Benjamin Goodwin

MSDS6371

HW4

1. Read Chapter 4 from Statistical Sleuth and answer the conceptual problems at the end of the chapter. Note: You do not need to type these up and turn them in. The answers are at the very end of the chapter.
2. A. Perform a complete analysis using a rank sum test in SAS. (Logging data)

Step 1: H\_0: The two populations are equal (forest recovery)

Step 2: H\_a: The two populations are not equal (forest recovery)

Step 3: alpha: 0.05

Step 4: test statistic: 2.4875

Step 5: p-value: 0.0129

Step 6: Reject H\_0!

Step 7: Based on the evidence above we will reject the null hypothesis and conclude that there is evidence supporting there is a difference between forest recovery.

SAS input:

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

A screenshot of a social media post

Description automatically generated

B. Verify the p-value and confidence interval by running the rank sum test in R

R Input:



R Output:

A picture containing bird, flower

Description automatically generated

A screenshot of a cell phone

Description automatically generated

From the above, we can confirm that both SAS and R have produced the same output/

3) A. State the problem, address the assumptions. Be sure to support with your knowledge of theory (CLT) as well as with histograms, box plots, q-q plots, etc

Problem Statement: A study was conducted concerning income levels of people in two groups, the first was non-college education people, and the second was college educated people. The purpose of the study was to determine if there was a difference in income levels between the groups. We would like to statistically determine if there is any difference in income levels between the two groups of people. In further detail the researchers conducted a study of subjects between ages 41 and 49 years old in the year 2006. They are hoping to answer their initial question using this population, which was, “is there a difference in mean income level between subjects in the group who went to college vs those who did not.”

Assumptions: Check normality of data, independence of samples, and equal variances among groups.

Normality of data seems to be violated, in both the histograms and QQ-plots

Histograms are right skewed

A close up of a map

Description automatically generated

QQ-plots show evidence of non-normal distribution

A close up of a device

Description automatically generated

Additionally, we can see that the equal variances among groups does not hold

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A) A study was conducted concerning income levels of people in two groups, the first was non-college education people, and the second was college educated people. The purpose of the study was to determine if there was a difference in income levels between the groups. We would like to statistically determine if there is any difference in income levels between the two groups of people. In further detail the researchers conducted a study of subjects between ages 41 and 49 years old in the year 2006. They are hoping to answer their initial question using this population, which was, “is there a difference in mean income level between subjects in the group who went to college vs those who did not.”

B) Show all 6 steps, including a thoughtful, thorough, yet non-technical conclusion. Include a confidence interval.

Step 1: H\_0: Mu\_educ\_12 = Mu\_educ\_16

Step 2: H\_A: Mu\_educ\_12 != Mu\_educ\_16

Step 3:critical value: -9.98

Step 4: P-value: <0.0001

Step 5: Reject H\_0

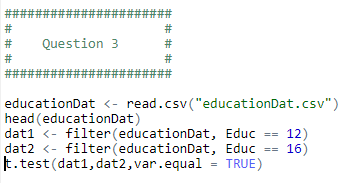
Step 6: Since we rejected the null hypothesis we can conclude that there is a difference between the two groups. A 95% confidence interval is (-13197.61,-8807.313)

Conclusion and confidence interval:

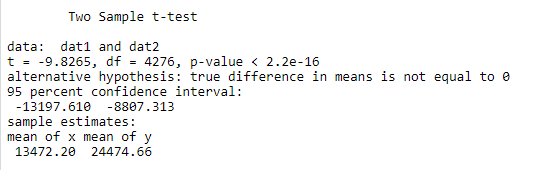
C) Scope of inference: The data is a subset of National Longitudinal Survey of Youth (NLSY79). The question did not indicate random sampling. However, the sample size was quite large, all subjects were between 41 and 49. I believe this data can be inferred om the population of subjects sampled.

D) Verify the Welch’s t statistic and p-value with R (using R function t.test).

R input:



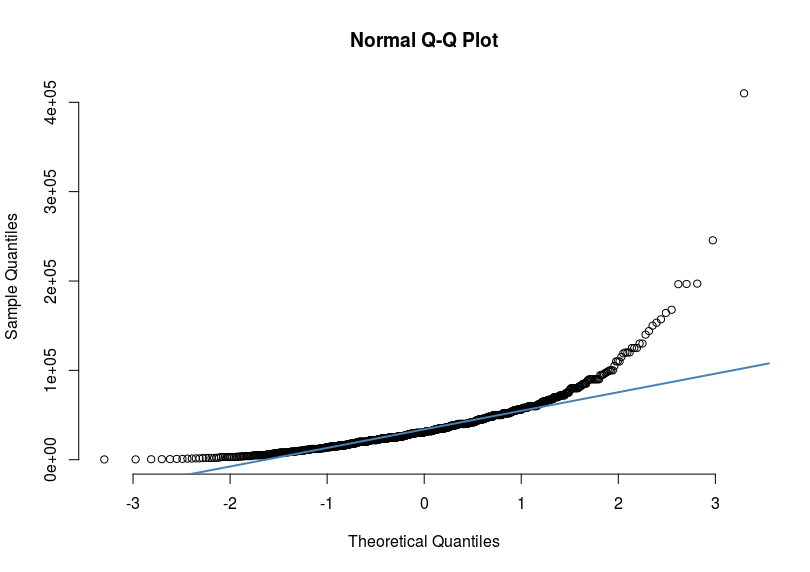
R output:



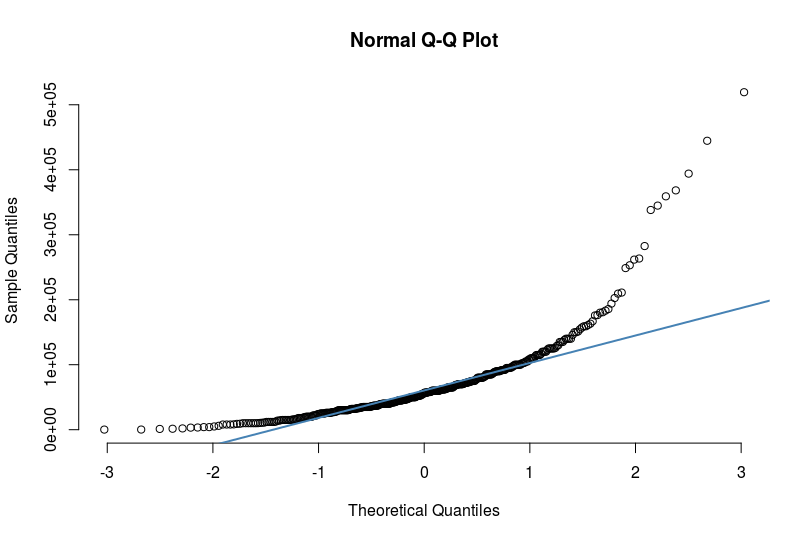
E) Would you prefer to run the log transformed analysis you ran in HW3, or do you feel this analysis is more appropriate?

I think that the log transformed analysis is more appropriate, this is because I feel the log transformed model better conforms to model assumptions of normality of data, equal variances, and independent data. I have attached the plots below of the transformed data and non-transformed data.

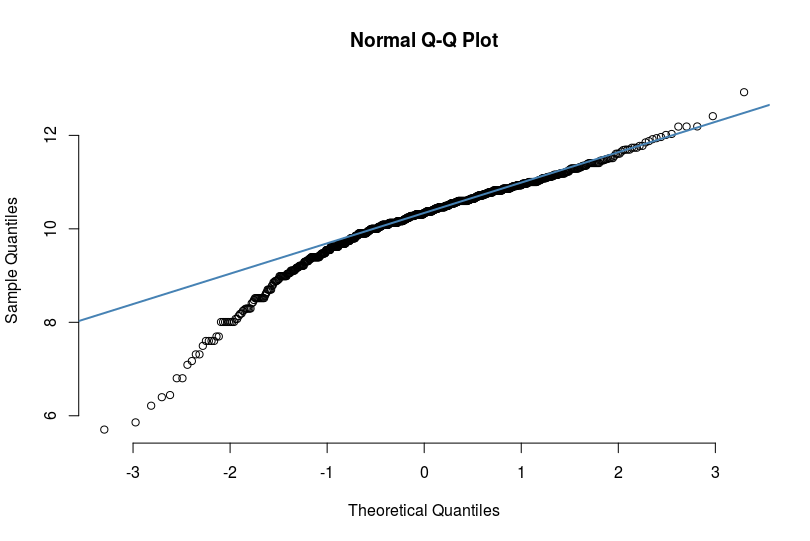
Plot 1: (QQ-plot for non-transformed data (12-year education data))

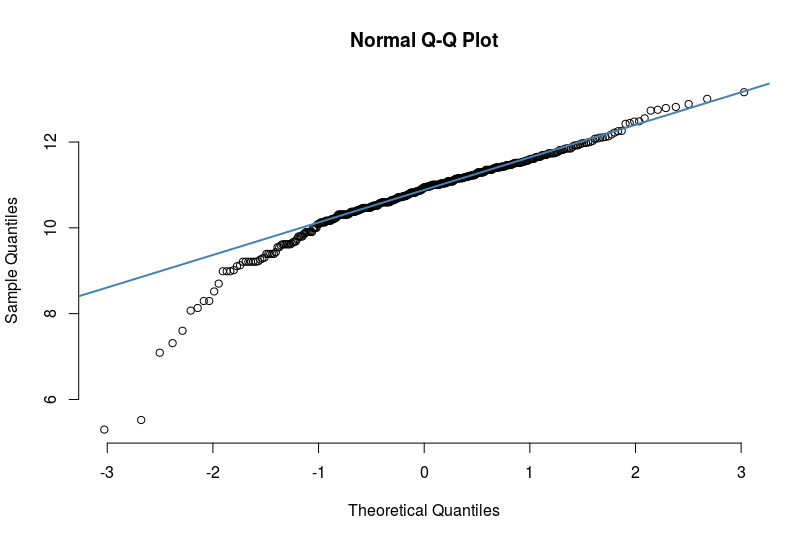


Plot 2:(QQ-plot for non-transformed data (16-year education data))



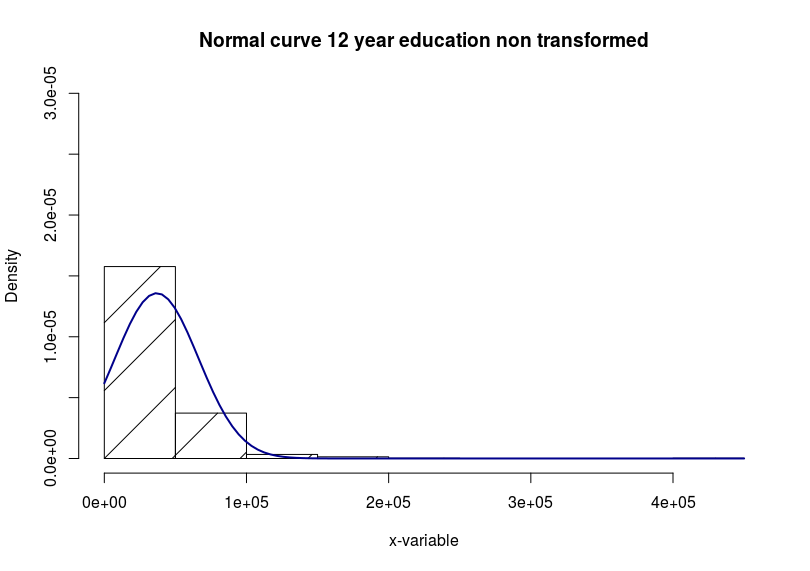
Plot 3:(QQ-plot for transformed data (12-year education data))



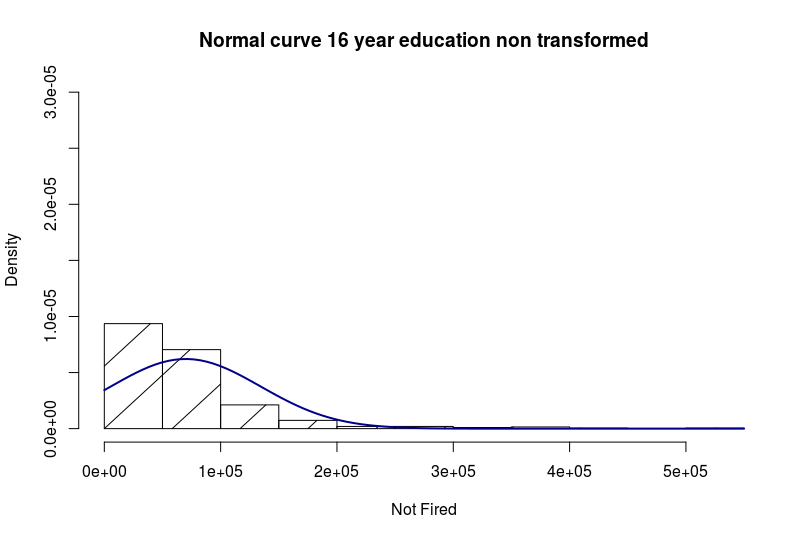
Plot 4: (QQ-plot for non-transformed data (12-year education data))

Based on the qq-plots, it seems that log transformed data better adheres to the assumption of normally distributed data

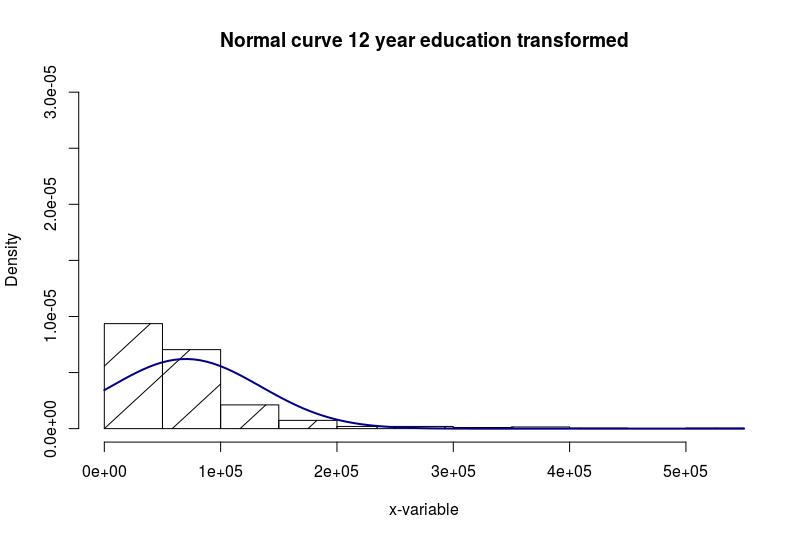
Histogram 1: (Plot of normal curve overlaid by non-transformed data (12-year education data))



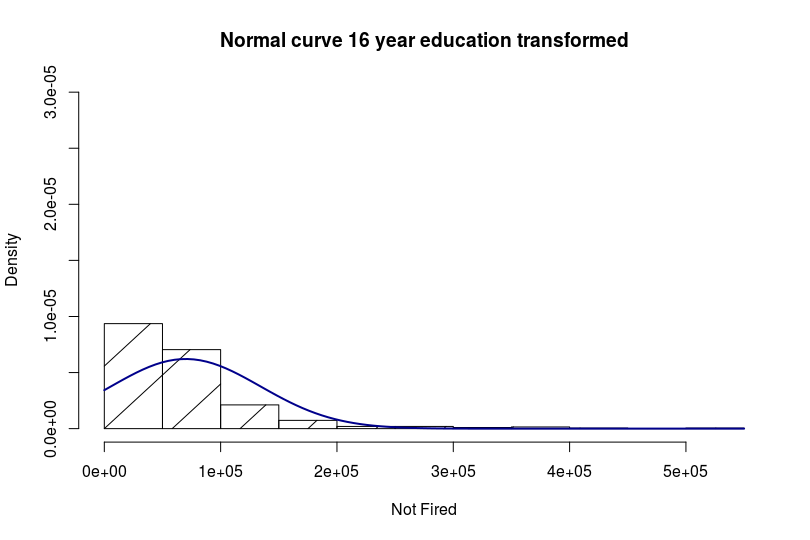
Histogram 2: (Plot of normal curve overlaid by non-transformed data (16-year education data))



Histogram 3: (Plot of normal curve overlaid by transformed data (12-year education data))



Histogram 4: (Plot of normal curve overlaid by non-transformed data (16-year education data))



Based on all the model assumptions (plots above) and general subject knowledge, we can see that the transformed model more correctly follows our assumptions for a t-test. Transforming the data has the interesting effect of changing the conclusion, instead of rejecting the null, we will now fail to reject the null on the transformed data.

4) Chapter 4, Problem 20 from the text. Show all work. “By hand” here means actually by hand. Simply take a picture of your work and include it in your pdf/doc file. Include your sorted, labeled, and ranked data; your calculations of the mean and standard deviation of the assumed distribution of the rank sum statistic under Ho; your calculation of the Z statistic with a continuity correction; your p-value, and conclusion. (No confidence interval necessary here.)

A) A close up of text on a white background

Description automatically generated

Conclusion: Reject H\_0! We can conclude that is there is a difference between the trauma and non-trauma groups.

B) SAS output:

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

No, my p-value is slightly different than the SAS p-value. However, the conclusion is still the same, and the difference is very small. Additionally, I did not use the continuity correction in my SAS code. This can however be specified. The z scores are on opposite sides, however if you take the absolute value of the z-score, they are nearly identical.

C)

i) Statement of problem: We are concerned with the metabolic expenditure between patients admitted to the hospital for non-trauma related causes and trauma related causes. We are interested in determining if there is a difference between the two groups.

II) Assumptions:

Equal Variance, based on the data, there is evidence to suggest the two groups don’t have equal variance:

A screenshot of a cell phone

Description automatically generated

Normality of data, appears to be violated. The QQ-plots don’t support this and neither do the histograms.

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

The samples however appear to be independent.

III)

Step 1: Trauma\_distribution = non\_trauama\_distribution

Step 2: Trauma\_distribution != non\_trauama\_distribution

Step 3: Z-value: 3.0116

Step 4: P-value = 0.0013

Step 5: Decision: Reject H\_o

Step 6: Based on the evidence above I will reject the null hypothesis and conclude that there is evidence for a difference in distribution between the trauma and non-trauma patients.

Hodges-Lehmann Confidence interval from SAS:

A screenshot of a social media post

Description automatically generated

IV) Scope of inference: Based on the model assumptions and small dataset, results from this model may not be accurate. In addition, the authors did not mention if the patients were randomly sampled. Finally, this is an extremely small dataset, so we cannot guarantee this will extend to the general population of interest.

5 A)

A close up of text on a white background

Description automatically generated

B) R

Input:A screenshot of a cell phone

Description automatically generated

Output:

A picture containing knife, bird

Description automatically generated

SAS

Input:

A screenshot of a cell phone

Description automatically generated

SAS output:

A screenshot of a cell phone

Description automatically generated

C)

Six Step test

Step 1: H\_0: Median\_before\_time = Median\_after\_time

Step 2: H\_a: Median\_before\_time != Median\_afrer\_time

Step 3: Z-value: -2.1919

Step 4: P-value =0.01618

Step 5: Reject H\_0!

Step 6: We rejected the null hypothesis and conclude that there is a difference in the median puzzle doing time between the children before and after doing yoga.

D

Step 1: H\_0 mu\_before =mu\_after

Step 2: H\_a mu\_before != mu\_afrer

Step 3: t-value: 1.99

Step 4: P-value: 0.0636

Step 5: Reject H\_0!

Step 6: We rejected the null hypothesis and conclude that there is a difference in the mean puzzle doing time between the children before and after doing yoga.

SAS input:

A screenshot of a social media post

Description automatically generated

SAS output:

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

We can see that the results are significant, and the assumptions of the t-test are met.

E) R output: Results match

A picture containing bird, flower

Description automatically generated

F)

Complete Analysis:

Problem Statement: A study was performed to test a new treatment for autism in children. In order to test the new method, parents of children with autism were asked to volunteer for the study in which 9 parents volunteered their children for the study. The children were each asked to complete a 20 piece puzzle. The time it took to complete the task was recorded in seconds. The children then received a treatment (20 minutes of yoga) and were asked to complete a similar but different puzzle. We are interested in seeing if there is a difference in the amount of time it takes before and after doing yoga, and if this is a “real” result.

Assumptions:

A screenshot of a cell phone

Description automatically generated

We can see that the variances are approximately equal, and that this assumption is met.

A close up of a map

Description automatically generated

From these QQ-plots we can see that the data is fairly well approximately normal. Although the histograms below require the QQ-plots to verify this assumption.

A close up of a map

Description automatically generated

We will assume our samples are independent.

Thus all assumptions of the two-sample t-test are met.

Results: In both a signed rank test and a t-sided t-test we have p-values far below the 0.05 threshold. We will reject the null in both situations (difference in mean and median) and conclude that there is a difference between puzzle doing times before and after yoga.

Hypothesis test for two tests:

Six Step test

Step 1: H\_0: Median\_before\_time = Median\_after\_time

Step 2: H\_a: Median\_before\_time != Median\_afrer\_time

Step 3: Z-value: -2.1919

Step 4: P-value =0.01618

Step 5: Reject H\_0!

Step 6: We rejected the null hypothesis and conclude that there is a difference in the median puzzle doing time between the children before and after doing yoga.

Step 1: H\_0 mu\_before =mu\_after

Step 2: H\_a mu\_before != mu\_afrer

Step 3: t-value: 1.99

Step 4: P-value: 0.0636

Step 5: Reject H\_0!

Step 6: We rejected the null hypothesis and conclude that there is a difference in the mean puzzle doing time between the children before and after doing yoga.

Scope of inference: The study used volunteers, and so we should keep this in mind when evaluating results. A random sample was not used, and therefore I do not believe we can draw inference on the population of intertest from this study.