Name: James Tsai
Section: MSDS6371-401
Date: 10/04/15

Simply Answer Question 25 pg. 147:.*Assume that from a prior study we may assume that the standard deviations of each group are equal … not matter what the histograms look like.* You should use this fact when you address the assumptions.

1. **State the problem.**

We are examining the strength of the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others. We will also determine how many dollars and what percent does the mean or median for each of the last four categories exceed that of the next lowest category.

2. **Address the assumptions.**

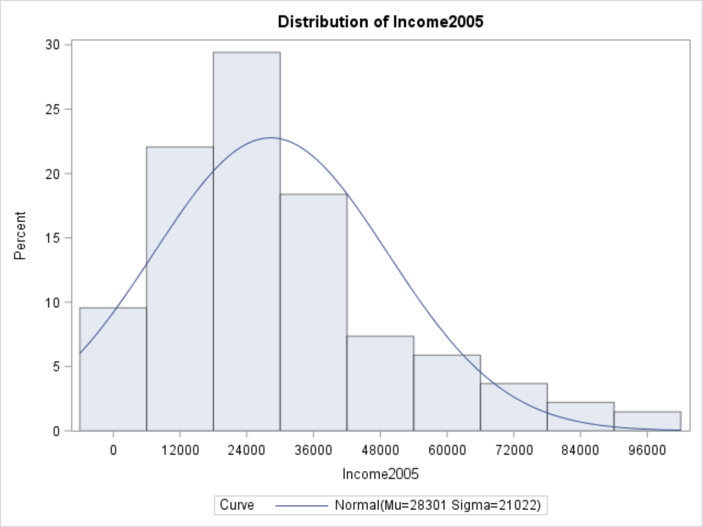
In order to run the ANOVA test, we must address the 3 assumptions.

Normality
We will assume that the samples are normally distributed due large sample size (central limit theorem). Also, the ANOVA test is somewhat robust to violations of the normality assumption, provided the violations are not too severe. Judging from the histograms (pictured below), we do notice that the distributions are all positively skewed, caused by some outliers in each of the respective groups.

The SAS System

The UNIVARIATE Procedure

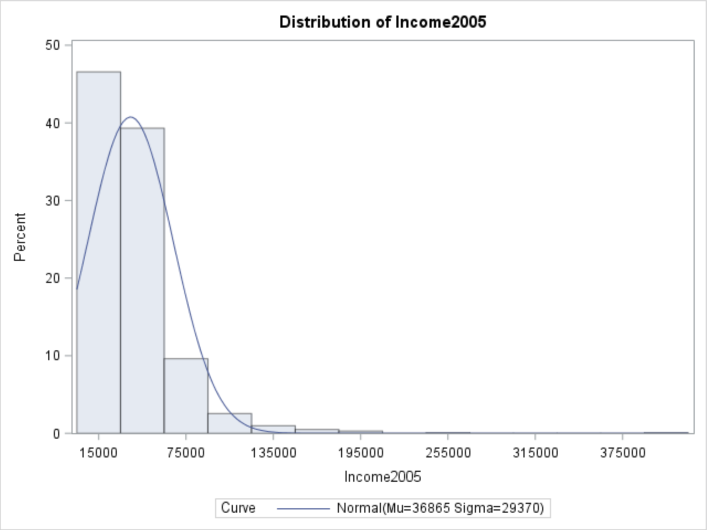
Educ=<12



The SAS System

The UNIVARIATE Procedure

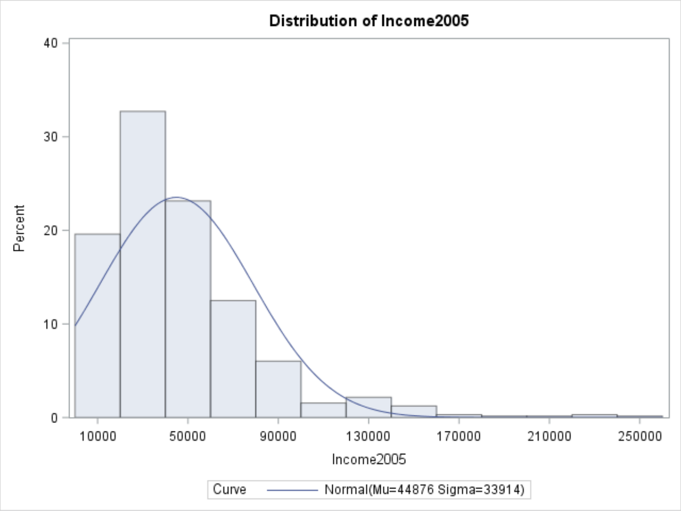
Educ=12



The SAS System

The UNIVARIATE Procedure

Educ=13-15

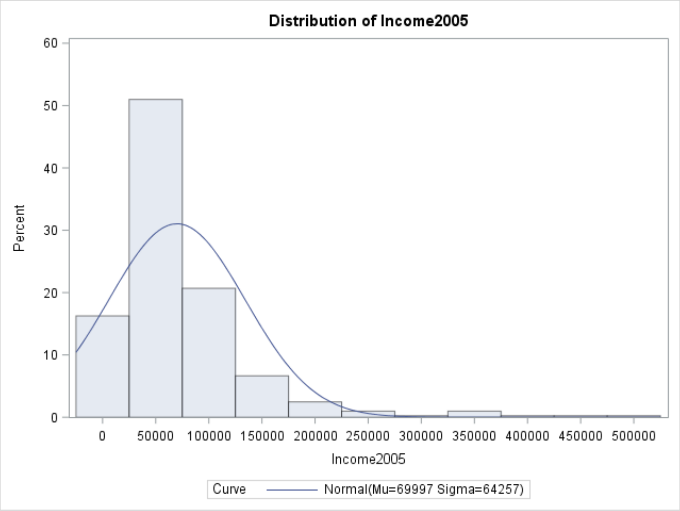


Distribution of Income2005

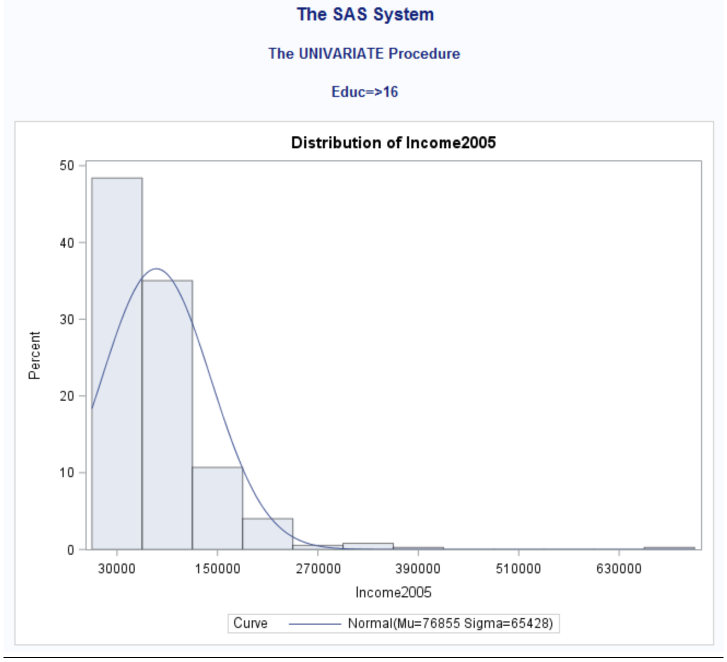
The SAS System

The UNIVARIATE Procedure

Educ=16



Distribution of Income2005



The SAS System

The UNIVARIATE Procedure

Educ=>16

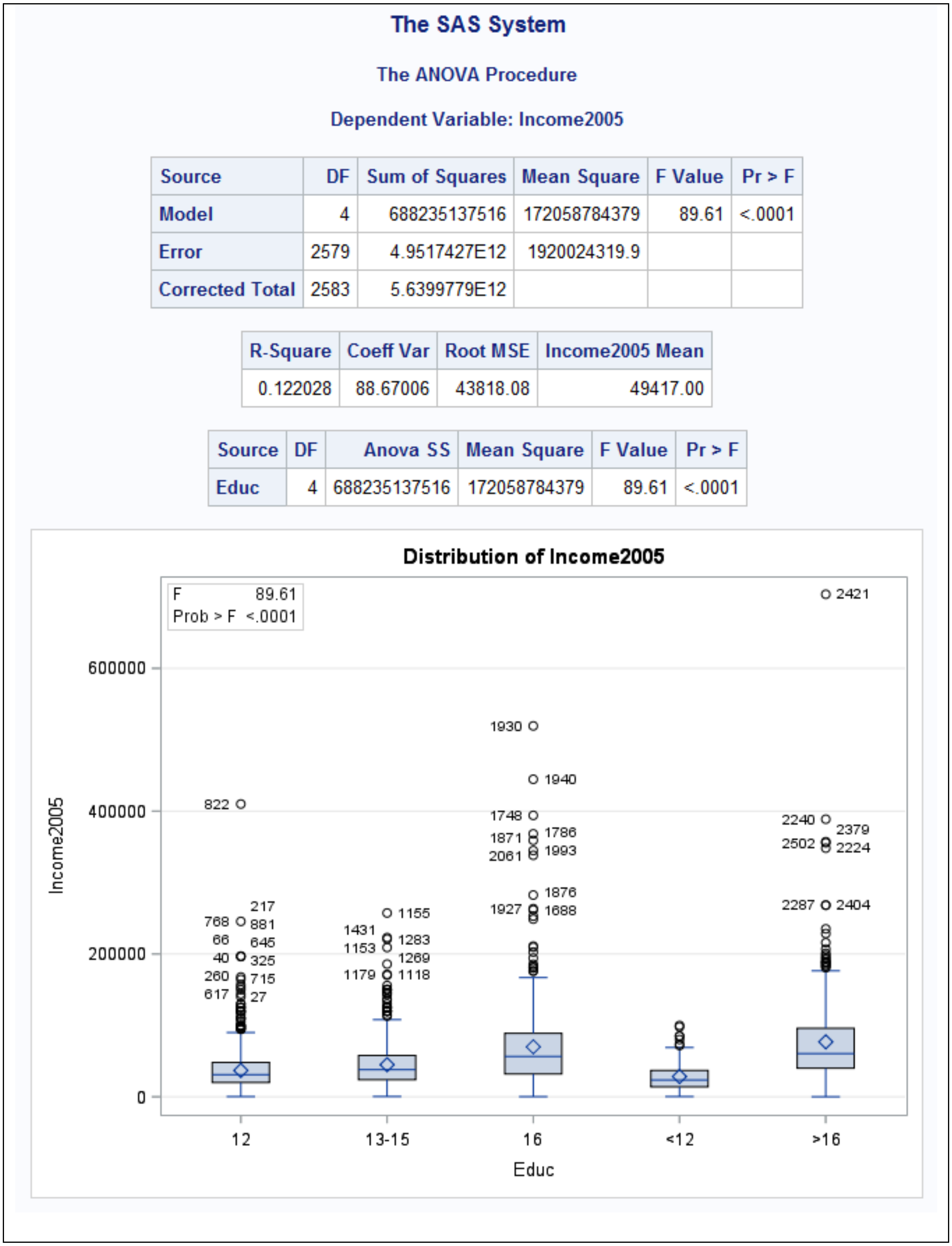
Variance
We are assuming the standard deviations of each group are equal per the instructions above. Therefore, the variances of each group are equal as well. As a side note, we notice the first 3 groups have standard deviations that differ quite a bit from the last 2 groups.

Independence
The 2,584 samples were randomly drawn and are independent of each other.

3. Conduct the test.

1. H_0: $u_{<12} = u_{12} = u_{13-15} = u_{16} = u_{>16}$
   H_A: At least 1 pair are different

2. F-Statistic = 89.61
3. P-Value < .0001
4. Reject H_0
5. The evidence suggests that at least 1 pair of the group means are different.



The SAS System

The ANOVA Procedure

Dependent Variable: Income2005

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

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Educ	4	688235137516	172058784379	89.61	<.0001

Distribution of Income2005

4. **Write a conclusion.**

There is strong evidence at alpha = .05 to support the fact that at least one pair of the five group means of education and future income data are different.

5. **State the scope. (Can we generalize to the entire population or just the sample that was taken?)(… The next step will be to look at these pairwise if we reject the Ho to discover WHICH pairs have evidence of different means.**

We cannot generalize this to the entire population since this study only covers the individuals who were chosen in 1979 and who had paying jobs in 2005. It is relevant only to this sample that was taken. Furthermore, our conclusion only supports the fact that at least one pair of the five group means are different, and does not contain information regarding comparisons of any specific pair.

ADDITIONAL THINGS TO INCLUDE

a. **Please also identify R^2**

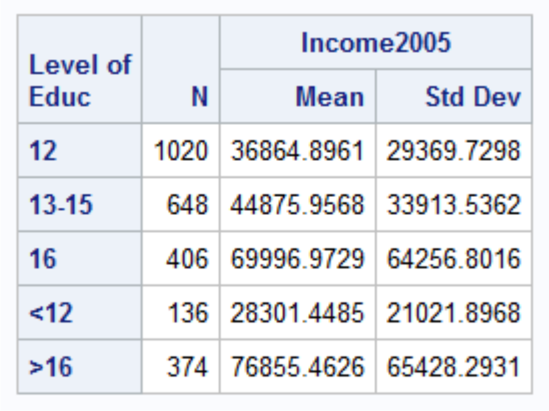
The R^2 value is 0.122028.

b. **Also specify the mean square error and how many degrees of freedom were used to estimate it.**

The MSE is 1920024319.9 and the degrees of freedom to calculate it is 2579.

Part II of the Question 25: 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?

Using the SAS output, we have the following summary of the means for each group:



Level of Educ	N	Income2005	
		Mean	Std Dev
12	1020	36864.8961	29369.7298
13-15	648	44875.9568	33913.5362
16	406	69996.9729	64256.8016
<12	136	28301.4485	21021.8968
>16	374	76855.4626	65428.2931

From this output, we summarize the mean differences into 4 categories, and calculate the dollar and percentage differences starting with the difference of means between 12 years of education and less than 12 years of education:

Category	Dollar Increase	Percent Increase
1	8563	30%
2	8011	22%
3	25121	56%
4	6859	10%

From this summary, we can see that the most dramatic jump in average salary in both dollar and percentage terms is in category 3. This category shows the difference of mean salary of those with 16 years of education vs. 13-15 years of education.