Cory Nichols Experimental Statistics I

Education and Future Income

**Research Question:**

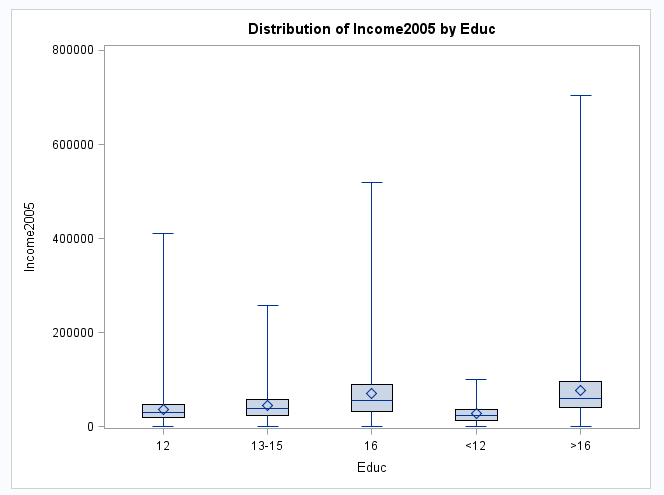
2584 Americans with paying jobs were selected at random for the National Longitudinal Survey of Youth in 1979. How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others? By how many dollars or by what percent does the mean OR median for each of the last four categories exceed that of the next lowest category?

**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.

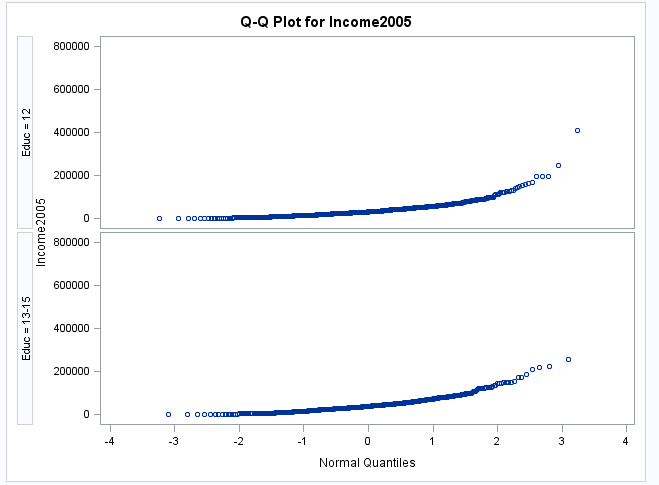
**Assumption Checks:**

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

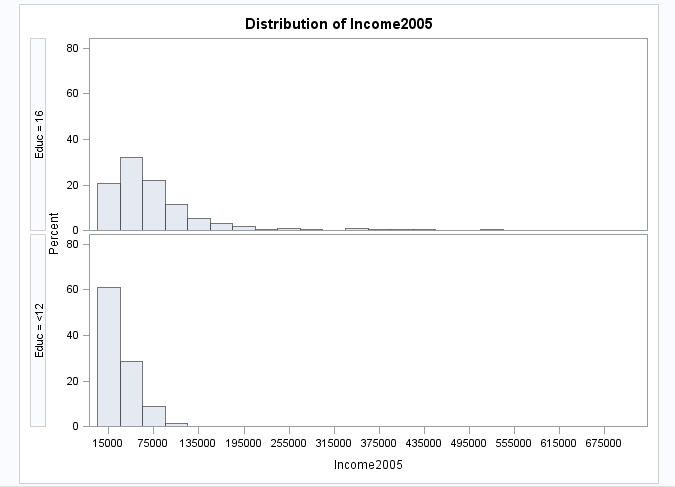


QQ plots confirm the non-normality of each education level category’s distribution, below is an example of Education = 12 and Education 13-15:

Cory Nichols Experimental Statistics I

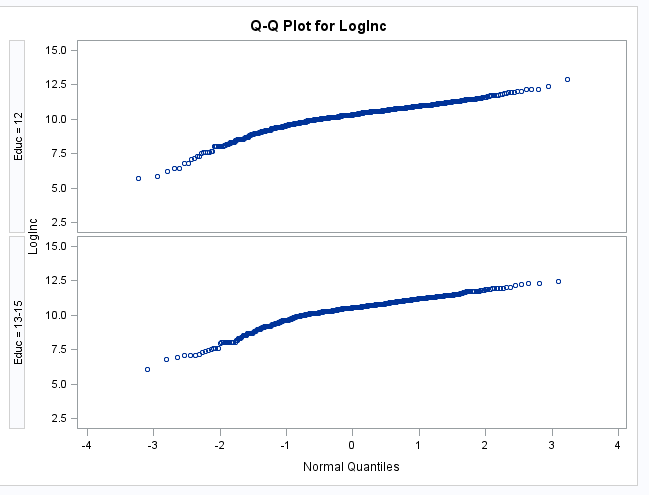


Histograms further confirm the non-normality, in this example Education level 16 and > 12 are displayed:



Cory Nichols Experimental Statistics I

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 previous for the original data set, now log transformed:



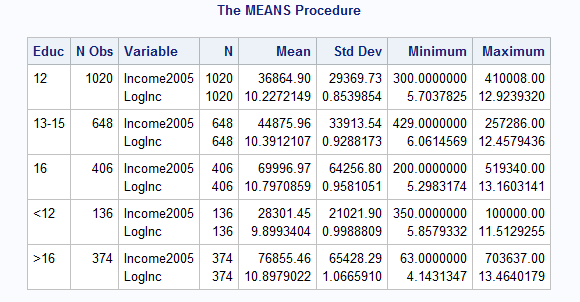
Log transformed data in box plot format show that our spreads are more acceptable, with less skewness and differences in spread between each category:

Cory Nichols Page 1 6/4/15

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

Cory Nichols Experimental Statistics I

**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.

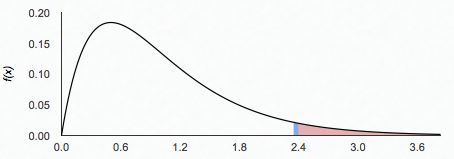
Cory Nichols Experimental Statistics I

**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

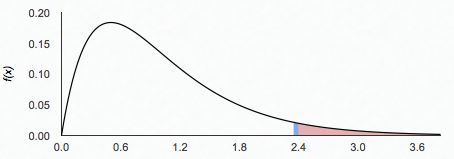


2.37

**Step 3: Identify Test Statistic**

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





62.87

Cory Nichols Experimental Statistics I

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 are equal.

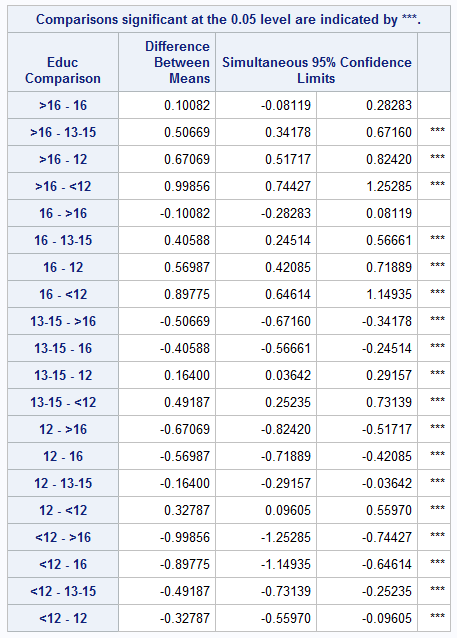
**Step 6: Make a Human Conclusion**

There is evidence to show that the means of income across all of the education groupings are not equal. Further analysis is required to identify between group differences, provided below by Tukey’s pairwise comparison method.

Cory Nichols Experimental Statistics I

**Additional Analysis:**

Based on Tukey’s method for comparison of means across all pairwise groupings, we see that all differences in means are significant except for >16 : 16 years of education pairwise comparison. All values are the natural log of their original values. When back converted, we see that bachelor’s degree holders have roughly a 96% higher income than high school graduates (log value of .67069). In our most extreme example, we can see that a person with post-baccalaureate education earns 171% (natural log .99856) more income annually than a person without a high school diploma with a 95% confidence interval of 110% to 250%. Education does not CAUSE higher income. This is an observational study therefore no causation can be prescribed. There are many other confounding variables to consider. Inference can be made to the education populations represented here with much caution because subjects were randomly selected.



Cory Nichols Experimental Statistics I

**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);

**RUN**;

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

by Educ Income2005 LogInc;

**RUN**;

**proc** **boxplot**;

plot Loginc\*Educ;

**RUN**;

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

CLASS EDUC;

VAR Loginc;

QQPLOT;

**RUN**;

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

CLASS EDUC;

VAR LogInc;

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**;

ODS RTF;

ODS GRAPHICS ON;

**PROC GLM** data=income;

class EDUC;

model LOGINC = EDUC;

OUTPUT OUT = FITDATA P = YHAT R = RESID;

**PROC** **GPLOT**;

PLOT RESID\*EDUC;

PLOT RESID\*YHAT;

**RUN**;

ODS RTF CLOSE;

ODS GRAPHICS OFF;