or moderate), then you will have three or more columns of quantitative data. Use the data you collected and Statcato to perform an ANOVA hypothesis test to see if there is a relationship between the categorical variables and the quantitative variable you collected at the beginning of the class. Check your assumptions for the ANOVA test. This includes graphs to check shape. Your hypothesis test should have the null and alternative hypothesis, the F test statistic, the critical value, the P-value, whether or not you reject the null hypothesis and the standard conclusion. You should also have sentences explaining the assumptions null and alternative hypothesis, test statistic meaning, critical value meaning, and the P-value meaning.

You will need to turn in your null and alternative hypothesis, and your F-test statistic (of T-test statistic), critical value, and P-value printout from Statcato and an essay with your sentences explaining the assumptions, F-test statistic, P-value and Conclusion. You should also include any graphs used to check assumptions.

Sample null and alternative hypothesis (add more µ's as needed)

Ho: $\mu 1 = \mu 2 = \mu 3$ (There is no relationship between the categorical and quantitative variables)
Ha: at least one μ is not equal (There is a significant relationship between the categorical and quantitative variables)