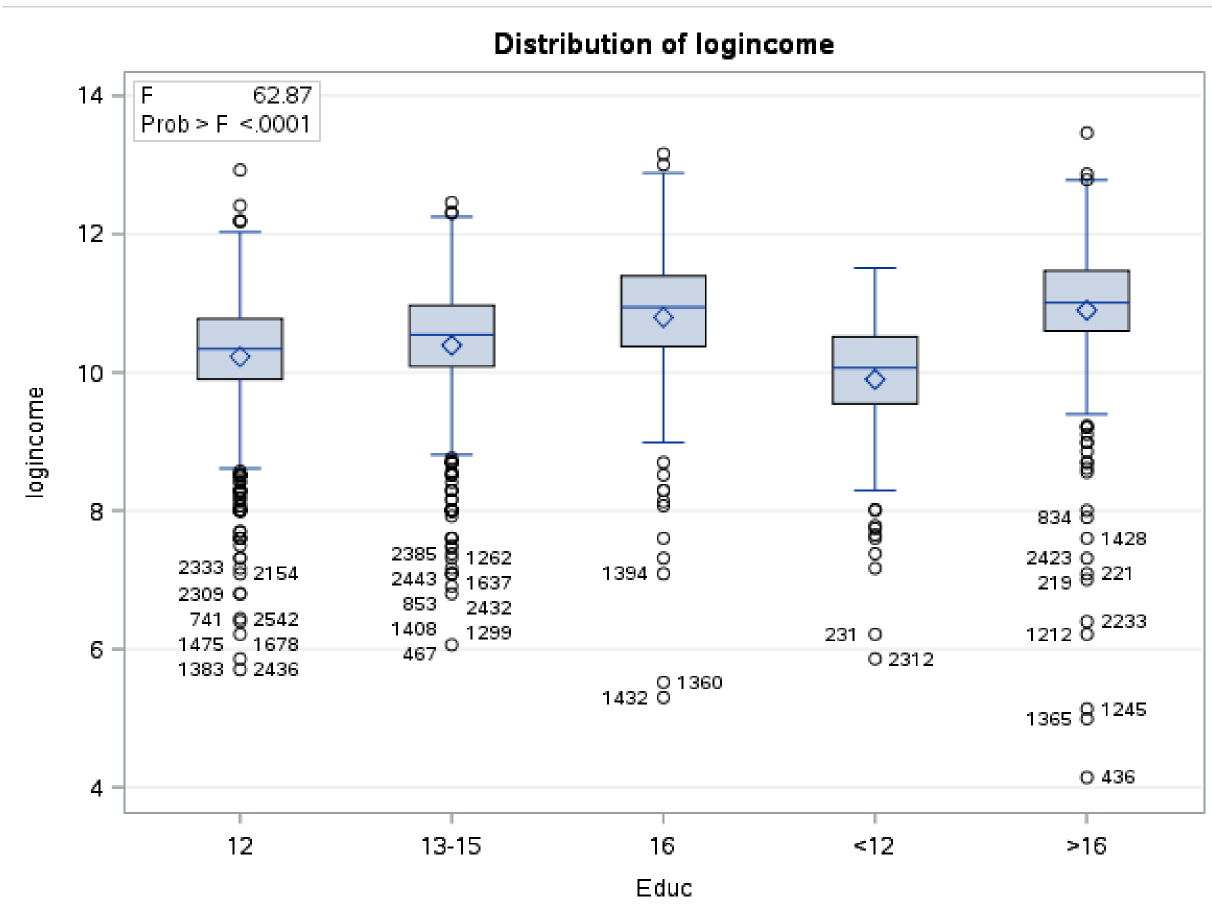
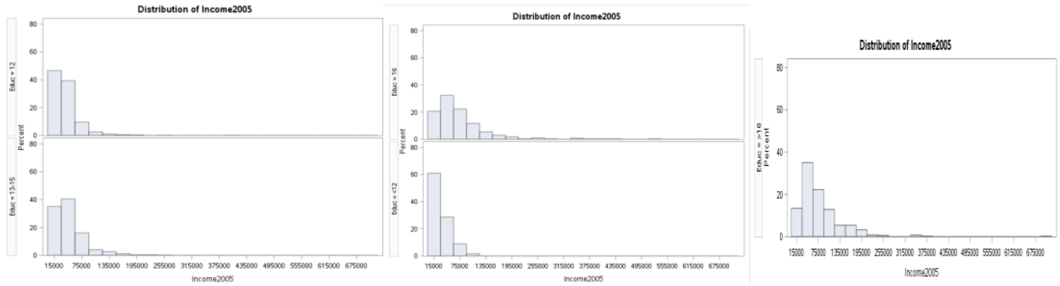
Unit 5 HW

1. **Simply Answer Question 25 on pg. 147 from the Statistical Sleuth (read it!):**
   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 or ex0525.xlsx. Perform this analysis in SAS. [Depending on where you find the data set, if you may see the value **<<12**. Note that **<<12 = 12**.]

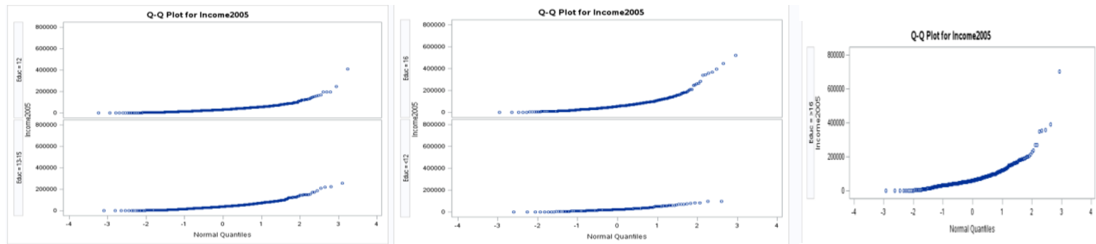
Regardless of whether the assumptions of the original data or log transformed data are met, please include a **complete analysis** on the **log transformed** data.

1. State the Problem.
   We need to test that at least one of the five population distributions (corresponding to the different years of education) is different from the others.
2. 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.

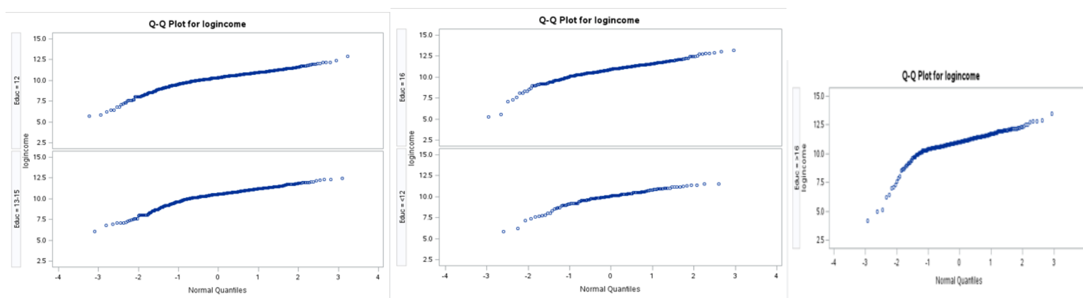




Regular QQ Plots



LOG QQ PLOTS



- Normality: We have a large sample size here. There is evidence for normality. We will proceed with caution under the assumption of normal distributions for each.
- Homogeneity of Variance: Judging from the box plots, there is some visual evidence of equal standard deviations
- Independence: We will assume the observations are independent both between and within groups.

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

The GLM Procedure

Dependent Variable: logincome

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			

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

Step1: $H_o: \mu_{<12}=\mu_{12}=\mu_{13-15}=\mu_{16}=\mu_{>16}$

H_a: *at least one pair* $\neq$

Step2: **Skip critical value for ANOVA**

Step3: **F= 62.87**

Step4: **p=.0001**

Step5: **Reject H_o**

Step6: **The evidence suggests that at least 1 pair of the group means are different (p=0.0001).**

4. Write a conclusion. (An example is in the UNIT 5 PowerPoint.)
   There is strong evidence at the α=0.05 level of significance (p<0.0001) to support the claim that the population distribution is different than that of the other distributions.
5. State the Scope. (Can we generalize to the entire population or just the sample that was taken? Is there a causal relationship present?) **This was an observational study; therefore, we cant' conclude causation and can only generalize to the sample of the data taken from the survey.**

<u>Looking to the future!</u> 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):

a. Please also identify R^2
   R^2 =0.88846
b. Also specify the mean square error and how many degrees of freedom were used to estimate it.
   Mean Square = 54.41 and 3 or 4 degrees of freedom?
c. Provide the code to perform the ANOVA in R and a screen shot of the output.

```
proc import datafile = '/home/chec0/New Folder/ex0525.csv'
 out = annual
 dbms = CSV
 ;

** log the data;
data annual2;
        set annual2;
        logincome = log(Income2005);
run;

proc glm data = annual2;
    class Educ;
    model logincome = Educ;
run;
```

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) income to the more than bachelor's degree group (>16) 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 (assuming equal standard deviations on logged data) yields a p-value of **.1648** (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.

```
proc import datafile = '/home/chec0/New Folder/ex0525.csv'
 out = annual
 dbms = CSV
 ;
proc print data=annual;
run;

data annual2;
        set annual2;
        logincome = log(Income2005);
proc print data=annual2;
run;

**Overall ANOVA;
proc glm data=annual2;
class Educ;
model logincome = Educ;
means Educ/ HOVTEST = BF;
run;

data annual3; set annual2;
if Educ in ('>16' '16') then groupedover16='a';
if Educ in ('<12') then groupedover16='b';
if Educ in ('13-15') then groupedover16='c';
if Educ in ('12') then groupedover16='d';
run;



proc glm data = annual3;
   class groupedover16;
   model logincome = groupedover16;
run;
```

The GLM Procedure

Dependent Variable: logincome

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			

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

The GLM Procedure

Dependent Variable: logincome

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	215.675158	71.891719	83.02	<.0001
Error	2580	2234.099010	0.865930		
Corrected Total	2583	2449.774168			

R-Square	Coeff Var	Root MSE	logincome Mean
0.088039	8.915315	0.930554	10.43770

Source	DF	SS	MS	F	Pr > F
Model	1	1.98	1.98	2.29	0.130
Error	2579	2232.12	.866		
Corrected Total	2580	2234.10			

Step1: H_o:μ_<12=μ_12=μ_13-15=μ_16=μ_>16

H_a:at least one pair ≠

Step2: Skip critical value for ANOVA

Step3: F= 2.29

Step4: p=0.130

Step5: *Fail to* Reject H_o

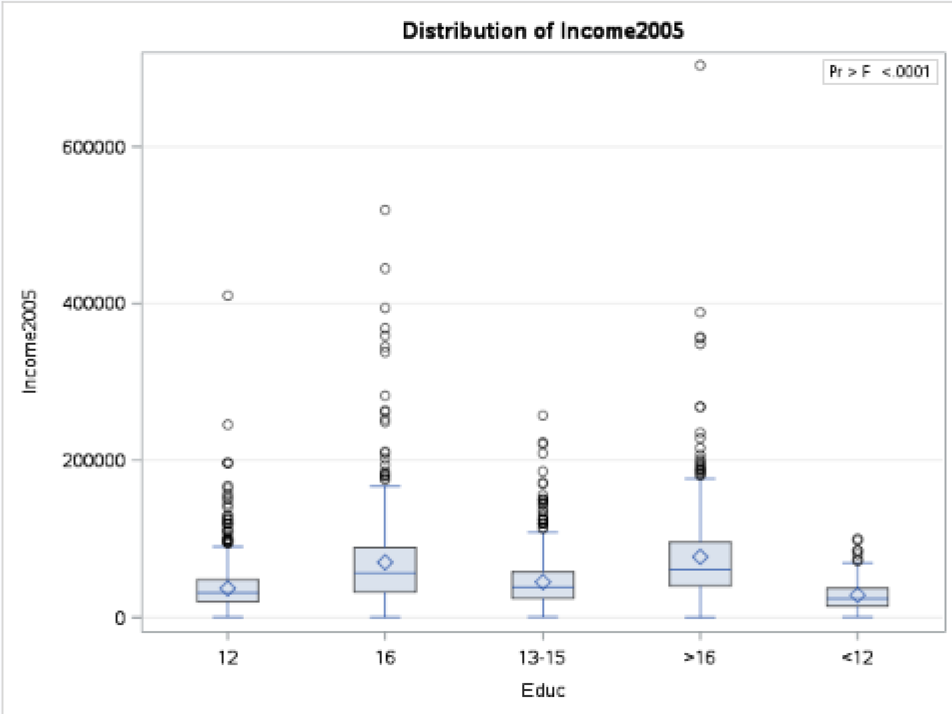
Step6: There is not sufficient evidence to suggest at the α=0.05 level of significance (p=0.130) that bachelor's degree group 16 and bachelor's degree group <16 have different mean depths.

This was an observational study; no causation and generalized to the incomes in the survey.

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.

State the Problem: **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?"**

Assumptions:



The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Income2005 Classified by Variable Educ					
Educ	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
12	1020	1097659.50	1318350.0	18536.1583	1076.13676
16	406	653168.50	524755.0	13800.4492	1608.78941
13-15	648	819191.00	837540.0	16437.7151	1264.18364
>16	374	654733.00	483395.0	13342.3770	1750.62299
<12	136	115068.00	175780.0	8467.9138	846.08824
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	349.4479
DF	4
Pr > Chi-Square	<.0001

- Normality: We have a large sample size here. There is evidence for normality. We will proceed with caution under the assumption of normal distributions for each.
- Homogeneity of Variance: Judging from the box plots, there is some visual evidence of unequal standard deviations
- Independence: We will assume the observations are independent both between and within groups.

Step1: $H_o: \mu_{<12}=\mu_{12}=\mu_{13-15}=\mu_{16}=\mu_{>16}$

H_a: *at least one pair* ≠

Step2: skip critical in kruskal test

Step3:

Step4: **p =<.0001**

Step5: **Reject the H_o**

Step6: **The evidence suggests that the group medians are different (p=<0.0001).**

There is sufficient evidence at the α=0.05 level of significance (*p*= <.0001 from Kruskal-Wallis Test) to suggest that at least two of the medins are different.

This was an observational study; therefore, we can't conclude causation and can only generalize to the sample of the data taken from the survey.