|  |
| --- |
| title: "Statistical Foundations for Data Science - Assignment5 " |
| author: "Travis Deason" |
| date: "Oct 1st 2017" |
| output: word\_document |

# MSDS 6371 HW 5

## 1. Simply Answer Question 25 pg. 147 from the Statistical Sleuth:. Plot the raw data and also plot the data after a log transform. Do the data after a log transform satisfy the assumptions better? The data is in ex0525.csv Perform this analysis in SAS. Please include a complete analysis on the log transformed data.

income\_df = read.csv('data/ex0525\_2.csv')  
names(income\_df) <- tolower(names(income\_df))  
income\_cats <- unique(income\_df$educ)  
edu\_sub\_12 = subset(income\_df$income2005, income\_df$educ == income\_cats[1])  
edu\_16\_plus = subset(income\_df$income2005, income\_df$educ == income\_cats[2])  
edu\_12 = subset(income\_df$income2005, income\_df$educ == income\_cats[3])  
edu\_13\_15 = subset(income\_df$income2005, income\_df$educ == income\_cats[4])  
edu\_16 = subset(income\_df$income2005, income\_df$educ == income\_cats[5])  
summary(income\_df)

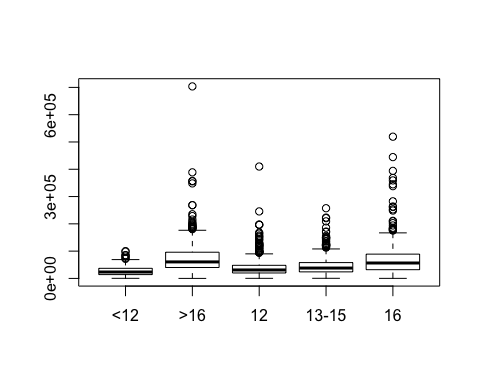
## subject educ income2005   
## Min. : 2 <12 : 136 Min. : 63   
## 1st Qu.: 1586 >16 : 374 1st Qu.: 23000   
## Median : 3108 12 :1020 Median : 38231   
## Mean : 3494 13-15: 648 Mean : 49417   
## 3rd Qu.: 4636 16 : 406 3rd Qu.: 61000   
## Max. :12140 Max. :703637

*1*. State the Problem

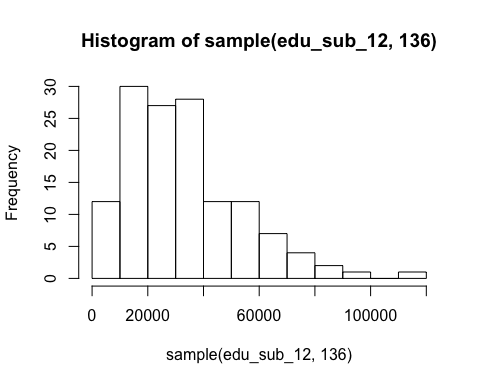
* Based on the 5 diffirent eduation levels provided in ex0525\_2.csv. How strong is the evidence that one or more of the populations is diffirent from the others.

*2*. Address the assumptions. (Comment on each assumption. Use the visual test as the Brown-Forstythe test will be overpowered due to the large sample size. This simply means that it is able to detect very small effect sizes (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.

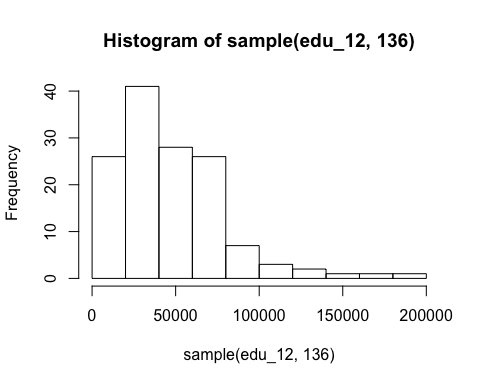
boxplot(income2005~educ , data= income\_df)



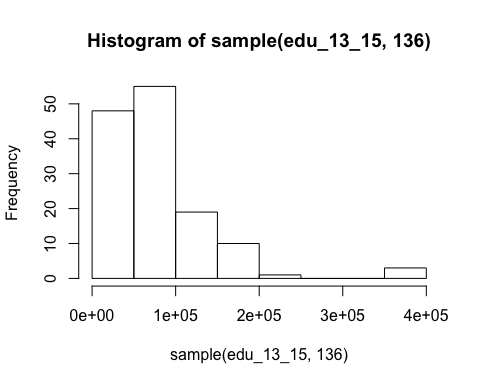
hist(sample(edu\_sub\_12, 136))



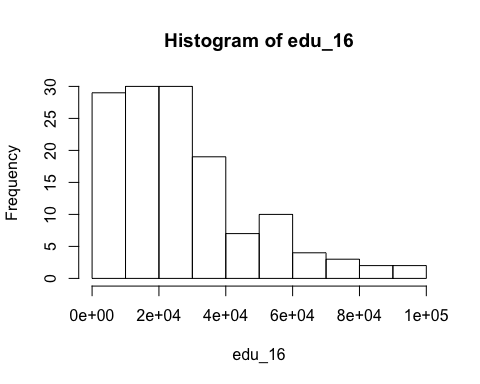
hist(sample(edu\_12, 136))



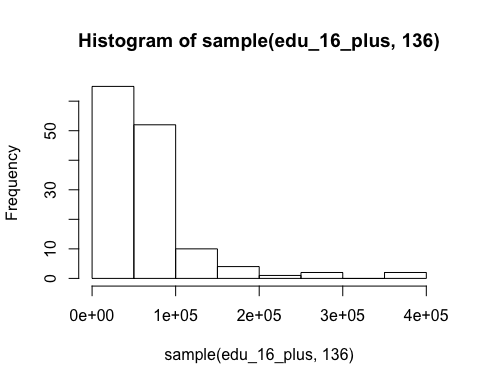
hist(sample(edu\_13\_15, 136))



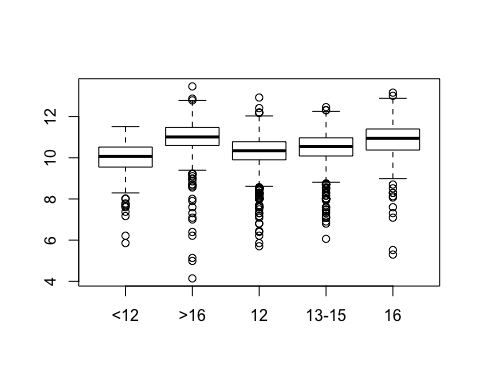
hist(edu\_16)



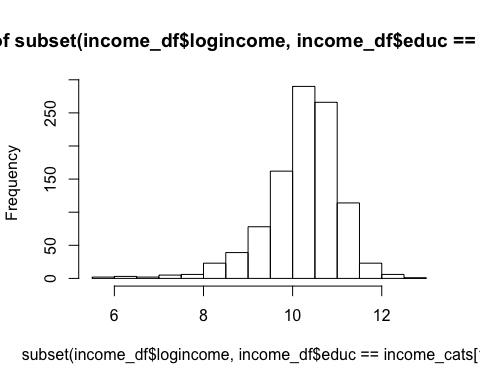
hist(sample(edu\_16\_plus, 136))



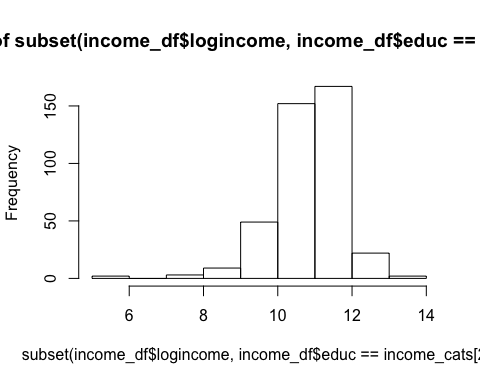
income\_df$logincome <- log(income\_df$income2005)  
  
boxplot(logincome~educ , data= income\_df)



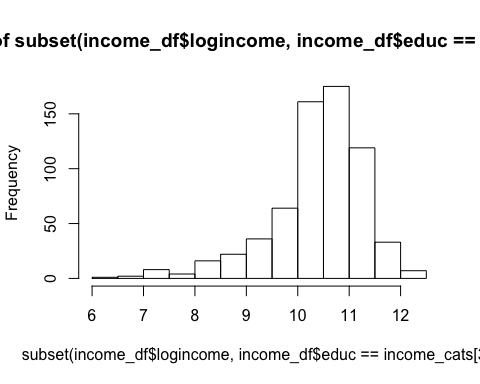
hist(subset(income\_df$logincome, income\_df$educ == income\_cats[1]))



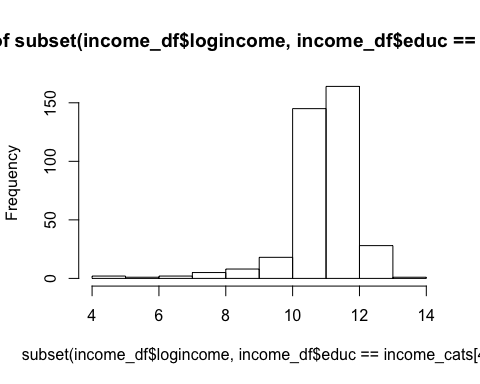
hist(subset(income\_df$logincome, income\_df$educ == income\_cats[2]))



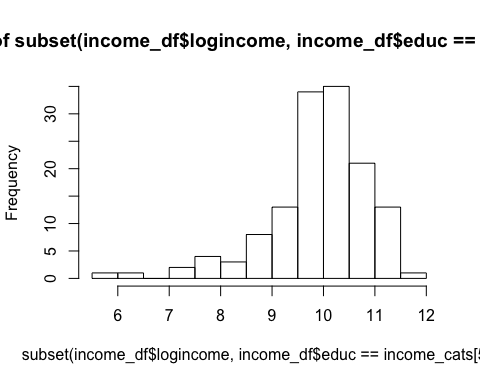
hist(subset(income\_df$logincome, income\_df$educ == income\_cats[3]))



hist(subset(income\_df$logincome, income\_df$educ == income\_cats[4]))



hist(subset(income\_df$logincome, income\_df$educ == income\_cats[5]))



* The data is not normal for most of the subsets. The data most similarly resemlbes an exponential curve.
* The plot variances seem similar for most of the groups, with the exception of the less then 12 years of education group. Since neither the data is normal or with similar variance, we will apply a log transformation to the data, and this gives us normal data of equal variance
* Since the data was chosen at random, and there is no covariance, it is safe to assume the data is independent. An exception to this assumption would be if not all of the children initally selected were surveyed for income in 2005.

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

income\_anova <- aov(logincome~educ, data= income\_df)  
summary(income\_anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## educ 4 217.7 54.41 62.87 <2e-16 \*\*\*  
## Residuals 2579 2232.1 0.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

income\_r2 = (2232.1) / (2232.1+217)  
print(income\_r2)

## [1] 0.911396

*4*. Write a conclusion. (An example is in the UNIT 5 PowerPoint.)

* Based on the ANOVA analysis of the Five income groups, we reject the null hypothesis that income is independent of education level with a p\_value of virtually zero.

*5*. State the Scope. (Can we generalize to the entire population or just the sample that was taken?)

* Since the test group was taken from a random sample of youth in 1979, we can assume there is a coorelation between education level and income for youth who are in grade school in 1979. We cannot neccessarily assume that coorelation extends to youth from a diffirent generational group, or that the relationship bewtween education levels and income is causational. There could be counfounding variables, and the cirmustances which led to the incomes observed in 2005 could have changed over the course of 26 years.
* The R^2 value for this model is .9113, and the mean

\*Looking to the future! 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

*a*. Please also identify R2

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

*c*. Provide the code to perform the ANOVA in R and a screen shot of the output.

## 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) mean income to the graduate degree group (>16) mean income. Show your final ANOVA table and your 6 step complete analysis. You will need to assume that the standard deviations are gain equal to proceed here. A two sample t-test between these two groups yields a p-value of .1403 (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.

library(stringr)  
income\_df$grp16s = gsub('>', '', income\_df$educ)  
sub\_anova = aov(logincome~grp16s, data= income\_df)  
print(summary(sub\_anova))

## Df Sum Sq Mean Sq F value Pr(>F)   
## grp16s 3 215.7 71.89 83.02 <2e-16 \*\*\*  
## Residuals 2580 2234.1 0.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

print(summary(income\_anova))

## Df Sum Sq Mean Sq F value Pr(>F)   
## educ 4 217.7 54.41 62.87 <2e-16 \*\*\*  
## Residuals 2579 2232.1 0.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

f <- 71.89 / 54.41  
print(f)

## [1] 1.321264

dof\_total <- 2232.1  
dof\_residual <- 2234.1  
p\_val = pf(f, dof\_total, dof\_residual)

* Based on the sample data, comparing the test group who received only a bachelors degree when compared to the students who received a masters or doctrorite degree (16 plus years of education). Takes the same assumptions as with a complete ANOVA. Using Logarithmic transformed data-set to approximate normality, and at a .95 significance level, we fail to reject the null hypothesis that there is no difference in income between those with a bachelors degree and those with a higher degree. The p value for this test with 2232 and 2234 degrees of freedom is .999

*3*. Now assume that you cannot assume the standard deviations are the same. Conduct another complete analysis of the question in Chapter 5, 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 and should use the results from the complete analysis without the assumption of equal standard deviations. Perform the test in SAS or R.

* If we cannot assume equal variance after a numerical transformation of the income dataset, we can then utilize a non-parametric version of the ANOVA test called the Kruskal Wallis test. For this test, we assume, since the children were chosen at random, that the observations are independent. That all groups have a similar shaped distribution. In this case, it is well approximated by an expoential distribution, and that There are two or more indepdent groups. In this case there are five independent groups.

nparam\_test <- kruskal.test(income2005~educ, data= income\_df)  
print(nparam\_test)

##   
## Kruskal-Wallis rank sum test  
##   
## data: income2005 by educ  
## Kruskal-Wallis chi-squared = 349.45, df = 4, p-value < 2.2e-16

* Running tht Kruskal-Wallis rank sum test in R gives us a p-value of 2.2 e -16 which is very similar to the value we got off the log-transformation of the data. this is expected. In this case, since both tests give the same p-value, it would be better to use the log transform; because it does not result in information loss. In either case, we will reject the null hypothesis at a significence level of .95. It is highly unlikley that all of the data was selected from the same distribution.