Sean Kennedy Homework 3 – Statistical Foundations for Data Science

The 5 Assumptions:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The variances of the two populations are equal.
4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired t-test).
5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Question 1:

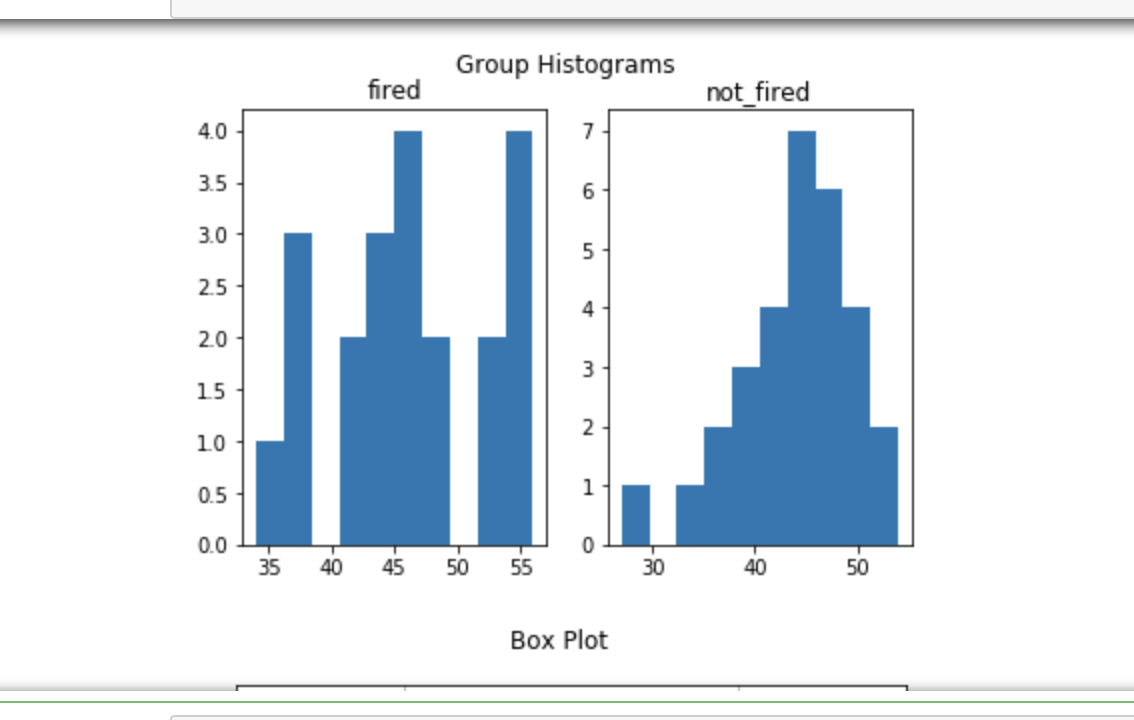
Alpha =0.05

**State The Problem:**

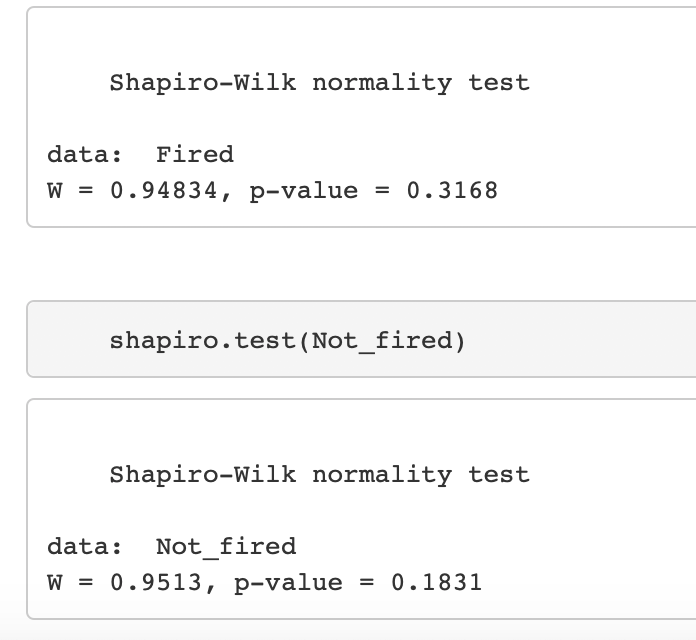
In the United States, it is illegal to discriminate against people based on age. Currently there is a lawsuit pending against the government of American Samoa alleging that older workers are being fired at a significantly higher rate than their younger counterparts. In order to test this hypothesis – we will compare the mean age of a random sample of two groups – those that were fired and those that were not. If we find in favor of the null hypothesis (H0: sample **medians** are equal) – we can assume that there is no evidence of discrimination based on the results of this observational study.

**assumptions of two-sided t-test:**

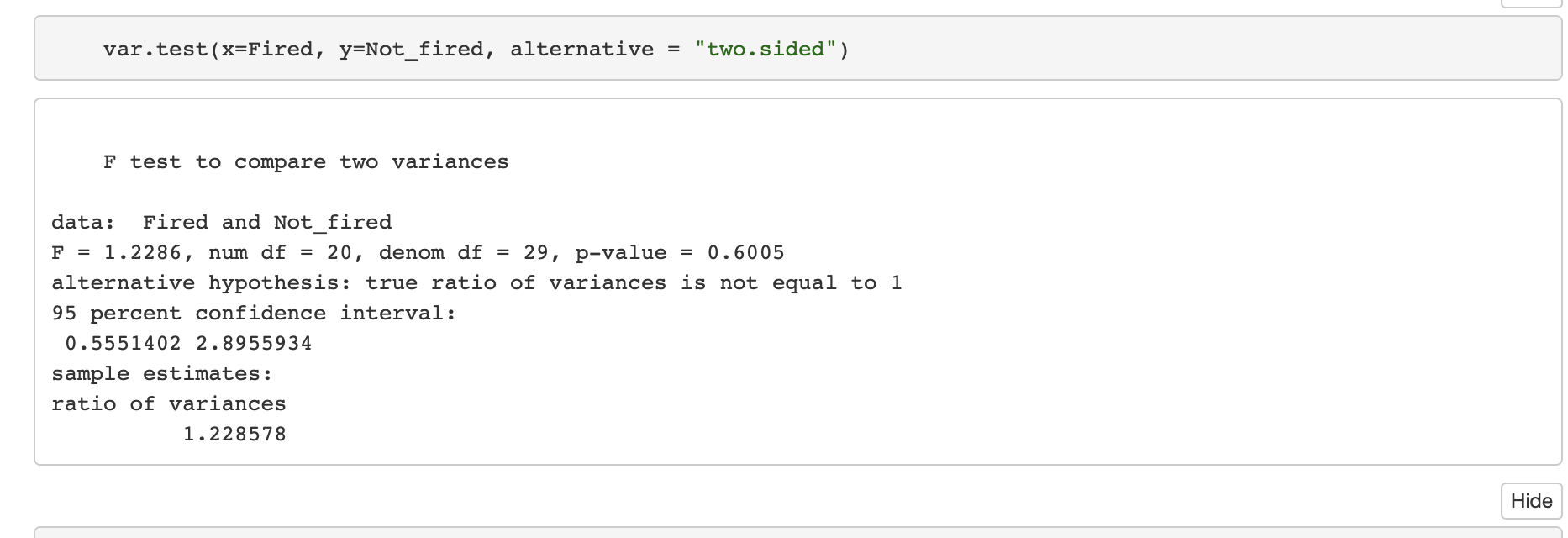
1. Data is generally continuous – the fired group seems to have a few coverage gaps but not significant enough to be considered sparse. This can be inferred from the histograms below:



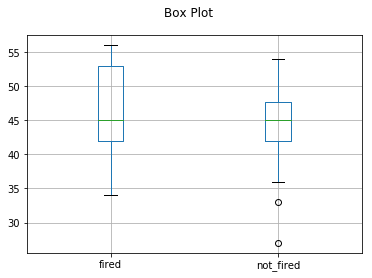
1. Performing a Shapiro Wilke-Test in R to check for normality shows that both samples are roughly normal since both p-values are > 0.05 (can also be inferred from histogram):



1. Performing an F-test in R to check for equal variance shows that both samples have roughly the same variance since the p-value is > 0.05

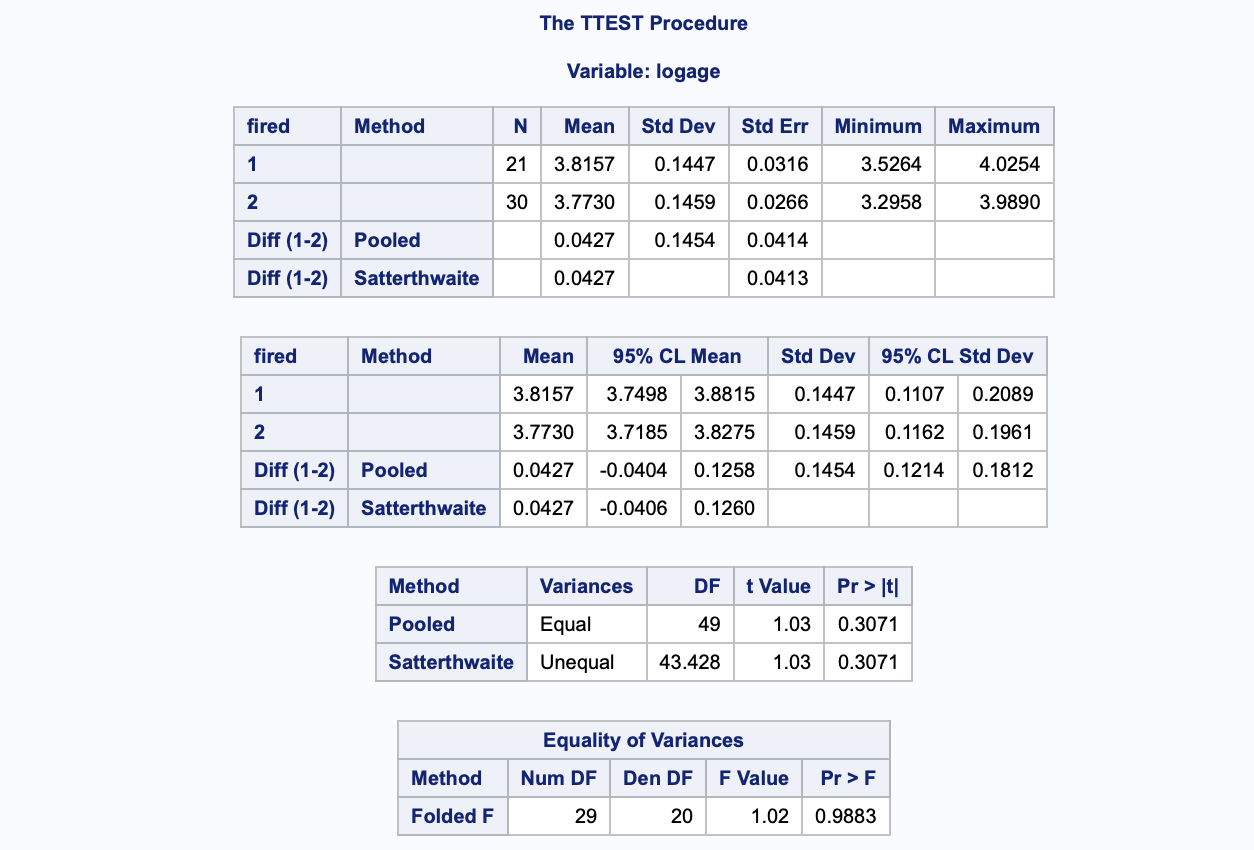


This can also be seen by the fact that the box plots have similarly overlapping variances and equivalent median though the fired group does show slight positive skew:



1. There is no reason to assume dependence of any type for the two samples.
2. Both samples were selected at random.

Because the data has identical medians – it is suggested that we use s two-sided t-test on the log transformed data set.



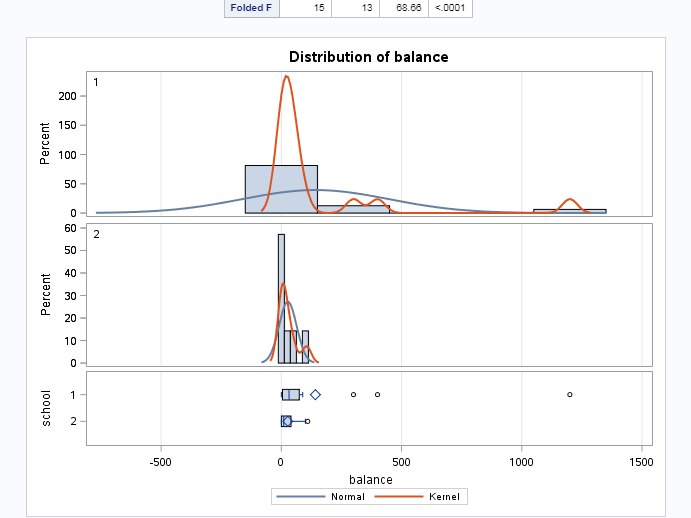
**It is estimated that the median age of fired employees was e0.0427=1.043 times as large as the age of employees that were not fired (p-value = 0.3071). A 95% confidence interval for this multiplicative effect on the median is 0.96 to 1.13 times. Due to the high p-value, we fail to reject the null hypotheses (equal medians). Since randomization was used to select the sample populations, it is safe to assume that amongst the population in American Samoa – there is not sufficient evidence to suggest employees are terminated as a result of their age. Because this was an observational study, direct causal inference cannot be made. The nature of the sample size (limited to American Samoa geographically) makes any extension to larger, more general populations, impossible. Confounding variables such as level of education and duration of employment are not considered at all and should be incorporated into future studies.**

Question 2:

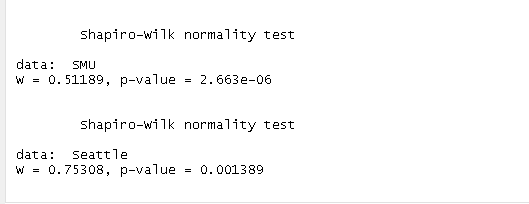
**assumptions of two-sided t-test:**

1. Data is highly skewed. This can be inferred from the histograms below:

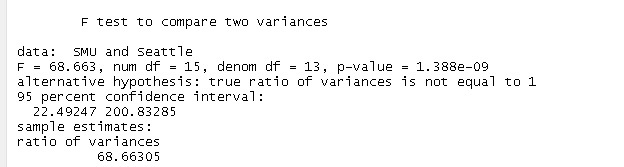
Both samples also suffer from small sample sizes (n << 30)



1. Performing a Shapiro Wilke-Test in R to check for normality shows that both samples are not normally distributed since both p-values are < 0.05 (can also be inferred from histogram):



1. Performing an F-test in R to check for equal variance shows that both samples have wildly different variances since the p-value is <<<< 0.05



1. There is no reason to assume dependence of any type for the two samples.
2. Both samples were selected at random.

Sample sizes are too small, variances are too different and neither distribution is normal. Log data is a possibility – but still suffers from small sample size.

Problem:

In order to determine if coin/paper accepting devices should be installed onto campus vending machines, a study of two populations of students was conducted on two separate campuses (SMU and SeattleU). Students were asked exactly how much cash was in their possession at the exact time of the survey (time variability?). Results were compiled and a permutation test was conducted to assess whether these two populations are representative of the general population of university students .

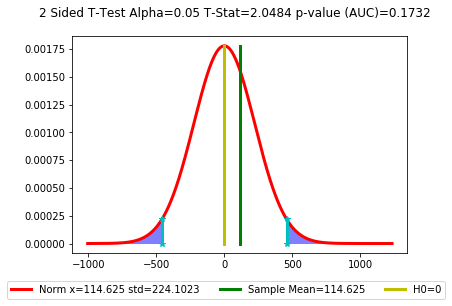
**Since T-Test Assumptions on non-transformed data are suspect, we will rely on a permutation test to calculate the p-value.**

H0: m1=m2 (m1-m2 = 0)

Ha: m1-m2 <> 0

Draw and Shade: Critical t = 2.0484

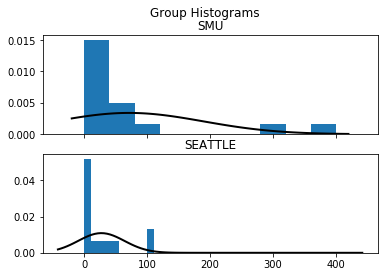
p-value = 0.1732 (t-test Alpha = 0.05)



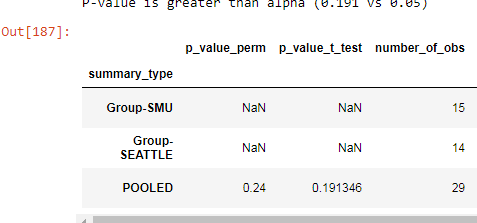
Performing a permutations test (n=1000) leads to a p-value of 0.176 with a corresponding t-statistic of 1.354

**Given that the calculated p-value in both instances was higher than our alpha cutoff of 0.05 – we find that there is very little evidence against the null hypothesis. A 95% CL for the mean diff of [-53.37, 282] indicates that this mean diff (114.6) falls within the center of the interval. This implies that the average cash balances in our study are not representative of the population and should likely be equal had we had better experimental design. Confounding variables such as age, time of day and employment status need to be considered in order limit the bias in this study. Since this is an observational study – no causal inference can be made.**

Re-running the exact same analysis with the SMU sample of 1200 removed yields similar results. Non-transformed data is till not normal (see below):



Both tests (t and permutation) yield similar results to those obtained with the outliers included:

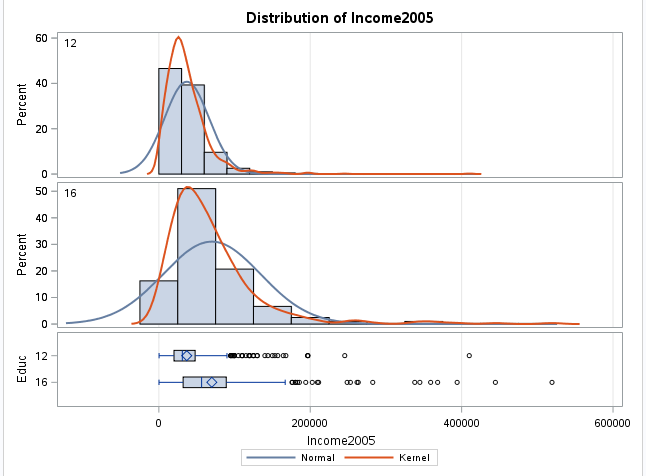
\

**Question 3:**

**Problem:** We wish to test the claim that adding an additional 4 years of education has a profound impact on the salary distribution of employees. To do so – 2 random(?) samples of data from NLSY79 were selected. All subjects were between 41 and 49 years old, had paying jobs in 2005 and had either 12 or 14 years of education at the time of the study. Their incomes for 2005 were recorded.

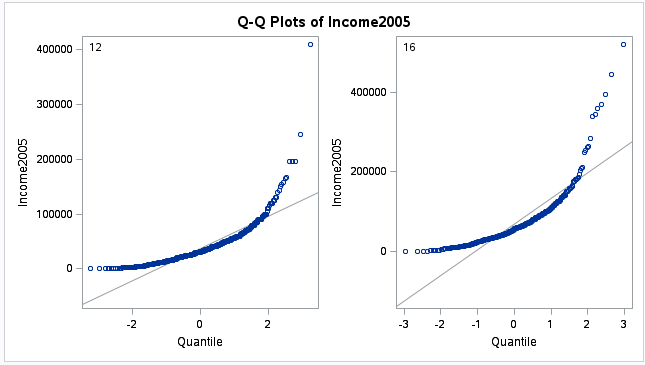
**T-Test Assumptions:**

1. **The data are continuous (not discrete).**



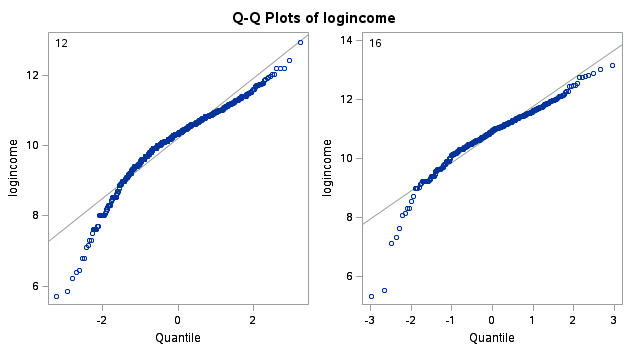
Continuity can be inferred from the histogram

1. **The data follow the normal probability distribution.**

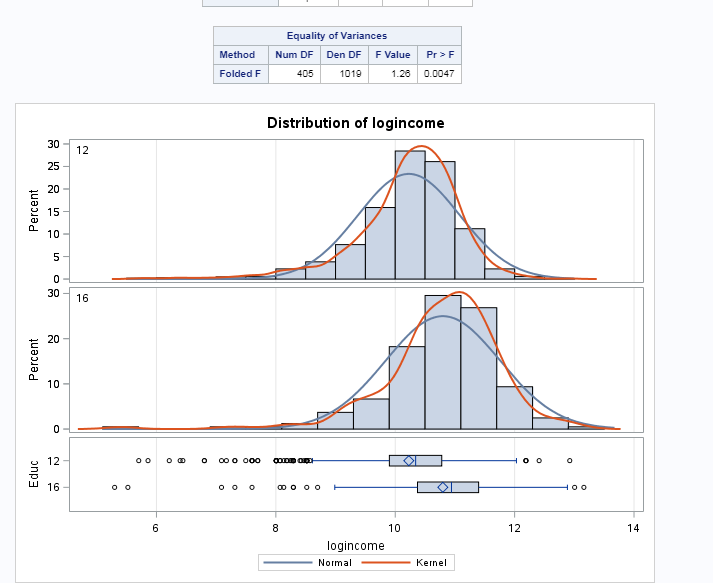


Visual inspection of the Q-Q plots indicate large skew in both samples.

Plots of log income are more normally distributed and mean centered:



1. **The variances of the two populations are equal.**



The log x-formed data exhibits greater variance symmetry than the raw data.

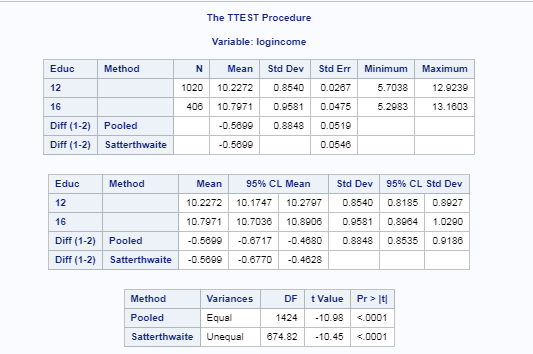
1. **The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired t-test).**

There is no reason to assume a link between any of the individuals in the samples.

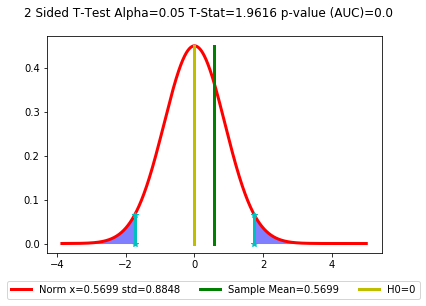
1. **Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.**

This was not explicitly stated – but it is assumed here.

**Two-sided t-test on log transformed data: Alpha = 0.05**



**Critical T = 1.962**



**It is estimated that the median income of the 12 year group was e-0.57=0.565 times as smaller than the income of the 16 year group (p-value = <.0001). A 95% confidence interval for this multiplicative effect on the median is 0.511 to 0.626 times. Due to the extremely low p-value, there is very strong evidence against the null hypothesis (equal medians). Since randomization was used to select the sample populations, it is safe to assume that amongst the population of individuals age 41-49 – there is sufficient evidence to suggest that the increased education led to higher income earned. Because this was an observational study, direct causal inference cannot be made. Though the sample size is very large – ideally large enough to account for confounding variables – there are other features that should be added to this set to enhance the analysis:**

* Starting salary
* Race
* State
* Graduate vs non-Graduates,
* Years on job
* Currently unemployed are excluded

**Of these – state, race and starting salary are the most important. Sample size should be sufficient to account for variations in gender.**