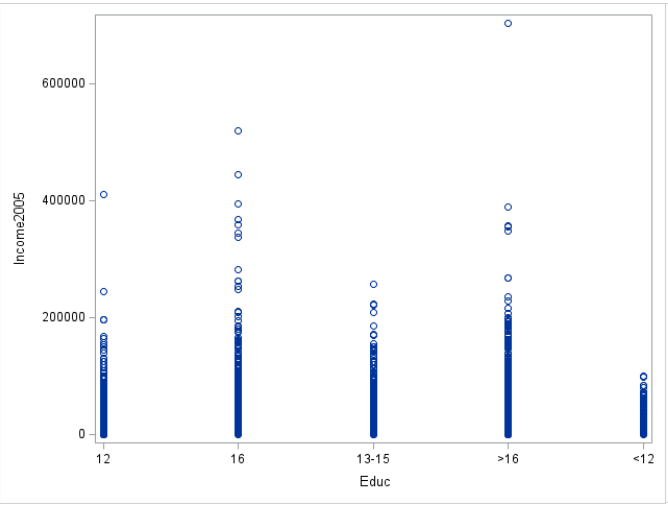
Duy Nguyen

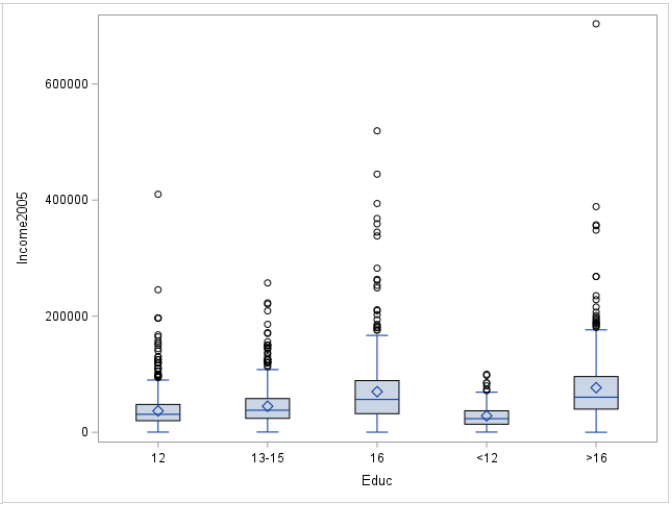
SAS Sample for ex0525.csv

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

```
proc sgplot data = ex0525;
scatter x = Educ y = Income2005;
run;

proc sgplot data = ex0525;
vbox Income2005 / category = Educ;
run;
```

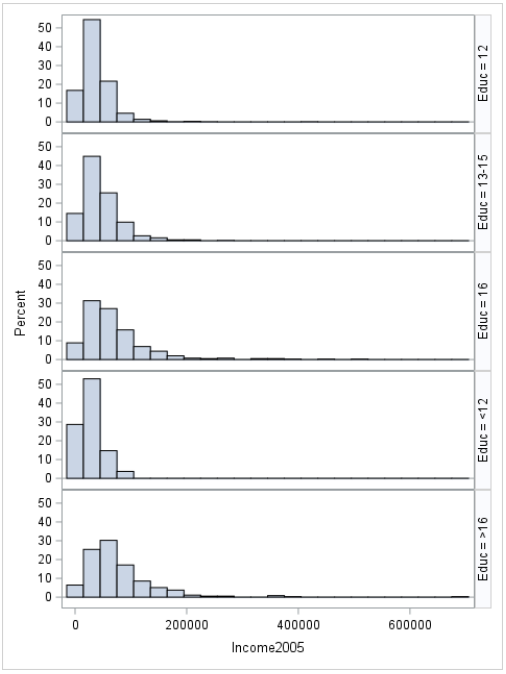


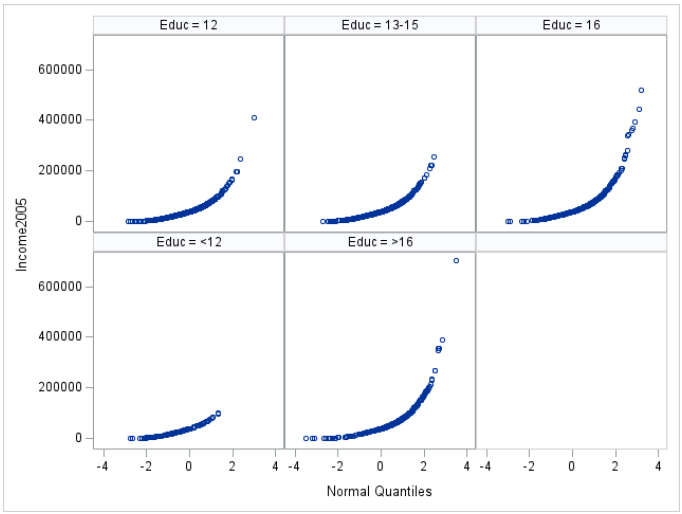


According to the histograms below, the data is strongly skewed to the right. Likewise for the QQ plots, which tells that the data does not have equal variances. Therefore we cannot use the raw data.

```
proc sgpanel data = ex0525;
panelby Educ / rows = 5 layout = rowlattice;
histogram Income2005;
run;

proc rank data = ex0525 normal = blom out = ex0525Quant;
var Income2005;
ranks Edu_Quant;
run;
proc sgpanel data = ex0525Quant;
panelby Educ;
scatter x = Edu_Quant y = Income2005;
colaxis label="Normal Quantiles";
run;
```



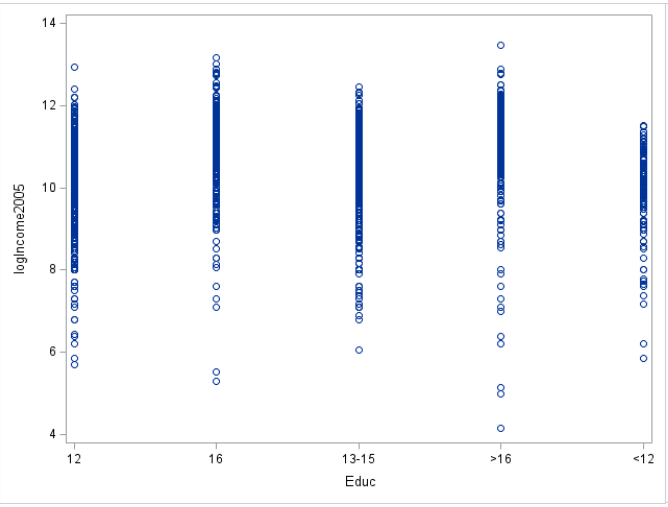


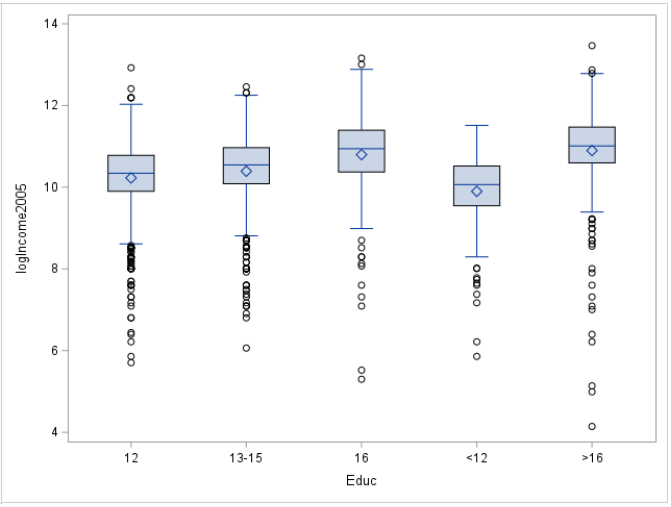
When the data is log transformed, both the plots below show that the groups now have similar sizes and closely related means and medians.

```
data logex0525;
set ex0525;
logIncome2005 = log(Income2005);
run;

proc sgplot data = logex0525;
scatter x = Educ y = logIncome2005;
run;

proc sgplot data = logex0525;
vbox logIncome2005 / category = Educ;
run;
```

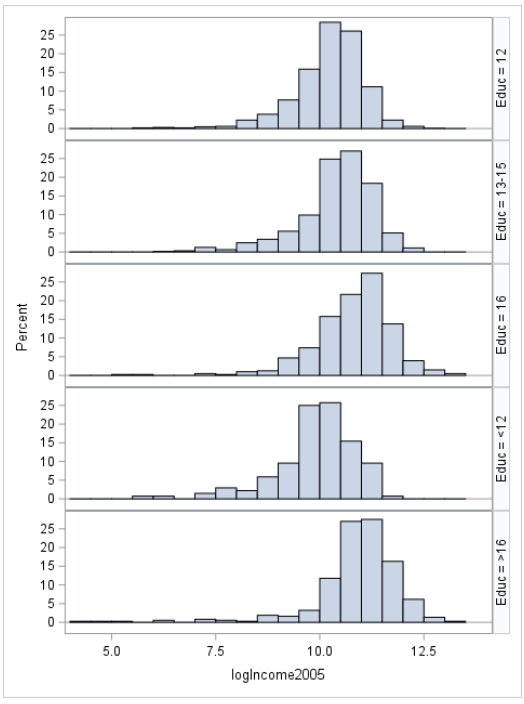


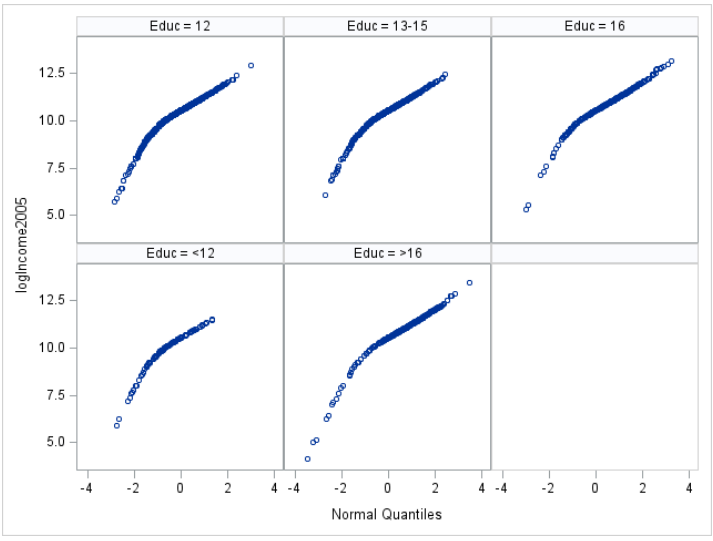


The data is now more normalized with a bigger center. Likewise for the QQ plots, the data is now more or less linear towards normal quantiles. Therefore we can assume that our data now has normality and equal variances.

```
proc sgpanel data = logex0525;
panelby Educ / rows = 5 layout = rowlattice;
histogram logIncome2005;
run;

proc rank data = logex0525 normal = blom out = logex0525Quant;
var logIncome2005;
ranks logex0525Quant;
run;
proc sgpanel data=logex0525Quant;
panelby Educ;
scatter x = logex0525Quant y = logIncome2005 ;
colaxis label = "Normal Quantiles";
run;
```





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

$Ho: Reduced\ Model: mean_{<12} = mean_{12} = mean_{13-15} = mean_{16} = mean_{>16}$

$Ha: Full\ Model:\ at\ least\ one\ mean\ is\ different$ alpha = 0.05

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

```
proc glm data = logex0525;
class Educ;
model logIncome2005 = Educ;
run;
```

Results from SAS

$F - value\ =\ 62.87$
$P - value\ <\ 0.0001$

Since p-value is less than α, 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.

```
proc means data = logex0525 nway;
class Educ;
var Income2005;
output out = incomesummary median = medianIncome;
run;
proc print data = incomesummary;
var Educ medianIncome;
run;
```

The MEANS Procedure

Analysis Variable : Income2005						
Educ	N Obs	N	Mean	Std Dev	Minimum	Maximum
12	1020	1020	36864.90	29369.73	300.0000000	410008.00
13-15	648	648	44875.96	33913.54	429.0000000	257286.00
16	406	406	69996.97	64256.80	200.0000000	519340.00
<12	136	136	28301.45	21021.90	350.0000000	100000.00
>16	374	374	76855.46	65428.29	63.0000000	703637.00

Obs	Educ	medianIncome
1	12	31000
2	13-15	38000
3	16	56500
4	<12	23500
5	>16	60500

<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

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

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.

$$R^2 = 0.088$$

R-Square	Coeff Var	Root MSE	logIncome2005 Mean
0.088846	8.913094	0.930322	10.43770

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)

$H0: median_{16} = median_{16}$

$Ha: median_{16} \neq median_{16}$ α = 0.05

```
data logex0525group;
set logex0525;
Others = Educ;
if Educ = "16" then Others = "a";
if Educ = ">16" then Others = "a";
run;

proc glm data = logex0525group;
class Others;
model logIncome2005 = Others;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	217.653784	54.413446	62.87	<.0001
Error	2579	2232.120383	0.865498		
Corrected Total	2583	2449.774168			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	205.017925	102.508962	117.86	<.0001
Error	2581	2244.756243	0.869723		
Corrected Total	2583	2449.774168			

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			

```
> pf(7.304, 2, 2579, lower.tail = FALSE)
[1] 0.0006868517
```

Results from SAS

$F - value = 7.304$
$P - value = 0.0006868517$

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.

Hypothesis (Extra Sum of Squares F-test)

$Ho: Reduced\ Model: mean_{<12} = mean_{12} = mean_{13-15} = mean_{16} = mean_{>16}$

$Ha: Full\ Model:\ at\ least\ one\ mean\ is\ different$ alpha = 0.05

Results from SAS

$F - value\ = 56.59$
$P - value\ < 0.0001$

```
proc glm data = logex0525;
class Educ;
model logIncome2005 = Educ;
means Educ / welch;
run;
```

Welch's ANOVA for logIncome2005

Source	DF	F Value	Pr > F
Educ	4.0000	56.59	<.0001
Error	673.9		

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.