Ben Goodwin

MSDS6371

HW 3

1. A:

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

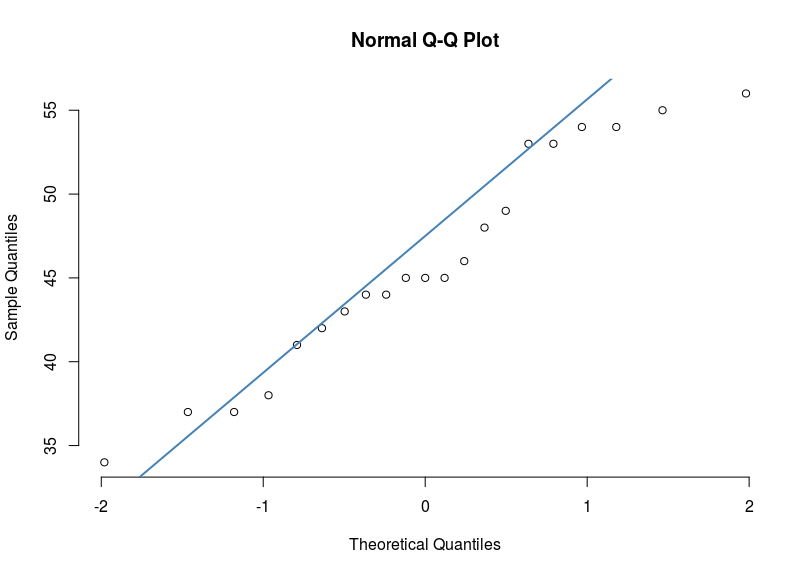
A screenshot of a cell phone

Description automatically generated

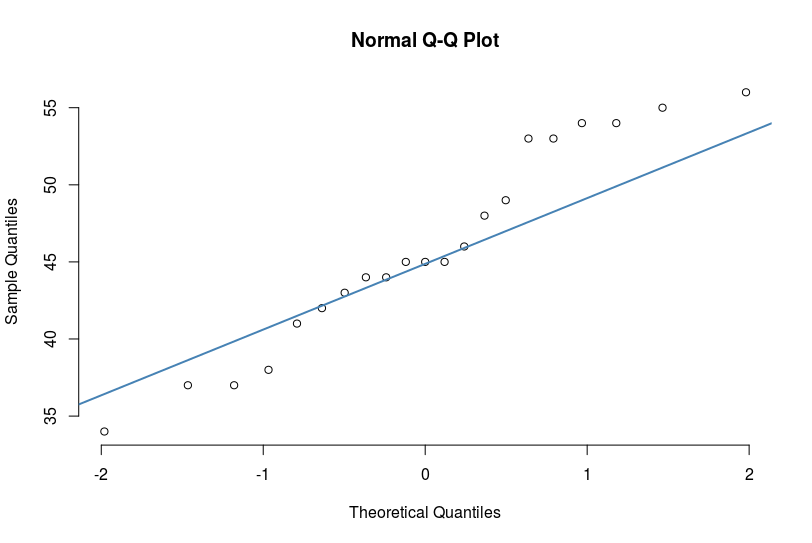
Based on the model t-test model assumptions of normally distributed data, and equal variances, we have approximately equal variances of 0.2771 and 0.2870, which we can say are roughly equal. Based on the histograms and QQ-plots, the data seems to violate the assumption of normality. Although the t-test is robust to this departure from normality, I think that a different test would result in “better” results.

B: R Code for assumption check

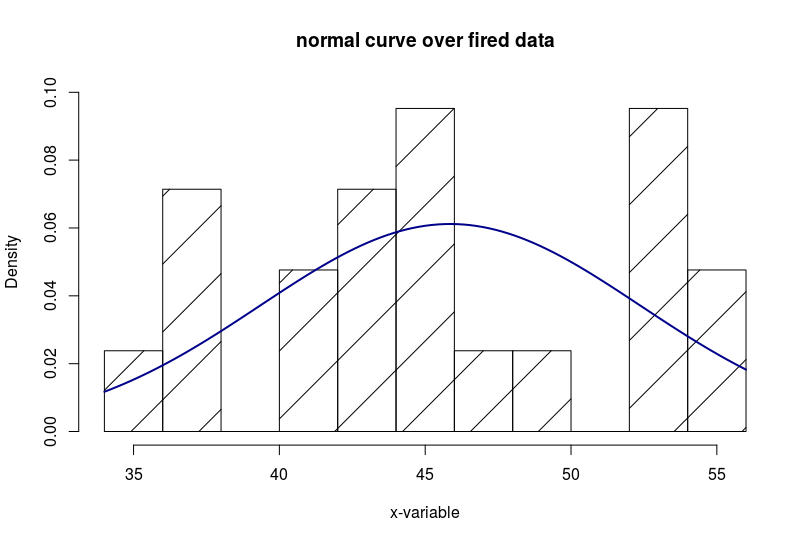
R output assumption plots:

Plot 1, QQ-plot for Fired: 

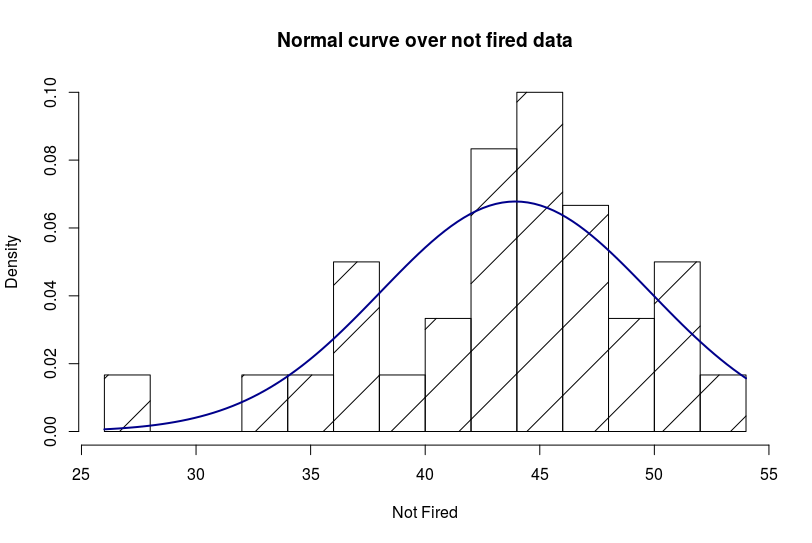
Plot 2, QQ-plot for not Fired:



Plot 3, Histogram for Fired data with overlay of normal curve:

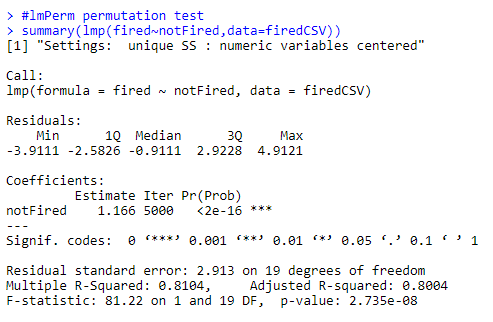


Plot 4, Histogram for not Fired data with overlay of normal curve:



C)

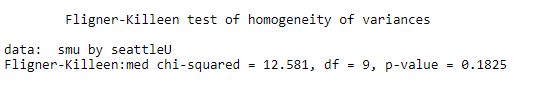
1. In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, suppose that a random sample of the ages of fired and not. We are interested to see if we can detect a difference between the mean ages of people who are fired and who are not fired to determine if there is evidence as to discrimination by preferentially firing older workers. I do not feel the t-test is appropriate.
2. For a T-test we are considered with the data being normally distributed and equal variance among the groups. From the plots, we can see the normality assumption is violated. Additionally, to test homogeneity of variance I used a Fligner test which is very robust against departures from normality. The test resulted in a p-value of 0.5916, and thus I will fail to reject the null hypothesis and conclude the variances differ.
3. I used a permutation test in R, with the LMperm package and verified my results by using the coin package to do an independence test. Below are my results from LmPerm:



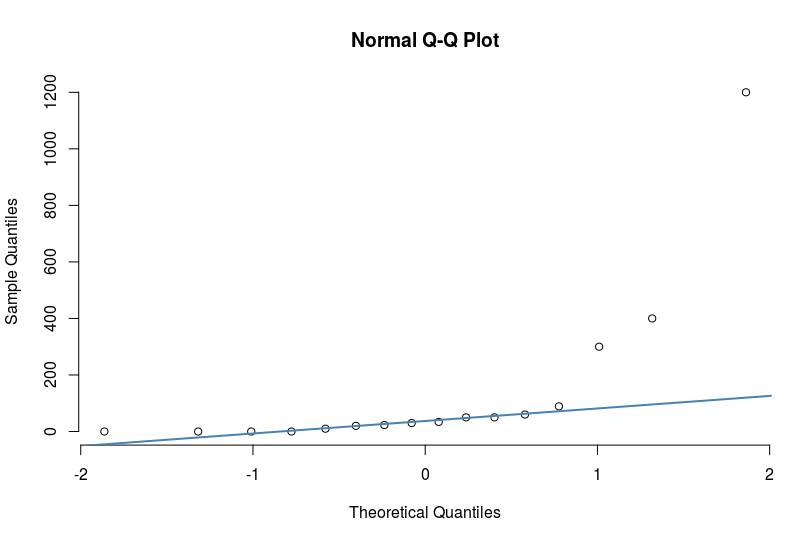
1. After using a permutation test, the resulting p-value is <0.00001 and very significant. Thus we can reject H\_0, and conclude there is sufficient evidence to conclude there is a difference between the groups. A 95% CI for the data (-1.678993,5.526612). ). If we repeatedly sampled and computed confidence intervals, 95% of the intervals would contain the true population mean.
2. Due to the relatively small amount of information presented in the question we cannot be sure if we can extend this sample to the general population of the bats. In general, we need to be careful as the sample may not be random.

Question 2

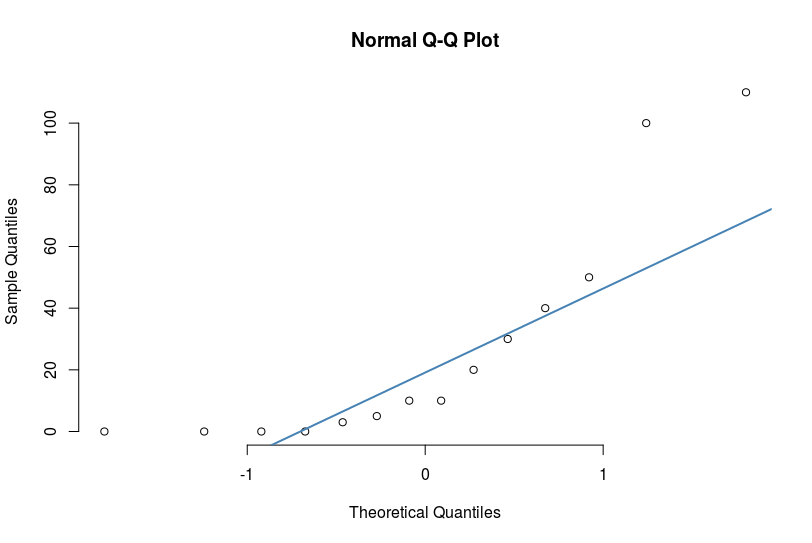
1. Image 1, test for homogeneity of variance



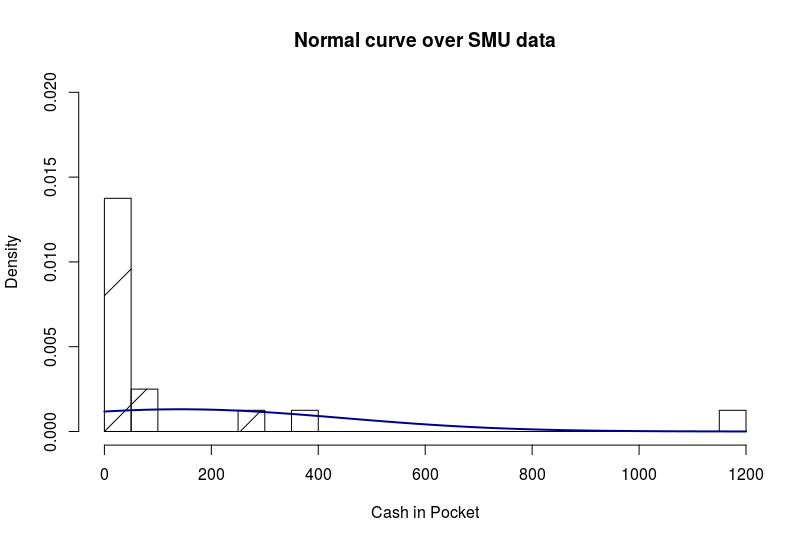
Plot 1, QQ-plot for SMU data



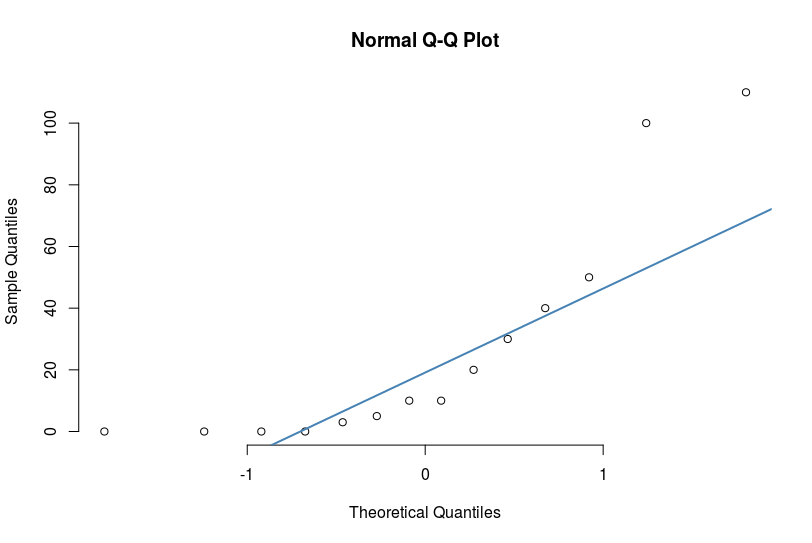
Plot 2, QQ-plot for Seattle U



Plot 3, Histogram of SMU data with overlay or normal curve



Plot 4, Histogram of Seattle U data



1. For a T-test we are considered with the data being normally distributed and equal variance among the groups. From the plots, we can see the normality assumption is violated. Additionally, to test homogeneity of variance I used a Fligner test which is very robust against departures from normality. The test resulted in a p-value of 0.5916, and thus I will fail to reject the null hypothesis and conclude the variances differ. Based on these factors, I do not feel the t-test is appropriate.

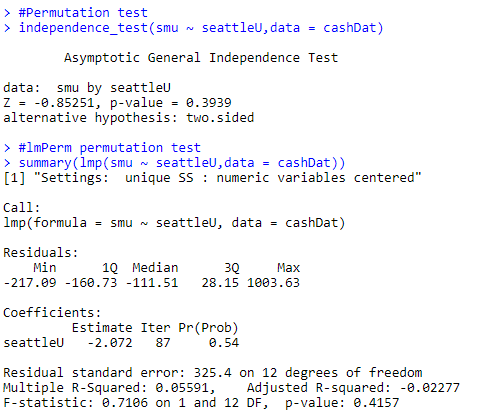
B) Permutation Test

B)

1) A business class was polled at both SMU and Seattle U to see if there was a difference in the amount of money they had in their pocket at the moment to determine if the vending machines on each campus should have a coin/cash acceptor or just a credit card reader. We are interested seeing if there is a difference in the mean cash amount carried by business students at the respective schools.

2) Based on the results of part A, I feel the assumptions of the t-test are not met in this problem. We will need to use an alternative test, such as a permutation test.

3) Permutation test below (alternative to T-test )

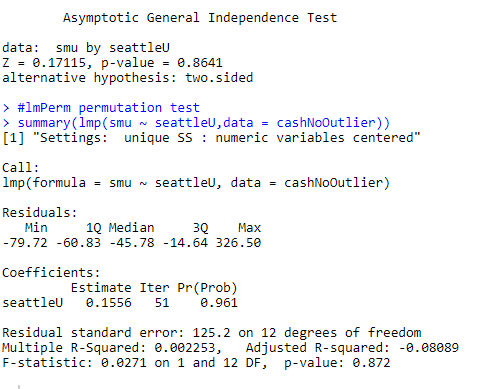


4) Conclusion, FTR h\_0 and conclude that there is insufficient evidence to conclude that the mean amount of cash carried by the respective business students differs.

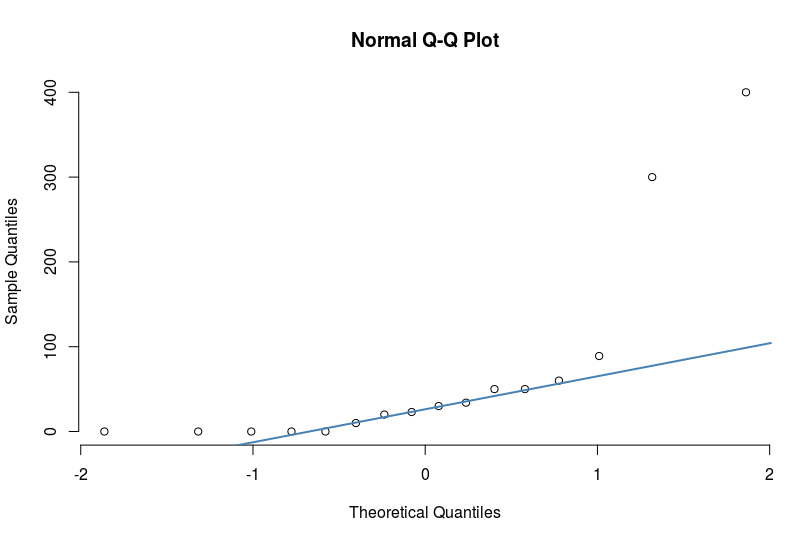
A 95% CI: (-48.39481,277.64481). Meaning, if we sampled repeatedly we can conclude the true population mean lies between these values. P-value = 0.4157.

1. The scope of inference on this population is limited, as the researchers did not employ random sampling, thus we cannot infer this result to the larger population. However, if our population of interest were a particular business class at SMU or Seattle U, these results would work!

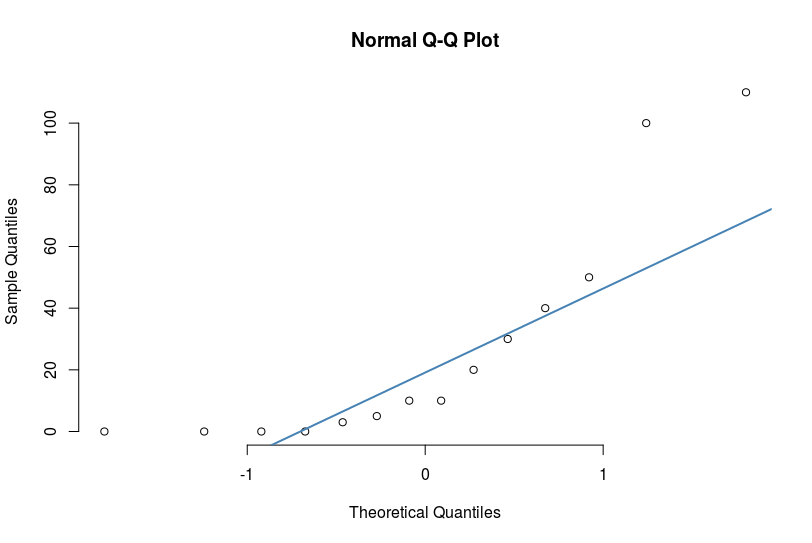
C) The plots and test after “outlier removed”



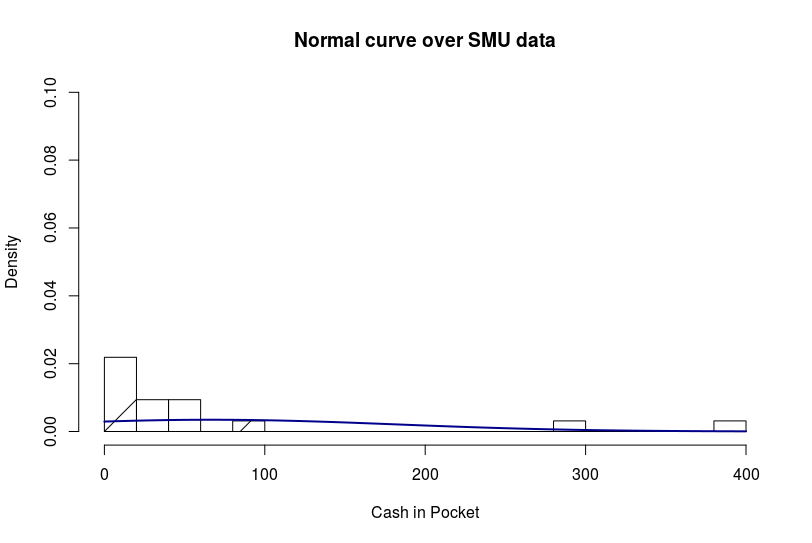
Plot 1,QQ-plot of SMU data with no outlier



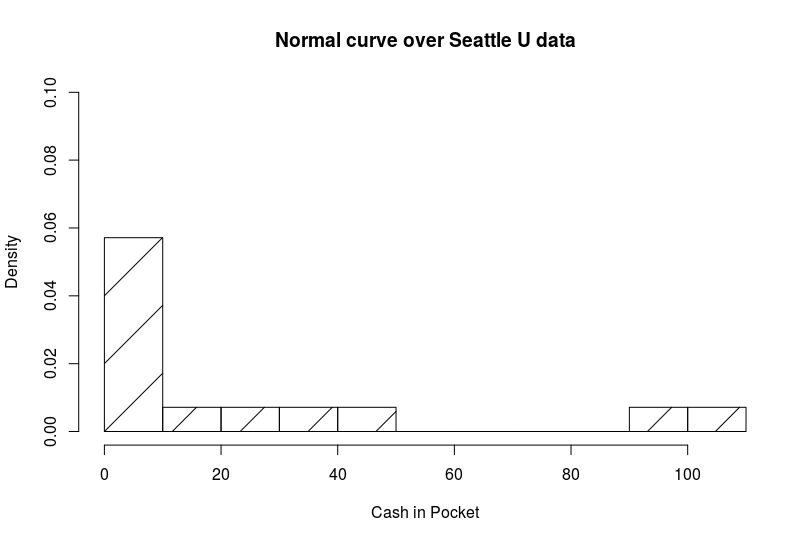
Plot 2, QQ-plot of Seattle U



Plot 3, Normal curve over SMU data with outlier removed



Plot 4, Histogram over Seattle U



P-value: 0.872 with outlier removed. Based on the analysis so far, seems like the outlier has no reason to be removed.

3) Six Step Test to test for a difference between incomes

1) H\_0: Mu\_12\_income = Mu\_16\_income

2) H\_A: Mu\_16\_income != Mu\_16\_income

3) T-Statistic: -7.4721

4) P-Value: P<0.0001

5) Reject H\_0!

6) Based on the evidence presented above we have evidence to conclude that there is a difference in the mean income level of those who graduated college and those who did not. Thus, we rejected H\_0. Based on the data, it seems as though the college educated group while having a higher mean income level has a higher standard deviation as well. This implies that in general college educated humans have a higher mean income, this is not always the case, as incomes are highly variable. However, not having a college educational almost certainly guarantees a lower income.

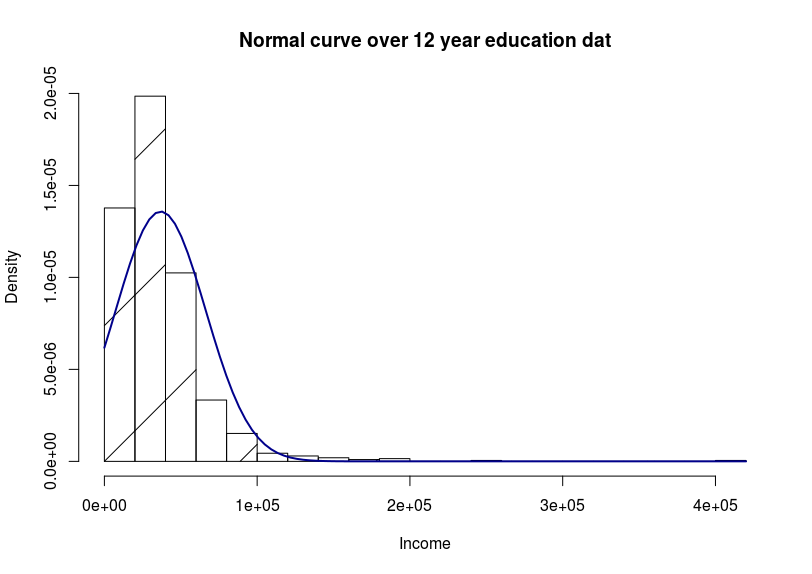
Questions continued below as a separate page, summary of statistical findings.

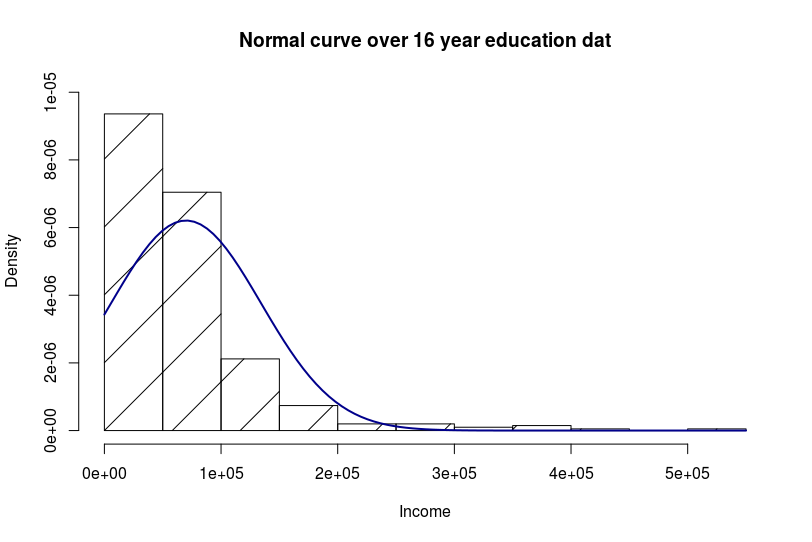
Summary of Statistical Findings on National Longitudinal Survey of youth (NLSY79)

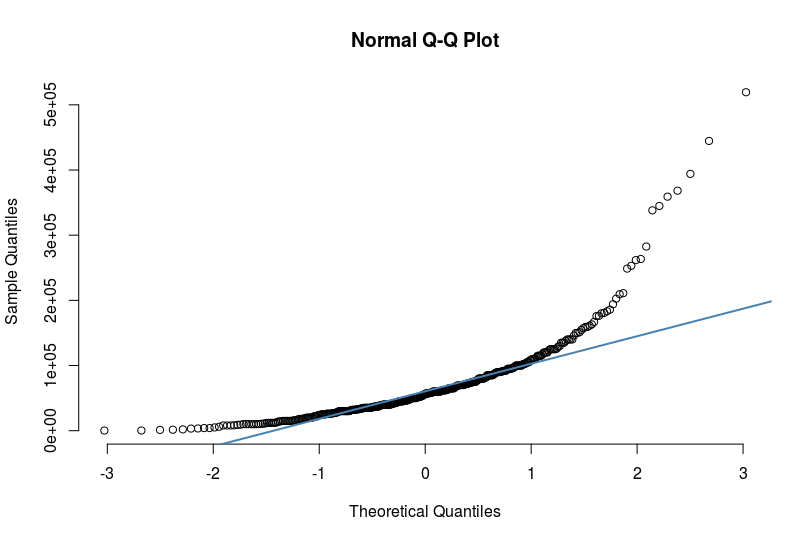
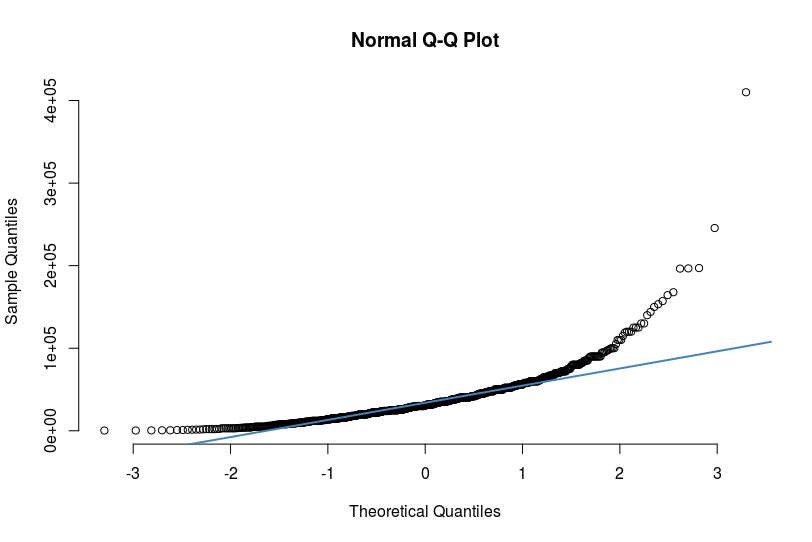
1. What is the question we would like to answer?

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.”

1. Ideally we would have normally distributed data and equal variance among groups. However this data does not meet all the assumptions. Below are plots supporting this conclusion.

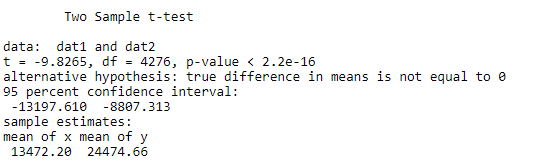






We can very clearly see that our model assumptions are violated in both normality, and variance.

1. Using a t-test (with pooled variance, var.equal=true)



1. Based on the evidence presented above we have evidence to conclude that there is a difference in the mean income level of those who graduated college and those who did not. Thus, we rejected H\_0. Our p-value is <0.001 (very small). A 95% CI (13472.20,24474.66)
2. 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.