**Unit 5 HW**

**1. Simply Answer Question 25 on pg. 147** **from the Statistical Sleuth:** *Plot the raw data and also plot the data after a log transform. After a log transform, do the data satisfy the assumptions better?* The data is in ex0525.csv. Perform this analysis in SAS.

Regardless of whether the assumptions of the original data or log transformed data are met, please include a complete analysis on the **log transformed** data.

1. State the Problem

## Problem

A random sample of 2,584 Americans who were selected for the NLS of youth in 1979 and who had paying jobs in 2005. These americans had different levels of education.

How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others?

By how many dollars or by what percent does the mean or median for each of the last four categories exceed that of the next lowest category?

1. Address the assumptions. Comment on each assumption. (Use the visual test, as the Brown-Forsythe test will be overpowered due to the large sample size. This simply means that it is able to detect very small effect sizes—here, differences in standard deviations—which may not be big enough to practically affect the test.) Comment on your thoughts of the assumptions, but, in the end, assume there is not enough visual evidence to suggest the standard deviations of the log transformed data are different.

## Code Used:

/\*\* Import an XLSX file. \*\*/

PROC IMPORT DATAFILE="/home/marinfamily1010/sasuser.v94/EDIncome/ex0525\_2\_2.xlsx"

OUT=WORK.EDINCOME

DBMS=XLSX

REPLACE;

RUN;

/\*\* Print the results. \*\*/

/\*

\* Received error that data must be in ascending order. Modifying data by Educ asc

\*/

proc sql;

create view work.edincomesort as

select \* from work.edincome

order by Educ asc;

quit;

proc univariate data = work.edincomesort;

by Educ;

histogram Income2005;

qqplot Income2005;

run;

proc glm data = work.edincome;

class Educ;

model Income2005 = Educ;

run;

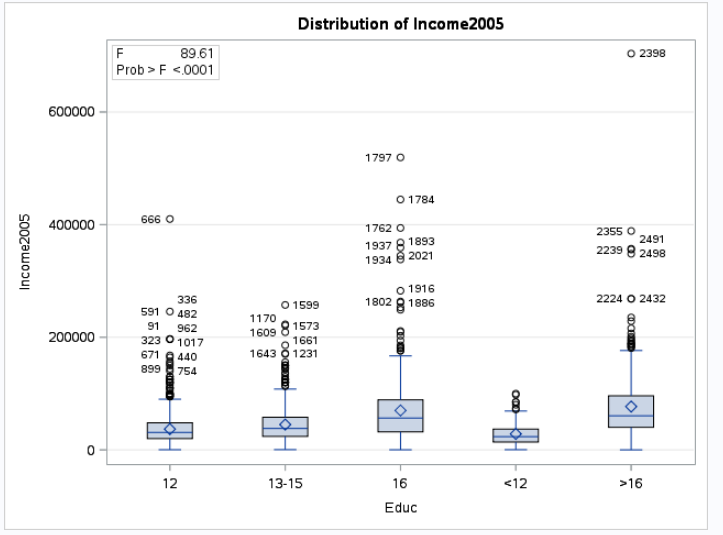
## Assumptions

Since we have 5 groups within the sample, I will compare the groups using Anova to see if at least one pair of means are different. To do so, the following assumptions must apply:

1. Normality: Similar to t-tools hypothesis testing, ANOVA is robust to this assumption. Extremely long-tailed distributions (outliers) or skewed distributions, coupled with different sample sizes (especially when the sample sizes are small) present the only serious distributional problems.
2. Equal Standard Deviations: This assumption is crucial, paramount, and VERY important.
3. The assumptions of independence within and across groups are critical. If lacking, different analysis should be attempted.

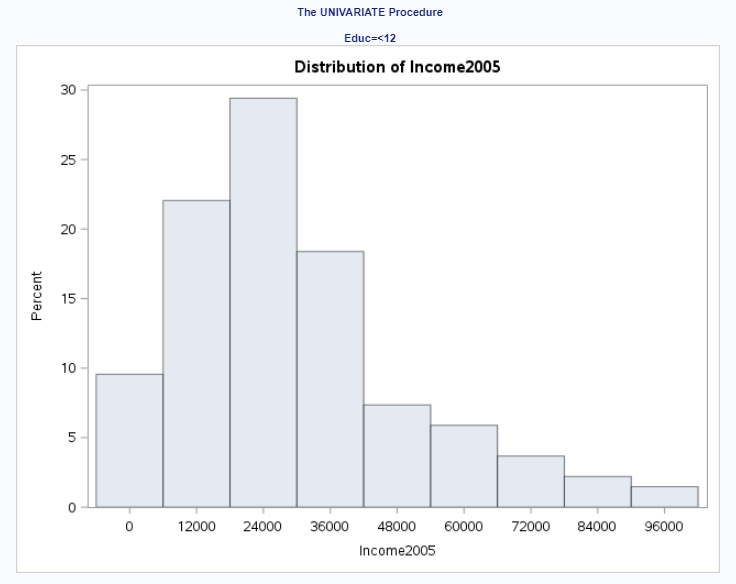
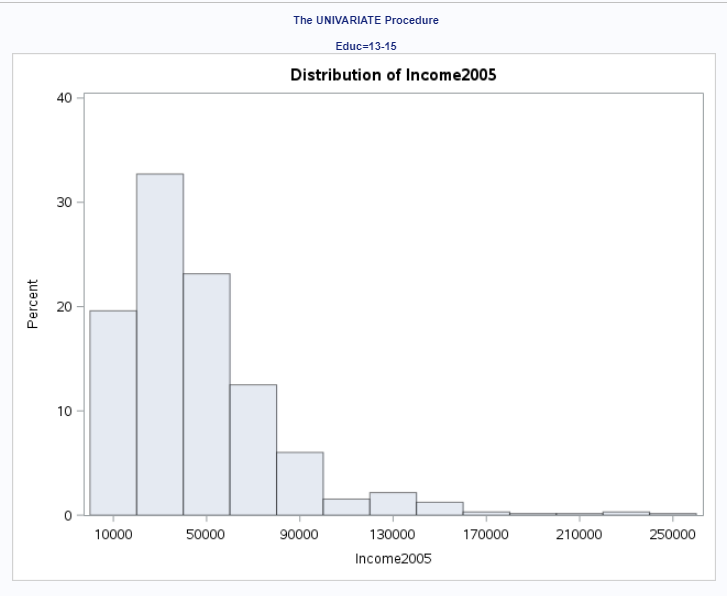
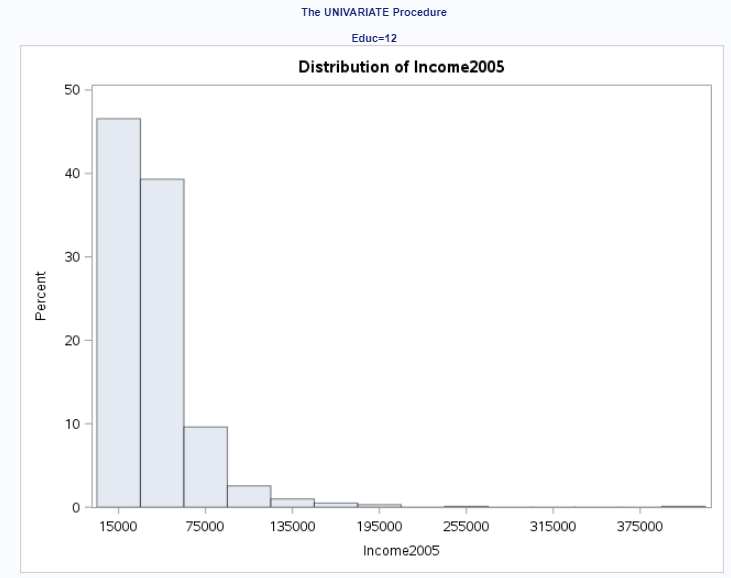
### Normality

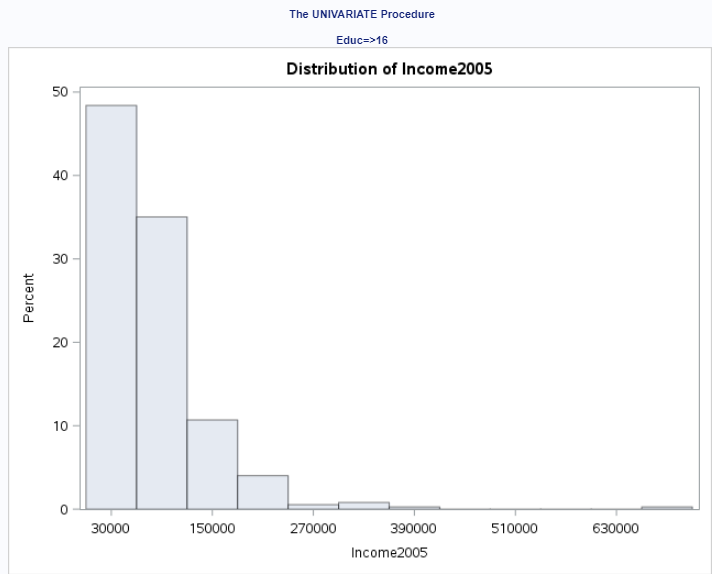
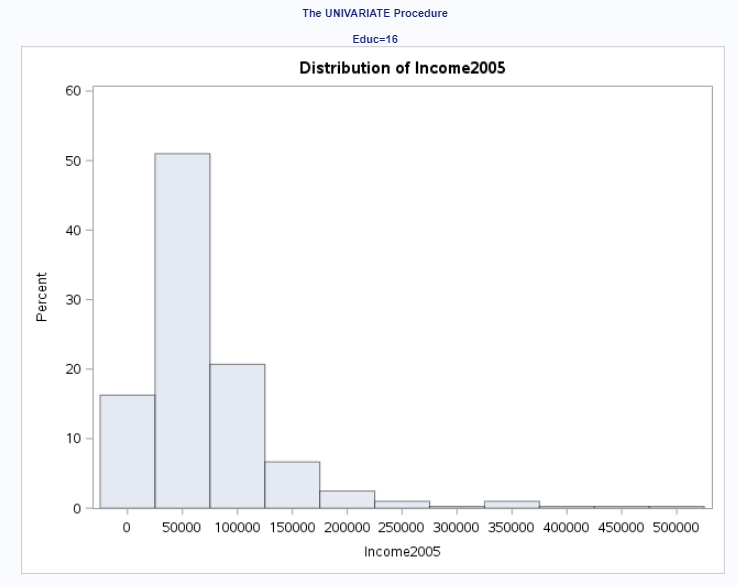
## Box and Whisker:



Interesting that the medians look pretty similar and boxplots do have a bit of symmetry.

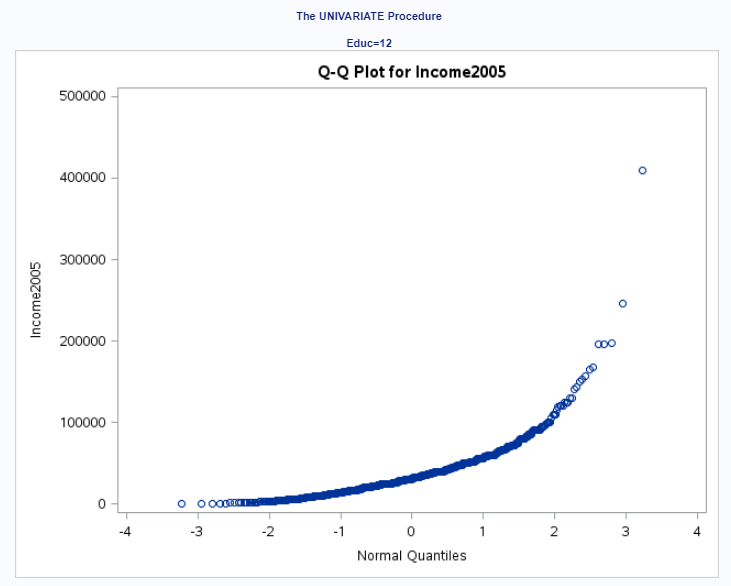
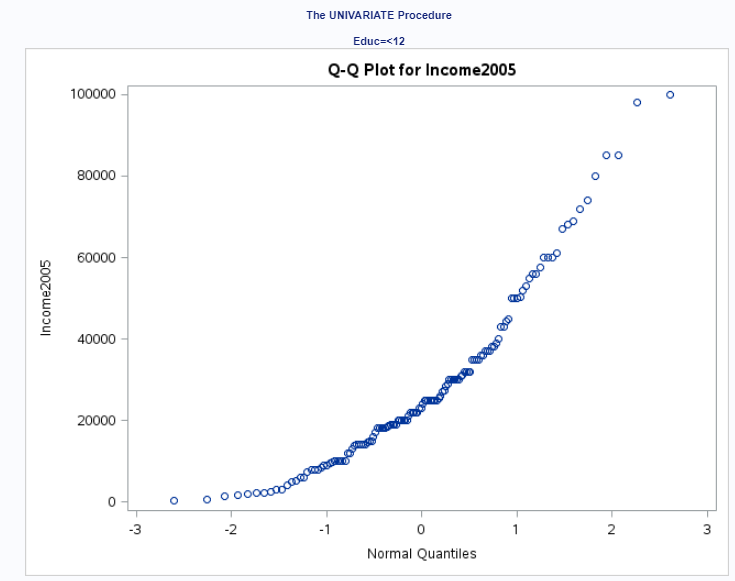
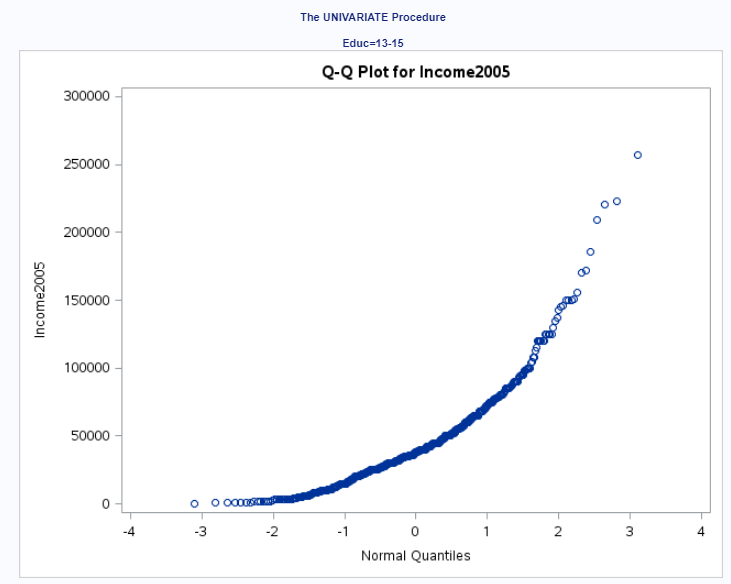
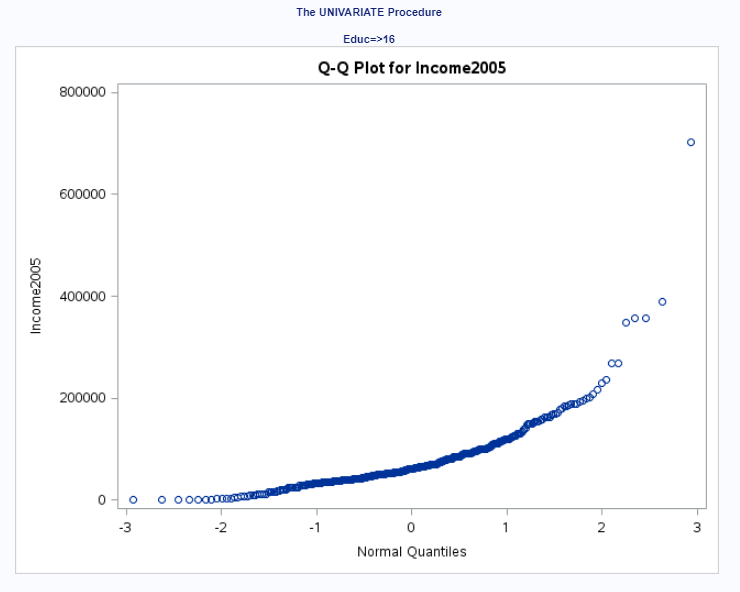
Histograms:

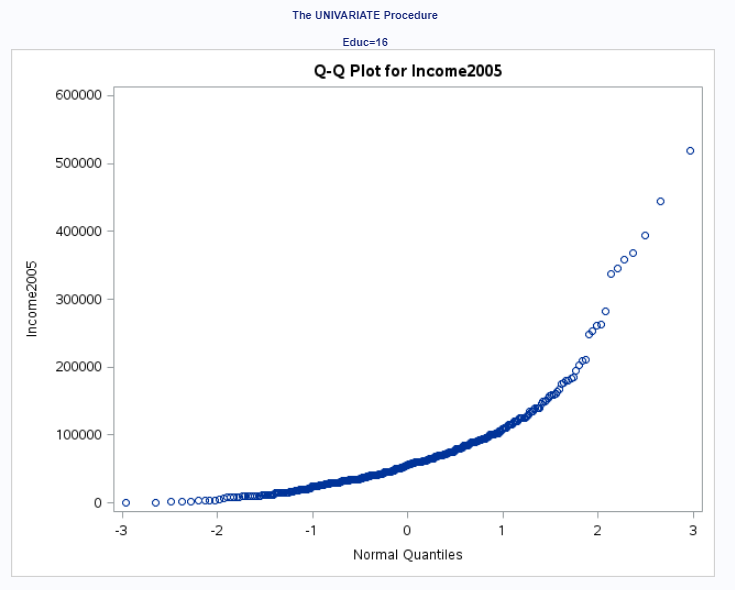
 



By looking at the charts above, many of the distributions don’t appear to be normal with Educ >= 16 having the most right skew while Educ<= 12 appearing to be the most normal compared to the other distributions that have a little right skew. We should transform the data to see if this fixes the normality assumption, but will look at QQPlots first before we do so.

QQPlots:



When looking at the QQPlots, the groups do not look normal for the most part when drawing a straight line. Many of the QQPlots data look like they are exponentially increasing.

Transforming Data:

Code:

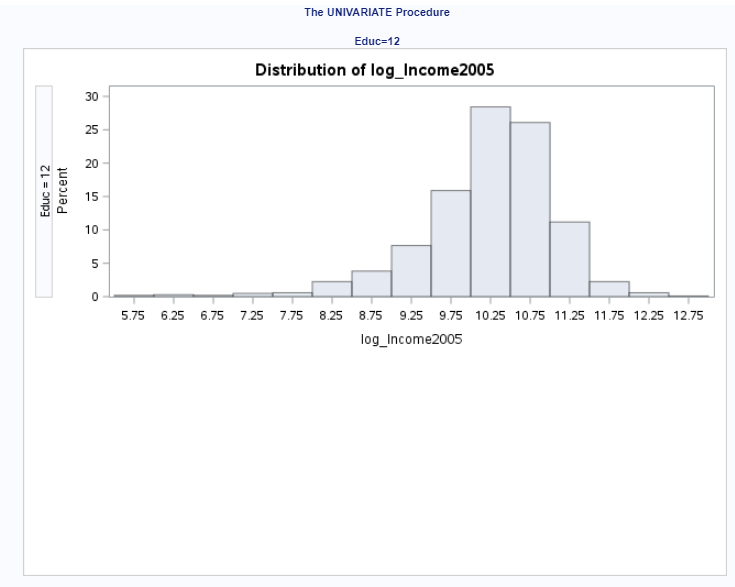
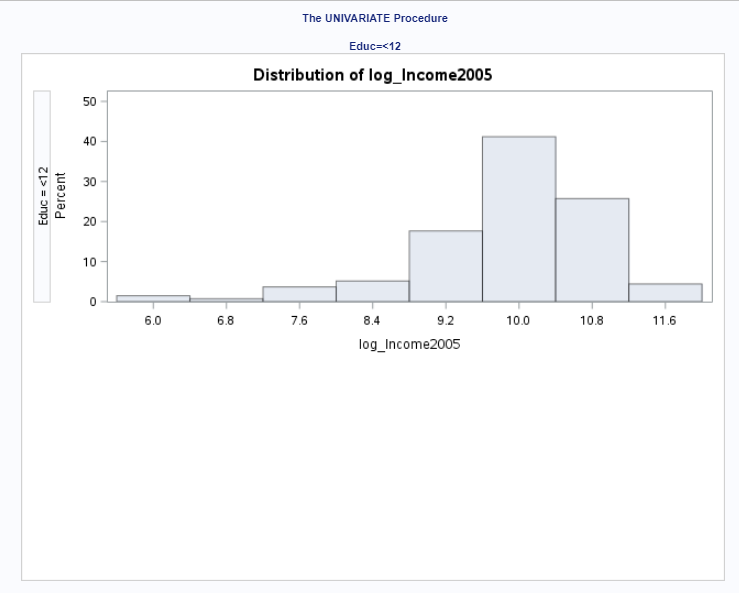
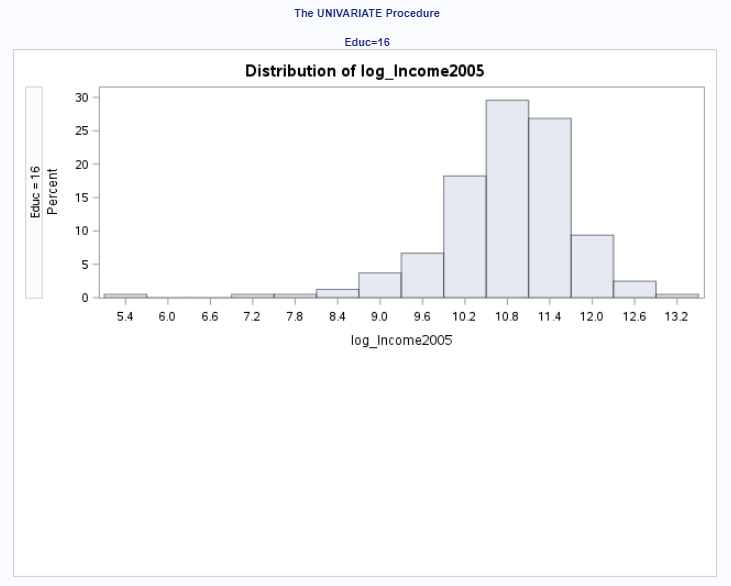
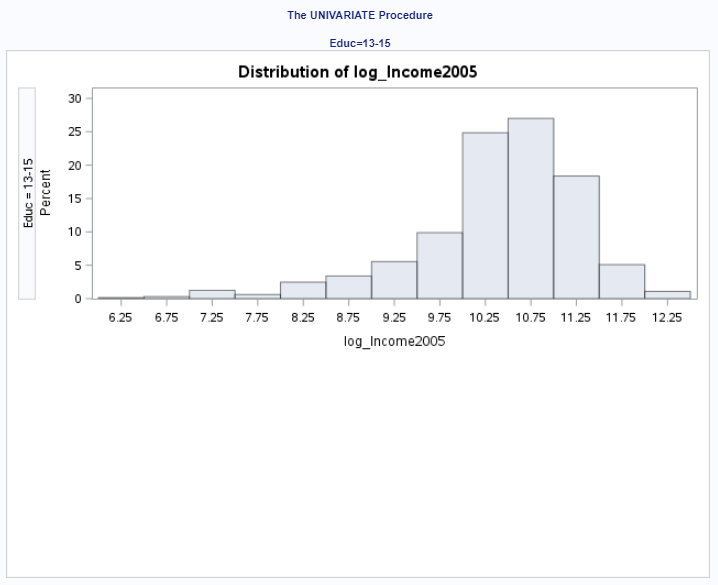
data WORK.transform;

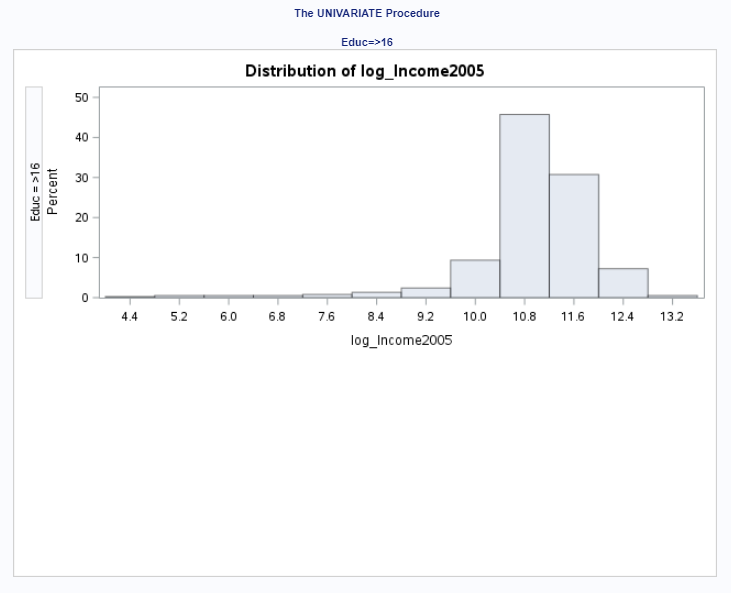
set work.edincomesort;

log\_Income2005=log(Income2005);

run;

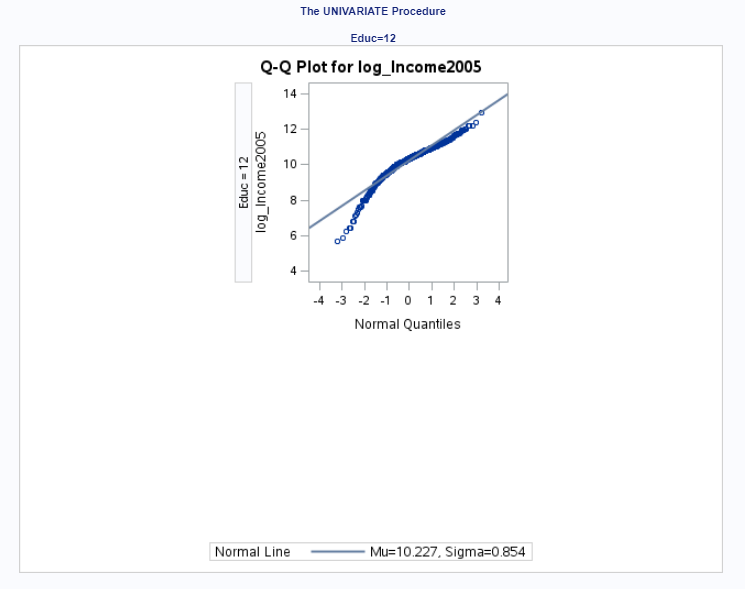
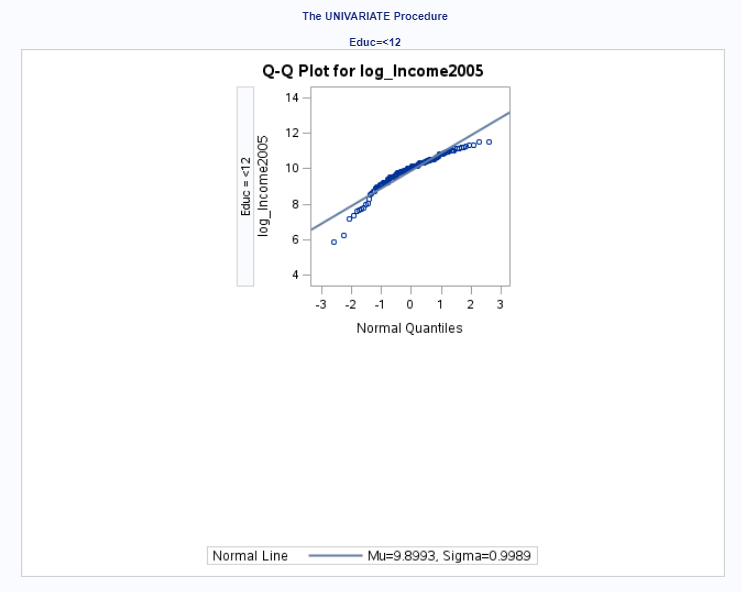
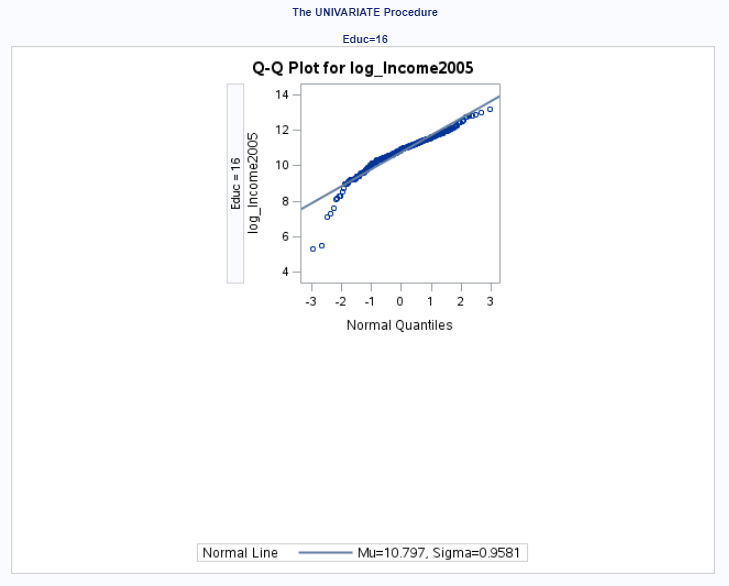
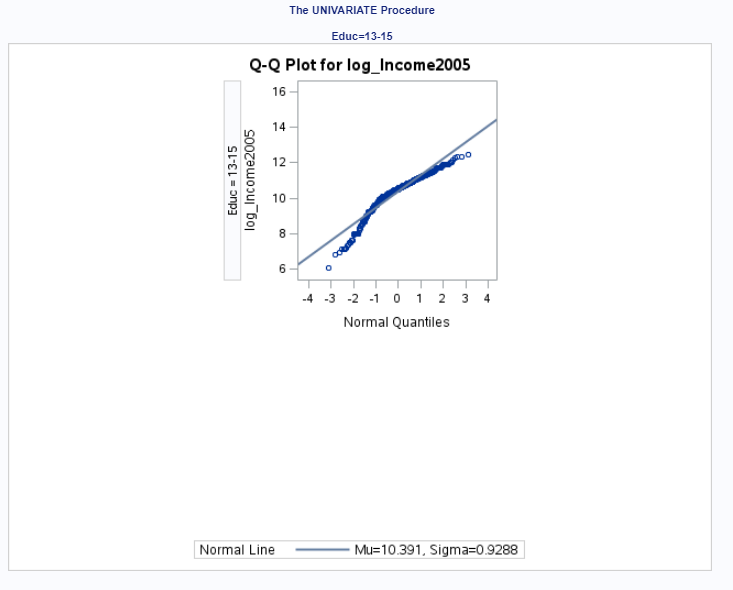
Histograms with log data:

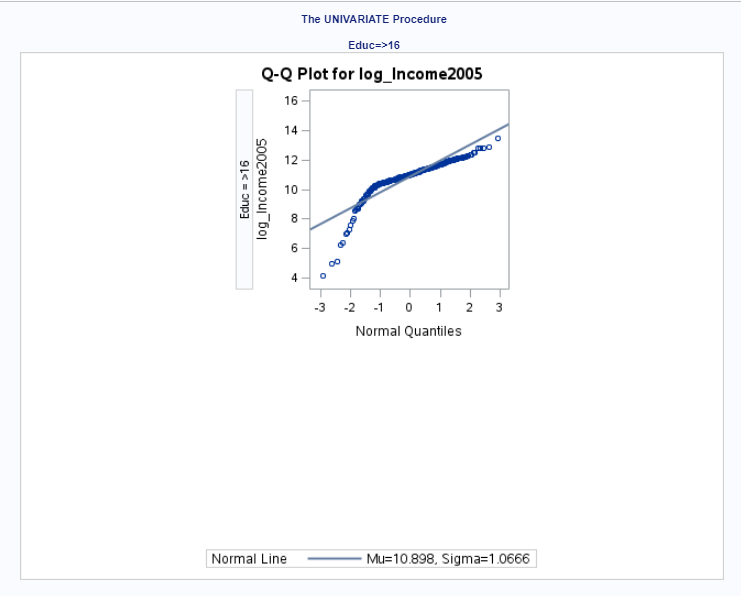
 



After doing the log transformation, it looks like a lot of the histograms are a bit more normal, but have too much left skew.

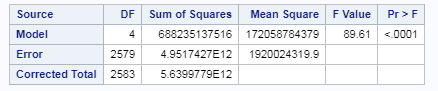
QQPlots:



These QQPlots show the issue with the left skew and don’t appear to be normal. Instruction said to proceed with Anova test anway. We will assume there is not enough visual evidence to suggest the standard deviations of the log transformed data are different.

## Equal Variances:



F –Test shows that there is a difference in the variances, but we will assume for now that variances are equal to finish this HW problem.

## Independence:

We will assume independence given that we don’t have any data to support that there are any dependencies with our observations.

1. Conduct the Test. (An example is in the UNIT 5 PowerPoint.)

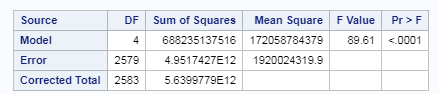
Step 1:

H0 = µ12 = µ<12= µ13-15 = µ16 = µ>16

Ha = At least one pair of the means is different

Step 2: Critical Value not needed for ANOVA

Step 3: F-Statistic = 89.61



Step 4: P-Value: <.0001

Step 5: We reject the null hypothesis.

Step 6: Conclusion -Write a conclusion. (An example is in the UNIT 5 PowerPoint.)

There is strong evidence that at least one of the Education level groups has a mean income that is different than the others (p-value <.0001 from an ANOVA)

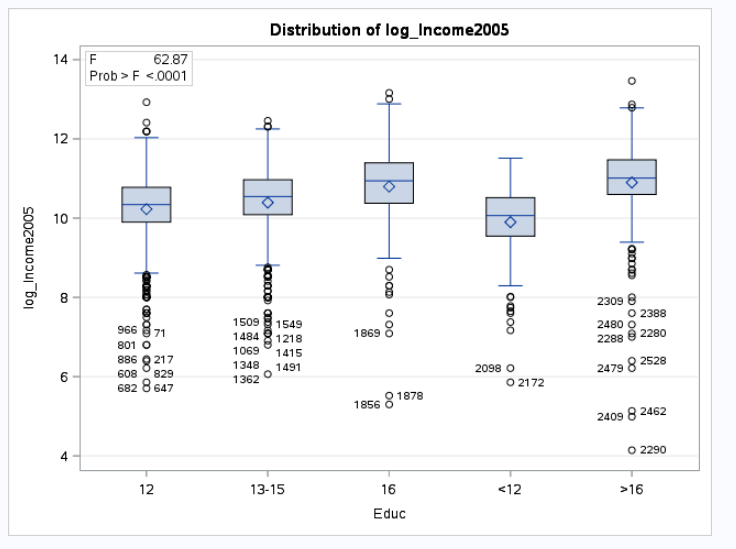
Code Again:

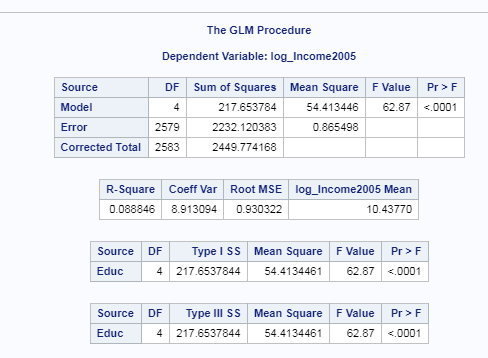
proc glm data = WORK.transform;

class Educ;

model log\_Income2005 = Educ;

run;





1. State the Scope. (Can we generalize to the entire population or just the sample that was taken? Is there a causal relationship present?)

This is an observational study where the members were selected by random from a very specific group. This study only applies to those youths in 1979 and who had paying jobs in 2005. Inference to the general population cannot be made.

*Looking to the future! This is not an additional problem. Just FYI: The next step will be to look at these pairwise if we reject the Ho to discover WHICH pairs have evidence of different means / medians.*

ADDITIONAL THINGS TO INCLUDE (for the logged data):

1. Please also identify R2

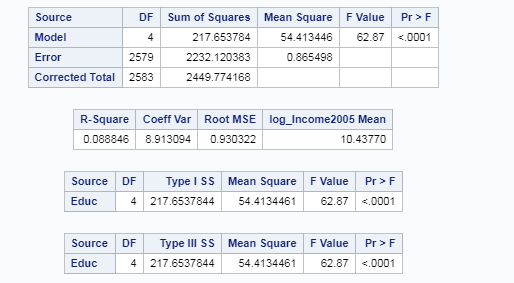
From code:

proc glm data = WORK.transform;

class Educ;

model log\_Income2005 = Educ;

run;



R-square = .088846

1. Also specify the mean square error and how many degrees of freedom were used to estimate it.



DF = 2575

MSE = .86549 = (.930322^2)

1. Provide the code to perform the ANOVA in R and a screen shot of the output.

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 5")

EducData <- read.xlsx("ex0525\_2\_2.xlsx", sheetIndex = "ex0525")

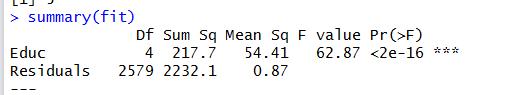
EducData$Income2005 <- log(EducData$Income2005)

library(sqldf)

EducData <- sqldf("select \* from EducData order by Educ asc")

fit <- aov(Income2005 ~ Educ, data=EducData)

summary(fit)



2.Use an extra sum of squares F-test (BYOA … Build Your Own ANOVA!) to use all the data (to increase the degrees of freedom and thus the power of the test!) to compare only the bachelor’s degree group (16) mean income to the graduate degree group (>16) mean income. Show your final ANOVA table and your 6-step complete analysis. You will need to assume that the standard deviations of the log-transformed data are again equal to proceed here. A two-sample t-test between these two groups yields a p-value of .1403 (try it!), but it only uses 778 degrees of freedom (from a pooled t-test). Make note again of how many degrees of freedom were used to estimate the pooled standard deviation in your extra sum of squares test. You may use SAS or R.

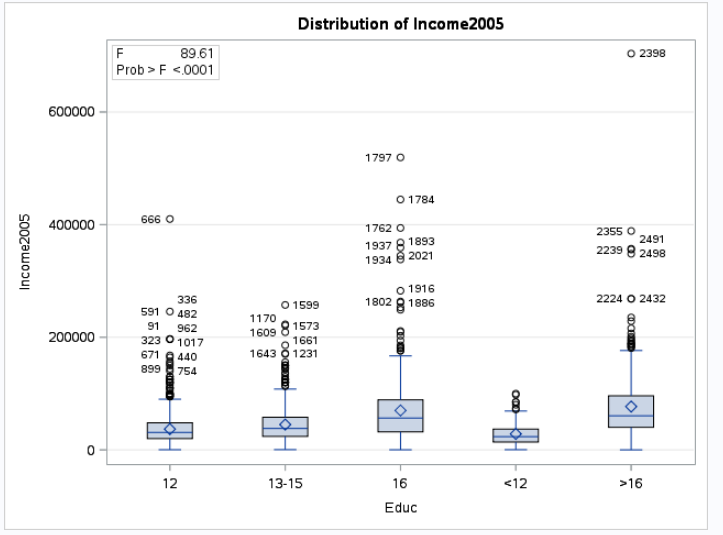
Very tricky. Was pulling my hair out trying to figure out where .1403 came from as I got a different value, but then tried it on the non-transformed data and got it.

Since we have 5 groups within the sample, I will compare the groups using Anova to see if at least one pair of means are different. To do so, the following assumptions must apply:

1. Normality: Similar to t-tools hypothesis testing, ANOVA is robust to this assumption. Extremely long-tailed distributions (outliers) or skewed distributions, coupled with different sample sizes (especially when the sample sizes are small) present the only serious distributional problems.
2. Equal Standard Deviations: This assumption is crucial, paramount, and VERY important.
3. The assumptions of independence within and across groups are critical. If lacking, different analysis should be attempted.

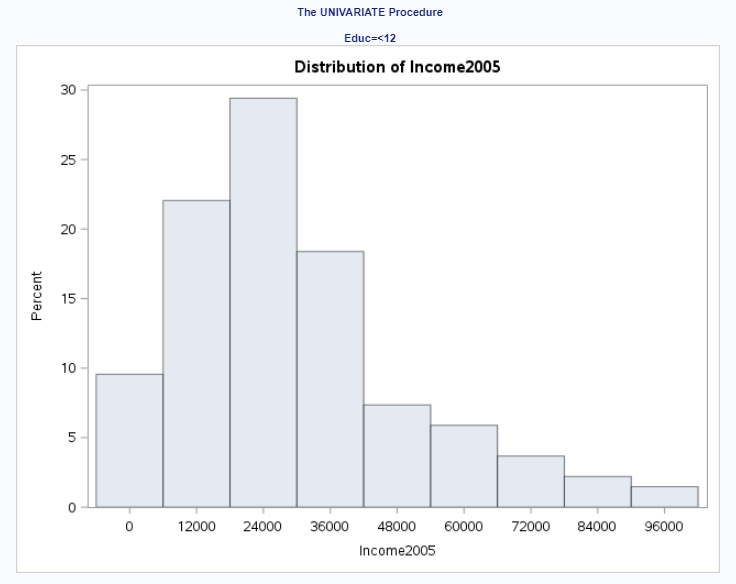
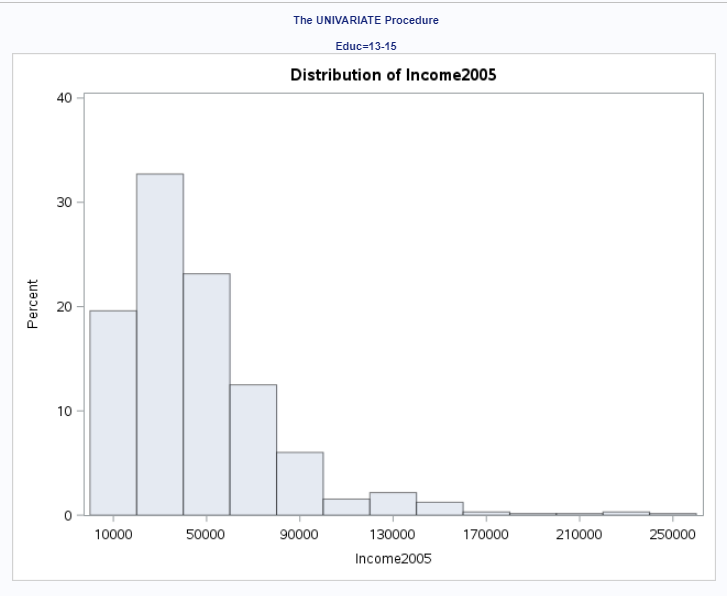
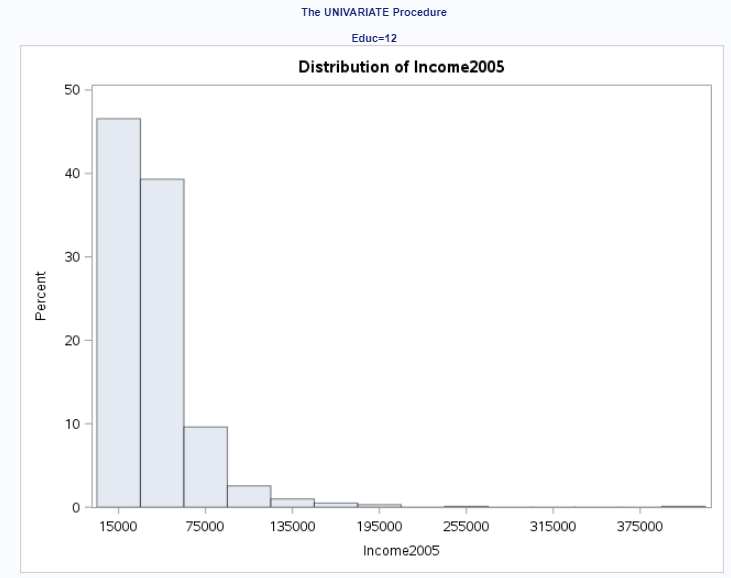
### Normality

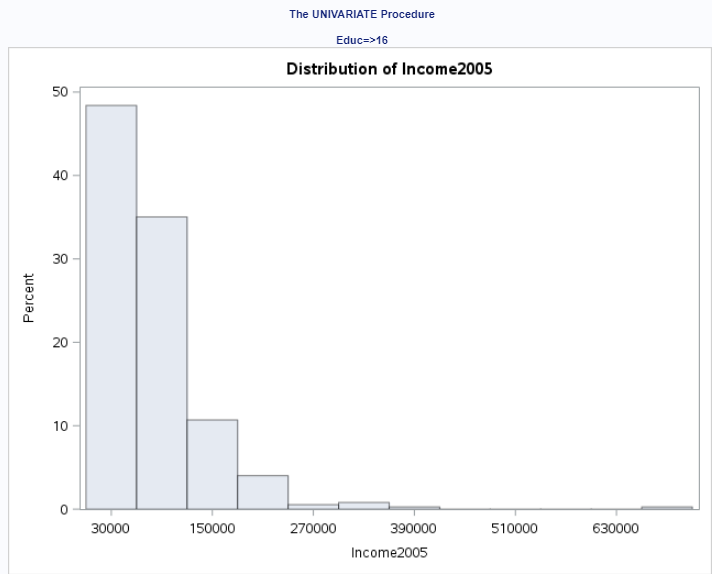
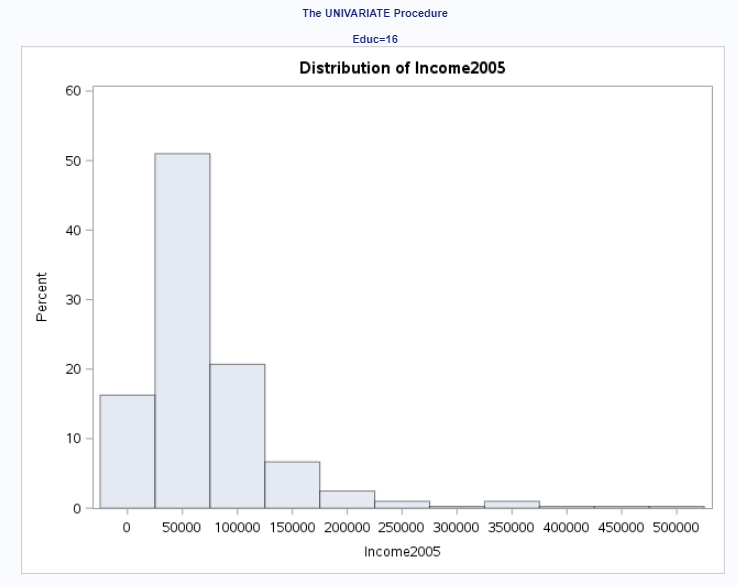
## Box and Whisker:



Interesting that the medians look pretty similar and boxplots do have a bit of symmetry.

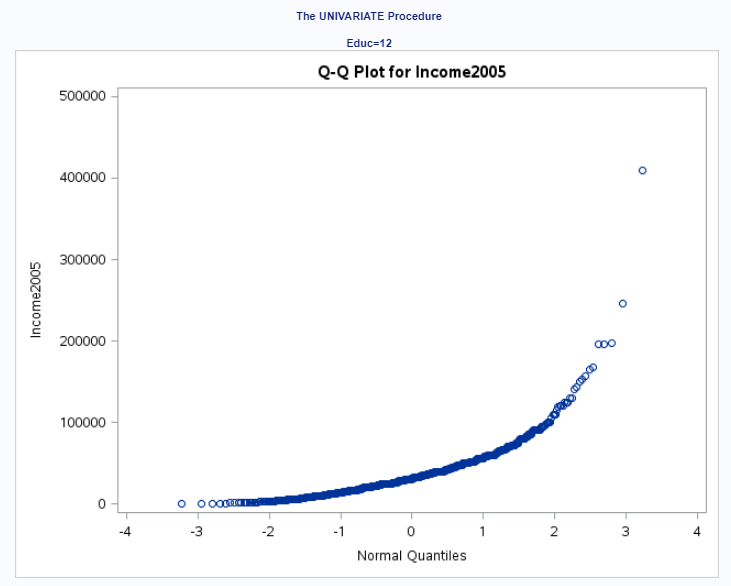
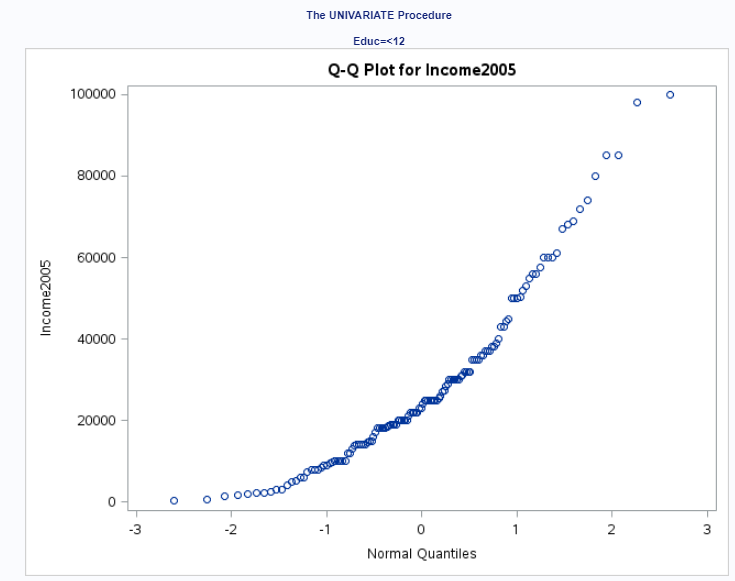
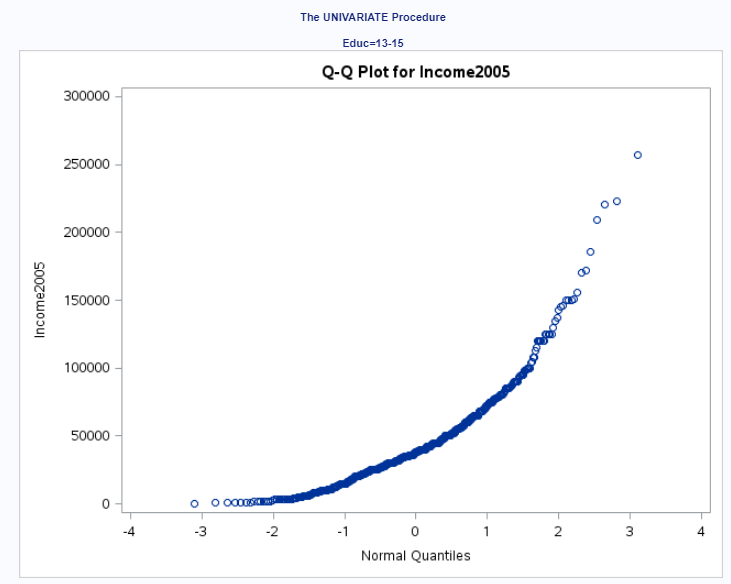
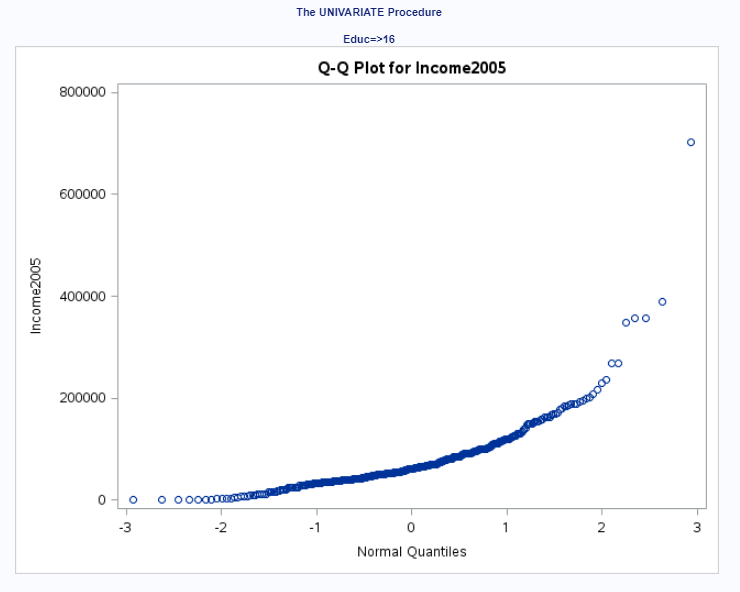
Histograms:

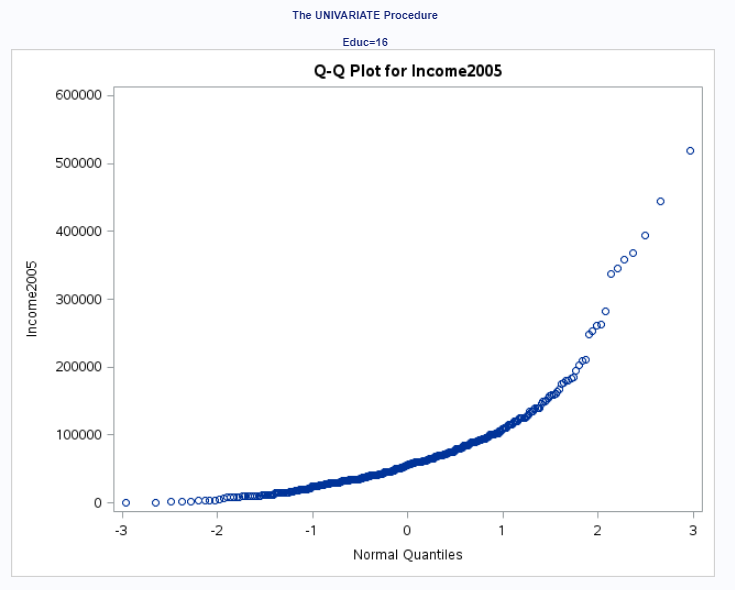
 



By looking at the charts above, many of the distributions don’t appear to be normal with Educ >= 16 having the most right skew while Educ<= 12 appearing to be the most normal compared to the other distributions that have a little right skew. We should transform the data to see if this fixes the normality assumption, but will look at QQPlots first before we do so.

QQPlots:



When looking at the QQPlots, the groups do not look normal for the most part when drawing a straight line. Many of the QQPlots data look like they are exponentially increasing.

Transforming Data:

Code:

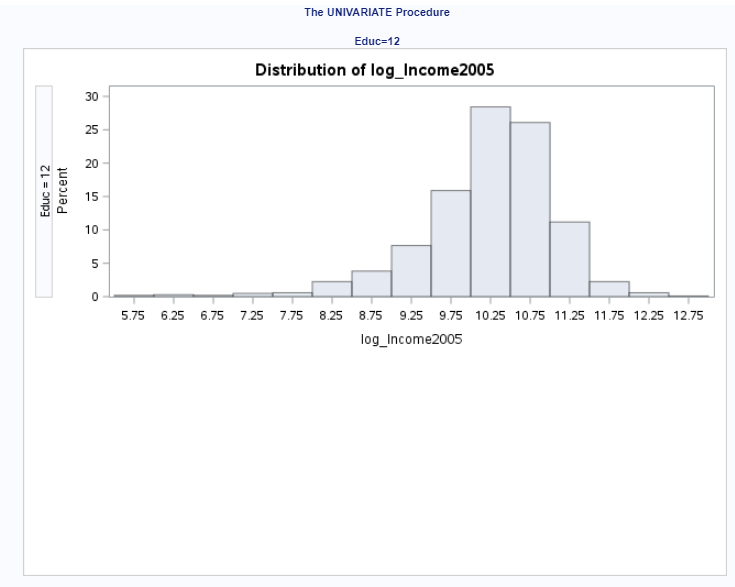
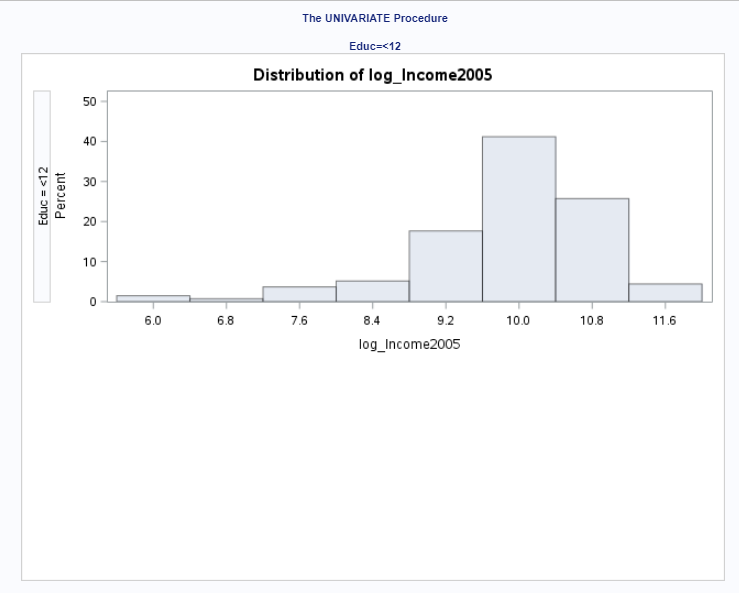
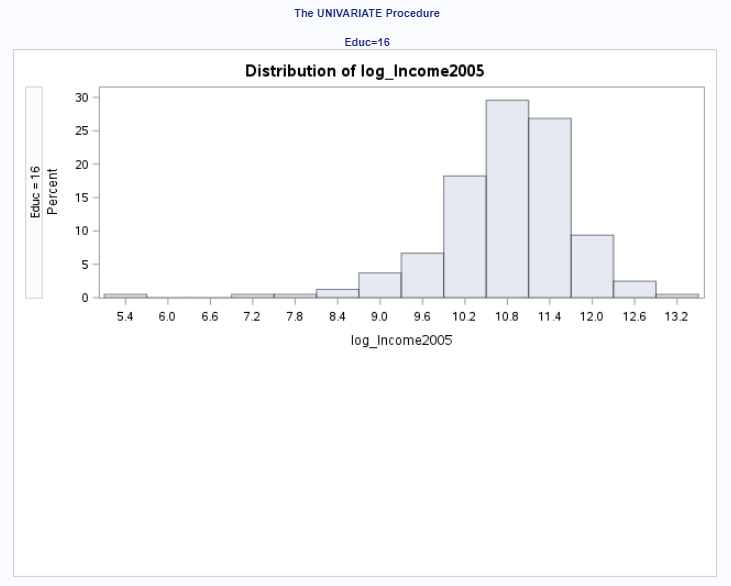
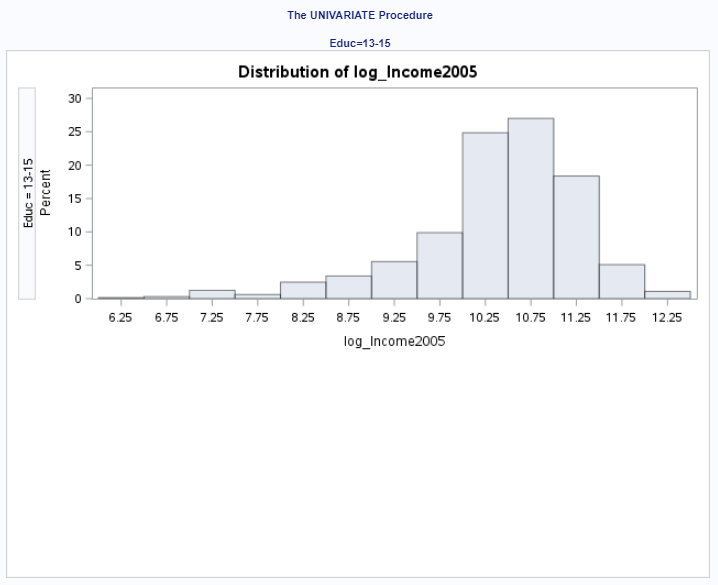
data WORK.transform;

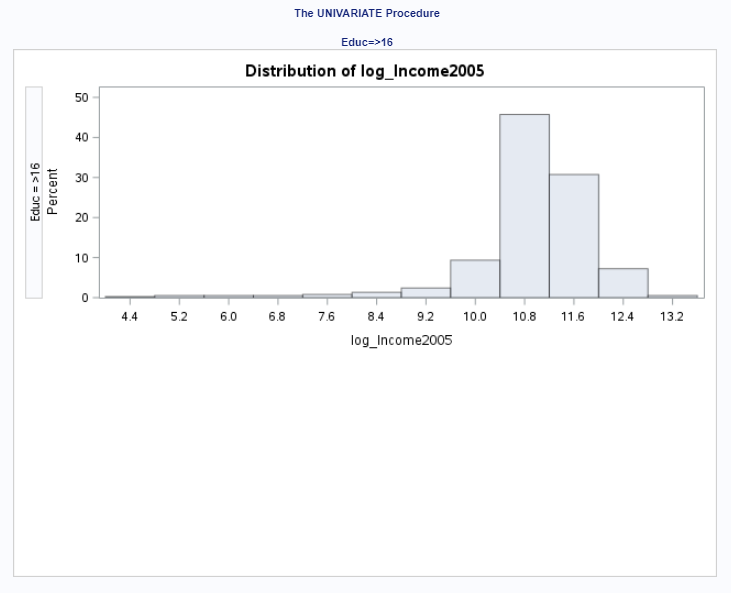
set work.edincomesort;

log\_Income2005=log(Income2005);

run;

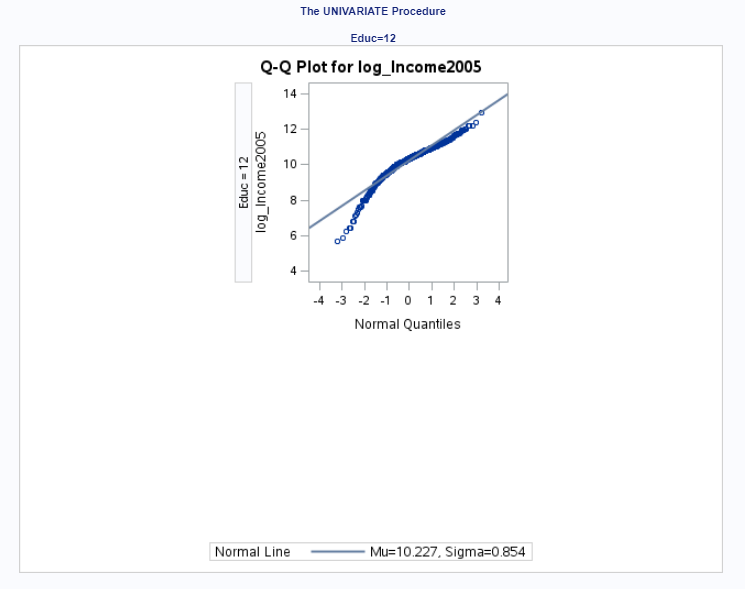
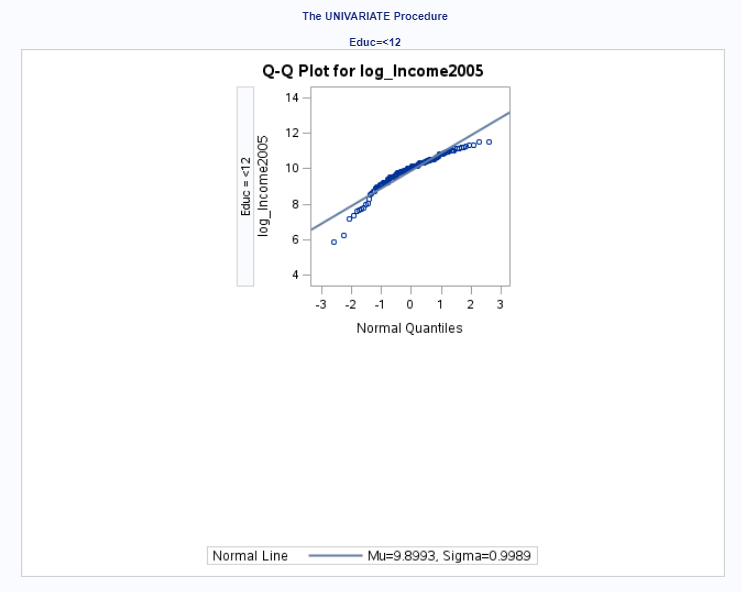
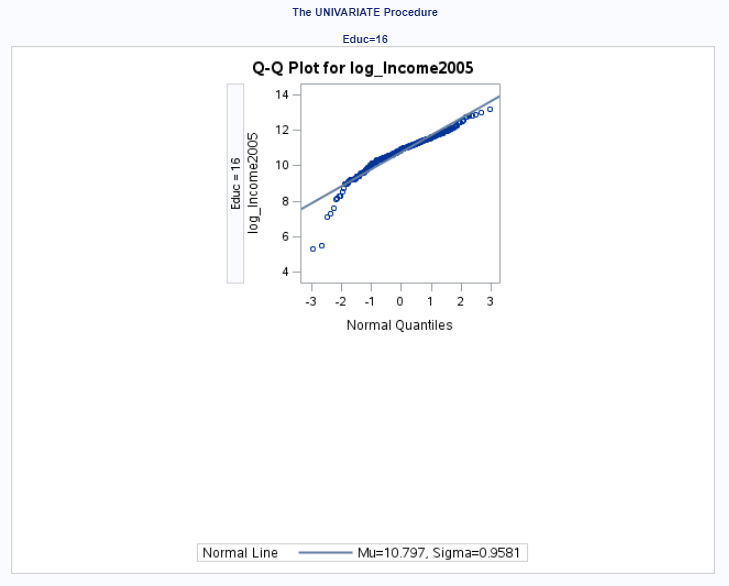
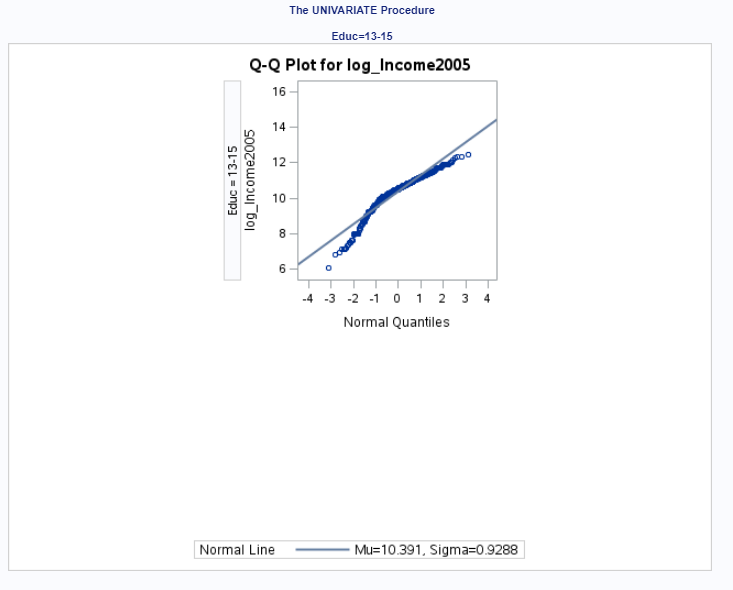
Histograms with log data:

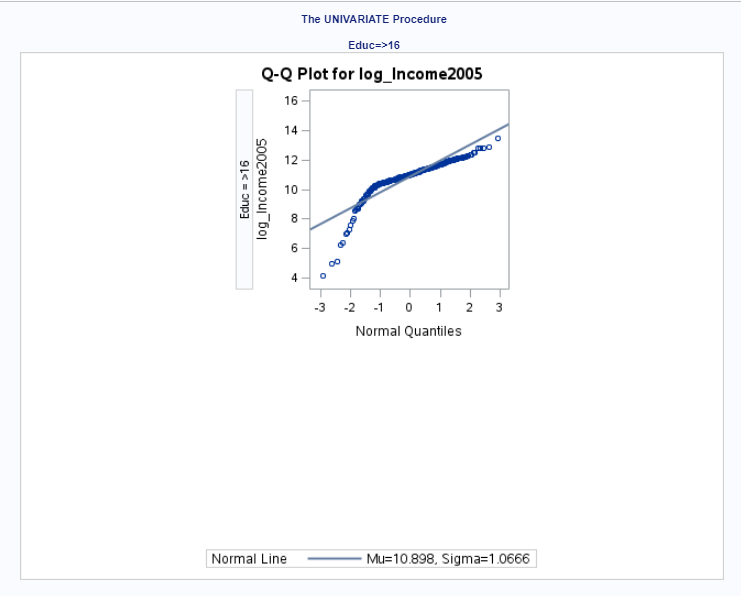
 



After doing the log transformation, it looks like a lot of the histograms are a bit more normal, but have too much left skew.

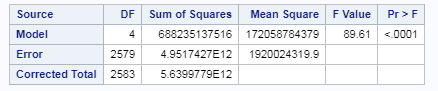
QQPlots:



These QQPlots show the issue with the left skew and don’t appear to be normal. Instruction said to proceed with Anova test anway. We will assume there is not enough visual evidence to suggest the standard deviations of the log transformed data are different.

## Equal Variances:



F –Test shows that there is a difference in the variances, but we will assume for now that variances are equal to finish this HW problem.

## Independence:

We will assume independence given that we don’t have any data to support that there are any dependencies with our observations.

## Step 1:

Ho: Reduced Model: u12 u>12 u13-15 uo uo

Ha: Full Model: u12 u>12 u13-15 u16 u>16

Code:

PROC IMPORT DATAFILE="/home/marinfamily1010/sasuser.v94/EDIncome/ex0525\_2\_2.xlsx"

OUT=WORK.EDINCOME

DBMS=XLSX

REPLACE;

RUN;

/\*\* Print the results. \*\*/

/\*

\* Received error that data must be in ascending order. Modifying data by Educ asc

\*/

proc sql;

create view work.edincomesort as

select \* from work.edincome

order by Educ asc;

run;

data WORK.transform;

set work.edincomesort;

log\_Income2005=log(Income2005);

run;

proc glm data = WORK.transform ;

class Educ;

model log\_Income2005 = Educ;

means Educ / HOVTEST=bf;

run;

proc sql;

create table work.problem2 as

select \* from work.transform order by Educ asc;

run;

proc sql;

update work.problem2 set Educ = 'o'

where Educ in ('>16','16');

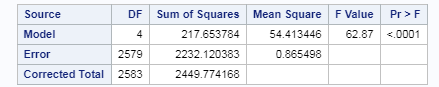
run;

## Step 2:

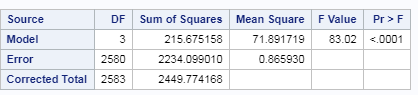
Critical Value not needed for ANOVA

## Step 3: 2.28

Full:



Reduced:



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DF | SS | MS | F | Pr>F |
| Model/Extra SS | 1 | 1.969717 | 1.969717 | 2.28 | .13 |
| Error/Residual/Full Model | 2579 | 2232.120383 | 0.865498 |  |  |
| Total (Reduced) | 2580 | 2234.099010 |  |  |  |

## Step 4:

P-Value: .13

## Step 5:

We fail to reject the null hypothesis

Step 6:

There is evidence to suggest (alpha = .05, p-value = .13) that the income means of those with Education of 16 years are not that different than those with 16 years or greater.

Scope:

This is an observational study where the members were selected by random from a very specific group. This study only applies to those youths in 1979 and who had paying jobs in 2005. Inference to the general population cannot be made.

3. Now, suppose that you cannot assume the standard deviations are the same (for both the original or log transformed data). Conduct another complete analysis of the question in Chapter 5, problem 25 in **Statistical Sleuth**. Answer the question, “How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others?” This question should be answered in at least 1 or 2 sentences after providing a **complete analysis** without the assumption of equal standard deviations for the logged data (or for the original data). Perform the test in SAS or R.

## Problem:

A random sample of 2,584 Americans who were selected for the NLS of youth in 1979 and who had paying jobs in 2005. These americans had different levels of education.

By how many dollars or by what percent does the mean or median for each of the last four categories exceed that of the next lowest category?

## Method:

In order to determine the question above, I’m going to group the 4 highest education brackets together and compare that group of 4 to the lowest education level.

Code to Merge Educ class together:

proc sql;

create table work.problem3 as

select \* from work.edincome

order by Educ asc;

run;

proc sql;

update work.problem3 set Educ = 'Other'

where Educ in ('>16','16','13-15','12');

run;

proc ttest data = work.problem3 alpha = .05 h0=0 sides = 2;

class Educ;

var Income2005;

run;

proc npar1way data = work.problem3;

class Educ;

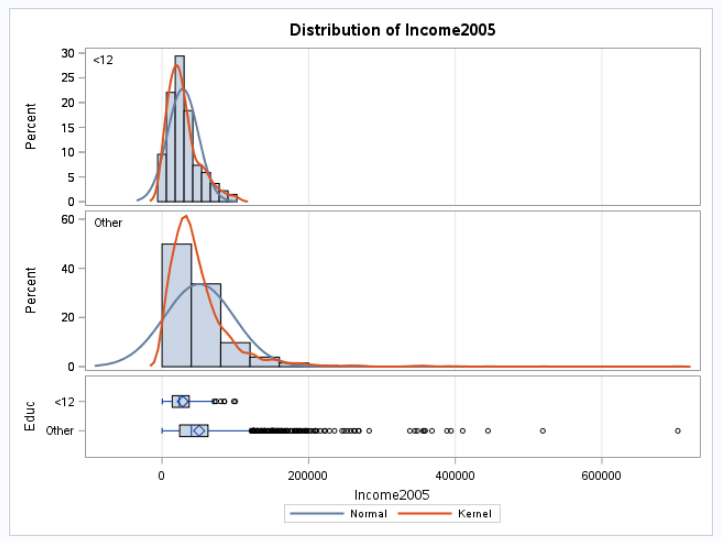
var Income2005;

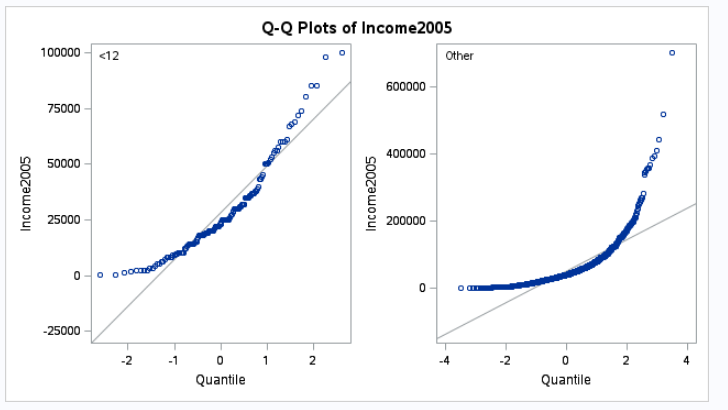
run;

## Normality Check:

First I’m going to do a normality check against the original data.

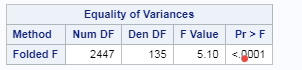
Original Data:



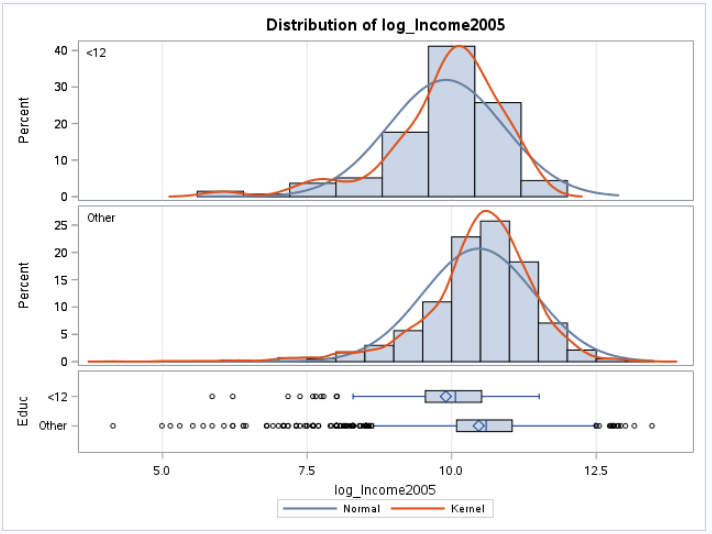


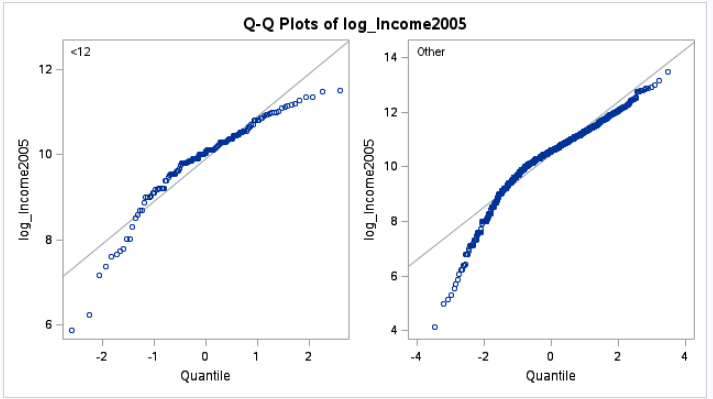
In the histograms, the data for <12 years appears to be normal, but the rest of the data (“Other”) representing higher eduction does not appear to be normal. The qqplots also show that the data do not appear to be normal.

Evidence shows that the variance is different.



Going to try this with the logged data as well.



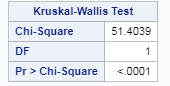


Histograms appear to be normal in center with log income data, but don’t appear to be that normal on the tails. QQPlots show this as well that beginning intervals are not normal. Since normaility doesn’t seem to pass for regular income data and log income data, I’m going to do a non-parametric test. Using a Kruskal-Wallis Test test with non-logged transformed data.

H0: Median12YearsEduction = Median>12YearsEducation

Ha: Median12YearsEduction  <> Median>12YearsEducation

P-value: <.0001



We reject the null hypothesis.

Conclusion: There is evidence to suggest to reject the null hypothesis that the median income between the two groups (12 Years of education and greater than 12 years of education) are not equal with a p-value of .001 using a Kruskal-Wallis Test.

Scope:

This is an observational study where the members were selected by random from a very specific group. This study only applies to those youths in 1979 and who had paying jobs in 2005. Inference to the general population cannot be made.