Unit 3 HW

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:

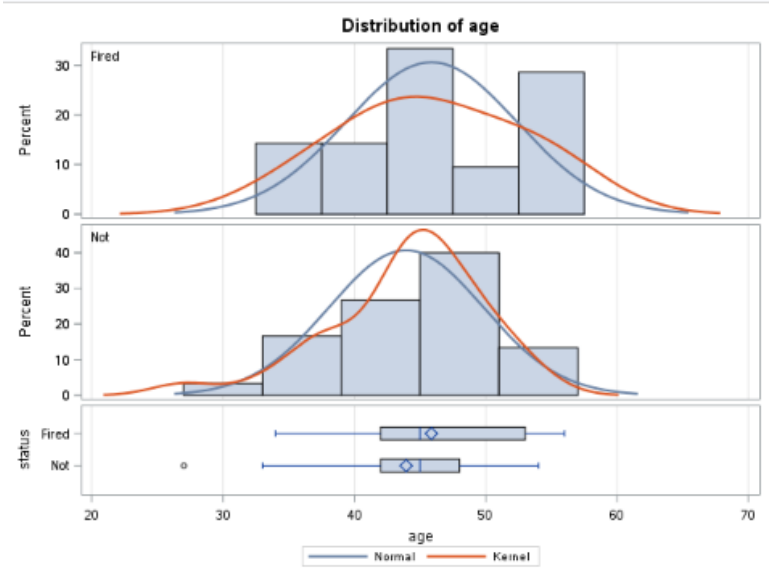
Fired

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

a. 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, q-q plots or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?



Normality: Judging from the histogram and QQ plots, there is little to no evidence that the population distribution of the fired and not fired workers are not normal. We will assume that this distribution is normal and proceed.

Independence: These subjects were randomly selected from the population thus we will assume that the observations are independent.

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

##Create data set

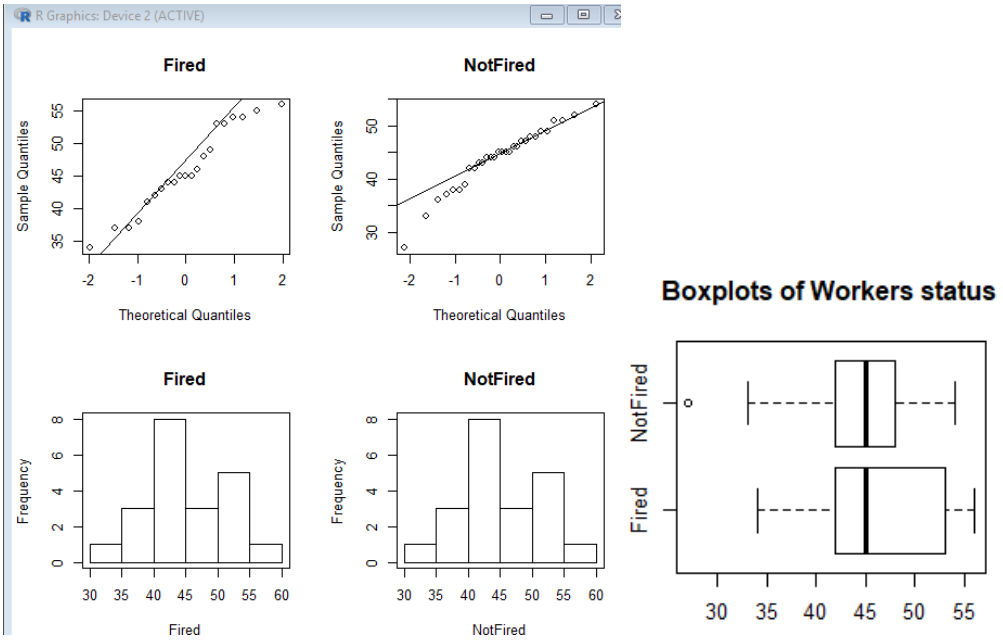
```
fired <- rep('Fired', 21)
not.fired <- rep('NotFired', 30)
fired2 <- c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)
not.fired2 <- c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)
fired <- data.frame(fired, fired2)
not.fired <- data.frame(fired=not.fired, fired2=not.fired2)
status.final <- rbind(fired, not.fired)
```

##Make QQ-plots and Histograms

```
par(mfrow=c(2,2))

qqnorm(status.final$fired2[status.final$fired=="Fired"], main='Fired')
qqline(status.final$fired2[status.final$fired=="Fired"])

qqnorm(status.final$fired2[status.final$fired=="NotFired"], main='NotFired')
qqline(status.final$fired2[status.final$fired=="NotFired"])

hist(status.final$fired2[status.final$fired=="Fired"], xlab='Fired', main='Fired')
box()
hist(status.final$fired2[status.final$fired=="Fired"], xlab='NotFired', main='NotFired')
box()
```

##Side-by-Side Boxplots

```
boxplot(fired2 ~ fired, data=status.final, horizontal=T,
main='Boxplots of Workers status')
```



c. 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.

We would like to test the claim that the mean of the Fired group(H_o) is different than the mean of the NotFired group(H_a).

$$H_o: \mu_{Fired} = \mu_{NotFired}$$
$$H_a: \mu_{Fired} \neq \mu_{NotFired}$$

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

Normally Distributed Populations & equal standard deviations
Visual inspection of the histograms and QQ-plots of each of the populations are consistent with the normality of each population. We assume normality.

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

##
Two Sample t-test
##
data: Fired and Not_fired
t = 1.0991, df = 49, p-value = 0.2771
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.593635 5.441254
##sample estimates:
mean of x mean of y

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

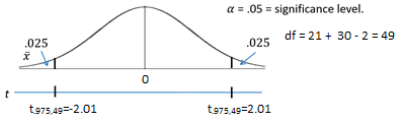
On the basis of this test, there is not enough evidence to suggest that the mean ages of the fired and not fired groups are different. In other words, there is not enough evidence to suggest that there is discrimination based on age ($p = 0.2771$ from a two-sample t-test). A 95% confidence interval for this difference is $[-1.60, 5.44]$ years.

5. Provide the scope of inference.

Step 1 - Hypotheses:

$$H_o: \mu_{Fired} = \mu_{NotFired}$$
$$H_a: \mu_{Fired} \neq \mu_{NotFired}$$

Step 2 - Identification of Critical Value using alpha = 0.05 and 30+21-2=49 degrees of freedom: ± 2.01 **(2-sided) or** 1.677 **(1-sided)**



Step 3 - Value of Test Statistic: $t = 1.10$
Step 4 - Give p-value: $p = 0.2771$ **(2-sided)**
Step 5 - Decision: Fail to Reject H_o
Step 6 – Conclusion: On the basis of this test, there is not enough evidence to suggest that the mean ages of the fired and not fired groups are different. In other words, there is not enough evidence to suggest that there is discrimination based on age ($p = 0.2771$ from a two-sample t-test, $p = 0.1385$ from a one-sample t-test). A 95% confidence interval for this difference is $[-1.60, 5.44]$ years. Since the subjects in this sample were randomly sampled, inference can be generalized to the population of all employees in the American Samoa Government.

2. In the last homework, it was mentioned that a Business Stats class here at SMU was polled and students were asked 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 they 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

   a. 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, q-q plots, or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

The TTEST Procedure

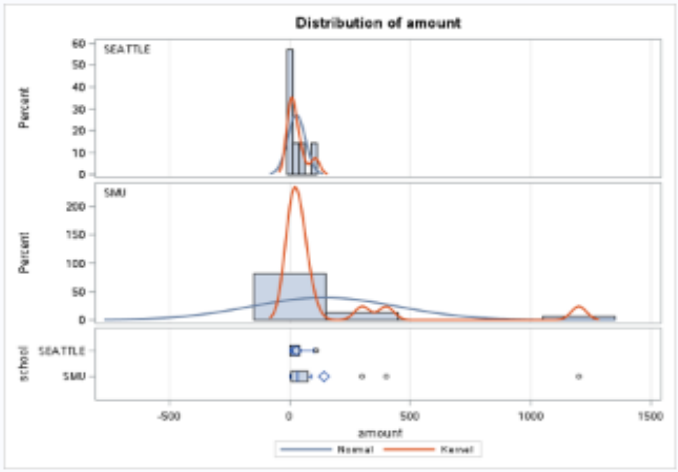
Variable: amount

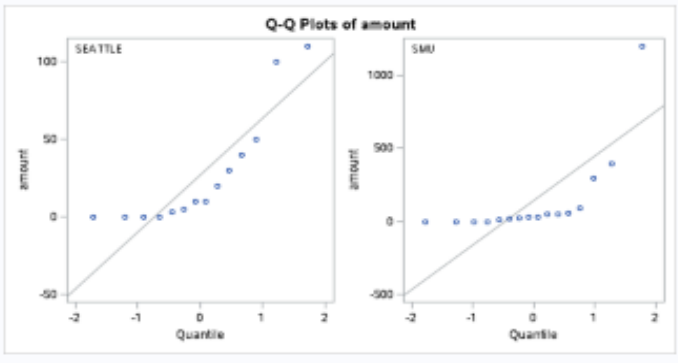
school	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
SEATTLE		14	27.0000	36.7193	9.8136	0	110.0
SMU		16	141.6	304.3	76.0870	0	1200.0
Diff (1-2)	Pooled		-114.6	224.1	82.0131		
Diff (1-2)	Satterthwaite		-114.6		76.6974		

school	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
SEATTLE		27.0000	5.7989	48.2011	36.7193	26.6198	59.1564
SMU		141.6	-20.5079	303.8	304.3	224.8	470.9
Diff (1-2)	Pooled	-114.6	-282.8	53.3711	224.1	177.8	303.1
Diff (1-2)	Satterthwaite	-114.6	-277.8	48.3848			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-1.40	0.1732
Satterthwaite	Unequal	15.499	-1.49	0.1551

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	15	13	68.66	<.0001





b. 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 the t-test (from part a)

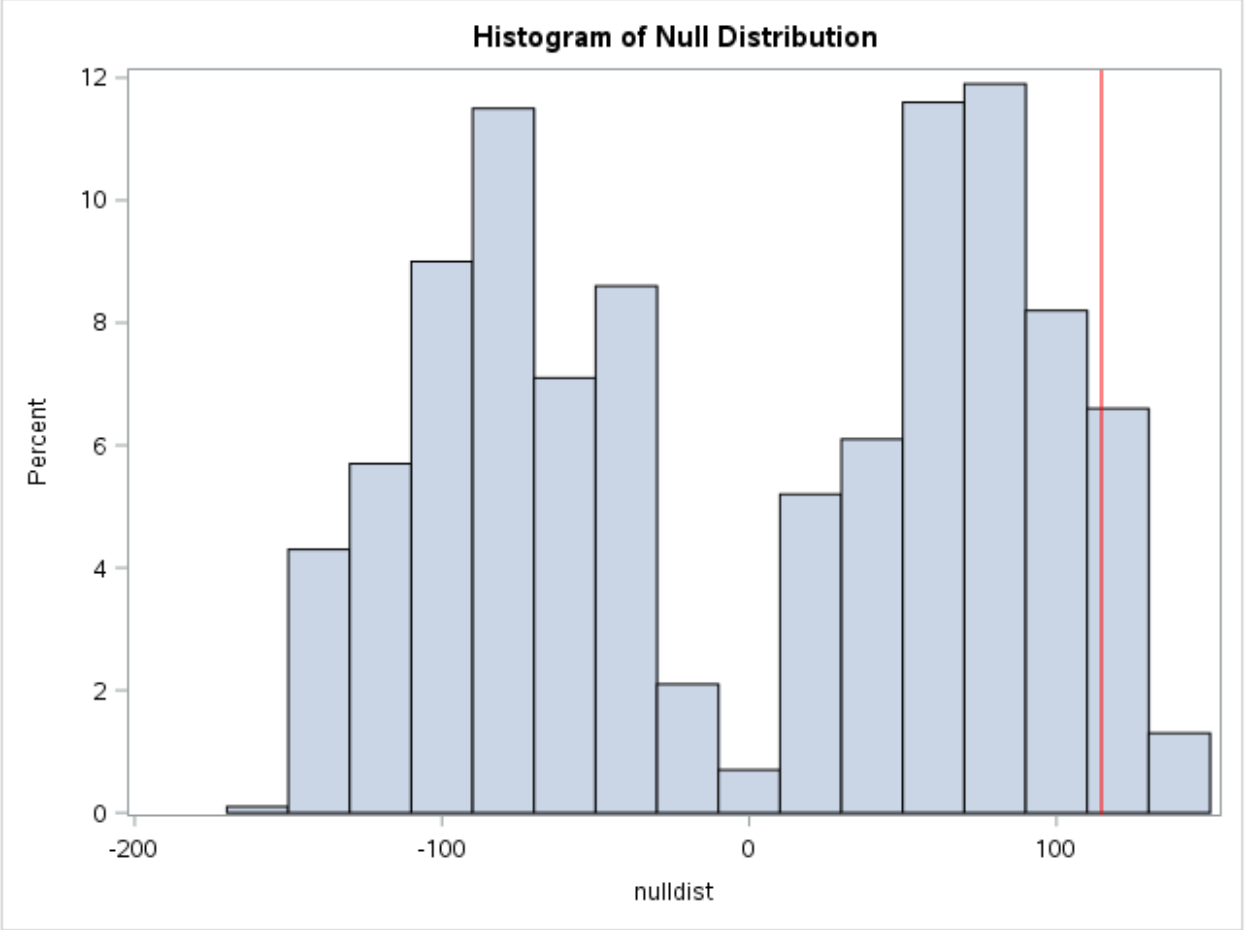
Normality of Distributions: Judging from the histograms and QQ plots, there is evidence of outliers in both the SMU and SEATTLE sets. The most pronounced outlier seems to be in the SMU data set thus there is significant visual evidence against these data being normally distributed. In addition, we are not satisfied that the t-test will be robust to this assumption since the sample sizes are so small.

Equal Standard Deviations: Judging from the histograms and box plots, there is significant visual evidence that the standard deviations are different. .

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

The two sample t-test is not appropriate here. Use permutation test.

obsdiff
114.625



Histogram of Null Distribution

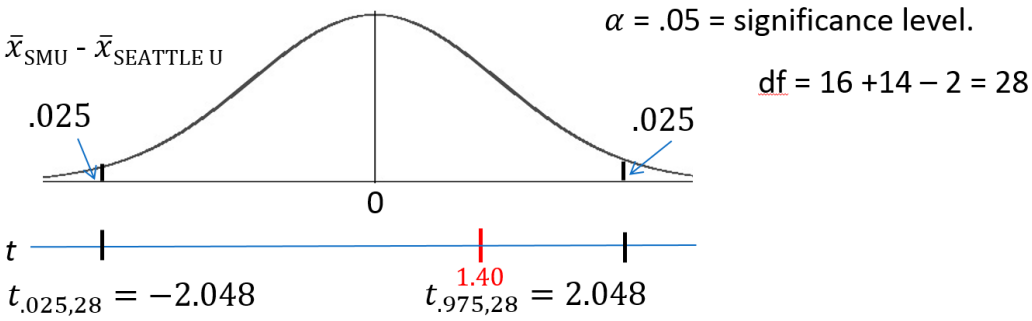
pval
0.1538462

4. Provide a conclusion, including the p-value and a confidence interval.
Based on this test, there is not enough evidence to suggest that the mean amount of pocket cash of the SMU students is different than that of the students from Seattle U ($p = 0.1732$ from a two-sided t-test). A 95% confidence interval for this difference is [-$53, $282].

5. Provide the scope of inference.
Step 1 - Hypotheses:

$$H_o: \mu_{SMU} = \mu_{SeattleU}$$
$$H_a: \mu_{SMU} \neq \mu_{SeattleU}$$

Step 2 - Identification of Critical Value (1 point for drawing, 1 point for value): ± 2.048



Step 3 - Value of Test Statistic: $t = 1.40$
Step 4 - Give p-value : $p = 0.1372$
Step 5 - Decision: Fail to Reject H_o
Step 6 - Conclusion: Based on this test, there is not enough evidence to suggest that the mean amount of pocket cash of the SMU students is different than that of the students from Seattle U ($p = 0.1732$ from a two-sided t-test). A 95% confidence interval for this difference is [-$53, $282]. Since the subjects in this sample were not randomly sampled, the results only generalize to the subjects in the study (no need to discuss causal conclusions for a non-significant result).

c. Note the potential outlier in the SMU data set. Re-check 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.
No my answer stays the same, The p-value without the outlier is 0.1913 and with the outlier 0.1732, so I would keep the outlier in the report results since it had little effect.

3. Find the "Education Data" data in the course materials. This data set includes 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: There is some SAS code in the course materials to help you download the data into SAS. It is a very large dataset… "datalines" is not a good idea here! You could also use the File/Import option.

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

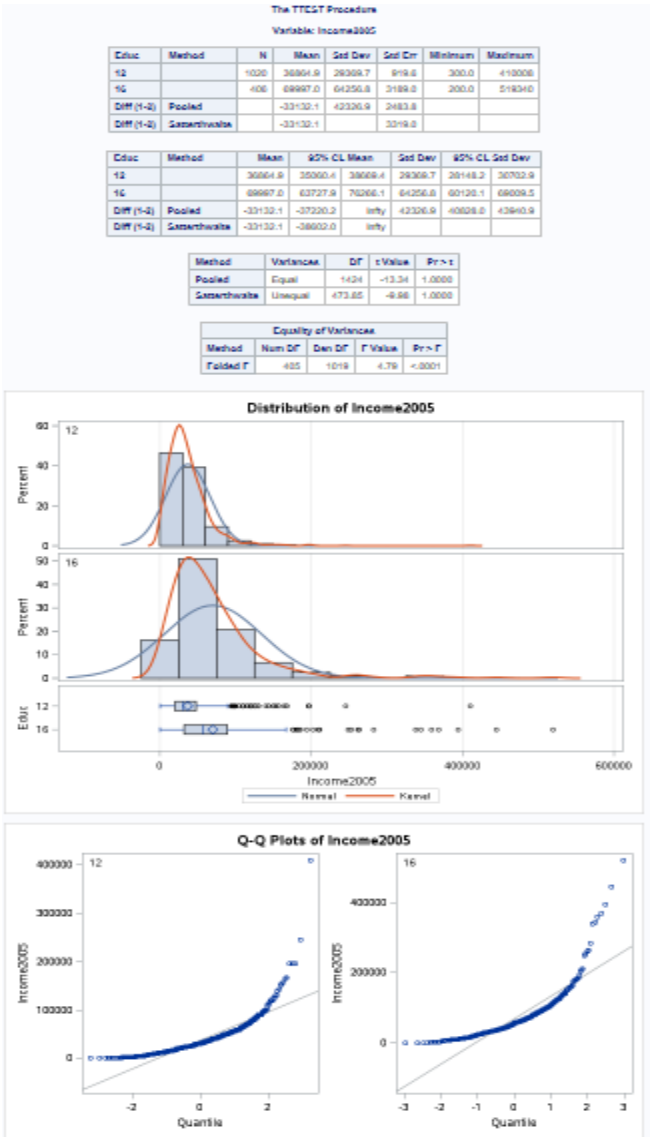
1. State the Problem.

We would like to test the claim that the mean of the 16 years of education group($\mu_{1)}$ is more than the mean of the 12 years of education group($\mu_{2)}$.

$H_{0:}\ \mu_{1=}\ \mu_{2}\ H_{a:}\ \mu_{1} > \mu_{2}$

2. Address the Assumptions (graphically and using words).

```
proc import datafile = '/home/chec0/New Folder/ex0330.csv'
 out = salary
 dbms = CSV
 ;
run;

proc ttest data= salary sides = upper
alpha = .05;
class Educ;
var Income2005;
run;
```



The TTEST Procedure

Variable: Income2005

Educ	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
12		1020	36864.9	29309.7	918.0	300.0	419000
16		406	69997.0	64256.8	3189.0	200.0	519340
Diff (1-2)	Pooled		-33132.1	42326.9	2483.8		
Diff (1-2)	Satterthwaite		-33132.1		3019.0		

Educ	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12		36864.9	35060.4	38669.4	29309.7	28148.2	30702.9
16		69997.0	63727.9	76266.1	64256.8	60128.1	69609.5
Diff (1-2)	Pooled	-33132.1	-37220.2	Infty	42326.9	40628.0	43940.9
Diff (1-2)	Satterthwaite	-33132.1	-38602.0	Infty			

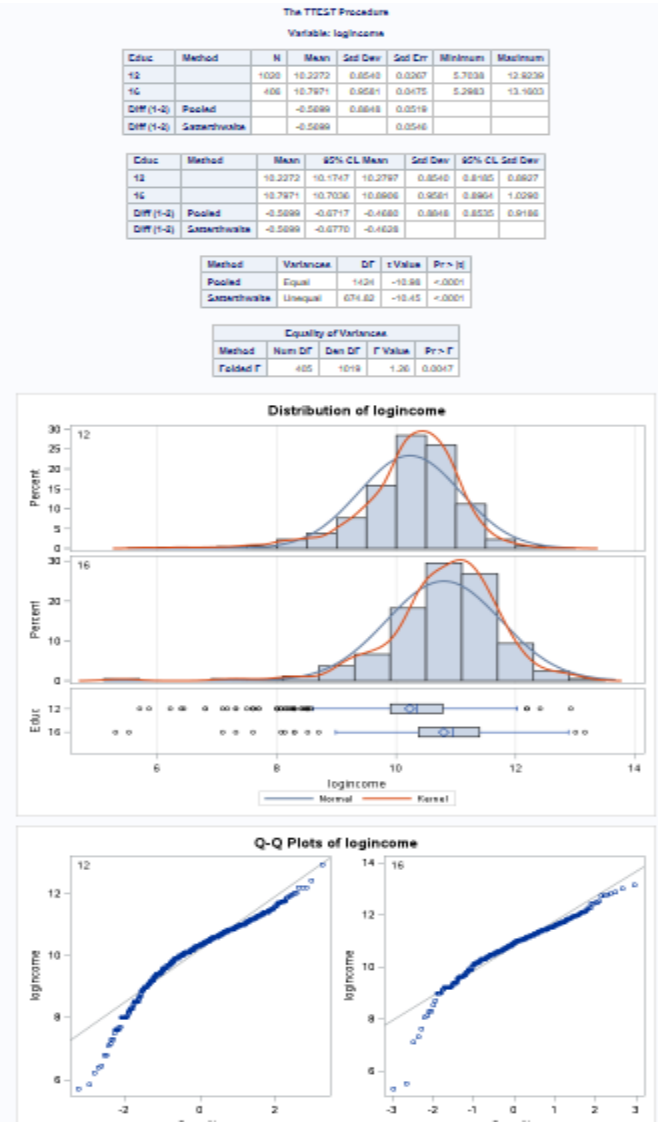
Method	Variances	DF	t Value	Pr > t
Pooled	Equal	1424	-13.34	1.0000
Satterthwaite	Unequal	473.65	-8.98	1.0000

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	405	1019	4.79	<.0001

Normality of Distributions: Judging from the histograms and QQ plots, there is significant visual evidence to suggest the data come from right-skewed distributions.
Equal Standard Deviations: Judging from the histograms and box plots, there is significant visual evidence that the standard deviations are different.

3. 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.)
Ran a log transformation on the data sets.
proc ttest data= salary sides = 2
alpha = .05;
class Educ;
var logincome;
run;



4. Provide a conclusion including a p-value and a confidence interval.
($p < .0001$)

5. Provide a scope of inference.

Bonus (5 pts): Create two q-q plots (by hand) for the original data in Chapter 3, question 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 ttest. (Run proc ttest 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?