MSDS 6371 HW 3:

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 fired people in the American Samoa Government are listed below:

**Fired**

34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56

**Not fired**

27 33 36 37 38 38 39 42 42 43 43 44 44 44 45 45 45 45 46 46 47 47 48 48 49 49 51 51 52 54

* + - * 1. Check the assumptions (with SAS) of the two sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make sure and copy and paste the histograms, qq-plots or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

title 'Histogram Compare of Ages for Fired vs Not Fired';

ods graphics off;

proc univariate

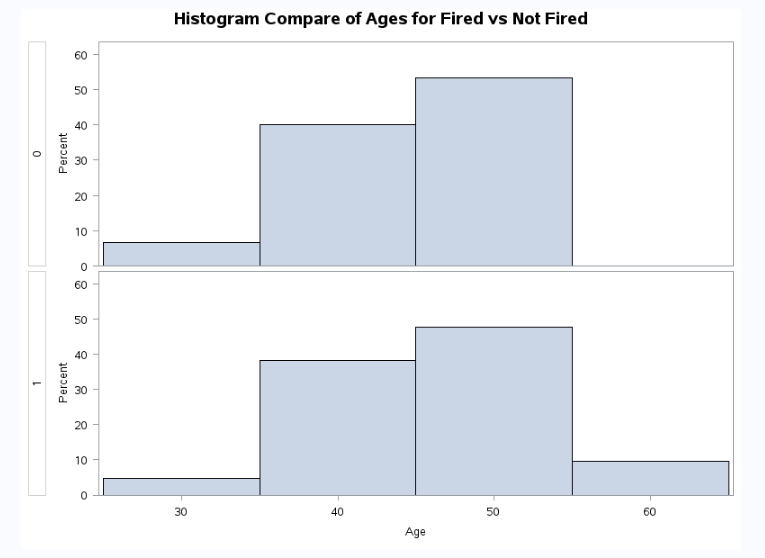
data = agedisctest noprint;

histogram Age / midpoints = 30.0 to 60.0 by 10;

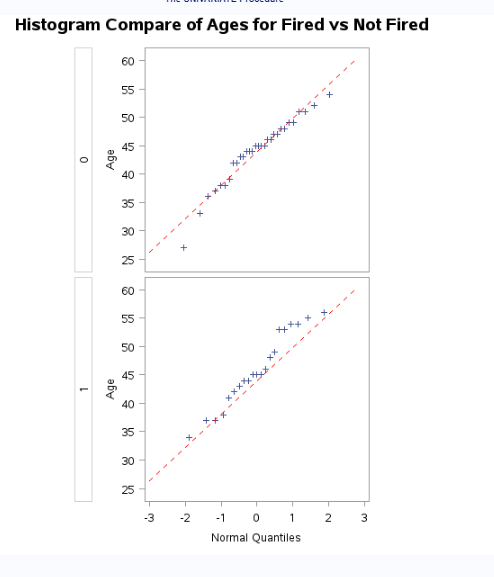
qqplot Age / normal(mu=est sigma=est color = red l=2) square;

class FiredInd;

run;



**Histogram bins are large, but both histograms look somewhat normal.**



The QQPlots between the fired and non-fired group look normal as the points look like they’re on a straight line.



A visual check was completed by comparing histograms and qqplots between the Fired and Not fired data sets and I have concluded that both data sets appear to be normal distributions and it appears they have equal variances.

Also, given that the p-value is greater than our significance level of alpha = 0.05, we fail to reject the null hypothesis of equality of variances (p-value = 0.6005) and conclude that there is not enough evidence to suggest the variances are different.

Given the above, I think we can continue appropriately with a t-test.

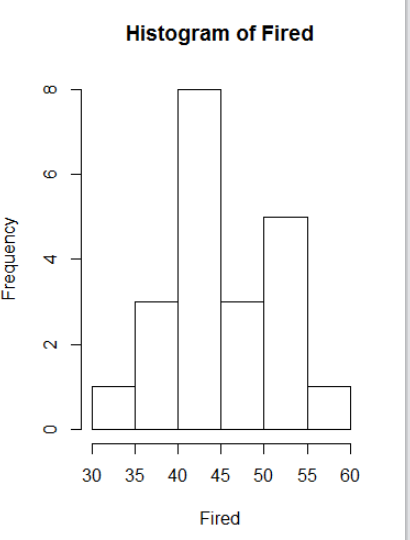
* + - * 1. Check the assumptions with R and compare them with the plots from SAS.

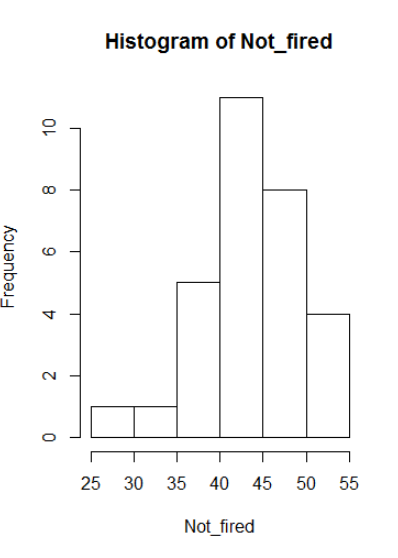
hist(Fired)

hist(Not\_fired)

qqnorm(Fired)

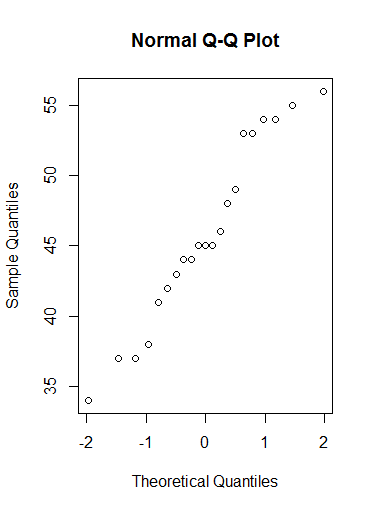
qqnorm(Not\_fired)



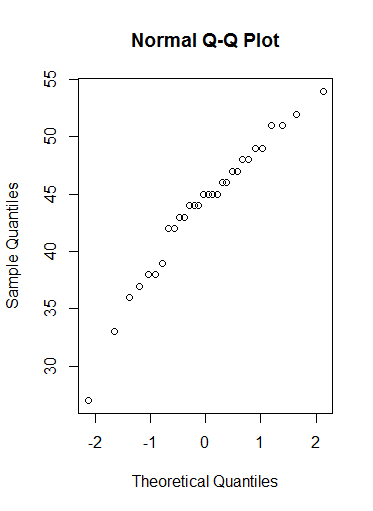


I used the default command in R to generate the histogram which have a different bin size and are not centered. Distributions are slightly shifted, but I believe that both can still pass as being normal.

Fired:



Not Fired:



Both qqplots suggest a normal distribution like the qqplots in SAS. Concluding that we can proceed with a ttest after looking at both the histograms and qqplots between SAS and R.

* + - * 1. Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis be sure and cover all the steps of a complete analysis.

1. State the problem.

2. Address the assumptions of t-test (from part a).

3. Perform the t-test if it is appropriate and a permutation test if it is not (judging from your analysis of the assumptions).

4. Provide a conclusion including the p-value and a confidence interval.

5. Provide the scope of inference. (Steps 3-5 are from your previous HW; you are just putting everything together.) NOTE: THIS QUESTION SHOULD BE EASY AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 or 2 above.) (You can basically copy and paste the rest. We are simply putting everything together to make a complete report.)

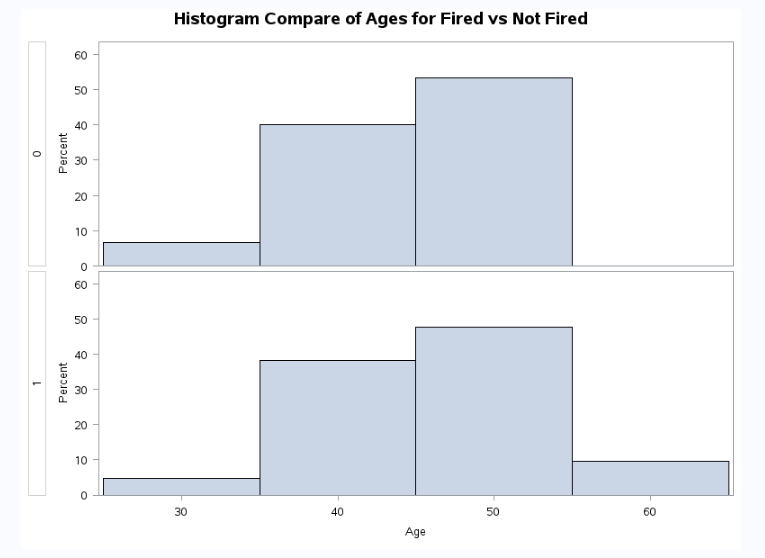
# Report:

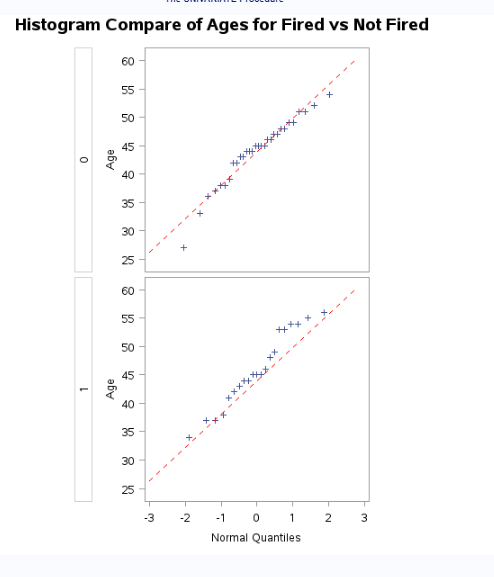
Problem: In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. A random sample of the ages of fired and not fired people were provided in the American Samoa Governement. We would like to determine if the means between the two groups are equal.

## Check Assumptions:

After reviewing the histograms and qqplots for both the fired group and not fired group, I have concluded that both data sets appear to be normally distributed.

1. Normally distributed populations:





1. Independent Observations

The employees or ex-employees for that matter were randomly selected. We will assume that the observations are independent.

## Identify the H0 and Ha:

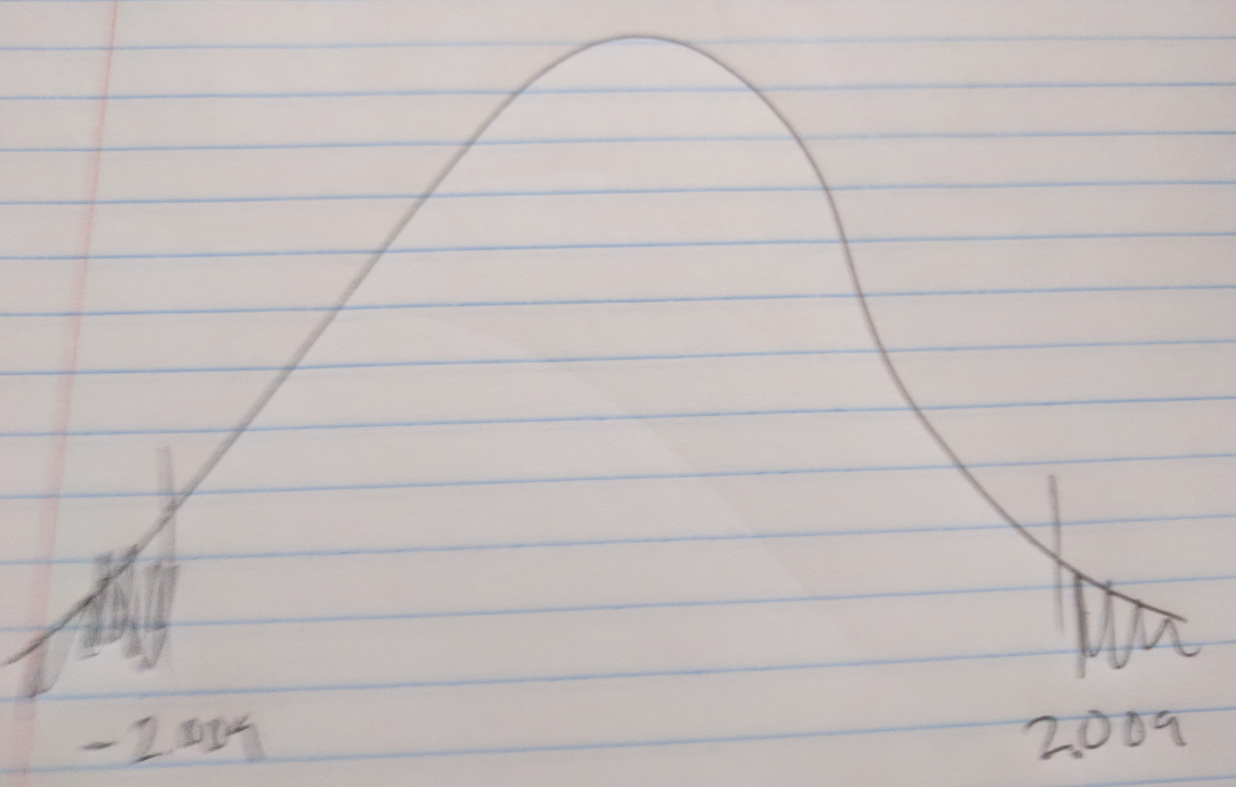
H0 = Average Fired Age = Average Not Fired Age | uF – uNF = 0

HA = Average Fired Age <> Average Not Fired Age | uF – uNF <> 0

## 

## Determine the critical value:

*Critical value: 2.009*



Find the t-value:

*-1.10*

Find the p-value:

*.2771*

## Step 5:

*Fail to reject H0*

## Step 6

*There is not sufficient evidence to suggest that the mean age of those who were fired is different from the mean age of those who were not fired (p-value = 0.2771).*

**Scope:**

Since the study is between fired and non-fired people, this is an observational study has both groups don’t have the opportunity to be both fired or not, so we can’t infer causality.

However, the sample was random sample and thus any results can be inferred back to the individuals of the American Samoa Governement.

**Conclusion:**

There is not sufficient evident to support the claim with an alpha =.05 level of signifigance (p-value = .2771) that the mean age of those who were fired is higher than those who were not fired. A 95% pooled confidence interval for this difference is (-5.4413, 1.5936).

Note: Perhaps you might be wondering at this point in the HW, “Why are we always testing the assumptions of the t-test? Is it the best test? Should we always run the t-test when we can?” These are very good questions and open questions that are up for debate! The one thing that is mathematically proven and not up for debate is that if the assumptions are met, the two sample t test is the most powerful (in terms of Power = 1 – beta) test in the universe at testing the claim of the difference of means. Two questions may arise here … 1. Do we ever really have the assumptions fully met in the real world and just how much power do we give up at varying degrees of violation of the assumptions? 2. Do we always want inference on the equality/difference of means? We will continue to answer these questions in Unit 4. Also note that we started to answer number two with a t-test of log transformed data. The inference there is on the equality (ratio) of medians which may be a better measure of center when dealing with right or left skewed data!)

1. In the last homework it was mentioned that a Business Stats class here at SMU was polled and students were asked them how much money (cash) they had in their pockets at that very moment. The idea was to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if it should just have the credit card reader. However, a professor from Seattle University polled her class with the same question. Below are the results of the polls.

**SMU**

34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0

**Seattle U**

20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

* + - * 1. Check the assumptions **(with SAS or R)** of the two sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make sure to copy and paste the histograms, qq-plots or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

title 'Histogram Compare of Pocket Money';

ods graphics off;

proc univariate

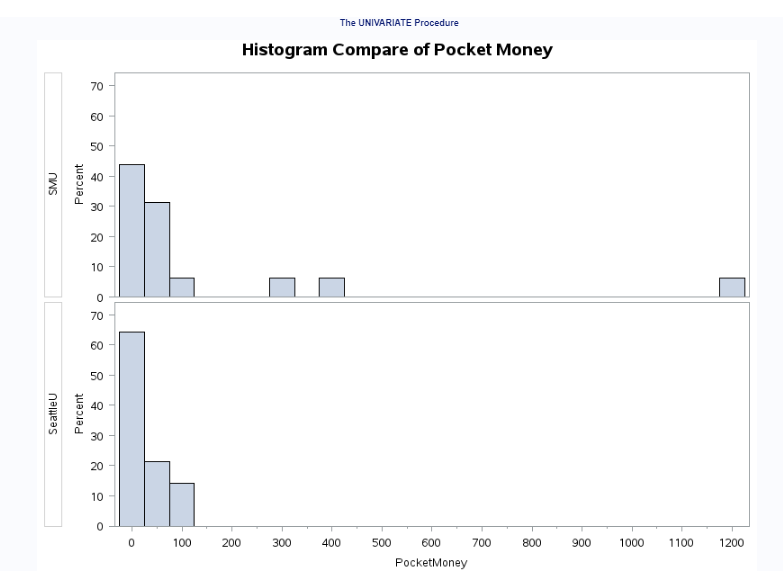
data = pocketmoney noprint;

histogram PocketMoney / midpoints = 0 to 1200.0 by 50;

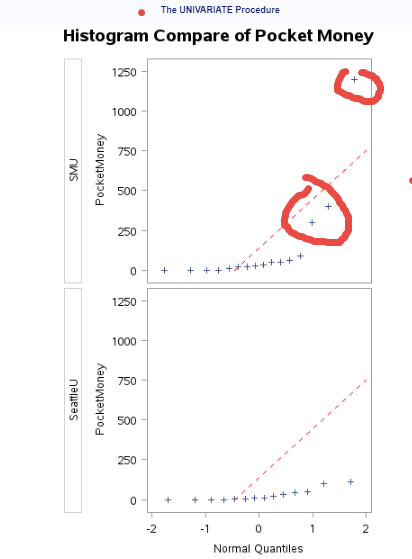
qqplot PocketMoney / normal(mu=est sigma=est color = red l=2) square;

class School;

run;



There is a lot of right skew above. This is mainly because of the outlier of 1200 in the SMU dataset. Also, with so many people not carrying cash, the distribution looks a little one sided given that students can’t have negative money here. Based on the above , picture, I can’t say that this data looks normally distributed.



When looking at the qqplot, You can see that there is somewhat of a linear relationship for SeattleU, but not SMU. Three points (circled in red) really standout in the SMU dataset that once again suggests that this data isn’t normally distributed.

Based on the above, I don’t think a t-test is appropriate for this data set.

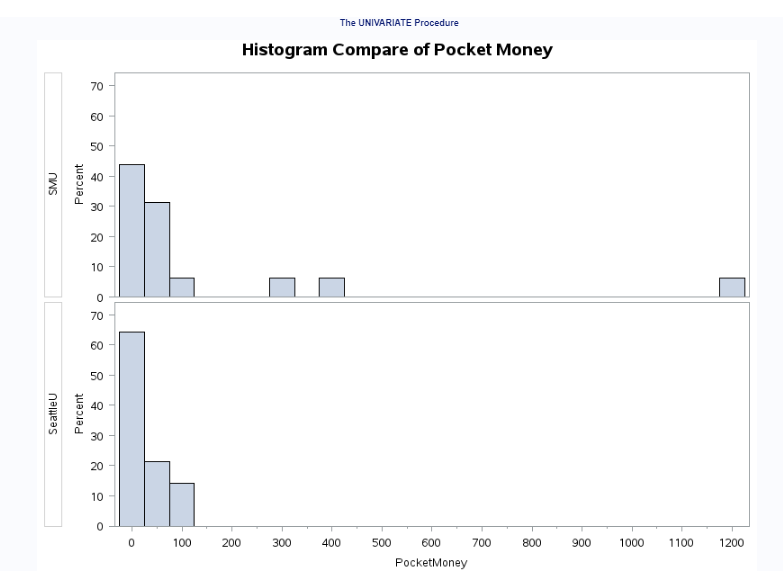
Independence:

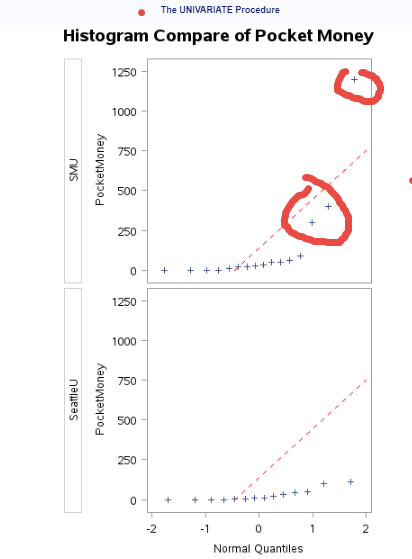
Independence can’t be established here given that the population is from the same class. We cannot infer that this population represents the entire school that will be accessing the vending machine.

* + - * 1. Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis, be sure to cover all the steps of a complete analysis. 1. State the problem. 2. Address the assumptions of t-test (from part a) 3. Perform the t-test if it is appropriate and a permutation test if it is not (judging from your analysis of the assumptions.) 4. Provide a conclusion including the p-value and a confidence interval. 5. Provide the scope of inference. (Steps 3-5 are from your previous HW; you are just putting everything together.) NOTE: AGAIN, THIS QUESTION SHOULD BE EASY AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 or 2 above.) (You can basically copy and paste the rest. We are simply putting everything together to make a complete report.)

Problem: Should those in charge of school vending machines include expensive bill/coin acceptor or should it have a credit card reader. We will examine students from two business classes to see if they have any money in their pockets and we would like to see if the mean amount from SMU equals the mean amount from Seattle U.

Assumptions:





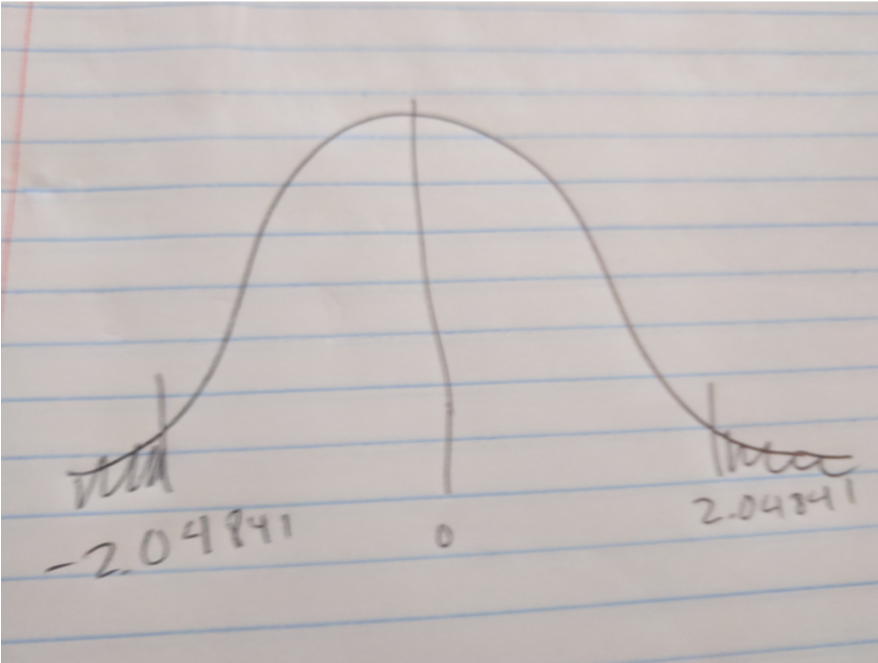
Based from the histogram and qqplots above, it doesn’t look like this data is normal. In red, there are 3 values that appear to be outliers for SMU.

Step 1:

H0 = uSMU = uSeattleU | uSMU – uSeattleU = 0

HA = uSMU<> uSeattleU

Step 2:



Step 3:

t-value = 1.40

Step 4: P-Value = .1732

Step 5:

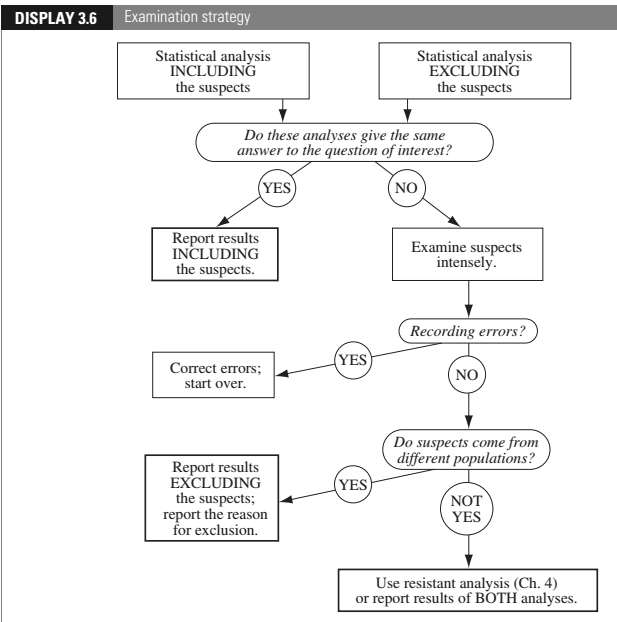
Fail to reject the null hypothesis

Step 6:

This is an observational study where students from two classes were observed on how much money they had in their pockets. Casualty cannot be inferred and the results only apply to the population within these classes.

There is not sufficient evidence to suggest that the mean dollars of those from SMU are different than the mean dollars from SeattleU. The p-value is high at .1732 and the two-side test cannot be rejected. . A 95% pooled confidence interval for this difference is (-53.3711, 282.6).

* + - * 1. Note the potential outlier in the SMU data set. Recheck the assumptions in SAS or R without the outlier. Does this change your decision about the appropriateness of the t tools? Compare the p-value from the t-test with and without the outlier. Based on your analysis so far, what should we do with this outlier? Consult the outlier flowchart in Section 3.4.

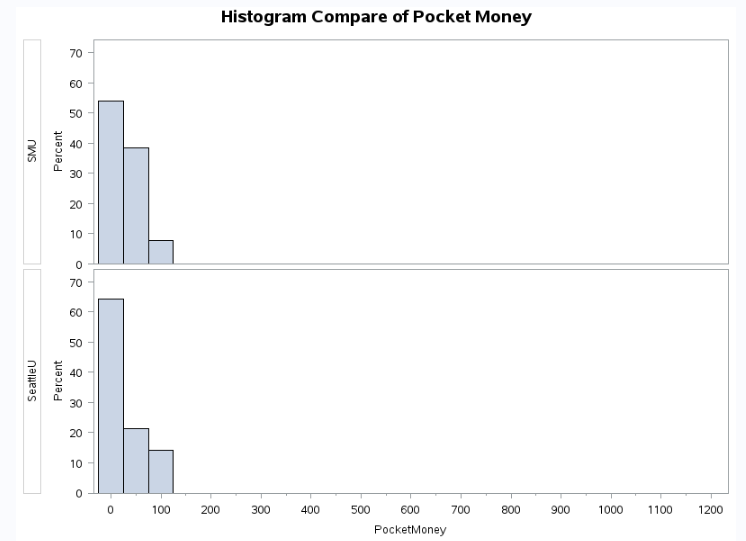
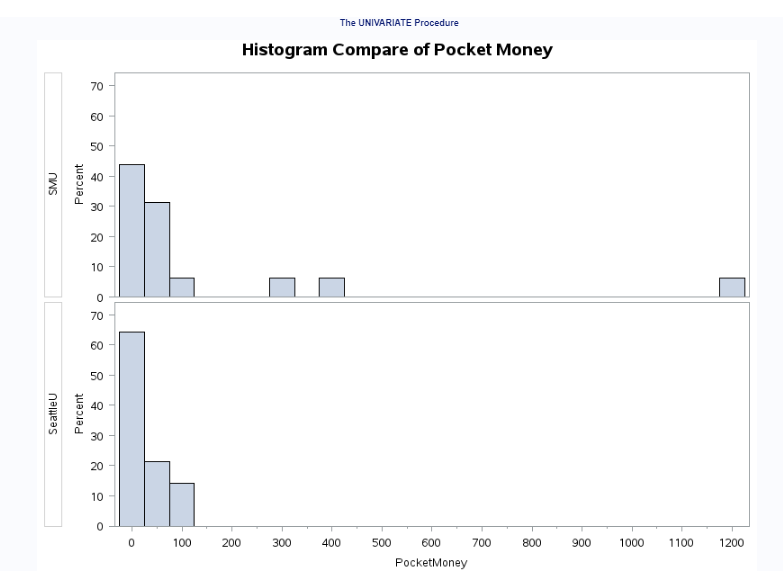
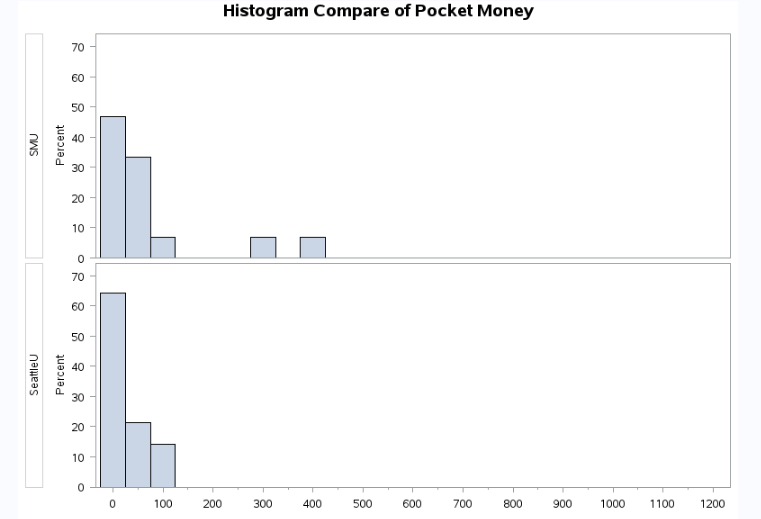


Source:

Ramsey, Fred; Schafer, Daniel. The Statistical Sleuth: A Course in Methods of Data Analysis (Page 69). Cengage Textbook. Kindle Edition.

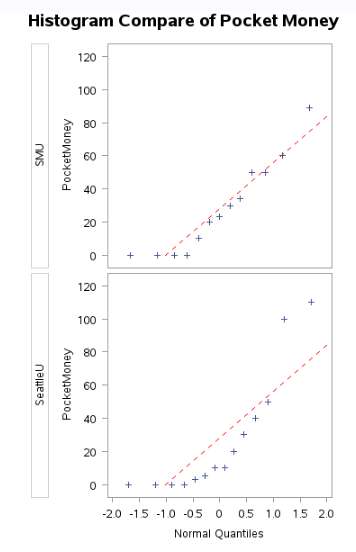
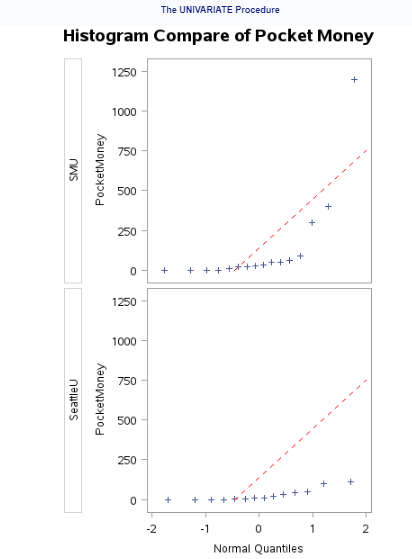
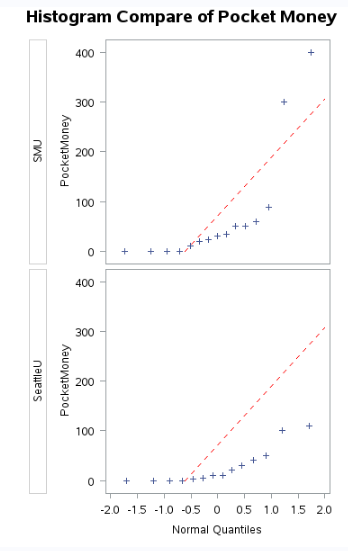
So, at this point we are not excluding outliers by a particular method like 2SDs exclusions, however, there are extreme values in the data set that I circled above for SMU in the qqplot. For starters, I’m going to exclude 1200 from SMU first to compare the two results.

Outlier Excluded: 1200 Outlier Not Excluded: 1200,400,300 Excluded:



When Looking at the three histograms, I guess you can argue that without the outlier(s), the histogram is closer to being a normal distribution, but it is still right skewed. When excluding the 3 outliers from SMU, I would argue that this histogram is more normally distributed and a t-test is now appropriate.

1200 Outlier Excluded: No Outliers Excluded SMU$: 1200,300,400 Excluded



You can tell that by removing the 3 outliers from SMU, that the line is better fitting the data on the QQPlot suggesting that the data is more normally distributed than before removing the outliers. SeattleU has two pretty high points that probably can be excluded as well, but I will leave them in for now.

Going to compare the p-values between the 3 sets.

No Outliers Excluded 1200 Outlier Excluded 1200,300,400 SMU Dollars Excluded

When excluding outliers, the p-value increases suggesting a higher percentage that the means will equal. However, all three results still fails to reject the null hypothesis with an alpha of .05. Given that the result is the same whether we exclude the outliers or not in regards to failing to reject H0, we should include the outliers in the analysis.

1. Find the “Education Data” data in the ***Files*** Folder. In it is a data set with annual incomes in 2005 of the subset of National Longitudinal Survey of youth (NLSY79) subjects who had paying jobs in 2005 and who had completed either 12 or 16 years of education by the time of their interview in 2006. All the subjects in this sample were between 41 and 49 years of age in 2006. Test the claim that the distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education. (Hint: pay careful attention to the ratio between the largest and smallest incomes in each group … also …. Is the bigger mean associated with the bigger standard deviation? … Transformation?) ***You may use SAS or R for this problem but be sure and include your code!***

Note: It is a very large dataset…. “datalines” is not a good idea here! OR … you could also use the Import Wizard depending on what source of SAS you are using.

Finally, make sure you present your findings as you would to a client:

So, we can set up the null hypothesis a couple of ways by doing either a one sided test or a two side test.

H0: uEduc12 = uEduc16

HA uEduc12 <> uEduc16

Or (one sided)

H0: Income @16 years <= Income @12 years

HA: Income @16 years > Income @12 years

Going with one sided t-test.

Code:

PROC IMPORT OUT= WORK.EducationData

DATAFILE= "/home/marinfamily1010/eddata1/EducationData.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

proc univariate

data = WORK.educationdata noprint;

histogram Income2005;

qqplot Income2005 / normal(mu=est sigma=est color = red l=2) square;

class Educ;

run;

proc sort data=work.transform;

by Educ descending Income2005;run;

proc ttest data = WORK.educationdata alpha= .05 h0=0 sides=U;

class Educ;

var Income2005;

run;

data test1;

alpha = .05;

p = 1-alpha; /\*/2;\*/

df = 1425;

CritVal = TINV(p, df);

proc print data = test1;

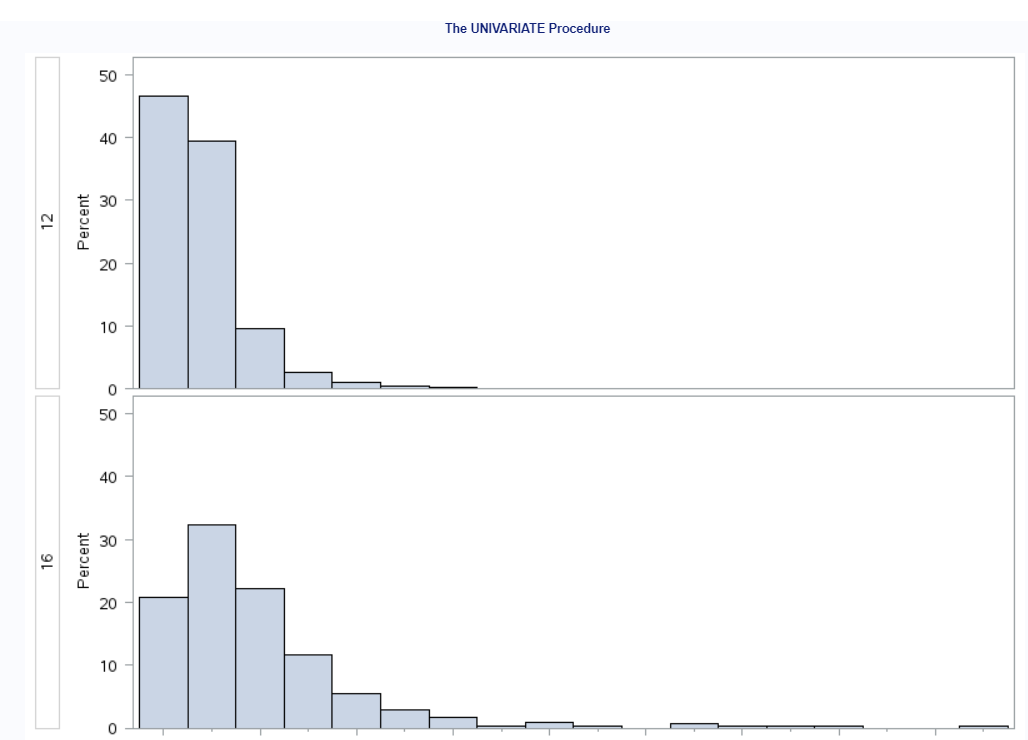
run;

1. State the Problem

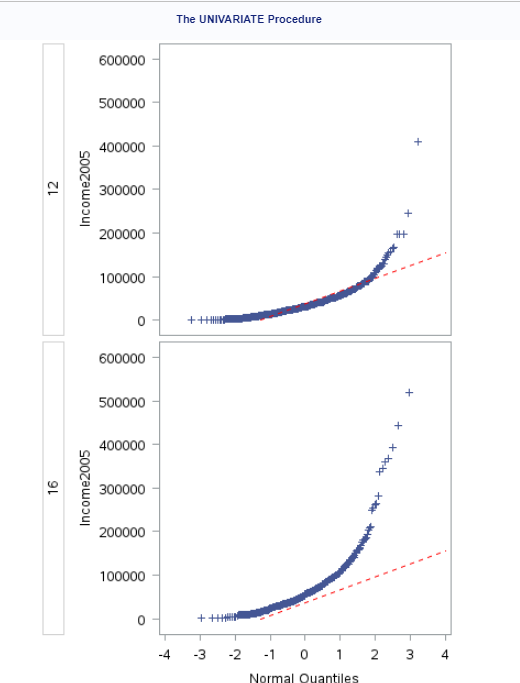
A survey was conducted for subjects who had 12 or 16 years of education. We would like to know if the income of those with 16 years of education exceed those with 12 years of education.

1. Address the Assumptions (Graphically and using words).

Normally Distributed Populations:

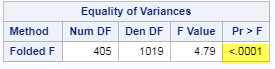


When comparing the two histograms, I can see that both have quite a bit of right skew and are not quite normal.



When looking at the qq plots, you can tell that the points don’t exactly resemble a straight line, but more of an exponential growth. By looking at the images alone, I would say a t-test is not appropriate.

Equal Standard Deviations:



According to the f-test of .0001 (highlighted above), there is sufficient evidence that the variance is different since .0001 is less than .05 which is my assumed alpha.

Independence:

We don’t know who surveyed the data and were not told that they were part of a group or what they did for a living, but filled out a particular survey. I’m going to assume that the members are independent of each other.

1. Perform the Most Appropriate (Powerful) Test (in reality this may be a pooled t-test on the original data, a t-test on the log transformed data, or a permutation test on the original data since these are the ones we have studied so far. For now, assume you must choose between the pooled t-test on the original data or on the log transformed data.)

Going to try to do a log transformation of the Income2005 variable.

data WORK.transform;

set WORK.EDUCATIONDATA;

log\_Income2005=log(Income2005);

run;

proc univariate

data = TRANSFORM noprint;

histogram log\_Income2005;

qqplot log\_Income2005 / normal(mu=est sigma=est color = red l=2) square;

class Educ;

run;

proc ttest data = TRANSFORM alpha= .05 h0=0 sides=U;

class Educ;

var log\_Income2005;

run;

data test1;

alpha = .05;

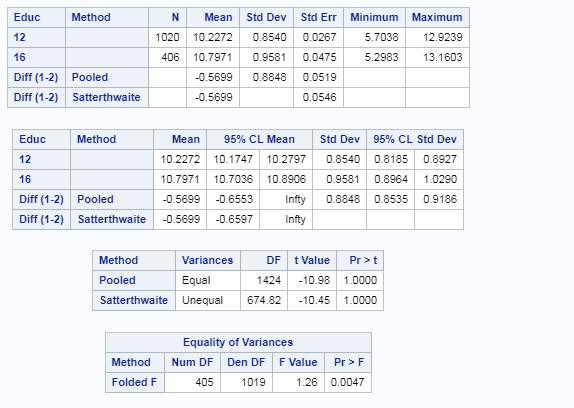
p = 1-alpha; /\*/2;\*/

df = 1425;

CritVal = TINV(p, df);

proc print data = test1;

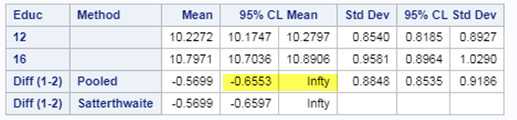
run;



The p-value is 1.00

1. Provide a conclusion including a p-value and a confidence interval.

There is sufficient evidence to support the claim with an alpha of .05 (p-value 1.00) that the mean income for those with 16 years of education are higher than those with 12. A 95% pooled confidence interval for this difference is (.519,infinity).



(Needed to the e power to the confidence interval, but upper is Infinity…

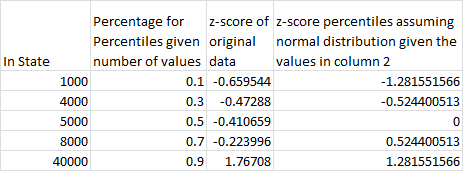
1. Provide a scope of inference.

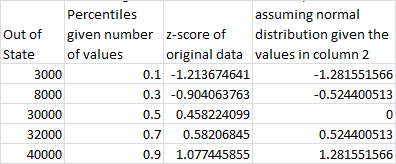
This is an observational study since we can randomize the income between the subjects and it appears that this survey was a voluntary survey so we might not have all incomes. Also, we can only infer the results to be true for those who participated in the survey with ages between 41-49.

Bonus (5 pts): Create 2 Q-Q plots (by hand) for the original data in question Chapter 3 number 20 of the text book. A Q-Q plot for the In-State and a Q-Q Plot for the Out-Of-State data. Show all work by filling in a table like the one below (one for In-State and one for Out-of-State):

|  |  |  |  |
| --- | --- | --- | --- |
| Original Data | Percentage for percentiles given number of values | Z-score of original data | Z-score percentiles assuming normal distribution given the values in column 2. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

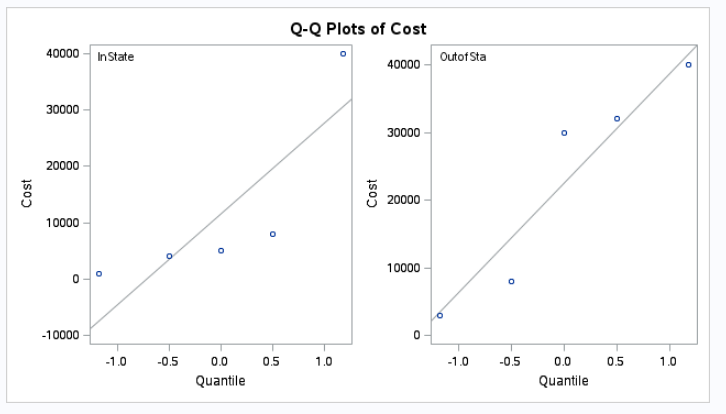
Check your Q-Q plots by comparing them with the ones from proc t-test. (Run proc t-test but just for the Q-Q plots. You do not need to run a full hypothesis test.) What would you conclude about the normality of the distributions these data came from?





In State:

Out of State:



Both qqplots (mine and SAS) show non-normal as two lines can be drawn for both instate and out of state. One difference between the two graphs is the negative values in my chart.