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?

***1. Independence: The problem states the samples are random; we will assume that the samples are also from the population. Therefore, we will assume the observations are independent.***

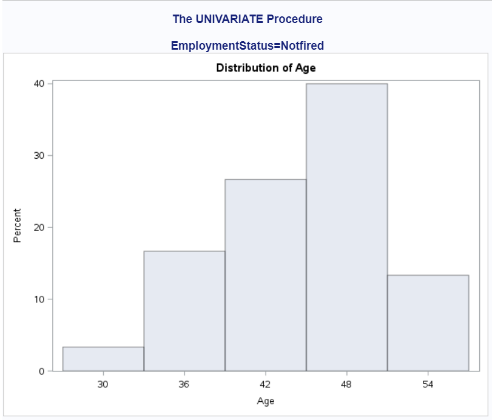
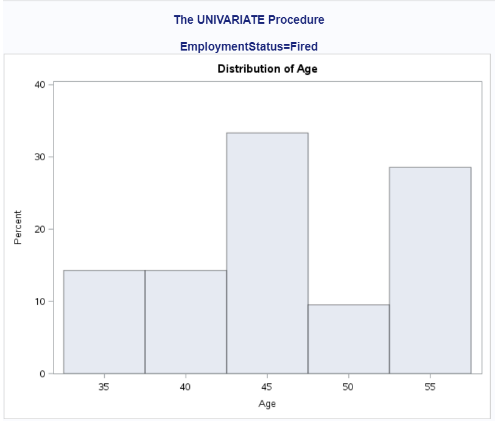
***2. Normality:***

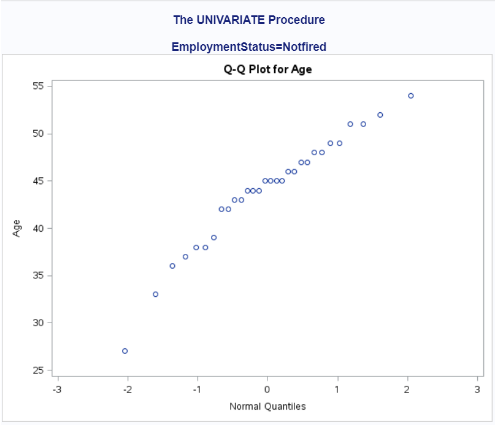
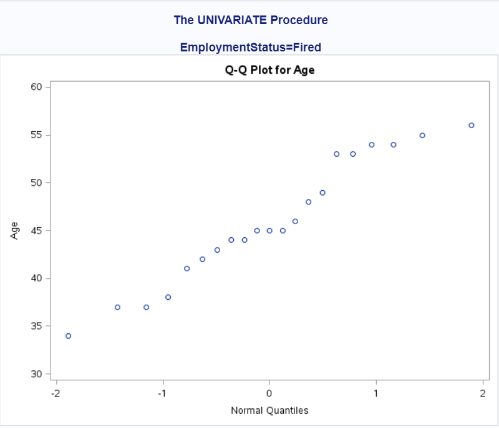
|  |  |
| --- | --- |
| *proc univariate data = SamoaEmployees;*  *by EmploymentStatus;*  *histogram;*  *qqplot Age;*  *run;* | *proc boxplot data=SamoaEmployees;*  *plot Age\*EmploymentStatus;*  *run;* |

***The following plots indicate that the distributions are mildly skewed; however, there are no extreme or unrealistic outliers.***

***The t-test is robust with regard to non-normal distributions as long as the sample size is sufficient (thanks to central limit theorem). Our sample sizes are 30 and 21. Therefore I would consider a t-test appropriate for this test but proceed with caution.***

******

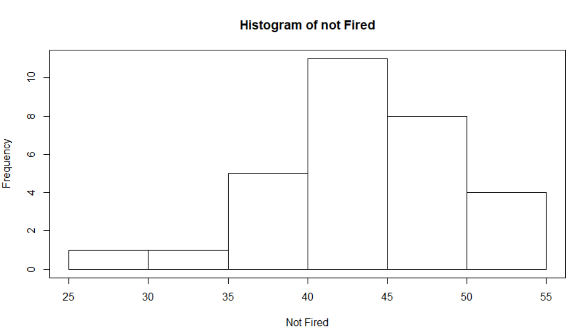
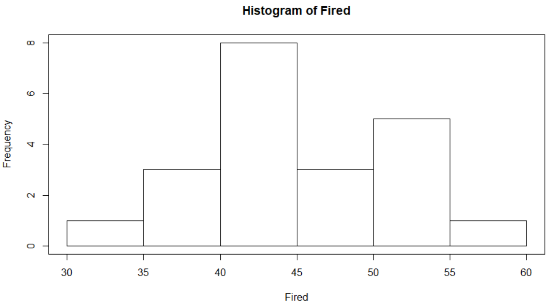
******

******

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

***Interestingly R is showing more detail than SAS (using default scales; I didn’t manipulate the plot scales in either tool) The histograms look a bit different but by looking at the scales carefully the assumptions are the same.***

|  |  |
| --- | --- |
| *fired <- c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)*  *hist(fired, xlab='Fired', main='Histogram of Fired')*  *box()* | *not.fired <- c(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)*  *hist(not.fired, xlab='Not Fired', main='Histogram of not Fired')*  *box()* |



* 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.)

***1. Problem: Test the claim that the Samoa government is discriminating by preferentially firing older workers.***

***2. Assumptions***

***Independence: The problem states the samples are random; we will assume that the samples are also from the population. Therefore, we will assume the observations are independent.***

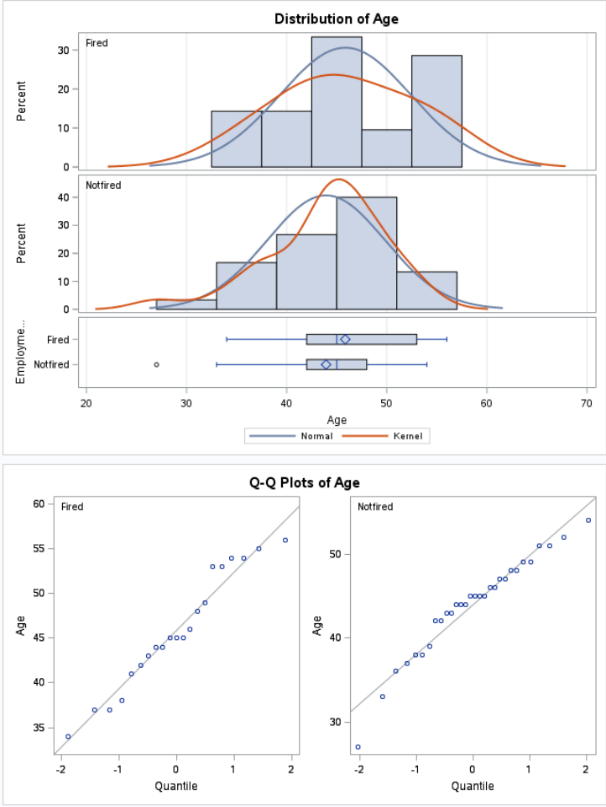
***Normality:***

|  |  |
| --- | --- |
| *proc univariate data = SamoaEmployees;*  *by EmploymentStatus;*  *histogram;*  *qqplot Age;*  *run;* | *proc boxplot data=SamoaEmployees;*  *plot Age\*EmploymentStatus;*  *run;* |

***The following plots indicate that the distributions mildly skewed; however, there no extreme or unrealistic outliers. The distribution of Fired is slightly non-normal.***

***The t-test is robust with regard to non-normal distributions as long as the sample size is sufficient (thanks to central limit theorem). Our sample sizes are 30 and 21 and the shapes are similar indicating that the variance is roughly equal.***

***I would consider a t-test appropriate for this test and proceed with caution.***

******

***3. H0: Ha:***

***4. Critical Value a = .05, df = 49,***

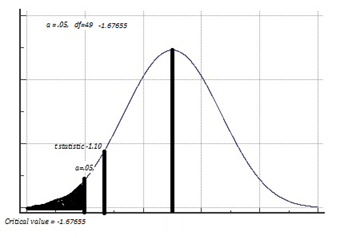
*data critval;*

*cv = quantile("T", .05, 49);*

*alpha = .05;*

*proc print data = critval;*

*run;*

******

***5. t statistic = -1.10***

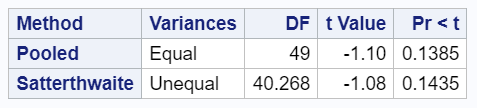
***6. p value = .1385***

***proc ttest data=SamoaEmployees sides=u;***

***class EmploymentStatus;***

***var Age;***

***run;***



***7. Because the t statistic is not in the critical value range and p value = .1385 we fail to reject .***

***8. Based on this test there is not enough evidence to suggest that the mean ages of the fired and not fired groups are different.***

***There is not enough evidence to suggest that there is discrimination based on age. (p-value=0.1385) from a one sided, two-sample t-test.***

***A 95% confidence interval for the difference in means is (-1.0780, Infinity) in years. Since we assume the subjects in this sample were randomly sampled, inference can be generalized to the population of all employees in the American Samoa Government.***

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?

***Assumptions***

***Independence***

***The problem states the samples are from 2 different geographic locations. There could be confounding variables (economics? crime rates?)***

***There is no information about whether the samples were random or the populations of each class and whether cash was shown or if the students volunteered verbally.***

***We will treat this as an observational study.***

***Normality:***

***Our sample sizes are 16 and 14, we have extreme outliers and significantly different standard deviations. SMU sd=304.3, Seattle sd=36.7193***

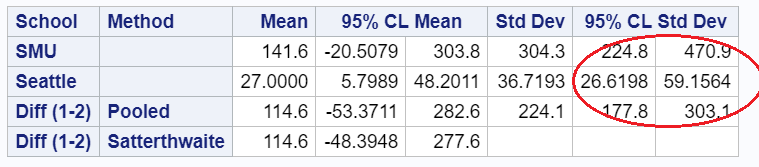
***The following plots indicate that the distributions are skewed, have extreme and possibly erroneous outliers.***

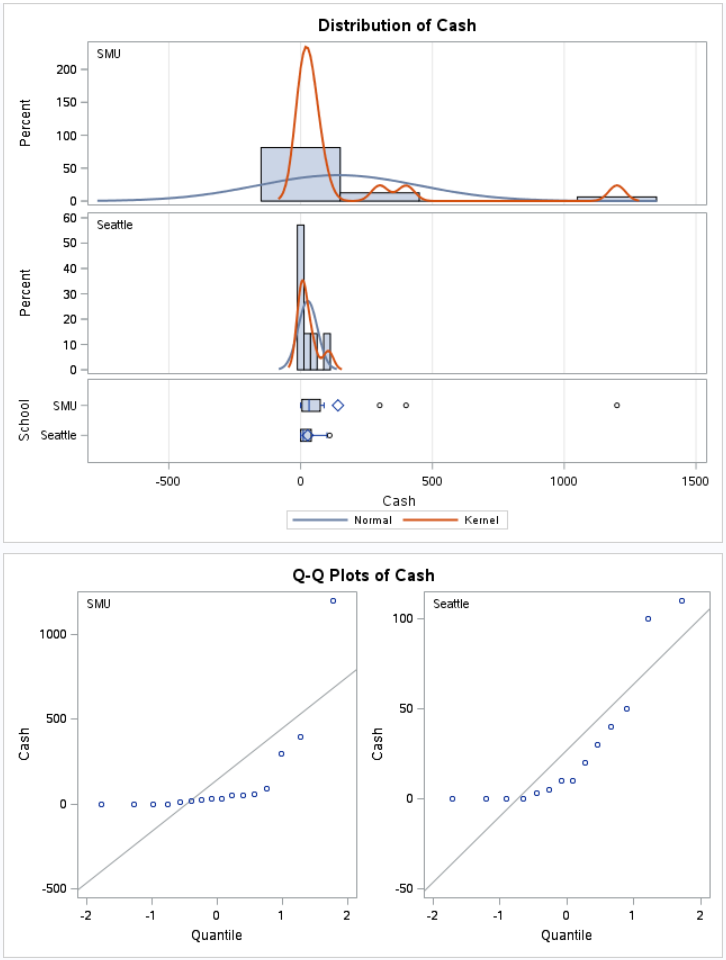
*proc ttest data=StudentCash sides=2;*

*class School;*

*var Cash;*

*run;*

******

******

***I would consider a t-test inappropriate for this test.***

* + - * 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.)

**1. Problem: Test to see if there is 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.**

**2. Assumptions:**

***Independence:***

***The problem states the samples are from 2 different geographic locations. There could be confounding variables (economics? crime rates?)***

***There is no information about whether the samples were random or the full populations of each class and whether cash was shown or if the students volunteered verbally.***

***We will treat this as an observational study.***

***Normality:***

***Our sample sizes are 16 and 14, we have extreme outliers and significantly different standard deviations. SMU sd=304.3, Seattle sd=36.7193. (See my plots in part a)***

***Though a t-test is robust with normal distributions given large enough sample size; I would consider a t-test inappropriate for this test due to skewedness and sample standard deviations.***

***I would prefer to use logarithm or provide results for both with and without outliers; however, I will use permutation as directed in the problem instructions.***

***3. Hypotheses***

***4. Critical Value = 1.64***

data critval;

cv = quantile("T", .95, 998);

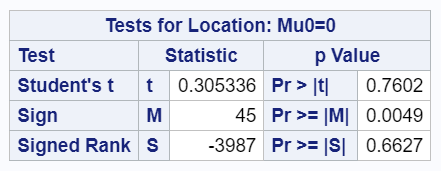
alpha = .05;

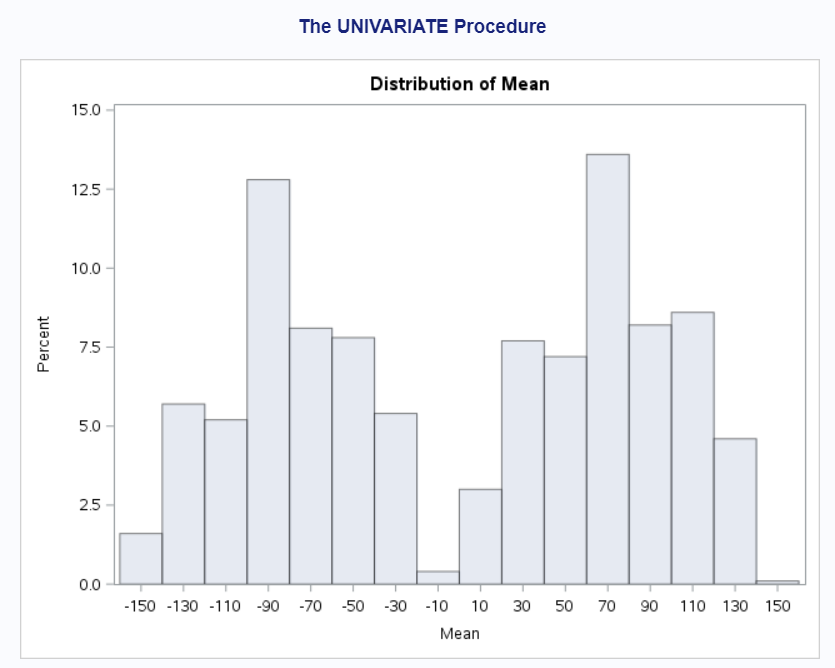
proc print data = critval;

run;

| **Obs** | **cv** | **alpha** |
| --- | --- | --- |
| **1** | 1.64638 | 0.05 |

***5. Test Statistic t = .305336***





***6. p-value = .7602***

***7. Fail to Reject .***

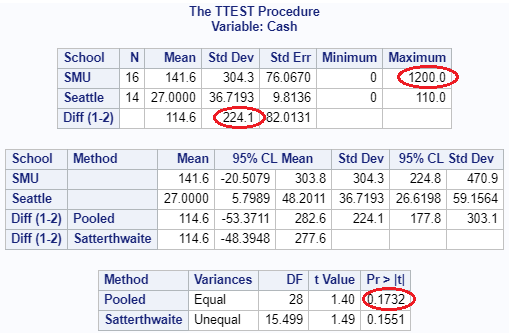
***8. Based on this test there is not enough evidence to conclude that there is a difference between the cash carried by SMU and Seattle students with 95% confidence interval -53.3711, 282.6. (p .7602).***

***Causal inference cannot be drawn about the populations of students as SMU vs Seattle U because this data is only from stats students and the data from Seattle is a year older. In addition, there are extreme outliers in the SMU data which are suspicious.***

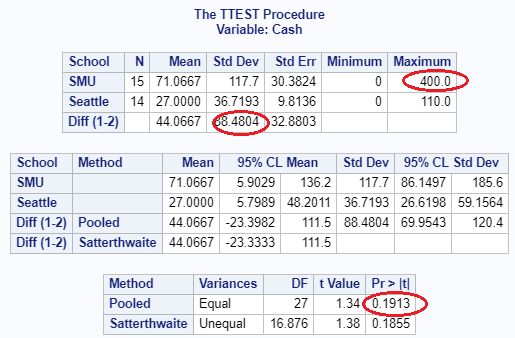
***Since it cannot be determined if the subjects in this sample were randomly sampled from a population the results only apply to the subjects in the study.***

* + - * 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.

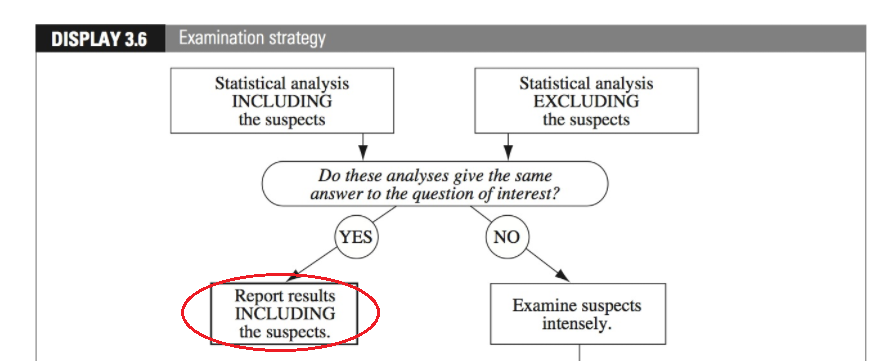
**With outlier**

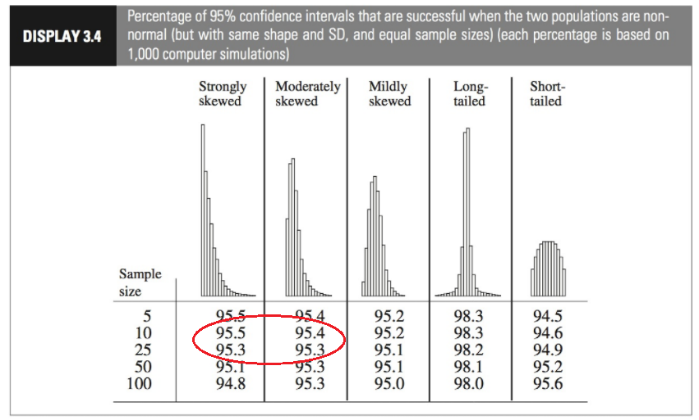


**Outlier SMU 1200 removed**



**Because the conclusion is the same this enforces my confidence in t-tools. I would report the results including the suspects.**





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:

1. State the Problem

***Test the claim that the distribution of incomes for those with 12 years of education is less than is less than the distribution for those with 16 years of education.***

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

***Independence:***

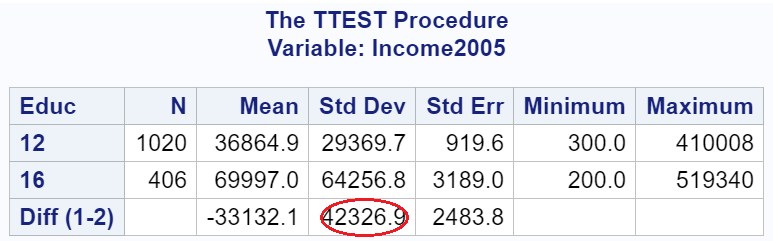
***The problem states the samples are from interviews. It is likely that the subjects volunteered therefore we have confounding variables.***

***Though it does say this is a subset; there is not enough information about how the samples were gathered to consider them random.***

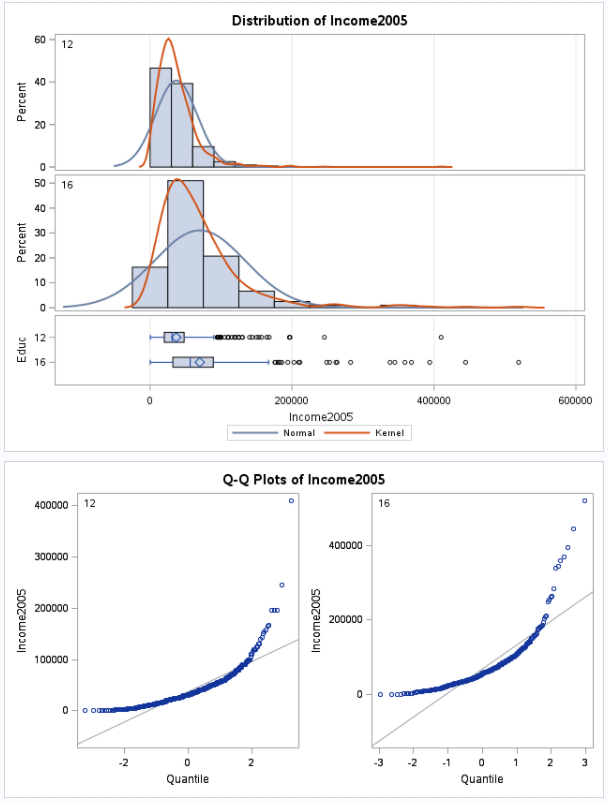
***For purposes of this homework we will treat this as an observational study.***

***Normality:***

***Our sample sizes are 1020, 406 and there is significant difference in sample standard deviations = 42326.9***

******

***Our histograms have roughly the same shape but QQ plots are showing significant curves***

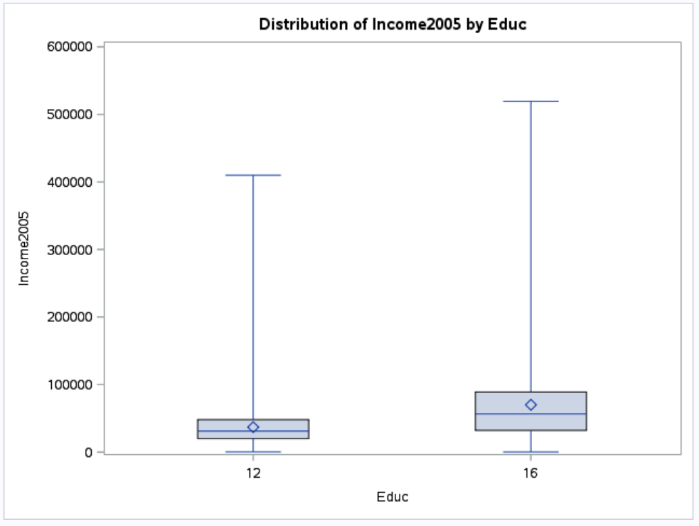
**

*proc boxplot data=EducationData;*

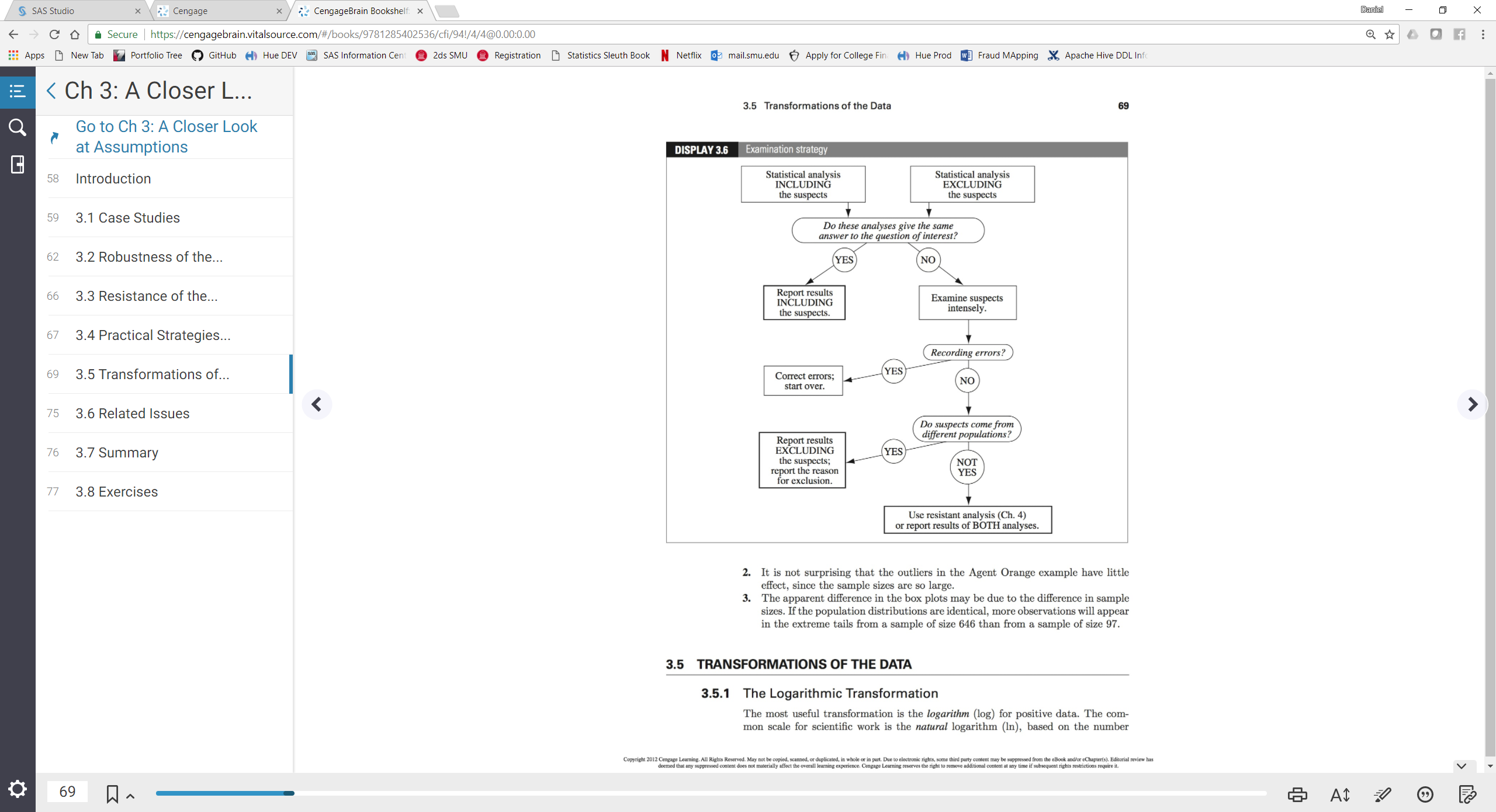
*plot Income2005\*Educ;*

*run;*

***We have extreme outliers***

**

***Given the skewedness and sample standard deviations I would consider a t-test inappropriate for this test.***



***We will use log transformation to lessen the skew and spread, then t-test the transformed data.***

*proc ttest data = EducationData sides=2;*

*class Educ;*

*var Income2005;*

*run;*

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.)

***Hypotheses***

***Critical Value = 1.64, df = 1424***

data critval;

cv = quantile("T", .95, 1424);

alpha = .05;

proc print data = critval;

run;



***Test Statistic t = -10.45 (Using Satterthwaite because of uncertainty about population standard deviations)***

*data lEducationData;*

*set EducationData;*

*loggedIncome = log(Income2005);*

*proc Sort data=lEducationData;*

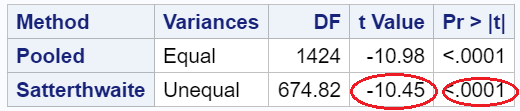
*by Educ;*

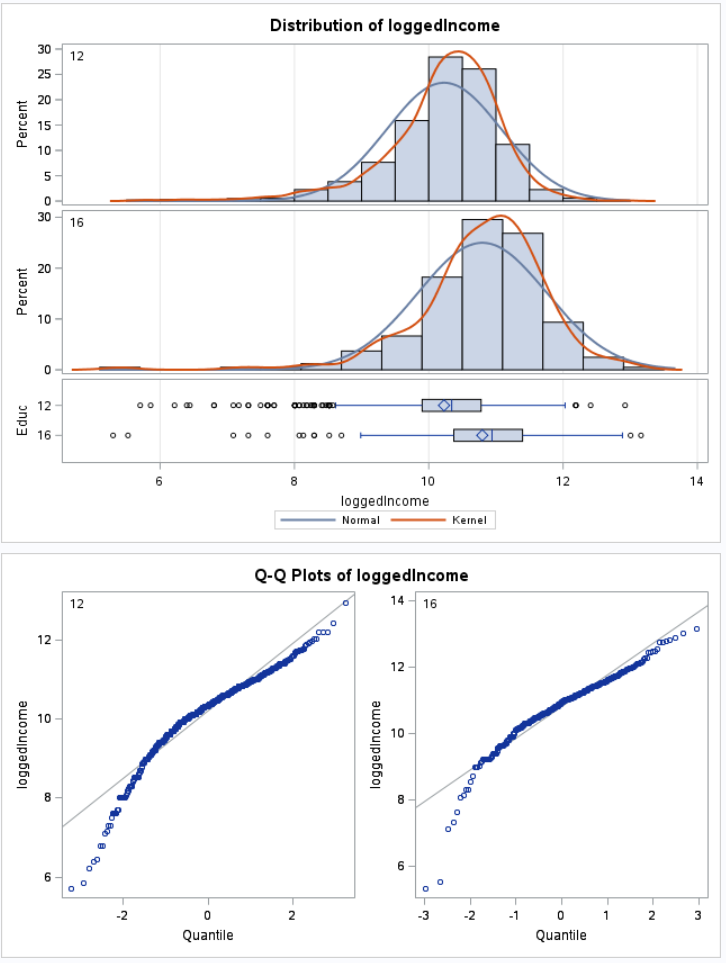
*proc ttest data= lEducationData sides=2;*

*class Educ;*

*var loggedIncome;*

*run;*

**

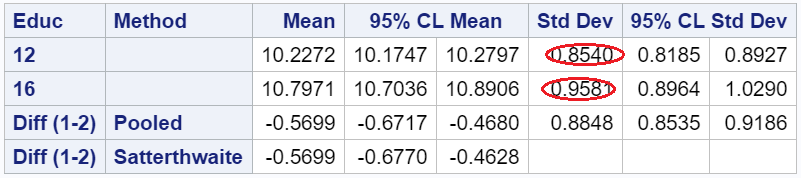


1. Provide a conclusion including a p-value and a confidence interval.

***Reject p-value = .0001 with 95% Confidence interval - .67 -.46***

***We made the distributions more normal and sample standard deviations closer to a 1:1 ratio by using log. This made the p value much smaller.***

***Original S2/S1=2.18, Transformed S2/S1=1.12***



1. Provide a scope of inference.

***It is estimated that the median income of subjects in this study with 12 years of education is .5699 times less than the median income of subjects with 16 years of education. p-value = .0001***

***We cannot make any causal inference and must limit our conclusions to differences of the groups in this study.***

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?

***I would definitely have done this but ran out of time.***