Lab 7 (100 points): Two-Sample Independent and Two-Sample Paired t Procedures

Objectives: Confidence interval and significance tests for two samples.

A. (10 points) Online Prelab

The statistical procedures introduced in this course depend on the assumption of normality. Sometimes, if the data are not well approximated by a normal distribution, we apply a transformation of the data. In this lab, you will need to apply the log transformation (natural log, or "ln", which is called "log" in both R and SAS), to make the data more normally distributed. When interpreting transformed data, remember to change the data back before looking at any practicality. This does make the problem more difficult to interpret so you should only transform the data if absolutely necessary. You must determine which question it is appropriate for. You can ask your instructor or TA for more information.

Remember that there are two types of two-sample inference used in this Lab, two-sample independent and two-sample paired. Though the code is similar for these two procedures, there are not identical.

B (45 points) Is the median household income Graduates (MedianIncome) significantly different between the Northeast and North Central regions (Region)? (Data Set: Clean US Data)

A researcher was interested in determining the differences in and influencing factors of the median household income for the individuals living in different areas across the United States. As a preliminary analysis, the researcher asked a basic question: "If I compare two regions of the U.S., will there be a statistically significant difference in the median household income on a county level?" The researcher selected the Northeast and North Central regions, and you will be testing whether the means of the median household income in the two regions are significantly different at the 5% level.

1. (5 points) Code. Please be sure to subset the data to only include the Northeast ("NE") and North Central ("NC") regions for this question. Create a new dataset, for example, USDataSubset, to store the selected data.
2. (5 points) Should you use a two-sample independent or two-sample paired procedure to analyze the data? Please explain your answer by discussing the statistical issues related to the analysis, instead of using the format of the dataset. In real studies, you will need to know what method you will use to analyze the data – paired or independent – to know what data to gather. If this is a paired situation, please state the common characteristic that makes these data paired. Do this part before you do any coding.
3. (5 points) Should you use a one-sided or two-sided alternative for this analysis? Explain your decision. Do this part before you do any coding.

4. (10 points) Create the three diagnostic plots (boxplot, histogram with the estimated kernel density and the estimated normal density curves, and QQ plot) for the appropriate variable(s). Do you think these data are normally distributed? If not, please apply a log transformation, show the diagnostic plots of the transformed data, and comment on the normality of the transformed data. **If you apply a transformation, use the transformation for the remainder of Part B**. Write a short summary of your findings being sure to comment on each graph and providing the answer on whether the appropriate variable is normal or not.
5. (5 points) No matter how you answered parts 3/4, determine and interpret the 95% confidence interval of the population parameter of interest (which depends upon whether it is a paired or independent sample situation).
6. (10 points) No matter how you answered parts 3/4, test the hypothesis that the average median household income is different between the two regions. Assume a 0.05 significance level for the test. Remember to use the full four-step process.
7. (5 points) Are the conclusions from parts 5 and 6 consistent? Please explain your answer.

C (45 points) Is the average amount of money spent on each pupil in period 1 (EducationSpending) different from that in period 2 (EducationSpendingP2)? (Data Set: Clean US Data) A psychology professor studying relationships wants to know whether the average amount spent on each pupil in period 2 is at least 70 dollars greater than that of period 1 (that is, larger by $70 or more) on average in the United States.

1. (5 points) Code.
2. (5 points) Should you use a two-sample independent or two-sample paired procedure to analyze the data? Please explain your answer by discussing the statistical issues related to the analysis, instead of using the format of the dataset. In real studies, you will need to know what method you will use to analyze the data – paired or independent – to know what data to gather. If this is a paired situation, please state the common characteristic that makes these data paired. Do this part before you do any coding.
3. (5 points) Should you use a one-sided or two-sided alternative for this analysis? Explain your decision. Do this part before you do any coding.
4. (10 points) Create the three diagnostic plots (boxplot, histogram with the estimated kernel density and the estimated normal density curves, and QQ plot) for the appropriate variable(s). Do you think these data are normally distributed? If not, please apply a log transformation, show the diagnostic plots of the transformed data, and comment on the normality of the transformed data. **If you apply a transformation, use the transformation for the remainder of Part C**. Write a short summary of your findings being sure to comment on each graph and providing the answer on whether the appropriate variable is normal or not.
5. (5 points) No matter how you answered parts 3/4, determine and interpret the 95% confidence interval of the population parameter of interest (which depends upon whether it is a paired or independent sample situation).
6. (10 points) No matter how you answered parts 3/4, test the hypothesis that the average median household income is different between the two regions. Assume a 0.05 significance level for the test. Remember to use the full four-step process.

7. (5 points) In one English sentence, explain whether there is evidence that the average amount spent on each pupil in Period 2 is at least 70 dollars greater than that of Period 1. If so, explain whether the amount is greater than $70 by a magnitude of practical significance. If you needed to transform the data, keep this in mind when interpreting the result.