**Research Question:**

The file [ex0525](https://s3-us-west-2.amazonaws.com/smu-mds/prod/Experimental+Statistics+I/Assignments/Updated+4.2015/ex0525.csv) contains annual income in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of youth in 1979 and who had a paying job in 2005.

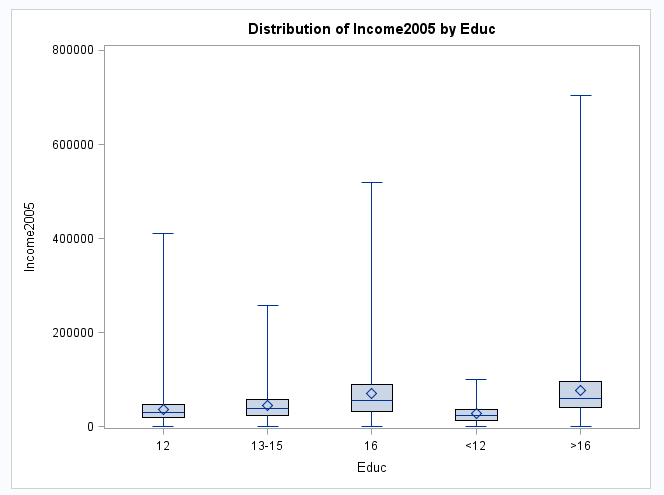
Examine these data with an ANOVA table. How strong is the evidence that at least one of the five population distributions is different from the others?

**Data Gathering and Treatments:**

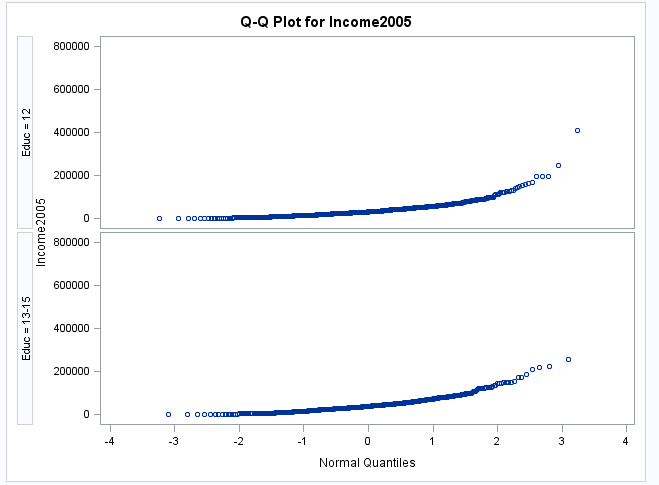
This is an observational study with pre-established groups of levels of education. A random sample was executed to select Americans with paying jobs and levels of education of <12 to >16. This is inclusive of bachelor’s degrees, master’s degrees, high school diplomas, associate degrees, and those who did not finish high school. There is no random effect considered for this analysis.

**Assumption Checks:**

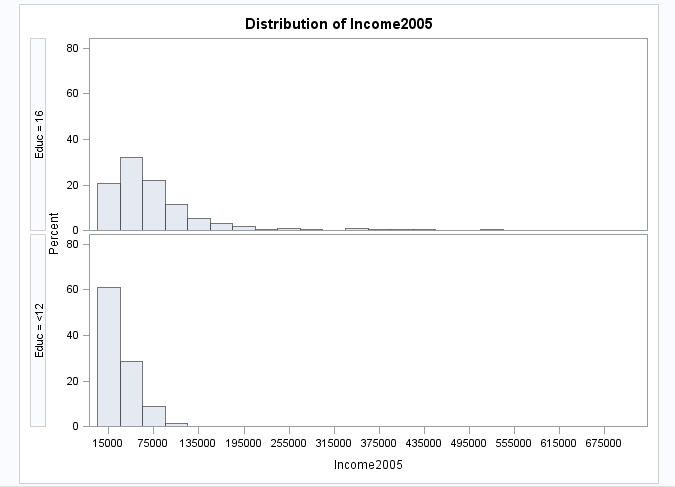
The original data in the National Longitudinal Study are skewed with a long right tail in most cases, as displayed by side-by-side box plots:



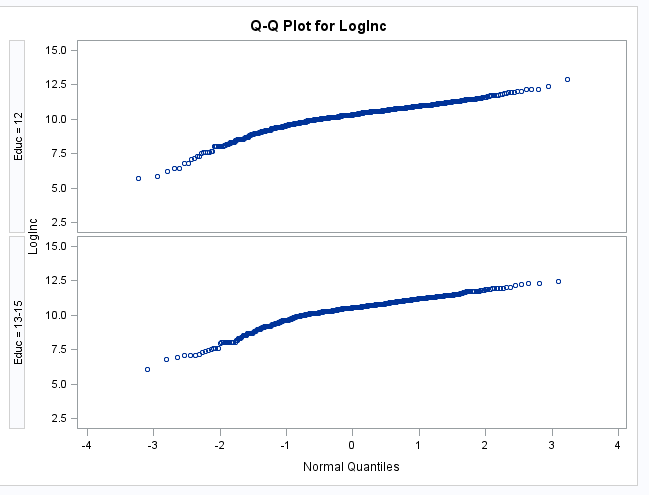
QQ plots confirm the non-normality of each education level category’s distribution, below is an example of Education = 12 and Education 13-15:



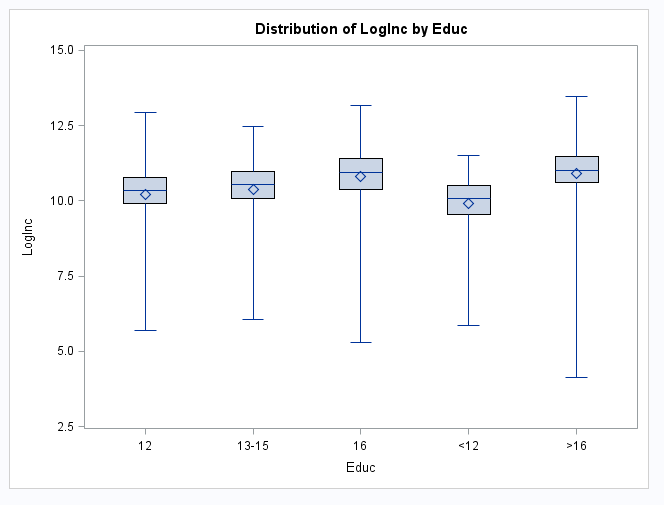
Histograms further confirm the non-normality, in this example Education level 16 and > 12 are displayed:



Due to the differences in spread and a lack of normality in their distributions, these data should be log transformed. Indeed, once log transformed, the data distributions are acceptable for further analysis using one-way ANOVA. Seen below are the same QQ plots mentioned previously for the original data set, now log transformed:

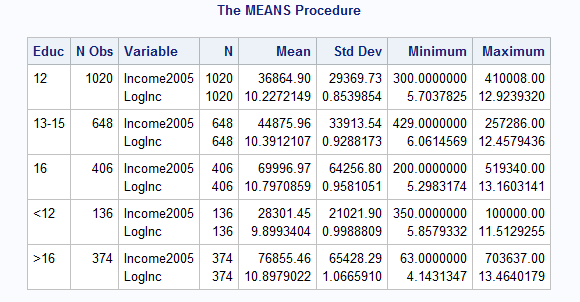


Log transformed data in box plot format show that our spreads are more acceptable, with less skewness and differences in spread between each category:



Given the data is now log transformed and visually meets the assumptions for one-way ANOVA, we are ready to begin our analysis.

**Descriptive Statistics:**

****

As evidenced by the log transformation, the standard deviation in each data set is greatly reduced. The original data set contains large ranges that required transformation. Early investigation reveals dramatic differences in means in the normal data set with potential outlier values at the bottom of each range. However, because our sample sizes are large, these outliers do not have large impact to our analysis, further, we cannot truly consider these values outliers or mistakenly recorded values.

Based on a visual inspection of means, we see subjects with more than 16 years of education outpace high school graduates by almost $40,000. The value of a bachelor’s degree based on means alone is $33,000 per year.

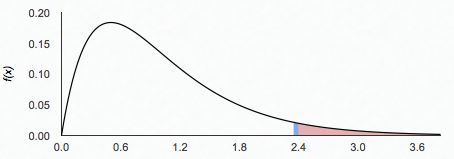
In order to confirm these observations, we need to test the mean variances using one-way ANOVA.

**Step 1: Hypothesis Test**



**Step 2: Identify Critical Values**

F critical for 4 and 2579 degrees of freedom with an alpha of .05 = 2.37

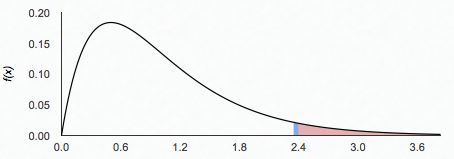


2.37

**Step 3: Identify Test Statistic**

**Based on one way ANOVA of log transformed education level and income, the f-statistic is :**





62.87

An F of 62.87 gives us an indication that our p-value is likely to be extremely small and we should be able to reject the null hypothesis that all means are equal.

**Step 4: Find the p-value**

The p-value resulting from the one-way ANOVA is



**Step 5: Make a Statistical Decision**

Based on a p value of <.0001, we reject the null hypothesis that the means of the groups are equal.

**Step 6: Make a Human Conclusion**

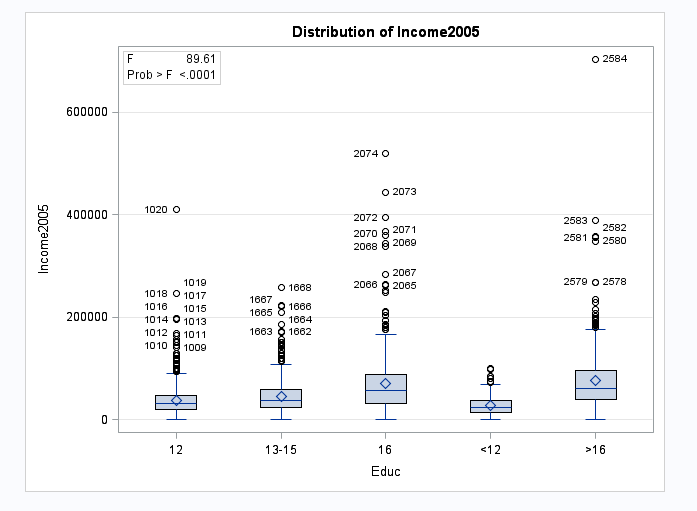
There is evidence to show that the means of income across all of the education groupings are not equal.

2) Now, instead of performing an ANOVA, use a permutation test. How strong is the evidence that at least one of the five population distributions is different from the others?

**We consider two scenarios here:**

If we consider the original data set, a permutation test would be more accurate. The one way ANOVA’s F test is dependent upon normal distributions, which, based on our exploratory data analysis, is NOT confirmed with box plots, QQ plots and histograms on the **original** data set. The resulting ANOVA is still significant with a p <.0001, on the **log transformed** data set using a Kruskal Wallis test and a standard One-Way ANOVA. However, when considering the original data set and its lack of normality, a permutation test is a better option to test for a difference in the groups because permutation tests are not distribution dependent.

As seen below, the **original** data set is highly variable between groups:



We would not want to use a t-test on this data set. Annual incomes in the 16 years of education group range from almost zero to over $500,000. That is a LARGE spread, in desperate need of log transformation or a distribution-free analysis. Below, we step through a permutation test using Kruskal Wallis and the original data set:

**Step 1: Hypothesis Test**



**Step 2: Identify Critical Values**

Chi Squared critical for 4 degrees of freedom with an alpha of .05 = 9.49.

However, we are not using a chi squared distribution, we are building a permutation distribution from regroupings of the data between five groups.

**Step 3: Identify Test Statistic**

Our chi squared test statistic, our test statistic value is 349.45.

**Step 4: Find the p-value**

We will now try and find values by regrouping the ranks 10,000 times and building a permutation distribution. Our p value will be the number of chi squared statistics greater than or equal to our test statistic of 349.45.

The p-value resulting from the Kruskal Wallis permutation test is:



**Step 5: Make a Statistical Decision**

Based on a p value of <.0001, we reject the null hypothesis that the means are equal. Under the null hypothesis in this study, there is a very small chance we’d obtain a value as extreme or more extreme than a chi squared statistic of 349.45.

**Step 6: Make a Human Conclusion**

There is evidence to show that the means of income across all of the education groupings are not equal.

**NOTE:**

I chose to use the original data set under the Kruskal Wallis test. Once the income data is log transformed, it is normal. We would not use a non-parametric test in this case as we can apply t-tools, which are more accurate under the normal curve.

The exact parameter in the KW test allows permutations; the chi squared statistic is used as part of the permutation test.

SAS Code:

**data** income (replace = yes);

infile '\\Client\C$\Users\patrickcorynichols\Desktop\Data Science\Stats\Data Sets\ex0525.csv' DLM = ',' FIRSTOBS = **2**;

INPUT Subject $ Educ $ Income2005;

LogInc = LOG(Income2005);

WHERE educ = '12';

**RUN**;

**PROC** **SORT** data = income;

by Educ Income2005 LogInc;

**RUN**;

**proc** **boxplot**;

plot Income2005\*Educ;

**RUN**;

**proc** **univariate** data = income;

CLASS EDUC;

VAR Loginc;

QQPLOT;

HISTOGRAM;

**RUN**;

**PROC** **MEANS** data = income;

CLASS educ;

**RUN**;

**PROC** **ANOVA** DATA = income ORDER = DATA;

CLASS EDUC;

MODEL Loginc=Educ;

MEANS EDUC/TUKEY CLDIFF;

**RUN**;

**QUIT**;

**PROC** **NPAR1WAY** WILCOXON data = income;

CLASS EDUC;

VAR Income2005;

EXACT / n = **10000**;

**RUN**;

**proc** **glm** data=income;

class EDUC;

model 2005Income = EDUC;

OUTPUT OUT = FITDATA P = YHAT R = RESID;

**PROC** **GPLOT**;

PLOT RESID\*EDUC;

PLOT RESID\*YHAT;

**RUN**;