# 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**

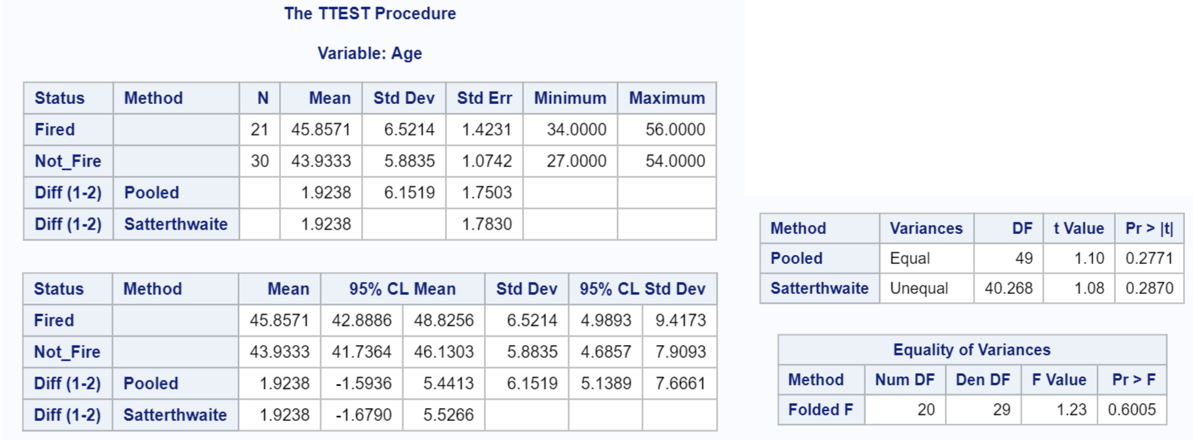
34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56

**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?

The evidence suggests there is insufficient evidence to indicate the not_fired sample is not normally distributed. It also suggests the fired sample is not normally distributed. However, it seems the fired sample has some outliers, which may be skewing the data to the left. The boxplot identifies an outlier in the fired sample. Considering the fired sample has a small size ($n_1 = 21$), the sample is not robust. In addition, the histograms show the samples do not have equal standard deviations. The fired sample size has larger standard deviations. Furthermore, the different sample sizes ($n_1 = 21$, $n_2 = 30$) indicate this set is not robust to the t-test based on the difference in sample means. While the true independence of the sample is unknown, it is known that the samples come from the same company. However, we don't know if they were all hired/fired at the same time, tenure, etc. Based on the information presented, a t-test is not an appropriate test for this sample.



The TTEST Procedure

Variable: Age

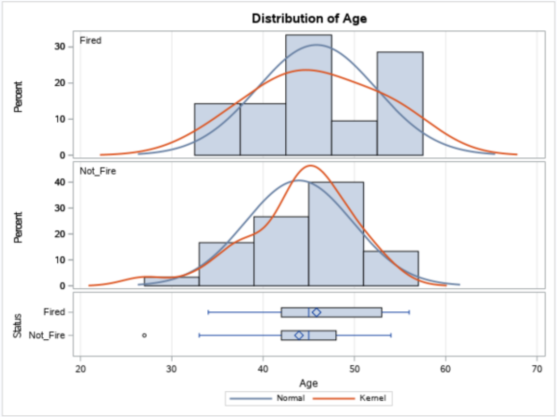
| Status | Method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| Fired | | 21 | 45.8571 | 6.5214 | 1.4231 | 34.0000 | 56.0000 |
| Not_Fire | | 30 | 43.9333 | 5.8835 | 1.0742 | 27.0000 | 54.0000 |
| Diff (1-2) | Pooled | | 1.9238 | 6.1519 | 1.7503 | | |
| Diff (1-2) | Satterthwaite | | 1.9238 | | 1.7830 | | |

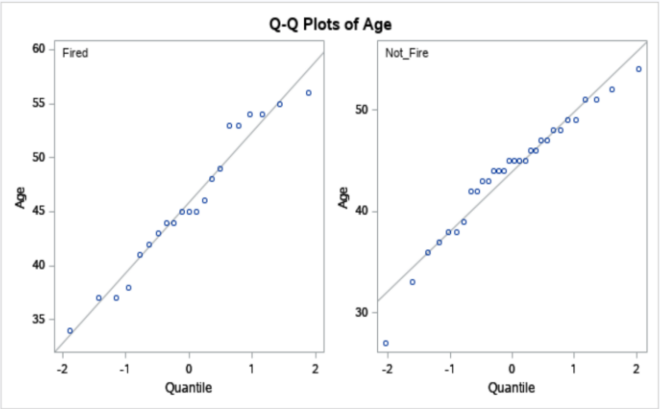
| Status | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|---|---|
| Fired | | 45.8571 | 42.8886 | 48.8256 | 6.5214 | 4.9893 | 9.4173 |
| Not_Fire | | 43.9333 | 41.7364 | 46.1303 | 5.8835 | 4.6857 | 7.9093 |
| Diff (1-2) | Pooled | 1.9238 | -1.5936 | 5.4413 | 6.1519 | 5.1389 | 7.6661 |
| Diff (1-2) | Satterthwaite | 1.9238 | -1.6790 | 5.5266 | | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | 49 | 1.10 | 0.2771 |
| Satterthwaite | Unequal | 40.268 | 1.08 | 0.2870 |

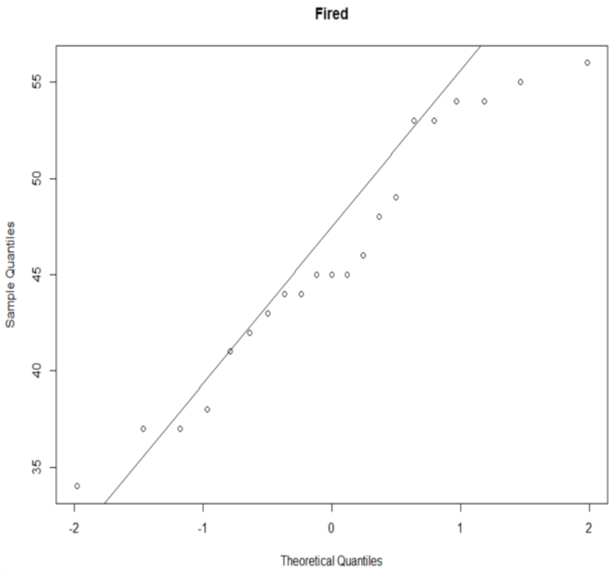
Equality of Variances

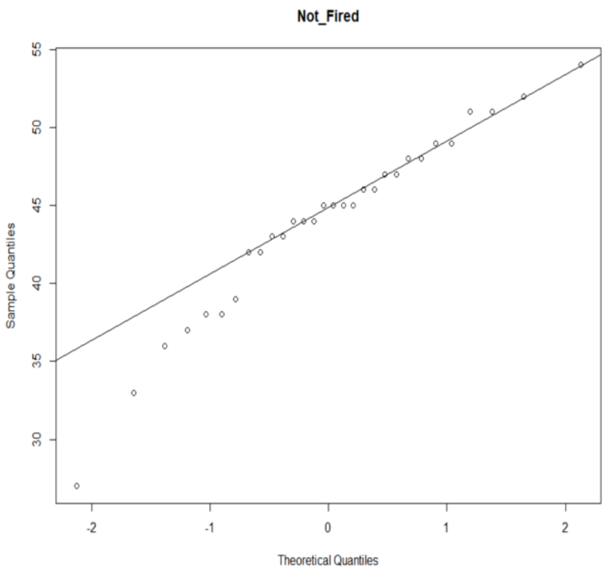
| Method | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| Folded F | 20 | 29 | 1.23 | 0.6005 |

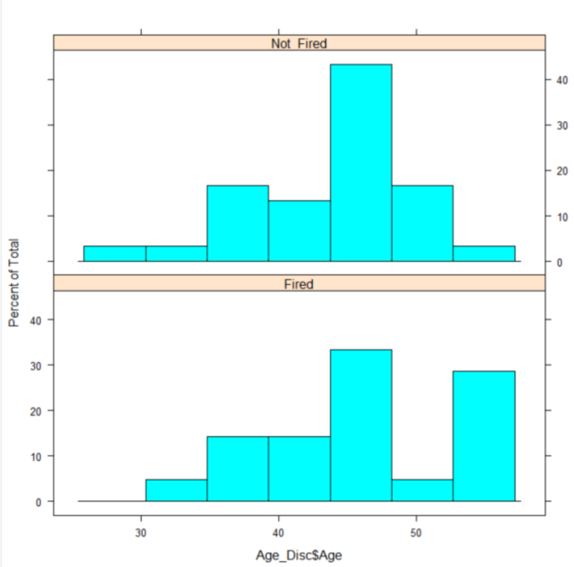


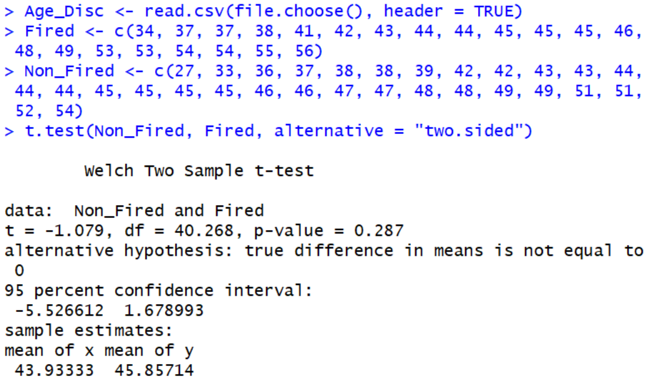


b. Check the assumptions with R and compare them with the plots from SAS.









```
> Age_Disc <- read.csv(file.choose(), header = TRUE)
> Fired <- c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46,
 48, 49, 53, 53, 54, 54, 55, 56)
> Non_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)
> t.test(Non_Fired, Fired, alternative = "two.sided")

        Welch Two Sample t-test

data:  Non_Fired and Fired
t = -1.079, df = 40.268, p-value = 0.287
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 -5.526612  1.678993
sample estimates:
mean of x mean of y
 43.93333  45.85714
```

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.

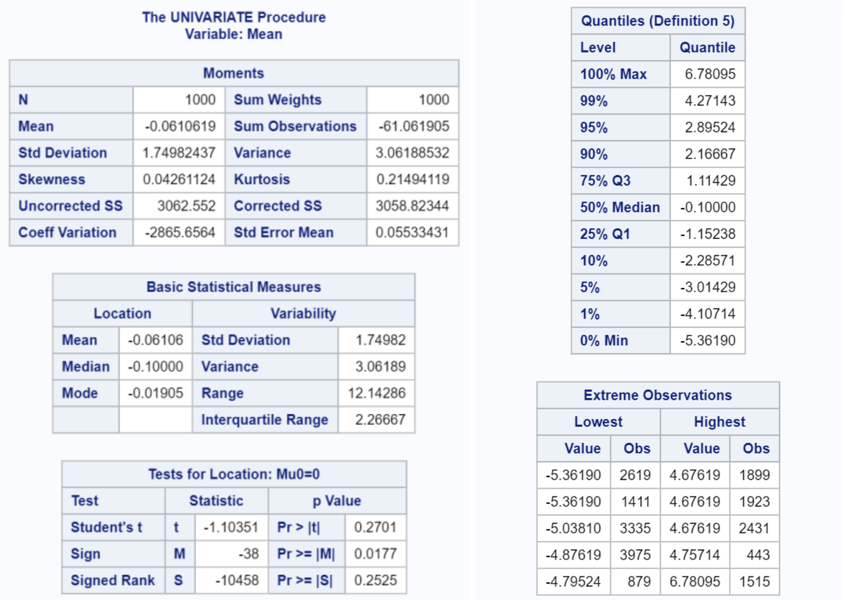
The individuals fired are alleging they were discriminated against based on age, which is illegal in the United States of America. We would like to test the claim of age discrimination by testing the mean of the fired group to determine if it is different from the mean of the currently employed group.

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

The evidence suggests there is insufficient evidence to indicate the not_fired sample is not normally distributed. It also suggests the fired sample is not normally distributed. However, it seems the fired sample has some outliers, which may be skewing the data to the left. The boxplot identifies an outlier in the fired sample. Considering the fired sample has a small size ($n_1 = 21$), the sample is not robust. In addition, the histograms show the samples do not have equal standard deviations. The fired sample size has larger standard deviations. Furthermore, the different sample sizes ($n_1 = 21$, $n_2 = 30$) indicate this set is not robust to the t-test based on the difference in sample means. While the true independence of the sample is unknown, it is known that the samples come from the same company. However, we don't know if they were all hired/fired at the same time, tenure, etc. Based on the information presented, a t-test is not an appropriate test for this sample.

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

Considering the t-test is not appropriate for this test, a permutation test was conducted.



The UNIVARIATE Procedure
Variable: Mean

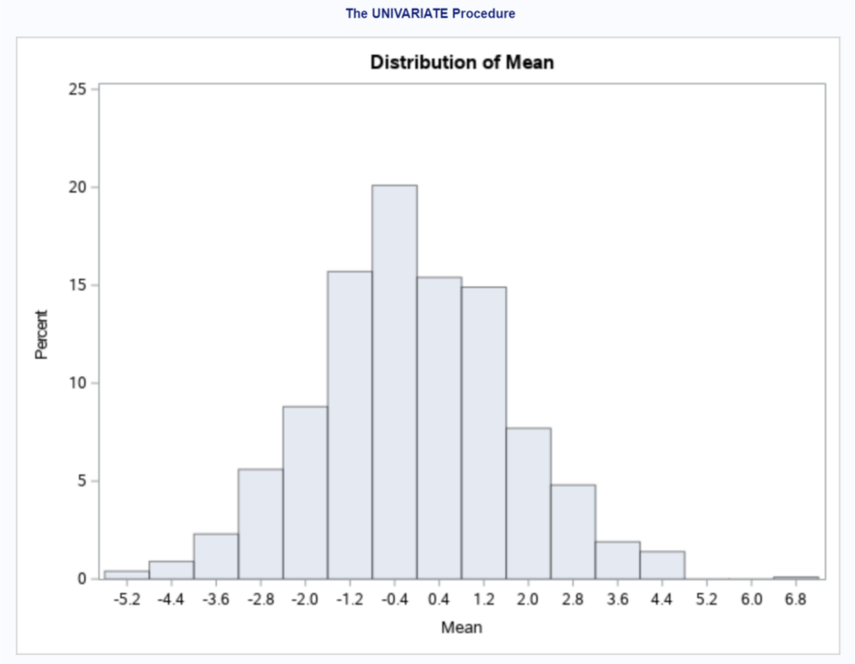
| Moments | | | |
|---|---|---|---|
| N | 1000 | Sum Weights | 1000 |
| Mean | -0.0610619 | Sum Observations | -61.061905 |
| Std Deviation | 1.74982437 | Variance | 3.06188532 |
| Skewness | 0.04261124 | Kurtosis | 0.21494119 |
| Uncorrected SS | 3062.552 | Corrected SS | 3058.82344 |
| Coeff Variation | -2865.6564 | Std Error Mean | 0.05533431 |

| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | -0.06106 | Std Deviation | 1.74982 |
| Median | -0.10000 | Variance | 3.06189 |
| Mode | -0.01905 | Range | 12.14286 |
| | | Interquartile Range | 2.26667 |

| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | -1.10351 | Pr > \|t\| | 0.2701 |
| Sign | M | -38 | Pr >= \|M\| | 0.0177 |
| Signed Rank | S | -10458 | Pr >= \|S\| | 0.2525 |

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 6.78095 |
| 99% | 4.27143 |
| 95% | 2.89524 |
| 90% | 2.16667 |
| 75% Q3 | 1.11429 |
| 50% Median | -0.10000 |
| 25% Q1 | -1.15238 |
| 10% | -2.28571 |
| 5% | -3.01429 |
| 1% | -4.10714 |
| 0% Min | -5.36190 |

| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| -5.36190 | 2619 | 4.67619 | 1899 |
| -5.36190 | 1411 | 4.67619 | 1923 |
| -5.03810 | 3335 | 4.67619 | 2431 |
| -4.87619 | 3975 | 4.75714 | 443 |
| -4.79524 | 879 | 6.78095 | 1515 |



The UNIVARIATE Procedure

Value of t-statistic = -1.10
P value = 0.2701 (two sided)
Decision: Fail to reject H0
Confidence intervals: -3.01429, 2.89524 (90%)

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

The results suggest there is insufficient evidence to support those who were fired, were done so because of their age (two sided p-value = 0.2701). A 90% confidence interval for this set is (-3.01429, 2.89524) points.

5. Provide the scope of inference.

Since this was an observational study, we are unable to infer age was a result of the termination. Furthermore, since the individuals that came forth under the Fired sample are volunteers and it is unknown if the other sample is all of the employees, we can only infer these results to these 51 participants.

(Steps 3-5 are from your previous HW; you are just putting everything together.)

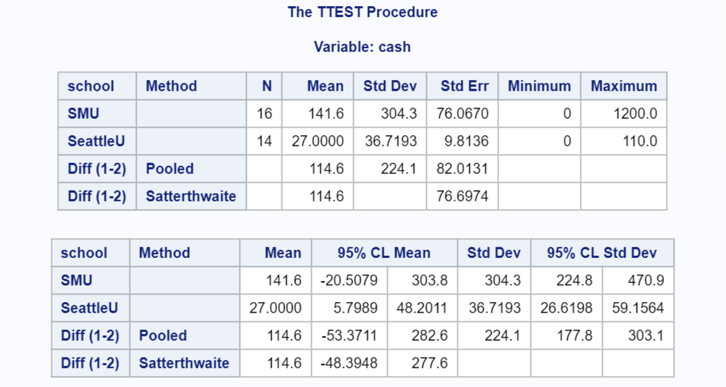
NOTE: THIS QUESTION SHOULD BE EASY AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. (Steps 3-5 are from your previous HW; you are just putting everything together.) IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 and 2 above). You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

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!)

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?

There is insufficient evidence to indicate the sample data is normally distributed. It seems both samples carry a number of outliers as identified by the boxplots. Considering the sample sizes are small ($n_{SMU}$ = 16; $n_{SeattleU}$ = 14), a t-test would not be robusts in this scenario Furthermore, the outliers also mean the samples don't share equal standard deviations. Yet, considering the sample sizes are close, their similarity in size helps the t-test robustness. In addition, the data is independent since the samples stem from different universities and are completely unrelated. However, in order to increase the robustness a log transformation would be appropriate in this instance to increase the robustness of the t-test.

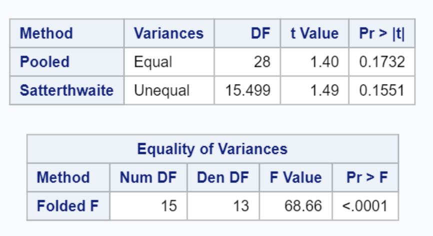


The TTEST Procedure

Variable: cash

| school | Method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| SMU | | 16 | 141.6 | 304.3 | 76.0670 | 0 | 1200.0 |
| SeattleU | | 14 | 27.0000 | 36.7193 | 9.8136 | 0 | 110.0 |
| Diff (1-2) | Pooled | | 114.6 | 224.1 | 82.0131 | | |
| Diff (1-2) | Satterthwaite | | 114.6 | | 76.6974 | | |

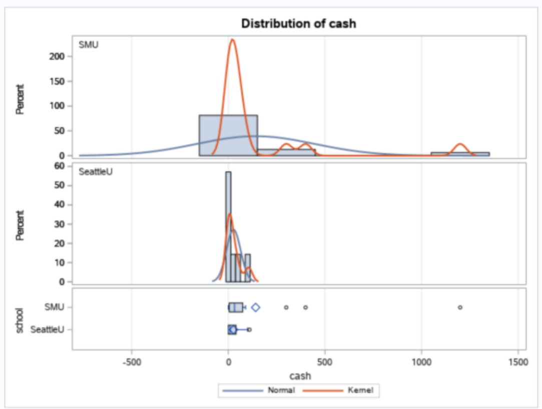
| school | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|---|---|
| SMU | | 141.6 | -20.5079 | 303.8 | 304.3 | 224.8 | 470.9 |
| SeattleU | | 27.0000 | 5.7989 | 48.2011 | 36.7193 | 26.6198 | 59.1564 |
| Diff (1-2) | Pooled | 114.6 | -53.3711 | 282.6 | 224.1 | 177.8 | 303.1 |
| Diff (1-2) | Satterthwaite | 114.6 | -48.3948 | 277.6 | | | |

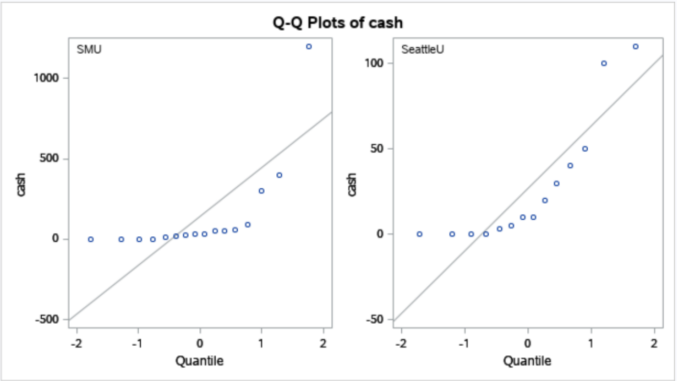


| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | 28 | 1.40 | 0.1732 |
| Satterthwaite | Unequal | 15.499 | 1.49 | 0.1551 |

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.

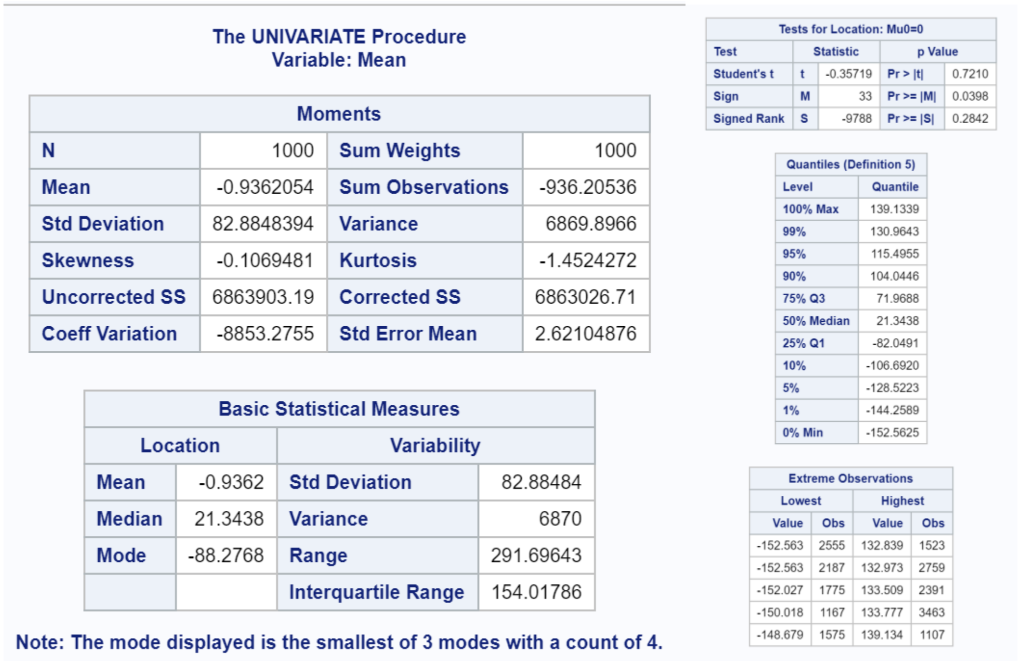
In an effort to determine whether a vending machine was most efficient in its current location, the question was raised whether students carried sufficient cash to utilize its services. Two samples were obtained from separate locations. Do the students from SMU carry the same amount of cash as the students from Seattle U?

$H_0$: $\mu_{SMU} = \mu_{SeattleU}$ $H_a$: $\mu_{SMU} \neq \mu_{SeattleU}$

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

There is insufficient evidence to indicate the sample data is normally distributed. It seems both samples carry a number of outliers as identified by the boxplots. Considering the sample sizes are small ($n_{SMU}$ = 16; $n_{SeattleU}$ = 14), a t-test would not be robusts in this scenario. Furthermore, the outliers also mean the samples don't share equal standard deviations. Yet, considering the sample sizes are close, their similarity in size helps the t-test robustness. In addition, the data is independent since the samples stem from different universities and are completely unrelated. However, in order to increase the robustness a log transformation would be appropriate in this instance to increase the robustness of the t-test. As a result, a permutation test would be better suited for this test.

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 UNIVARIATE Procedure
Variable: Mean

| Moments | | | |
|---|---|---|---|
| N | 1000 | Sum Weights | 1000 |
| Mean | -0.9362054 | Sum Observations | -936.20536 |
| Std Deviation | 82.8848394 | Variance | 6869.8966 |
| Skewness | -0.1069481 | Kurtosis | -1.4524272 |
| Uncorrected SS | 6863903.19 | Corrected SS | 6863026.71 |
| Coeff Variation | -8853.2755 | Std Error Mean | 2.62104876 |

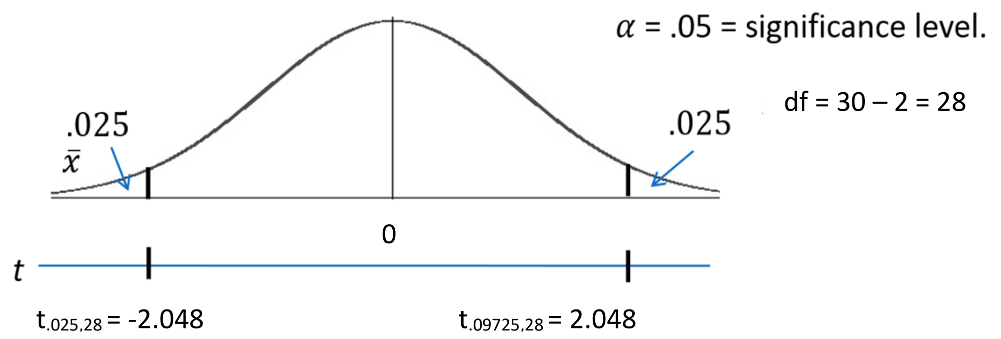
| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | -0.9362 | Std Deviation | 82.88484 |
| Median | 21.3438 | Variance | 6870 |
| Mode | -88.2768 | Range | 291.69643 |
| | | Interquartile Range | 154.01786 |

Note: The mode displayed is the smallest of 3 modes with a count of 4.

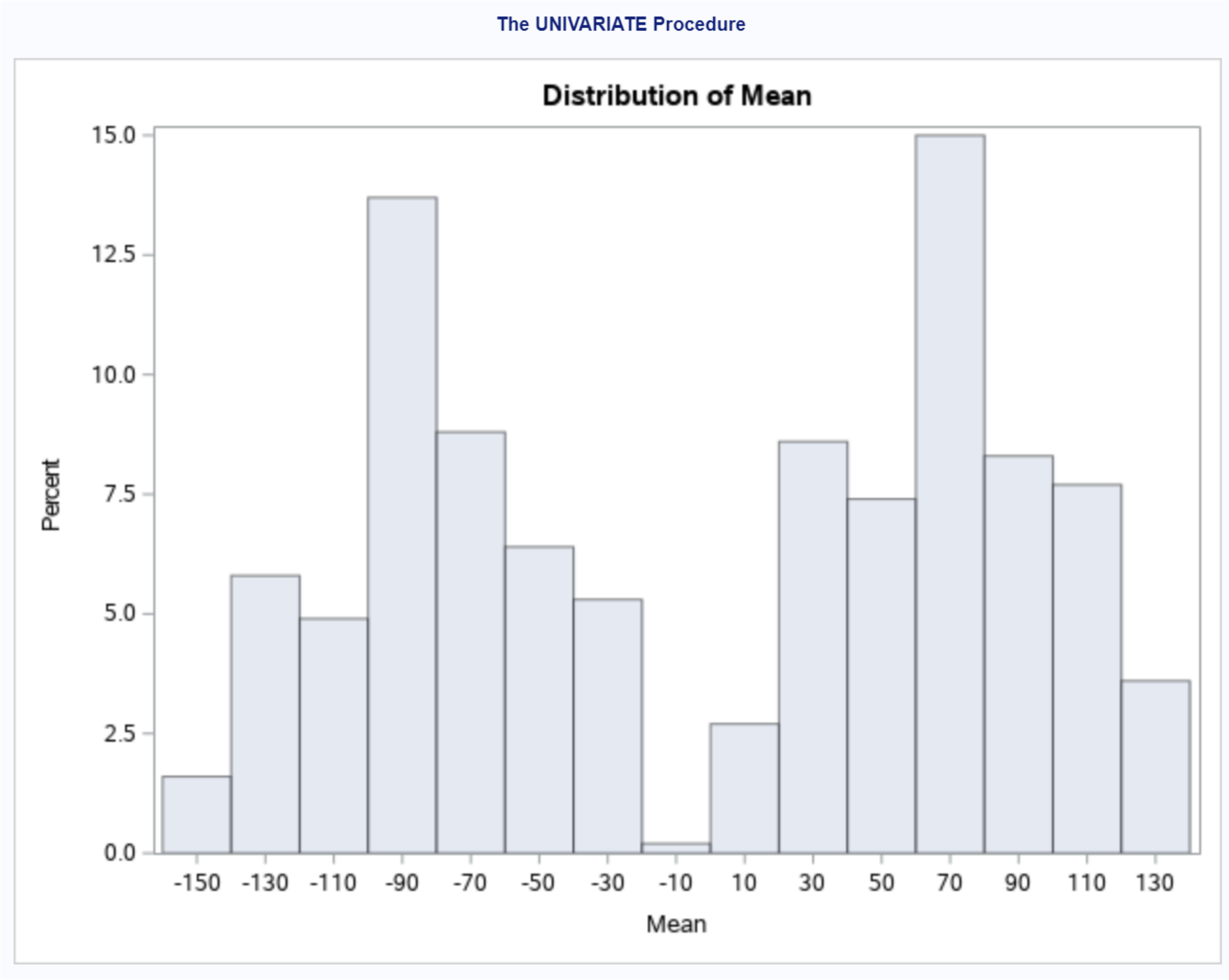
| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | -0.35719 | Pr > \|t\| | 0.7210 |
| Sign | M | 33 | Pr >= \|M\| | 0.0398 |
| Signed Rank | S | -9788 | Pr >= \|S\| | 0.2842 |

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 139.1339 |
| 99% | 130.9643 |
| 95% | 115.4955 |
| 90% | 104.0446 |
| 75% Q3 | 71.9688 |
| 50% Median | 21.3438 |
| 25% Q1 | -82.0491 |
| 10% | -106.6920 |
| 5% | -128.5223 |
| 1% | -144.2589 |
| 0% Min | -152.5625 |

| Extreme Observations | | | |
|---|---|---|---|
| Lowest | | Highest | |
| Value | Obs | Value | Obs |
| -152.563 | 2555 | 132.839 | 1523 |
| -152.563 | 2187 | 132.973 | 2759 |
| -152.027 | 1775 | 133.509 | 2391 |
| -150.018 | 1167 | 133.777 | 3463 |
| -148.679 | 1575 | 139.134 | 1107 |



P value = 0.7210
T statistic = -0.35719
Critical value = ±2.048


The UNIVARIATE Procedure

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

On the basis of this test, there is insufficient evidence to reject the claim that the SMU students and SeattleU students carry different means of cash (p = 0.7210 from two sided). A 95% confidence interval for the difference is [$115.50, -$128.53]. It seems as though the outliers may have skewed the data and permutation results, considering the boxplot identified multiple of them in the data set.

5. Provide the scope of inference.

Since the students were a convenience sample, the results may not apply to others outside of these sample groups. Since this was also an observation, there are clearly no causal results.

```
data criteval;
p = quantile("T",.975, 28);
proc print data=criteval;
run;

*ttest to get the observed difference of means for permutation test;
proc ttest data=Student_Cash;
    class school;
    var cash;
run;

ods output off;
ods exclude all;

*allows IML to use code in order to create matrixes. This allows for
randomization and counter code;
proc iml;
use Student_Cash; *Initiates data set desired to be used;
read all var{school cash} into x;   *Stores data set into variable x as a matrix;
p = t(ranperm(x[,2],1000));    *creates randomized permutation of x data set 1000
times. ;
paf = x[,1]||p;    *stores matrix results;
create newds from paf;  *creates new data set based on matrix results and stores
it as "newds" ;
append from paf;
quit;

*calculates differences and creates a histogram;
ods output conflimits=diff;
proc ttest data=newds plots=none;  *performs ttest on newly create matrix from
results;
  class col1;
  var col2 - col1001;
run;

ods output on;
ods exclude none;

proc univariate data=diff;   *Displays univeriate results and histogram of "diff"
variable stored above;
  where method = "Pooled";
  var mean;
  histogram mean;
run;
```

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.) Steps 3-5 are from your previous HW; you are just putting everything together. You can basically copy and paste the rest. We are simply putting everything together to make a complete report.

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.

The outlier (SMU student w/ $1200) appears to have a strong effect on the data set. The t-test p value prior to the removal of the outlier was 0.1732. Once the outlier is removed, we find that the p value only increases to 0.1913. This would indicate that the outlier was in fact not having that much of an impact in the data. The results produce the same conclusion, which is that the we fail to reject the null. Considering this information, we would not exclude the data point from the sample and leave it.

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:

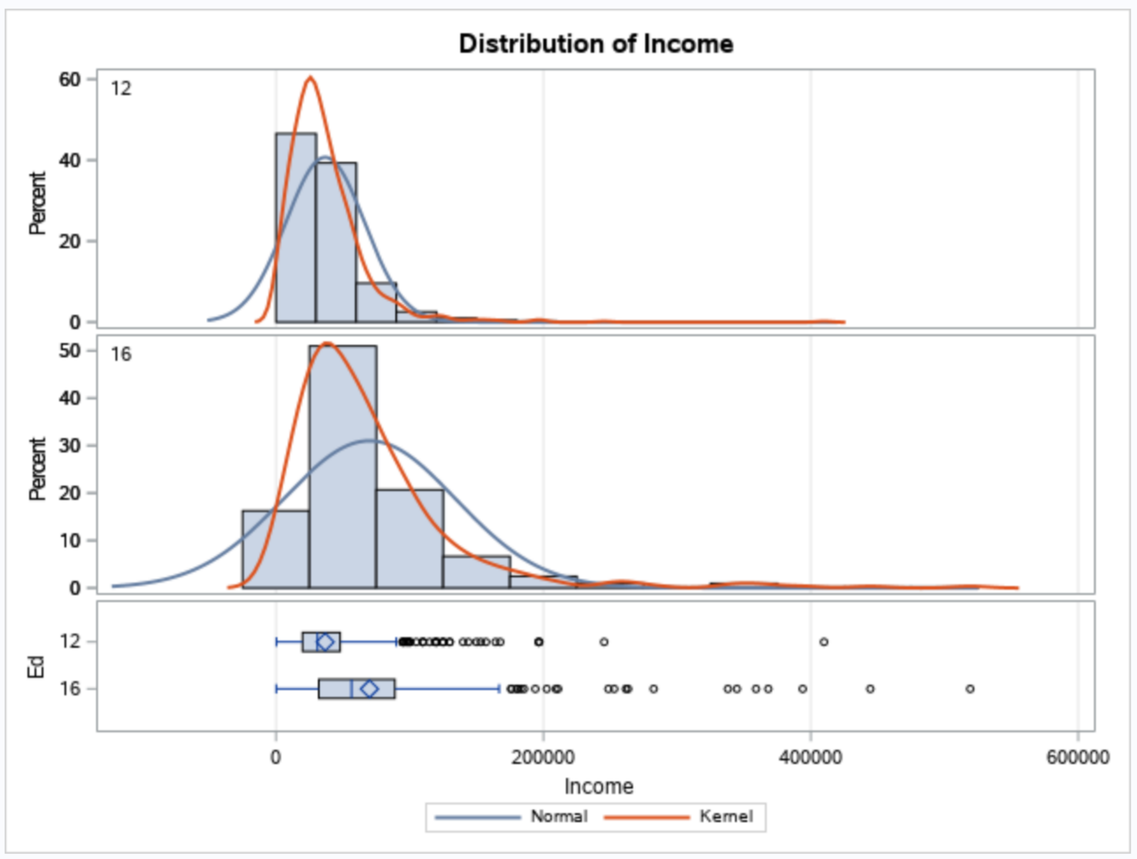
   1. State the Problem.

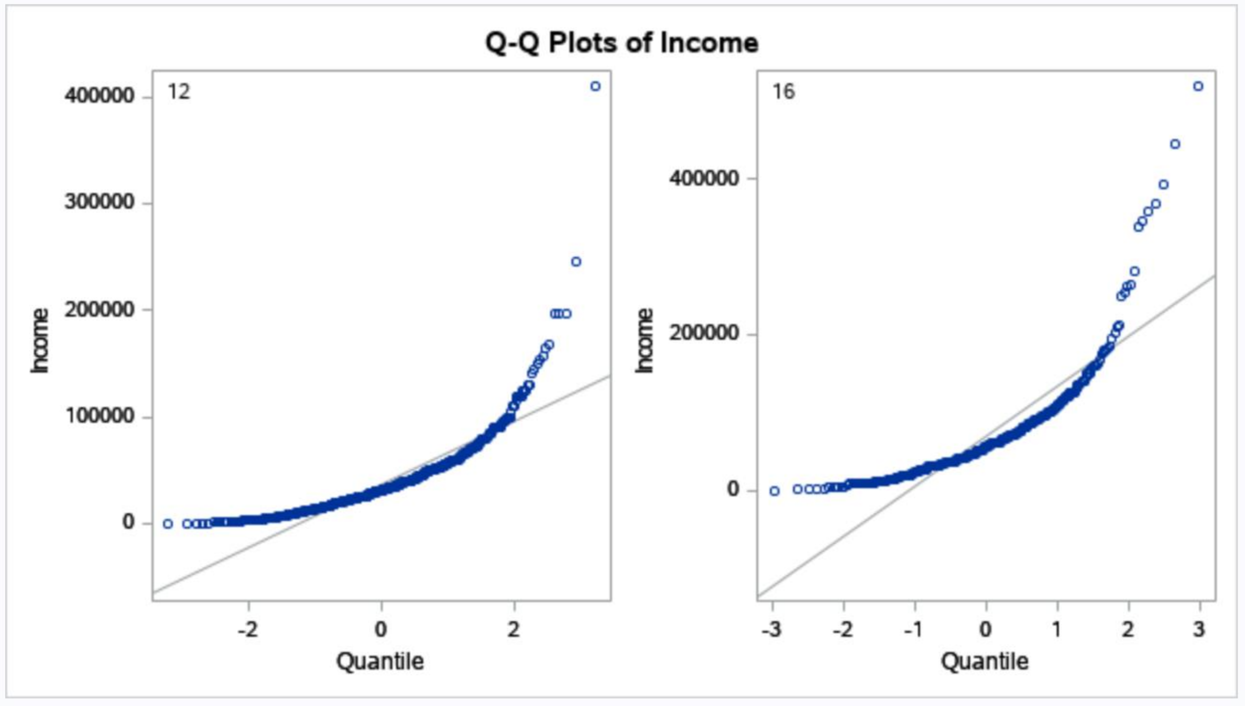
Over 1000 individuals’ salary and education levels were obtained. The claim is that those with a 16-year education earn more than those with a 12-year education. Do the individuals with 16 years of education have higher salaries than those with a 12-year education?

$H_0$: $\mu_{16\text{ years}} > \mu_{12\text{years}}$ $H_A$: $\mu_{16\text{ years}} \leq \mu_{12\text{ years}}$

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

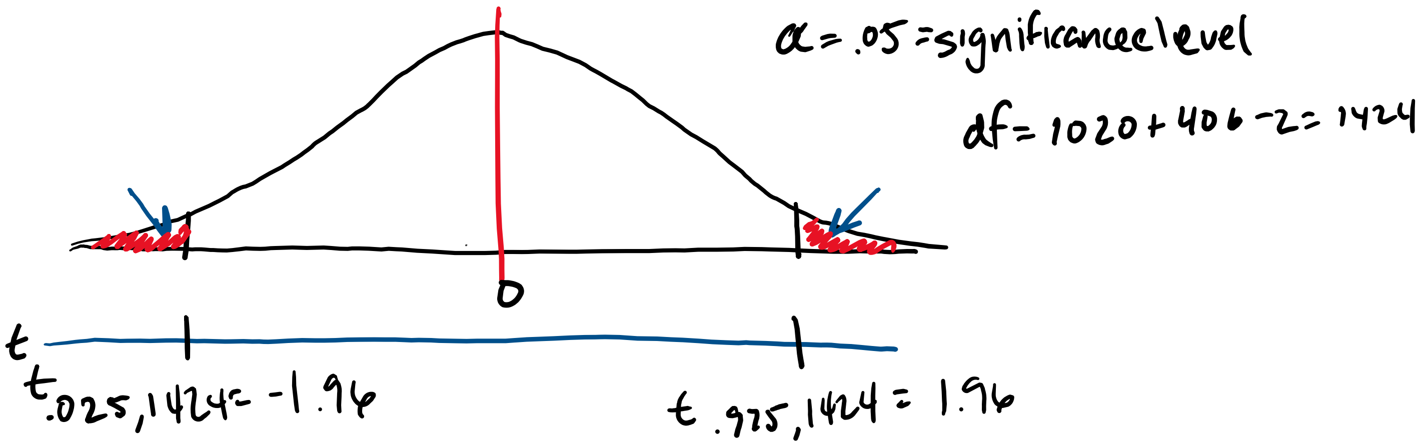
Based on the data, the evidence suggests there are outliers in both samples that are making the data samples skewed right. The Q-Q plot and the histogram indicates there is insufficient evidence to support the data is normally distributed. The sample size for both is large, thus making it robust. However, the outliers appear to be distorting the means of both samples. It looks as though the samples share similar variance, the different in sample sizes indicates a ttest would not be the ideal test for this data set. Considering we don’t know much about the sample selection aside from the fact that they were selected as individual with paying jobs, we can’t speak to whether this data is truly independent from each other. We would need to know more about the sample collection method.





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.)

Based on the assumptions, the data was transformed in order to gain greater normality. As illustrated below, the QQ plots clearly illustrate the normality of the data. In this instance, we performed a two sample one sided t-test on the transformed data.



Critical value = ±1.96

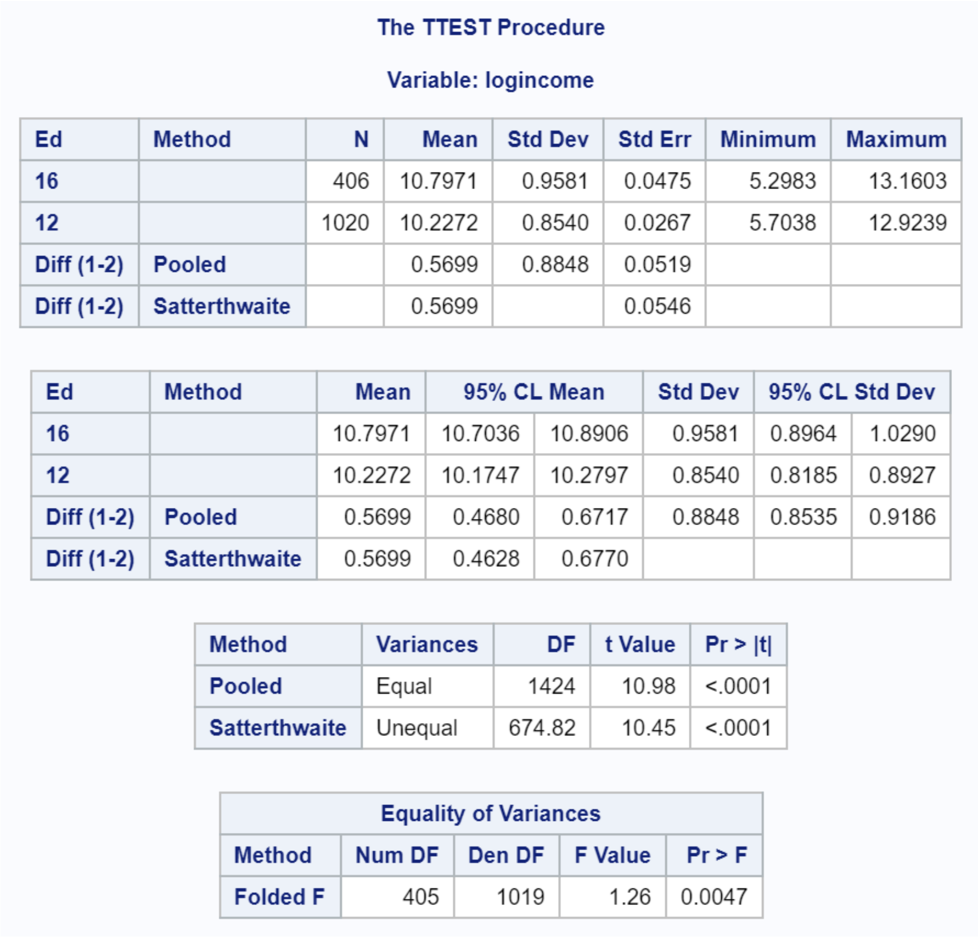
P value = < 0.001

T statistic = 10.98

Alpha = .05

Df = 1020 + 406 – 2 = 1424

CI = [$e^{0.4680}$, $e^{0.6717}$] = [1.599, 1.958]



The TTEST Procedure

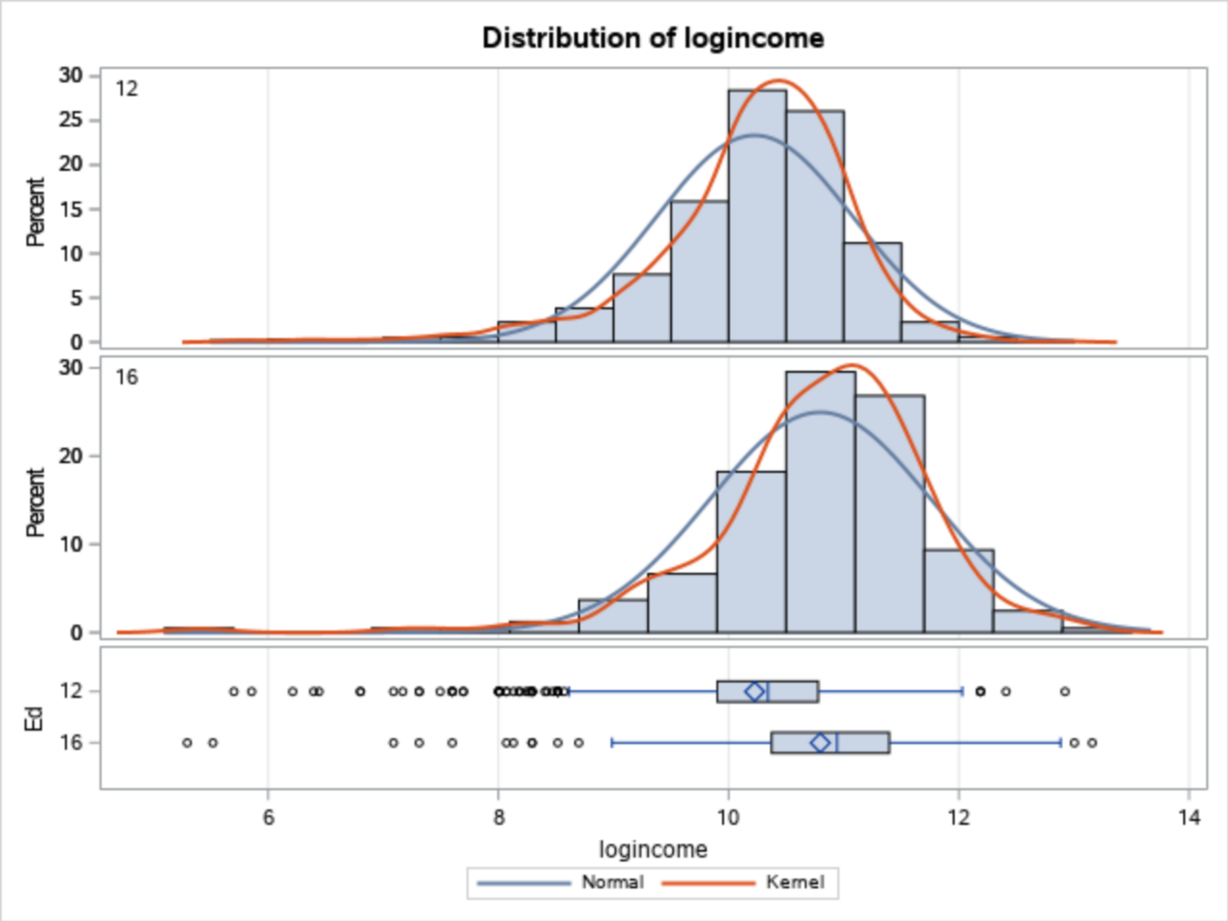
Variable: logincome

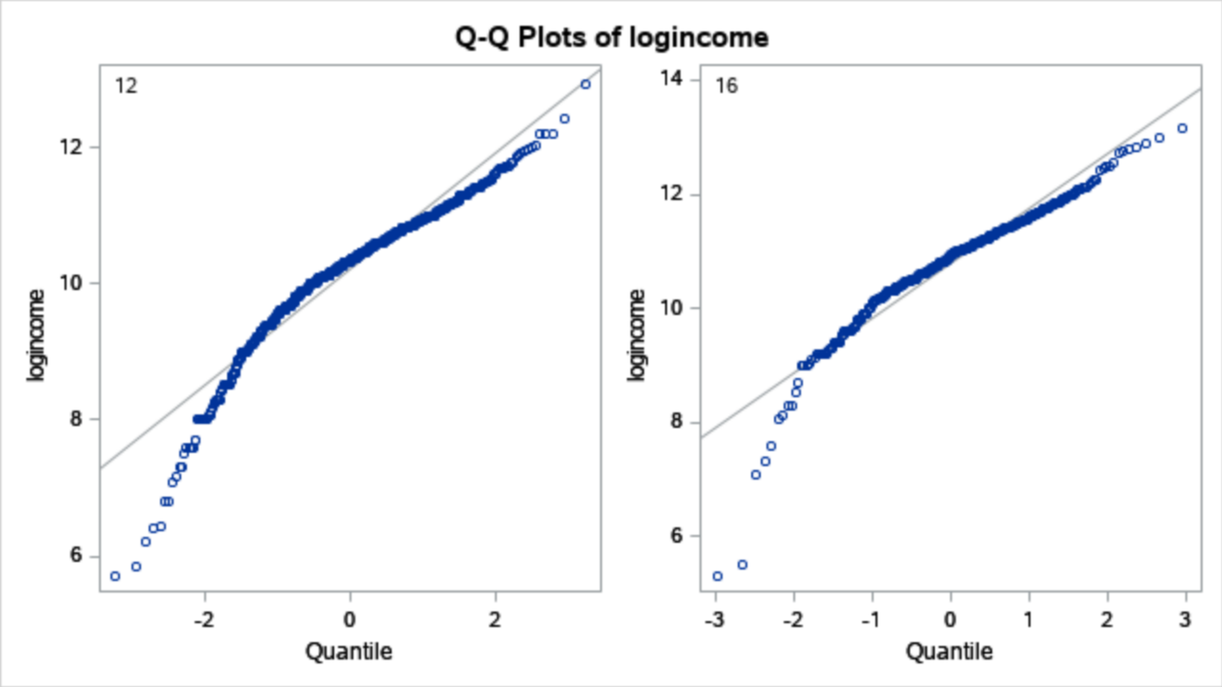
| Ed | Method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| 16 | | 406 | 10.7971 | 0.9581 | 0.0475 | 5.2983 | 13.1603 |
| 12 | | 1020 | 10.2272 | 0.8540 | 0.0267 | 5.7038 | 12.9239 |
| Diff (1-2) | Pooled | | 0.5699 | 0.8848 | 0.0519 | | |
| Diff (1-2) | Satterthwaite | | 0.5699 | | 0.0546 | | |

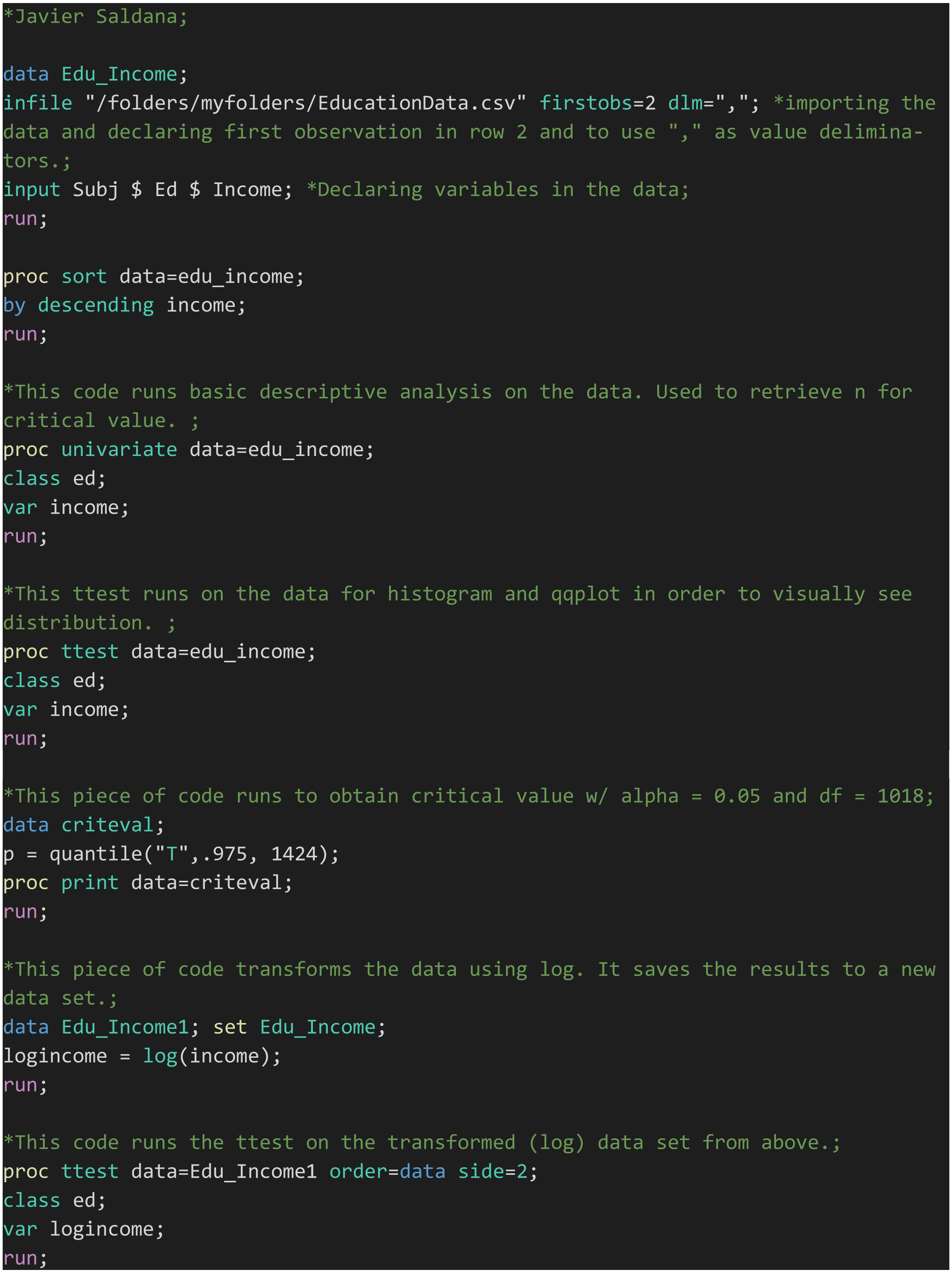
| Ed | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|---|---|
| 16 | | 10.7971 | 10.7036 | 10.8906 | 0.9581 | 0.8964 | 1.0290 |
| 12 | | 10.2272 | 10.1747 | 10.2797 | 0.8540 | 0.8185 | 0.8927 |
| Diff (1-2) | Pooled | 0.5699 | 0.4680 | 0.6717 | 0.8848 | 0.8535 | 0.9186 |
| Diff (1-2) | Satterthwaite | 0.5699 | 0.4628 | 0.6770 | | | |

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | 1424 | 10.98 | <.0001 |
| Satterthwaite | Unequal | 674.82 | 10.45 | <.0001 |

| Equality of Variances | | | | |
|---|---|---|---|---|
| Method | Num DF | Den DF | F Value | Pr > F |
| Folded F | 405 | 1019 | 1.26 | 0.0047 |


Distribution of logincome
Percent
12
16
Ed
logincome
Normal
Kernel
30
25
20
15
10
5
0
6
8
10
12
14


Q-Q Plots of logincome
12
16
logincome
Quantile
-3
-2
-1
0
1
2
3
6
8
10
12
14



```
*Javier Saldana;

data Edu_Income;
infile "/folders/myfolders/EducationData.csv" firstobs=2 dlm=","; *importing the
data and declaring first observation in row 2 and to use "," as value delimina-
tors.;
input Subj $ Ed $ Income; *Declaring variables in the data;
run;

proc sort data=edu_income;
by descending income;
run;

*This code runs basic descriptive analysis on the data. Used to retrieve n for
critical value. ;
proc univariate data=edu_income;
class ed;
var income;
run;

*This ttest runs on the data for histogram and qqplot in order to visually see
distribution. ;
proc ttest data=edu_income;
class ed;
var income;
run;

*This piece of code runs to obtain critical value w/ alpha = 0.05 and df = 1018;
data criteval;
p = quantile("T",.975, 1424);
proc print data=criteval;
run;

*This piece of code transforms the data using log. It saves the results to a new
data set.;
data Edu_Income1; set Edu_Income;
logincome = log(income);
run;

*This code runs the ttest on the transformed (log) data set from above.;
proc ttest data=Edu_Income1 order=data side=2;
class ed;
var logincome;
run;
```

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

There is sufficient evidence to suggest individuals with a 12-year education do not earn the same as those with a 16-year education (p value < 0.001, 2 side). As a result, we must reject the null hypothesis that both groups share the same mean salary. A 95% confidence interval suggests the true mean is anywhere between [1.599, 1.958] times as the other group.

5. Provide a scope of inference.

Considering the study was observational, it is difficult to attribute a causal relationship between education and salary. Furthermore, we don't know enough about the sampling methods so we can't attribute the results to individuals outside of this study.

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?