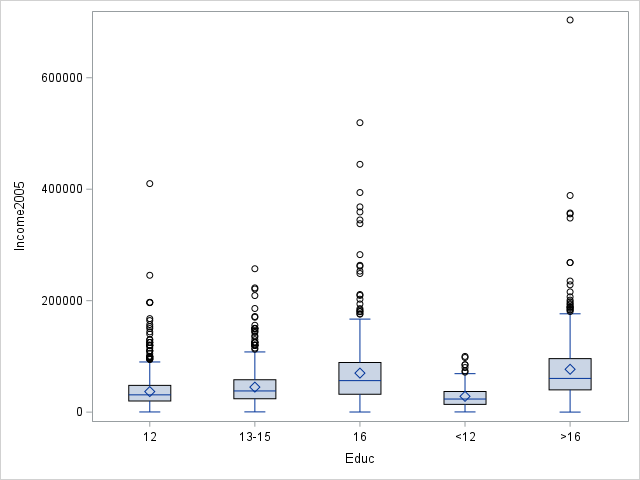
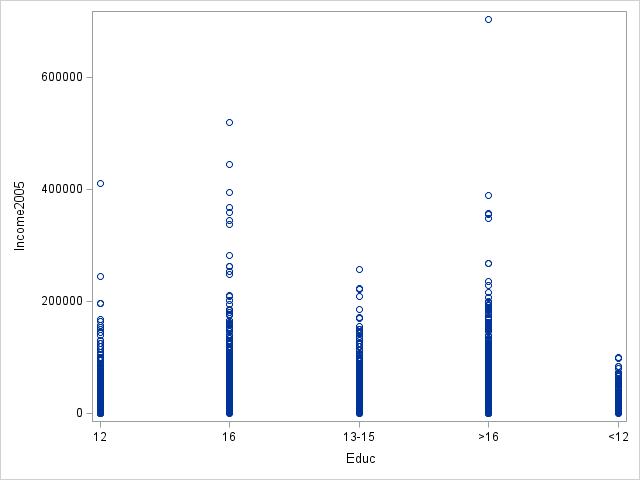
Text

Description automatically generatedSAS Sample for ex0525.csv

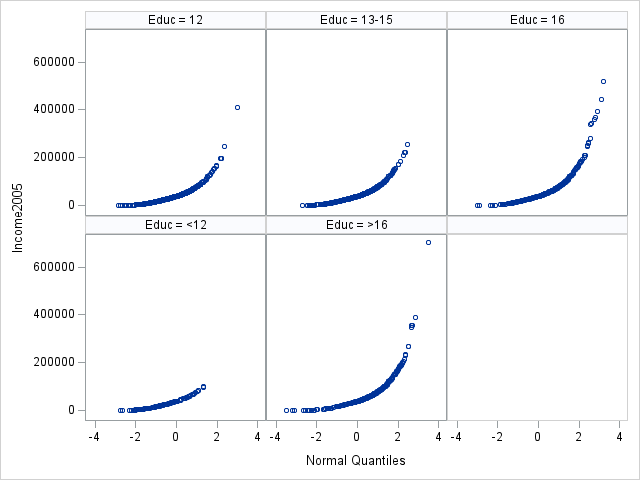
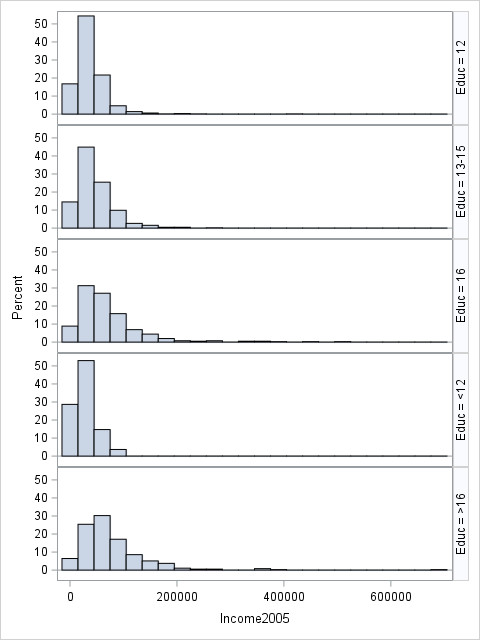
Both the plots below show that there are lots of data on only one side, meaning that the data is not normal.

**

Text

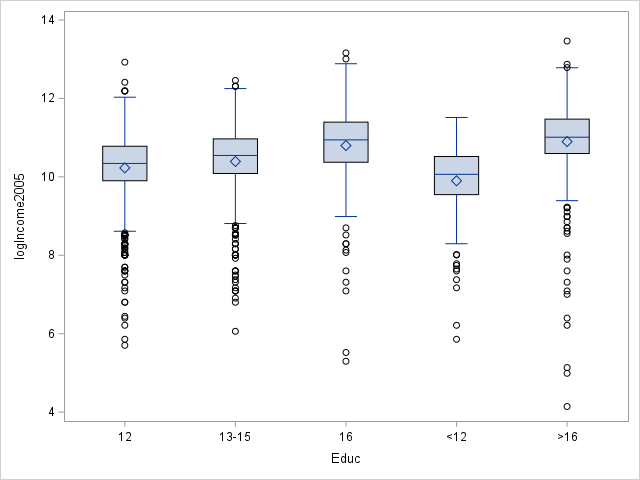
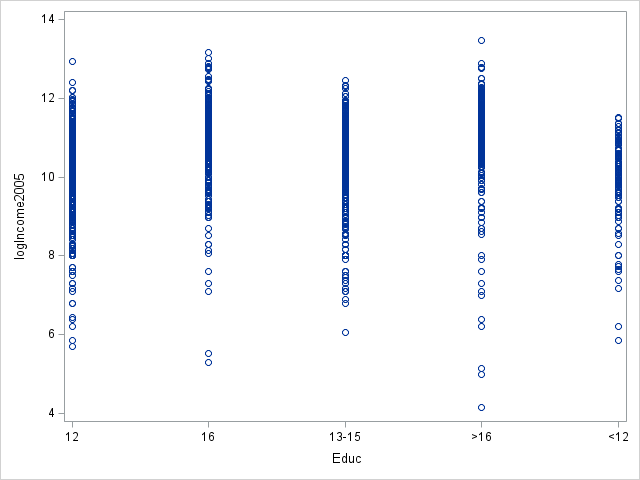
Description automatically generated

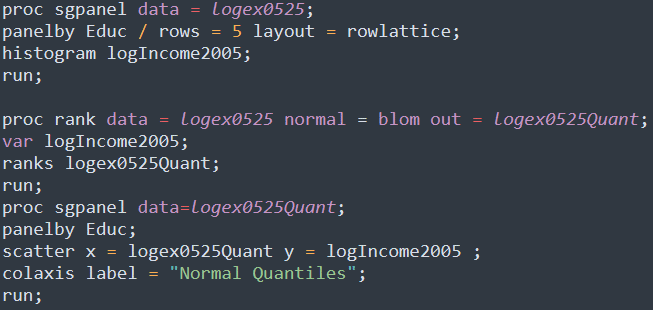
According to the histograms below, the data is strongly skewed to the right. Likewise for the QQ plots, which tells that the data does not have equal variances. Therefore we cannot use the raw data.

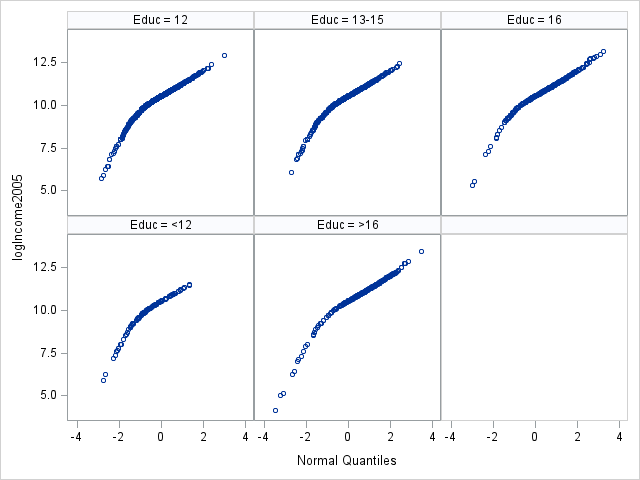
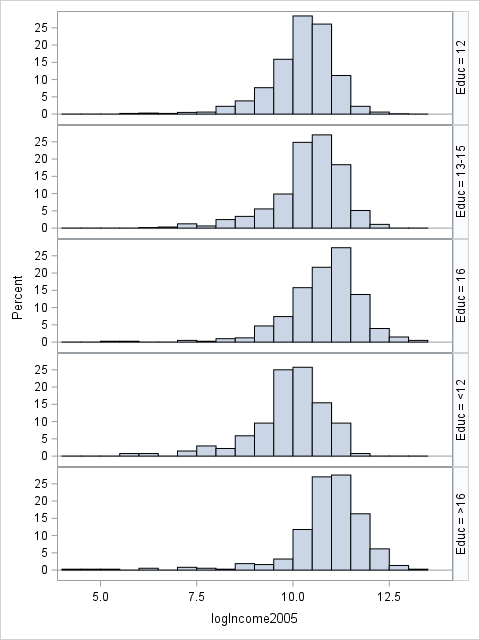
**

Text

Description automatically generatedWhen the data is log transformed, both the plots below show that the groups now have similar sizes and closely related means and medians.

**

The data is now more normalized with a bigger center. Likewise for the QQ plots, the data is now more or less linear towards normal quantiles. Therefore we can assume that our data now has normality and equal variances.

**

**Problem Statement** We need to test if at least one of the five distributions of people with different years of education is different from the others.

**Assumptions** By log transforming the data, it has more normality and equal variances. The data can also be assumed independent. Therefore, all the necessary assumptions for the ANOVA test is met by the log transformed data.

**Hypothesis**

Table

Description automatically generated

alpha = 0.05

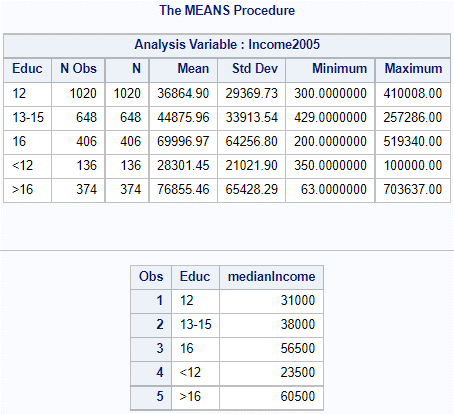
Text

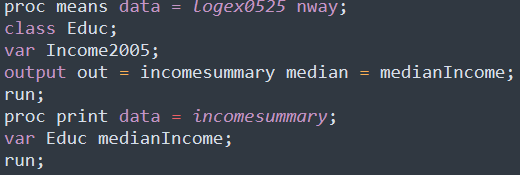
Description automatically generated with low confidence



**Results from SAS**

Since p-value is less than , we reject the null hypothesis.

**Conclusion**There is enough evidence to conclude that at least of the distributions is different from the others (p-value < 0.0001 from ANOVA). Therefore, we will now move on to question 2 and compare the medians between <12 and 12 years of school, 12 and 13-15 years of school, 13-15 and 16 years of school, and 16 and >16 years of school.



<12 and 12 years: (31000-23500)/23500 = 31.9% increase

12 and 13-15 years: (38000-31000)/31000 = 22.6% increase

13-15 and 16 years: (56500-38000)/38000 = 48.7% increase

1. and >16 years: (60500-56500)/56500 = 7.1% increase

**Scope**

Text

Description automatically generatedAs this was an observational study, we cannot make causal inferences about how higher education can mean higher income. However, we can make inferences about the sampled population because the NLSY is a random sample.

**Problem Statement**

We need to test if people with 16 years of education or more than 16 years of education have different distributions of income.

**Assumptions**

From the assumptions testing above with the log transformed data, the three necessary assumptions normality, equal variances and independence are met, therefore the ANOVA test can be used.

**Hypothesis (Extra Sum of Squares F-test)**

α = 0.05

Table

Description automatically generatedTable

Description automatically generatedText

Description automatically generated



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F | Pr > F |
| Model | 2 | 12.636 | 6.318 | 7.304 | 0.0006868517 |
| Error | 2579 | 2232.120 | 0.865 |  |  |
| Corrected Total | 2581 | 2244.756 |  |  |  |



**Results from SAS**

Since p-value is less than alpha, we reject the null hypothesis.

**Conclusion**

There is enough evidence to conclude that the distribution of income of people with 16 years of education is different than that of people with more than 16 years of income.

**Scope**

As this was an observational study, we cannot make causal inferences about how higher education can mean higher income. However, we can make inferences about the sampled population because the NLSY is a random sample.

**Problem**

We need to test if at least one of the five distributions of people with different years of education is different from the others, assuming that there is no equal standard deviation for the logged data.

**Assumptions**

From the assumptions testing above with the log transformed data, we can assume that normality and independence, but we cannot assume equal variances. Therefore the regular ANOVA test is not appropriate. We can instead use the Welch’s ANOVA test which does not assume equal standard deviations.

Table

Description automatically generated**Hypothesis**

alpha = 0.05

Text

Description automatically generated**Results from SAS**

Since p-value is less than alpha, we reject the null hypothesis.

**Conclusion**

There is enough evidence to conclude that there is at least one of the distributions is different from the others.

**Scope**

As this was an observational study, we cannot make causal inferences about how higher education can mean higher income. However, we can make inferences about the sampled population because the NLSY is a random sample.

**Problem Statement** We need to determine which pairs of education category means differ and by how much money or percentage.

We will continue to use the logged data, because a Tukey-Kramer test is most appropriate for this task since the test requires a normal distribution.

Table

Description automatically generatedText

Description automatically generated

Table

Description automatically generated

**Chart, line chart

Description automatically generated**

From the plot above, we can see that every other pairs is statistically significant at showing difference in means, except the pair between >16 and 16. Specifically, there is no statistical significant difference between people with a Bachelors and people with graduate-level education.

**Problem Statement** We need to test if there is a difference in income distributions between people with high school education (12 years) versus the other education categories using Dunnett's Test.

Text

Description automatically generated with medium confidence

Table

Description automatically generated

Using the group of people with high school education as control, every other education category is has a statistically significant difference.