Question 1)

1) Problem

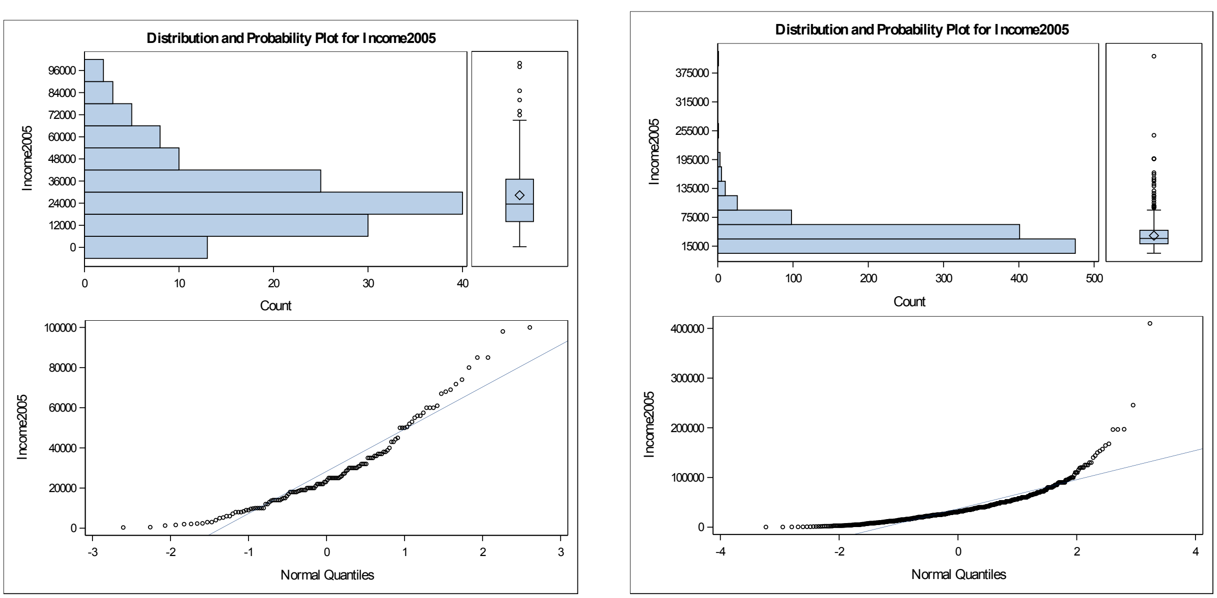
A one way analysis of variance test (ANOVA) was run to test the ***null-hypothesis (H_0)*** that: there is **no difference in mean income** for those with <12, 12, 13-15, 16, and >16 years of education ($\mu_{<12} = \mu_{12} = \mu_{13-15} = \mu_{16} = \mu_{>16}$).

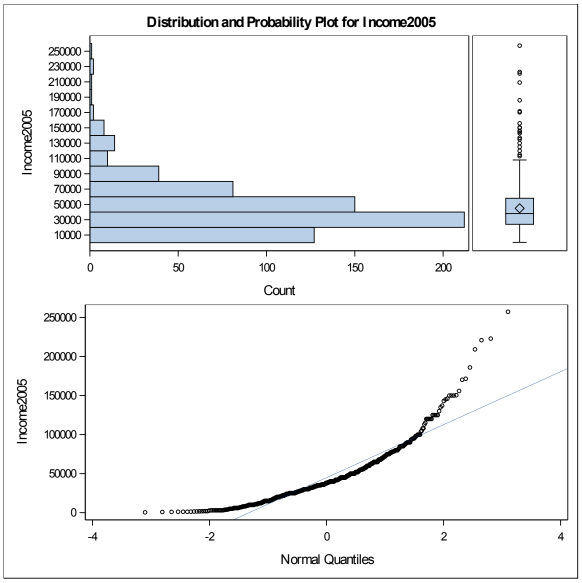
The ***alternative-hypothesis (H_A)*** would be that: for at least two of the groups, there is, in fact, a statistically significant **difference in the mean**: ($\mu_{<12} \neq \mu_{12}\ OR\ \mu_{13-15}\ OR\ \mu_{16}\ OR\ \mu_{>16}$).

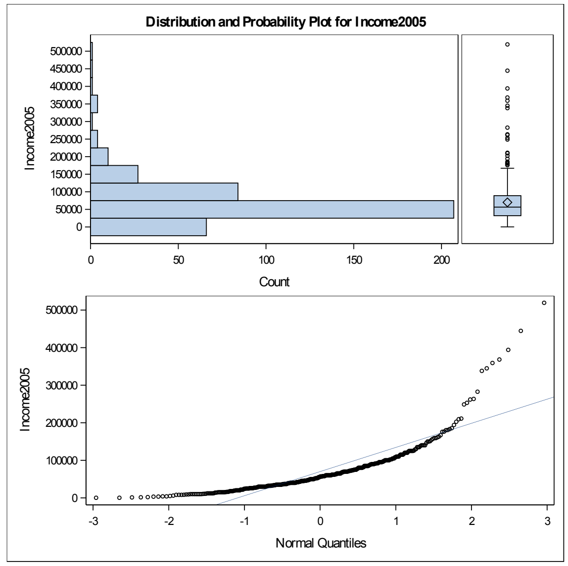
2) Assumptions

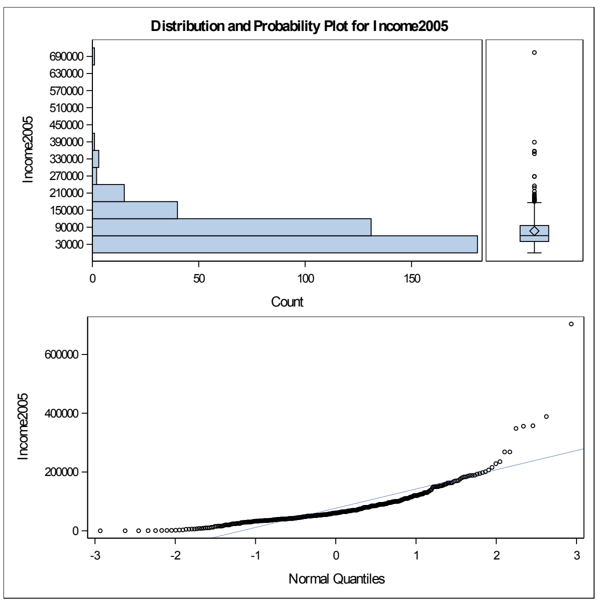
An initial assessment of this study design from the NLSY (which uses random probability sampling to estimate population means) and of the available data indicated that the assumption of **independence** should hold. To assess **normality** and **variance**, statistical software produced the following graphics (distribution and probability histograms and QQ-plots) and mean table (Figure 1) of the untransformed data by ascending educational level, left to right, indicating a general right-skewedness:

Figure 1:

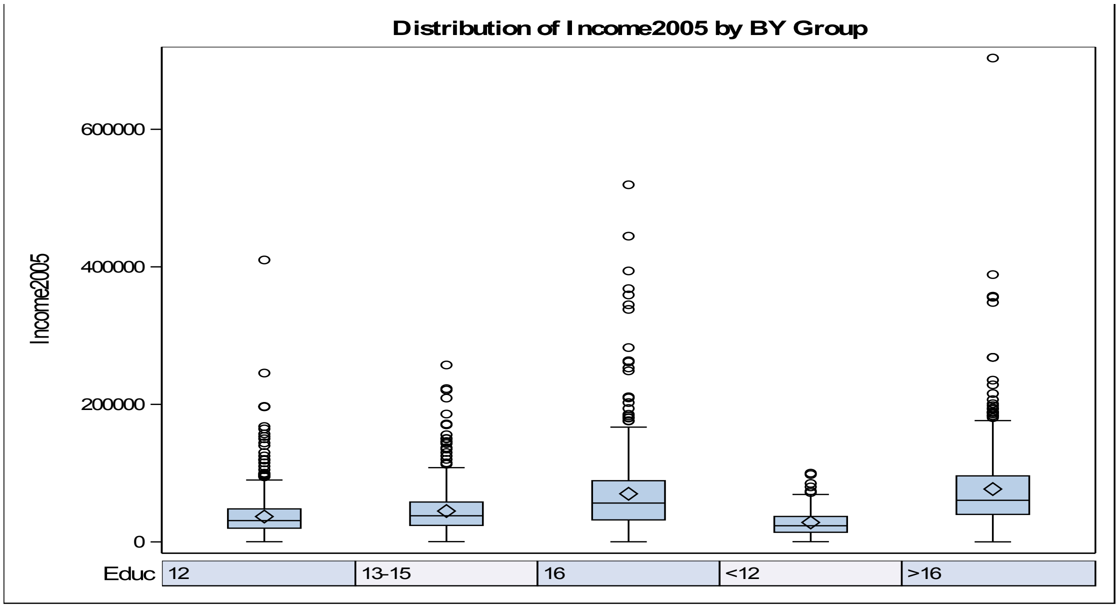








Analysis Variable : Income2005						
Educ	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
12	1020	1020	36864.90	29369.73	300.0000000	410008.00
13-15	648	648	44875.96	33913.54	429.0000000	257286.00
16	406	406	69996.97	64256.80	200.0000000	519340.00
<12	136	136	28301.45	21021.90	350.0000000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



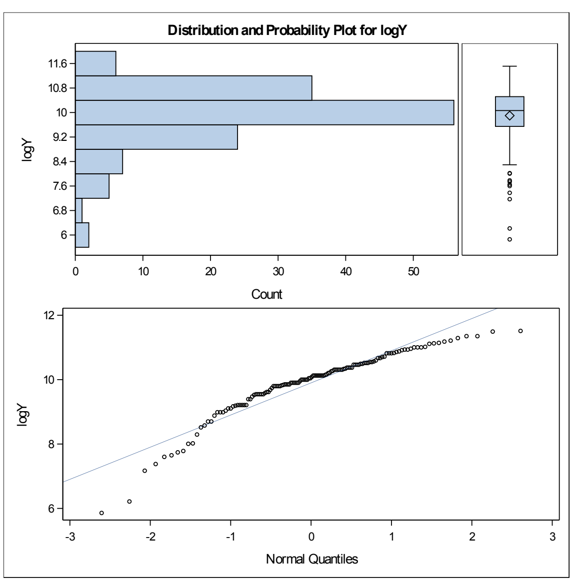
Using the mean table for the untransfored values, the sample differences between adjacent group categories were as follows:

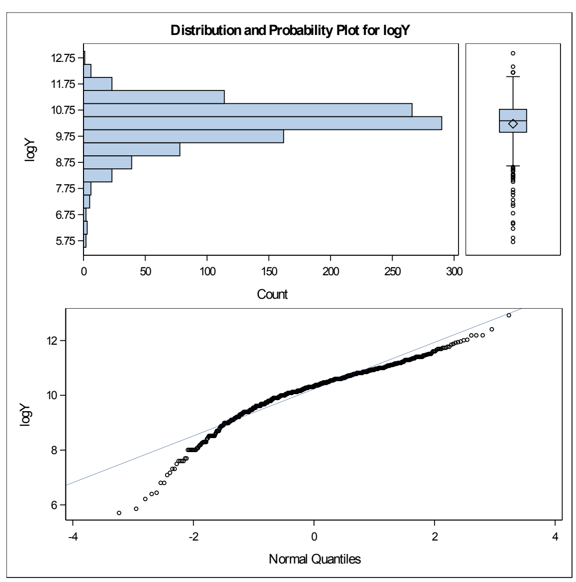
$(\bar{x}_{12} - \bar{x}_{<12} = 8563.45)$ $(\bar{x}_{16} - \bar{x}_{13-15} = 25121.01)$

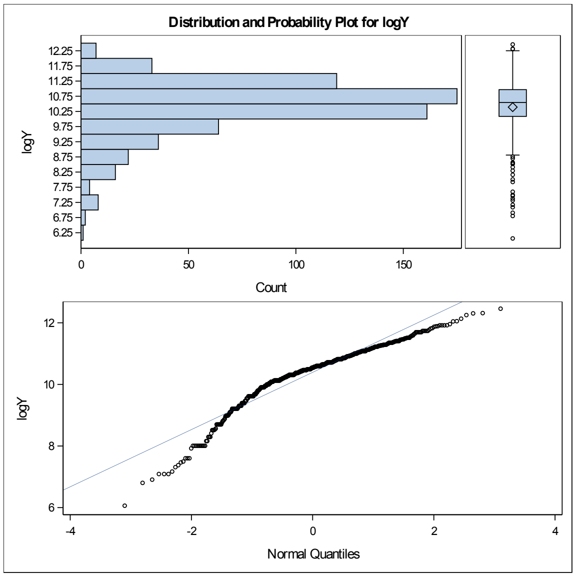
$(\bar{x}_{13-15} - \bar{x}_{12} = 8011.06)$ $(\bar{x}_{>16} - \bar{x}_{16} = 6858.49)$

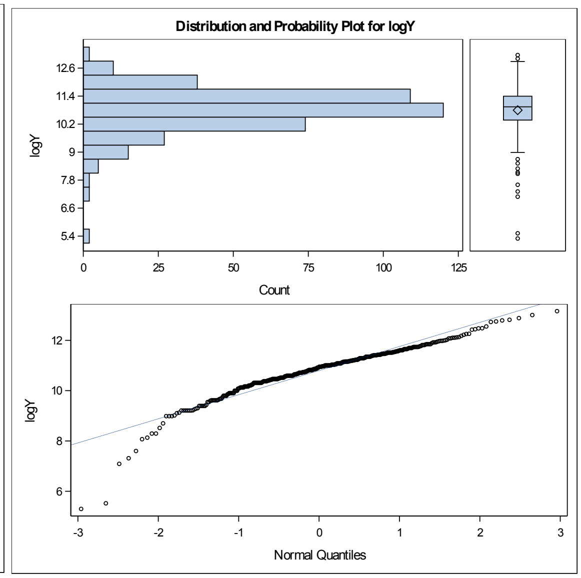
Since non-normality and group variance are evident, a log transformation of the data was run as a comparative assessment, figures for which are produced below, in ascending group order:

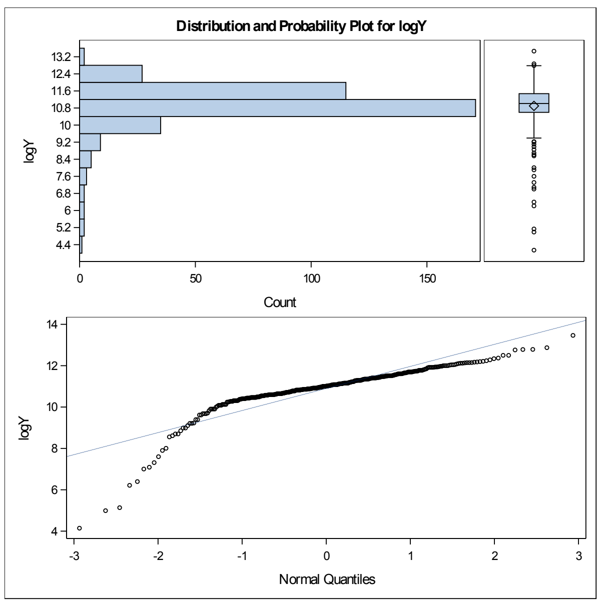
Figure 2:



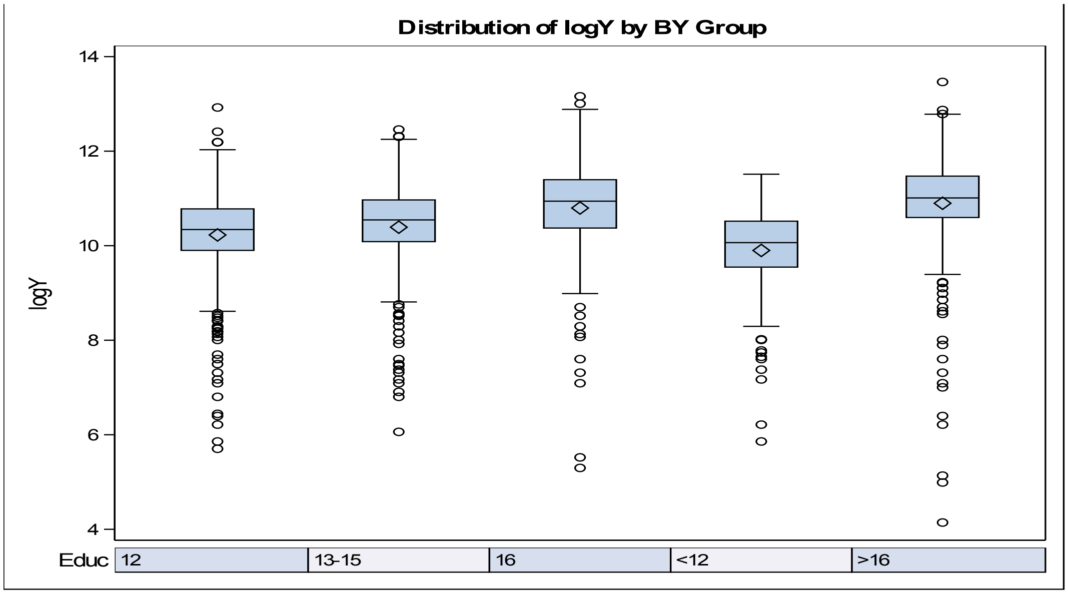








Analysis Variable : logY						
Educ	**N Obs**	**N**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
12	1020	1020	10.2272149	0.8539854	5.7037825	12.9239320
13-15	648	648	10.3912107	0.9288173	6.0614569	12.4579436
16	406	406	10.7970859	0.9581051	5.2983174	13.1603141
<12	136	136	9.8993404	0.9988809	5.8579332	11.5129255
>16	374	374	10.8979022	1.0665910	4.1431347	13.4640179



Since the transformation resulted in a relatively small (~5%) reduction in group variability, and since the sampe size was sufficiently large (n = 2584) to potentially overstate the results of the Brown-Forsythe test (Figure 3) and to unproblematically allow for some general right-skewedness, the ANOVA test proceeded with the original (untransformed) values.

Figure 3:

Brown and Forsythe's Test for Homogeneity of Income2005 Variance ANOVA of Absolute Deviations from Group Medians					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Educ	4	2.231E11	5.579E10	44.92	<.0001
Error	2579	3.203E12	1.242E9		

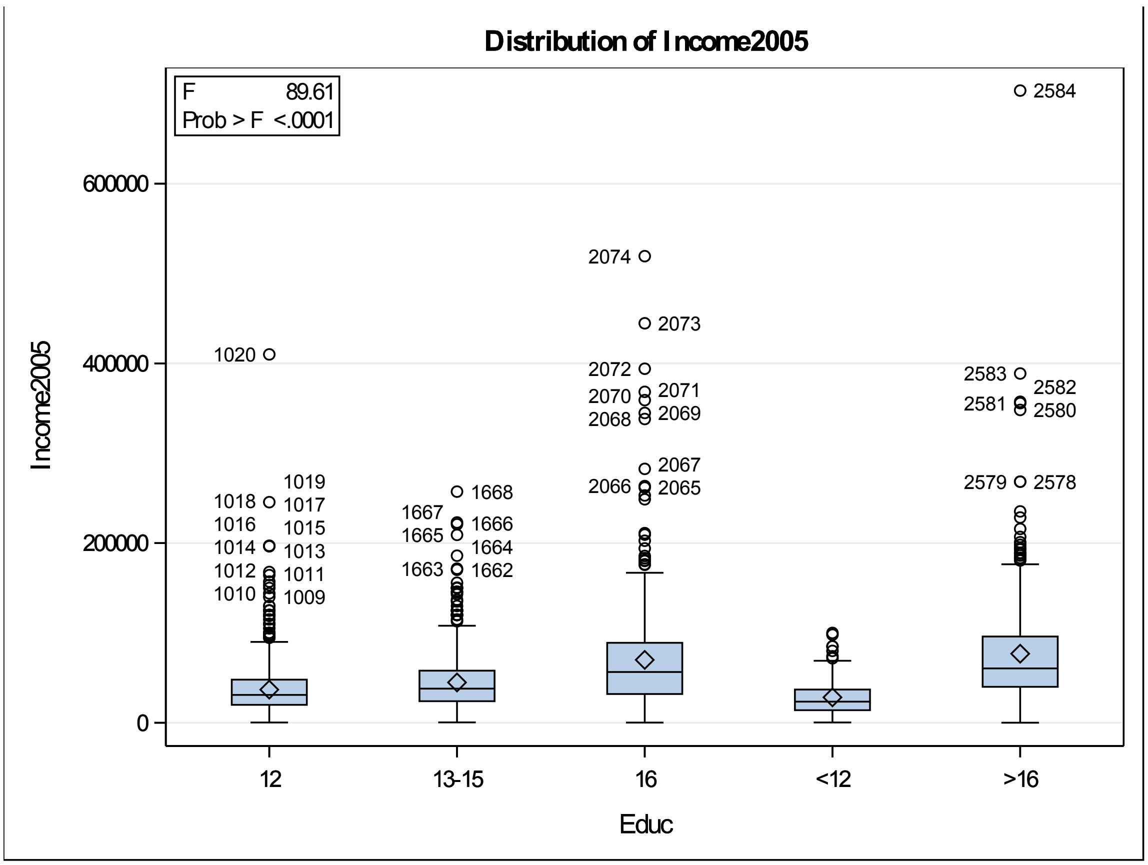
3) ANOVA test

Under the assumption of roughly equal group standard deviations, statistical software was used to run a generalized linear model testing the null hypothesis of equal group means, producing the following graphics and ANOVA tables for within-, between-, and total group variance:

Figure 4:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	688235137516	172058784379	89.61	<.0001
Error	2579	4.9517427E12	1920024319.9		
Corrected Total	2583	5.6399779E12			

R-Square	Coeff Var	Root MSE	Income2005 Mean
0.122028	88.67006	43818.08	49417.00



4) Conclusion

On the basis of the above information, one would **reject the null-hypothesis** (H_0), determining that: there is **statistically significant evidence** to suggest a difference in mean income between at least two of the groups observed ($\mu_{<12} \neq \mu_{12}\ OR\ \mu_{13-15}\ OR\ \mu_{16}\ OR\ \mu_{>16}$).

→ [**p_{val} < .0001**], [**R^2 = .122028**], [**Root MSE** = **43818.08**], [**F = 89.61**], [**df** = **4, 2579**]

5) Scope

Given the evidence for this randomly sampled observational study of existing distinct populations, while inferences regarding the difference in mean income **cannot be drawn** to the **entire population** of employed adults, they **can** be drawn to the **sample population** of similarly situated survey respondents, ages 41-49 and employed at the time of their interview in 2006.

Question 2)

1) Problem

An Extra Sum of Squares F-test was run to test the ***null-hypothesis (H_0)*** that: there is **no difference in mean income** for those with 16, and >16 years of education:
($[\mu_{16} = \mu_{>16}] = [\mu_{<12} = \mu_{12} = \mu_{13-15}]$).

The ***alternative-hypothesis (H_A)*** would be that: there is, in fact, a statistically significant **difference in the mean**: ($[\mu_{16} = \mu_{>16}] \neq [\mu_{<12} = \mu_{12} = \mu_{13-15}]$), and therefore (combining this evaluation with the results of Question 1):

($\mu_{16} \neq \mu_{>16}$).

2) Assumptions

Drawing on the results of Question 1, the initial assessments of this study design from the NLSY (which uses random probability sampling to estimate population means) and of the available data indicated that the assumptions of **independence**, **normality** and **variance** should hold sufficiently to use untransformed data values.

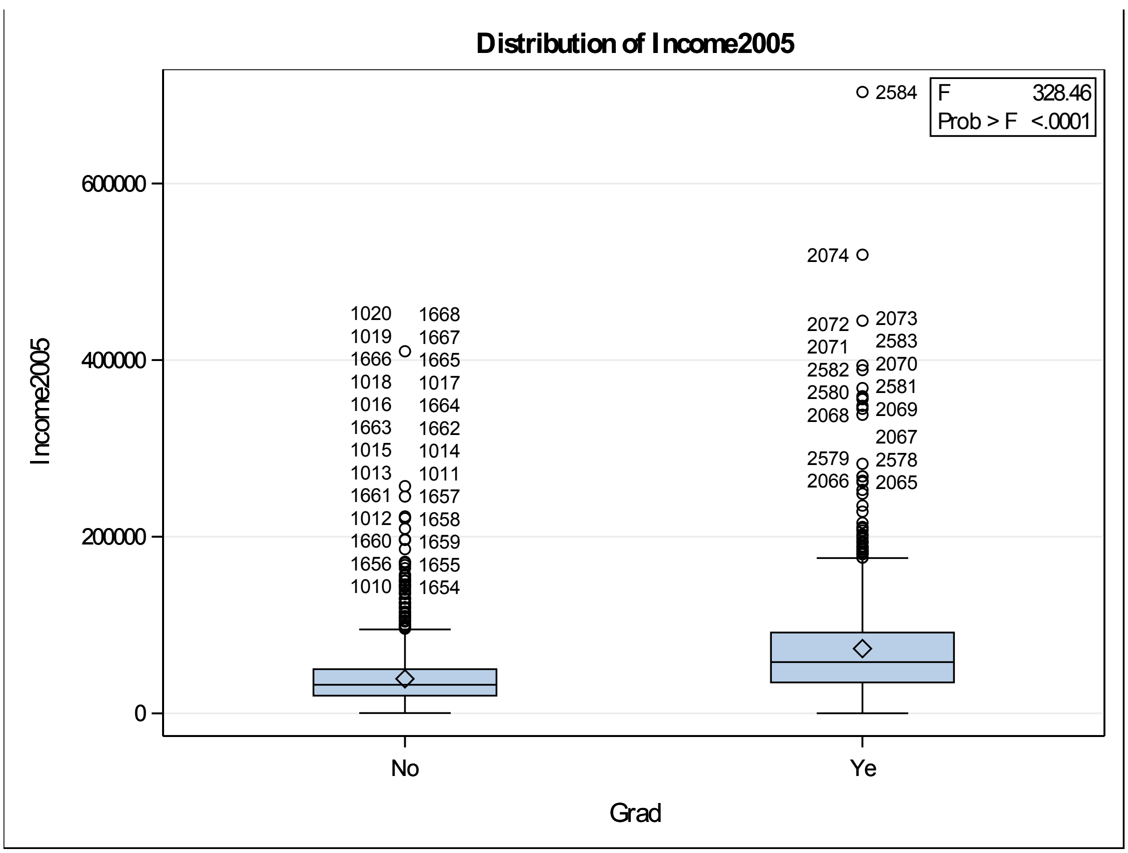
3) Extra Sum of Squares F-test

Under the assumption of roughly equal group standard deviations, statistical software was used to run a generalized linear model testing the null hypothesis of equal mean incomes for the two groups of interest, producing the following graphics and ANOVA tables for within-, between-, and total group variance (Figure 5), when all groups are assumed to have equal means aside from those with 16 and >16 years of education:

Figure 5:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	636505256541	636505256541	328.46	<.0001
Error	2582	5.0034726E12	1937828273.5		
Corrected Total	2583	5.6399779E12			

R-Square	Coeff Var	Root MSE	Income2005 Mean
0.112856	89.08022	44020.77	49417.00



The next step in the analysis involved creating a final ANOVA table based on the full model from Question 1, and the reduced model above, figures for which are produced below:

Figure 6:

Source	DF	SS	MS	F	P	R^2	Root(MSE)
Model	3	51729900000	17243300000	8.980772	0.00001	0.0103388	43818.08
Error	2579	4.95174E+12	1920024320				
Total	2582	5.00347E+12					

To gauge the relative power of the test, a comparison was made with the figures produced by a two sample t-test between the two groups of interest, figures for which are produced below:

Figure 7:

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	778	-1.48	0.1403
Satterthwaite	Unequal	770.26	-1.48	0.1406

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	373	405	1.04	0.7209

The figures of the t-test demonstrate the relative power of ANOVA, which pools all available data and degrees of freedom in making its determination.

4) Conclusion

On the basis of the above information, one would **reject the null-hypothesis** (H_0), determining that: there is **statistically significant evidence** to suggest a difference in mean income between the two groups of interest ($\mu_{16} \neq \mu_{>16}$).

→ [**p_{val} < .00001**], [**R^2 = .0103388**], [**Root MSE** = **43818.08**], [**F = 8.980772**], [**df** = **3, 2579**].

5) Scope

As for Question 1, given the evidence for this randomly sampled observational study of existing distinct populations, while inferences regarding the difference in mean income for the two groups of interest **cannot be drawn** to the **entire population** of employed adults, they **can** be drawn to the **sample population** of similarly situated survey respondents, ages 41-49 and employed at the time of their interview in 2006.

Question 3)

1) Problem

Presuming that one cannot assume equal standard deviations for the groups of interest, a Kruskal-Wallis nonparametric one-way ANOVA test was run to test the ***null-hypothesis (H_0)*** that: there is **no difference in the mean** income for those with <12, 12, 13-15, 16, and >16 years of education ($\mu_{<12} = \mu_{12} = \mu_{13-15} = \mu_{16} = \mu_{>16}$).

The ***alternative-hypothesis (H_A)*** would be that: for at least two of the groups, there is, in fact, a statistically significant **difference in the mean**: ($\mu_{<12} \neq \mu_{12}\ OR\ \mu_{13-15}\ OR\ \mu_{16}\ OR\ \mu_{>16}$).

2) Assumptions

Drawing on the results of Question 1, the initial assessments of this study design from the NLSY (which uses random probability sampling to estimate population means) and of the available data indicated that the assumptions of **independence**, **normality** and **variance** should hold sufficiently to use untransformed data values.

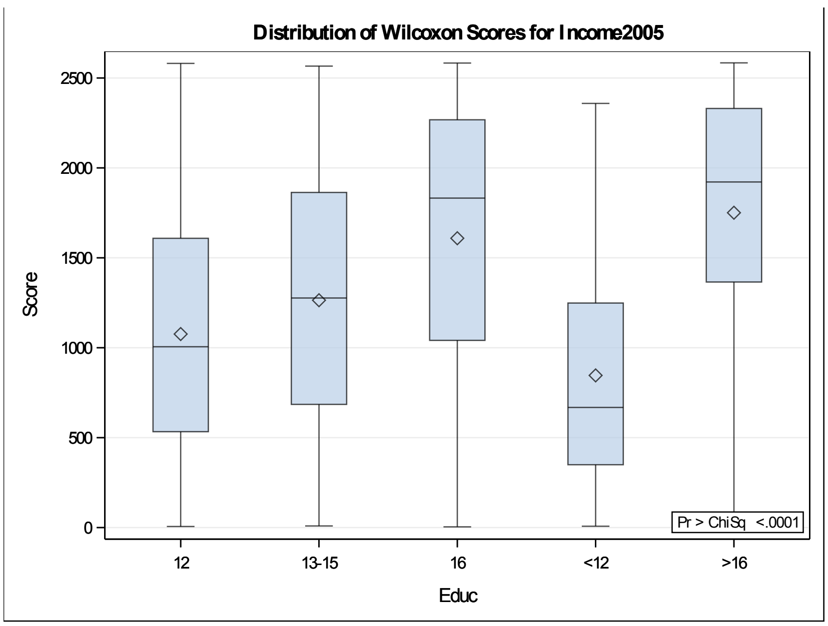
3) Kruskal-Wallis test

Presuming one cannot assume equal group standard deviations, statistical software was used to run Rank-sum, Chi-Square, and Monte Carlo Estimates for the null hypothesis of equal mean incomes for the groups of interest, producing the following figures and tables (Figure 8):

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
13-15	648	819191.00	837540.0	16437.7151	1264.18364
16	406	653168.50	524755.0	13800.4492	1608.78941
<12	136	115068.00	175780.0	8467.9138	846.08824
>16	374	654733.00	483395.0	13342.3770	1750.62299
Average scores were used for ties.					

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

Monte Carlo Estimate for the Exact Test	
Pr >= Chi-Square	
Estimate	<.0001
99% Lower Conf Limit	<.0001
99% Upper Conf Limit	0.0005
Number of Samples	10000
Initial Seed	12345



4) Conclusion

On the basis of the above information, one would **reject the null-hypothesis** (H_0), determining that: there is **statistically significant evidence** to suggest a difference in mean income between at least two of the groups observed $(\mu_{<12} \neq \mu_{12}\ OR\ \mu_{13-15}\ OR\ \mu_{16}\ OR\ \mu_{>16})$.

→ [**p_{val} < .0001**], [**χ^2 = 349.4479**], [**df = 4**], [**99% p-CI: <.0001, .0005**], for 10,000 samples.

5) Scope

As for Question 1, given the evidence for this randomly sampled observational study of existing distinct populations, while inferences regarding the difference in mean income for the groups of interest **cannot be drawn** to the **entire population** of employed adults, they **can** be drawn to the **sample population** of similarly situated survey respondents, ages 41-49 and employed at the time of their interview in 2006.