ADULT DATASET

Primary analysis on the influence of Education level on people's occupation and working hours

We are all comfortable with the work each group member has done.

Lingfeng Cao
Ni Zhuang
Shaoran Sun
Xiao Dong
Yilin Liu
Zhehao Li

2016-05-07

TABLE OF CONTENT

I. INTRODUCTION

This report is to study the influence of people's education level on their occupation, using data set "adult.csv" from UCI Machine Learning Repository. Most of the tools, methods, techniques, and theory studied in course STAT 3480 (Nonparametrics Statistics) will be used to facilitate our research. This report is written by Lingfeng Cao, Ni Zhuang, Shaoran Sun, Xiao Dong, Yilin Liu, and Zhehao Li collectively.

Our data source is the AdultUCI from UCI Machine Learning Repository. The AdultUCI data set contains the questionnaire data of the "Adult" database (originally called the "Census Income" Database) formatted as a data.frame. Extraction was done by Barry Becker from the 1994 Census database. The AdultUCI data set contains a data frame with 48842 observations on the following 15 variables. It was originally used to predict whether income exceeds USD 50K/yr based on census data.

We will be using methods learned in class to analyze the data set and find relationships between different attributes, and make hypothesis about the data points, eventually applying tests to study the actual relationships. Specifically, we are interested in the relationship between people's education level, occupation, and working hours. We chose the group of age 70 as our main interest, because they were the group born around the Great Depression and influenced by the WWII when they entered the workforce, two specific time periods that likely impacted the education, working habit, and occupation of the subject group. Our questions are mostly concerned with the relationship between education and occupation of this particular generation, due to the special background of this group.

i. Why Can't We Use T-Test

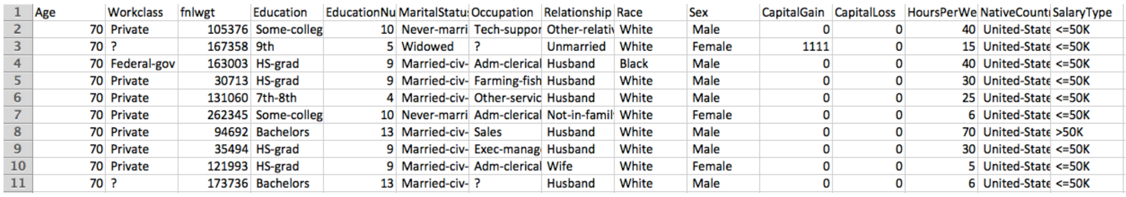
The use of student t-test requires the population of interest to be continuous and normally distributed. The violation of the population normality assumption will result in a great reduce in the power of the test.

ii. Parametric versus Nonparametric

Whereas parametric tests make assumptions about the population distribution, nonparametric tests require minimal assumptions about the population distribution. The specific form of our data's distribution is unknown and we do not have enough evidence to make strong assumptions such as population normality. Also our dataset contains a number of categorical variables of both types: ordinal and nominal. Thus nonparametric is more applicable and should yield more reliable results compared to student t-test and other parametric tests.

II. SUMMARY STATISTICS

The subgroup of data we will be using is extracted from the original "Adult" dataset, and all the observations extracted share a common value of 70 for the Age variable. The new dataset contains 89 observations for each of the original 13 variables, 5 of which are continuous. Below is a screenshot of the first ten rows of our data. For a clearer view of the data, please see appendix:



1	Age	Workclass	fnlwgt	Education	EducationNu	MaritalStatu	Occupation	Relationship	Race	Sex	CapitalGain	CapitalLoss	HoursPerWe	NativeCount	SalaryType
2	70	Private	105376	Some-colleg	10	Never-marri	Tech-suppor	Other-relati	White	Male	0	0	40	United-State	<=50K
3	70	?	167358	9th	5	Widowed	?	Unmarried	White	Female	1111	0	15	United-State	<=50K
4	70	Federal-gov	163003	HS-grad	9	Married-civ-	Adm-clerical	Husband	Black	Male	0	0	40	United-State	<=50K
5	70	Private	30713	HS-grad	9	Married-civ-	Farming-fish	Husband	White	Male	0	0	30	United-State	<=50K
6	70	Private	131060	7th-8th	4	Married-civ-	Other-servic	Husband	White	Male	0	0	25	United-State	<=50K
7	70	Private	262345	Some-colleg	10	Never-marri	Adm-clerical	Not-in-famil	White	Female	0	0	6	United-State	<=50K
8	70	Private	94692	Bachelors	13	Married-civ-	Sales	Husband	White	Male	0	0	70	United-State	>50K
9	70	Private	35494	HS-grad	9	Married-civ-	Exec-manag	Husband	White	Male	0	0	30	United-State	<=50K
10	70	Private	121993	HS-grad	9	Married-civ-	Adm-clerical	Wife	White	Female	0	0	5	United-State	<=50K
11	70	?	173736	Bachelors	13	Married-civ-	?	Husband	White	Male	0	0	6	United-State	<=50K

For the four questions related to education and occupation and the one extra question about gender and occupation, we will be using the following variables:

	Type	Minimum	Mean	Median	Max	Observations / Levels
Workclass	Categorical	-	-	-	-	7
HoursPerWeek	Quantitative	3	29.7	30	80	89
Education	Categorical	-	-	-	-	12
EducationNum	Quantitative	3	10	9.921	16	89
Sex	Categorical	-	-	-	-	2
CapitalGain	Quantitative	0	993.5	0	20050	89

In Figure 1.1, we can clearly see that the people's education level is not uniformly distributed. The majority of people in age group 70 have one of the following degrees:

High School degree, Bachelor degree or *degree from some College*. Most people spent around 9 to 10 years in school, with another significant spike in 13 years, which is approximately the time for regulation Bachelor degree, as shown in Figure 1.2.

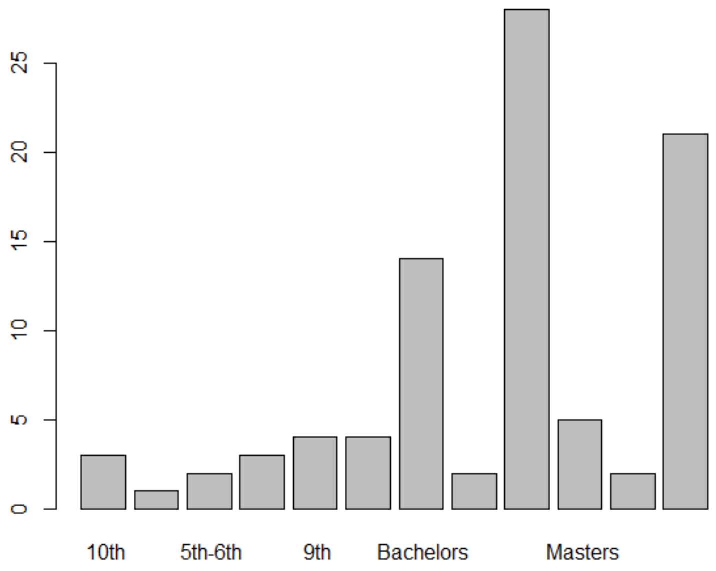


Figure 1.1

Histogram of the number of people in age group 70 based on education level

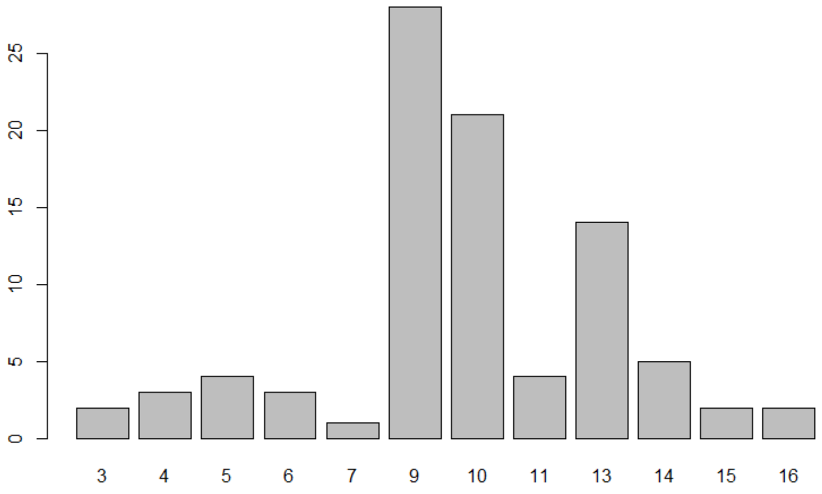


Figure 1.2

Histogram of the number of years people in age group 70 spent on education

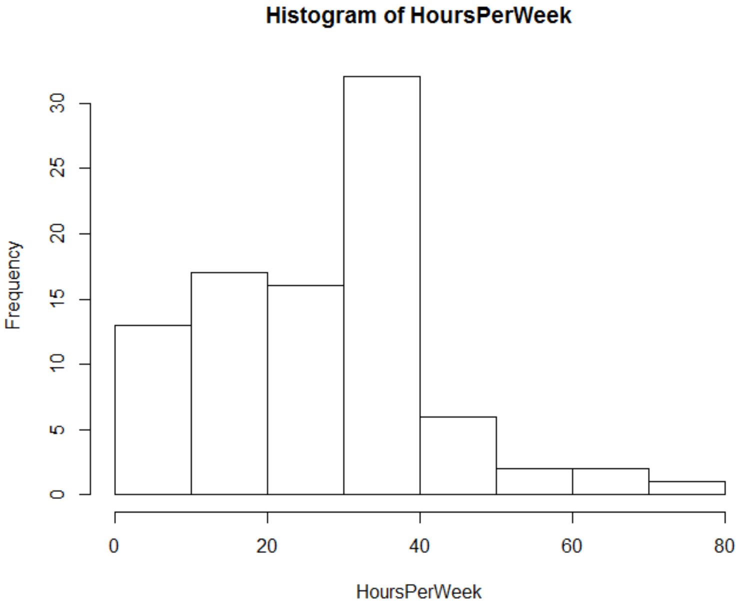


Figure 1.3

Histogram of working hours per week

The histogram above (Figure 1.3) of working hours per week shows that the distribution is not normal. The histogram shows a spike in 40 hours week working pattern, and significant less people working overtime than undertime. In Figure 1.4, despite few outliers the spread and location of working hours grouped by different occupation types are approximately the same.

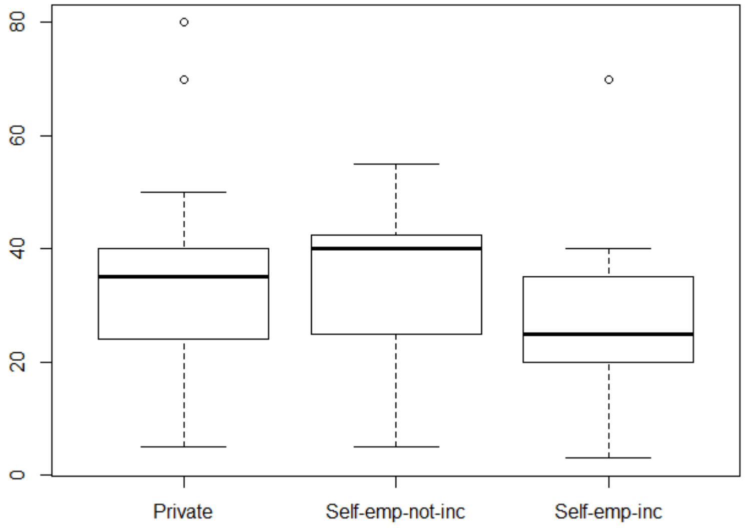


Figure 1.4

Boxplot of hours of working hours based on occupation types

III. STATISTICAL TESTS AND ANALYTICAL METHODOLOGIES

i. Any difference in Education Years (EducationNum) between Professional/Specialty and Executive Managerial?

We suspect that people who work as executive managerial could acquire their knowledge from real life experience and other sources, not necessarily dependent on their years of education. Education years may not be as important for executive managerial occupation as professional specialty.

We plan to use a two-sided Wilcoxon Rank-Sum (WRS) Test for this case. WRS is a two-sample permutation test based on W, the sum of the ranks of the observations from one of the treatments. It simply ranks each observation after combing the groups together, finding all possible permutations of the ranks and calculates the p-value based on the permutation results of the probability of how many number of rank sums are greater than the observed rank sum. As for the hypothesis, the null hypothesis is that H_0: $F_1(x) = F_2(x)$, that people who work as professional specialty and executive manager have same distribution in terms of education years. The alternative hypothesis is that H_a: $F_1(x) \neq F_2(x)$, that these two groups of people with different occupation have different distributions of education years.

However, WRS requires some assumptions, and the successful implementation of WRS depends on whether or not these assumptions are met. Firstly, the distribution needs to be continuous. Moreover, it should have identical population distributions. Thirdly, it requires equal variances among groups. The first three assumptions are met here. In order to apply the WRS, we need to test the equal variances.

Here we need to use the RMD test for median to test for the deviances. We do not know the normality of the sample variances and we do not know the location parameters. We will use the sample medians instead of means to obtain the deviances for each group.

The null hypothesis for RMD Test will be: H_0: $\sigma_1 = \sigma_2$. It means that the professional speciality and executive manager have the equal variances in education number. The alternative hypothesis for RMD Test will be: H_a: $\max(\sigma_1, \sigma_2) / \min(\sigma_1, \sigma_2) > 1$. This is a two-sided RMD Test and we will use it to test if the variances for two treatments are different.

ii. Working Hours and Work Class

We are interested in knowing if working hours per week differs among working classes. Since we have more than two treatments and we want to know if there is difference among these three treatments, we will use a Kruskal-Wallis test to test if at least one work class has different working hours than other classes.

Kruskal-Wallis Test is a nonparametric rank test that compares k treatments by replacing the original observations with ranks and performs the permutation F-test on those ranks. For large samples, KW test statistic follows chi-square distribution approximately with (k-1) degrees of freedom. Before applying Kruskal-Wallis test, we also need to check the Kruskal-Wallis test does require the assumption for homoscedasticity, which we have already checked using the boxplot in the summary statistics. The null hypothesis is that there is no difference in working hours among private, incorporated self-employment, and unincorporated self-employment working class. The alternative hypothesis is that at least one of the working classes has a different working hour from the others.

iii. Working Hours and Education

We anticipate that higher education lead to more working hours. We suspect that a higher education level will lead to better jobs, and better jobs will likely require more working hours. In this situation, we assume that people with bachelor degrees have longer working hours than those with some college degree than those with bachelor degree. We decide to use Jonchheere-Terpstra Test to test for an increasing trend. Under null hypothesis, the three different population distributions are the same. Therefore, we use a boxplot to check for equal variance cross these three groups before applying Jonchheere-Terpstra Test.

$$H_0 : F_1(x) = F_2(x) = F_3(x)$$

Our null hypothesis is that population distribution are the same for people with graduate education (HS-grad), some-college and bachelor degree.

$$H_a : F_1(x) \leq F_2(x) \leq F_3(x)$$

Alternative hypothesis is that working hours increase with education levels.

iv. Correlation between Gender and Occupation

Besides the relationship between education and occupation, we are also interested in knowing if gender plays a factor in determining the types of occupation a person of the 70-year-old generation chooses. For this question, we hope to determine if there exists a correlation between the gender of a person and the occupation he or she chooses. The jobs for the age group of people fall into thirteen categories, and observations for the gender only include “Male” and “Female,” another categorical variable. The contingency table is shown below.

	Male	Female
Unknown	15	9
Adm-clerical	2	6
Craft-repair	4	0
Executive-managerial	15	2
Farming-fishing	4	0

Handlers-cleaners	3	0
Machine-op-inspection	3	1
Other Services	2	6
Professional/Specialty	5	2
Protective Service	0	1
Sales	5	1
Technical Support	2	0
Transport-moving	1	0

To determine if an association between two categorical variables exists, a chi-square test should be used, and in our case we should be employing a permutation chi-square test, because of a violation of one of the two assumptions. A traditional chi-square test requires two assumptions: the independence assumption, that the observations are not dependent, and the sample size assumption, that most of the expected cell counts should be at least 5. The observations in our data violate the second assumption, as more than half of the cell counts are less than 5. Hence we will be using permutation chi-square test, an alternative method that solves the problem by creating a permutation distribution of the X^2 statistic. In our case, we will permute the data 1000 times. The p-value of this test is the fraction of the number of test statistics greater than the original chi-square test statistic over the number of all test statistics from our permutation (note that the chi-square test is always an one-tailed test). The null hypothesis is that there is no association between gender and occupation among the surveyees, and the alternative hypothesis is that there exists a correlation between gender and occupation among the 70 year olds.

v. The Effect of Education and Capital Gains on the Hours per Week for Working

We plan to use the regression to address this problem. Regressions are used when we want to find the scale effect of how variables can affect a dependent variable. In this case, we hope to see answers to two basic questions: (1) if people with higher education tend to work shorter and (2) if people with extra capital gains tend to affect their decisions of how long to work.

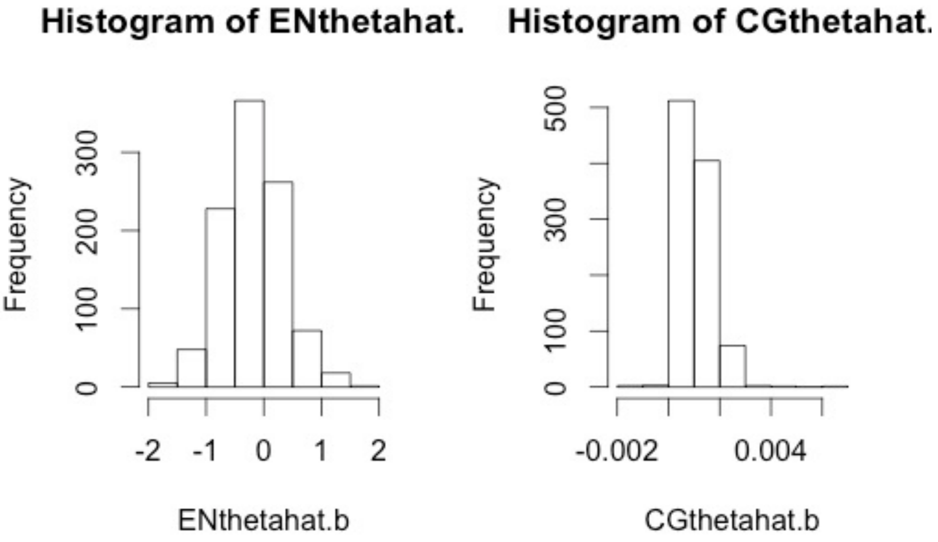
To be more specific, we plan to use the Bootstrapping Regression Method to address the problem. Bootstrapping is a nonparametric approach to statistical inference that substitutes computation for more traditional distributional assumptions and asymptotic results. As the simple multiple linear regression requires normal distribution for the data, we can not simply apply that to our analysis as the variables HoursPerWeek, EducationNum and CapitalGain are all skewed and not normally distributed due to our small data size and the nature of those data. The bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small. It is also possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically. Further, the bootstrap regression will provide us with a narrower confidence interval.

We are going to construct a bootstrap interval for the slope of the line for predicting HoursPerWeek from numbers of education and extra capital gains. We are performing the standard test of slope: $H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \beta_1 \neq 0, \beta_2 \neq 0.$

IV. RESULTS FROM THE TESTS AND APPLICATIONS

Question	**Test**	**Test statistic**	**Test statistic value**	**P-value**
I	RMD Test for Deviances	RMD	1.326797	0.6411
	Wilcoxon Rank-sum Test	W_{obs}	87.5	0.07334
II	Kruskal-Wallis Test	KW_X^2	2.7133	0.2575
III	Joncheere-Terpstra	JT	748	0.0777
IV	Chi-square Test	X^2	26.8	0.004
V	Bootstrap Regression			
	Independent Variable	*95% Lower Bootstrap Interval*	*95% Upper Bootstrap Interval*	*P-value*
	EducationNum	-1.141598	0.9127412	Greater than 0.05
	CapitalGain	0.0002935431	0.002403871	Smaller than 0.05

Figure 4.2 Histogram of bootstrap distributions (Question 5)



V. DISCUSSION OF RESULTS

i. Any difference in Education Years (EducationNum) between Professional/Specialty and Executive Managerial?

The RMD Test gave us the p-value of 0.64385 > 0.05, and hence we fail to reject the null hypothesis that there are equal variances between population distribution of education level for occupation professional specialists and executive managerial at a significance level of 0.05. We came to the conclusion that the two groups (Prof-specialty & Exec-managerial) have equal variances and we can proceed to use the WRS Test.

The Wilcoxon Rank Sum Test gave us the W test statistic as 87.5 with p-value 0.07334 > 0.05. We fail to reject the null hypothesis at a significance level of 0.05. We have weak evidence to show there is a difference between adult occupation professional specialty and executive managerial. We came to the conclusion that there is not much difference in distribution of education years (EducationNum) between occupation professional specialty and executive managerial.

ii. Working Hours and Work Class

The box plot shows that there is not significant difference in variances of working hours for these three work classes. The magnitude is about the same across these three different groups.

The Kruskal-Wallis test gives us the KW test statistic of 2.7133 with p-value 0.2575, which is higher than the significant level of 0.05. Hence we fail to reject the null hypothesis and conclude that there is not a significant difference in working hours among these three working classes.

iii. Working Hours and Education

From the boxplot, variances do not differ much cross these three population distributions. Thus, our assumption that population distributions for these three groups have equal variance is met and we could continue to run the Jonckheere-Terpstra Test.

Jonckheere-Terpstra Test gives JT test statistic 748 with p-value 0.0777 > 0.05. Even though p-value is small, we still fail to reject the null hypothesis at a significance level of 0.07. Therefore, we conclude that the distributions for working hours are the same among people with high school graduate education (HS-grad), some-college and bachelor degree. There is not an increasing trend in working hours with education levels.

iv. Correlation between Gender and Occupation

The permutation chi-square test yields test statistic of 26.8 and a p-value of 0.004. Under the conventional significance level of 0.05, we reject the null hypothesis in favor of the alternative hypothesis, and conclude that strong evidence suggests that there exists an association between gender and occupation choice.

One thing noticeable is that one of the occupation categories is “Unknown,” and it has the biggest cell counts in both genders. This category may be a result of the surveyee’s reluctance to answer or the surveyors’ failure to properly record the data. In either case, this category may negatively contribute to the inaccuracy of our test results. However, since they are only one of the thirteen categories and their cell counts are not huge outliers, this little uncertain element will not affect our conclusion and test results.

v. The Effect of Education and Capital Gains on the Hours per Week for Working

The 95% bootstrap confidence interval for the slope of Education Numbers is (-1.141598, 0.9127412). The p−value will be greater than 0.05 because 0 is in the confidence interval. We can make a conclusion that we have strong evidence that the slope may be equal to 0. Education may not affect working hours per week.

The 95% bootstrap confidence interval for the slope of Extra Capital Gains is (0.0002935431, 0.002403871). The p−value will be less than 0.05 because 0 is not in the confidence interval. Since the bootstrap interval is so close to 0, we can make a conclusion that we have weak evidence indicating that the slope may be equal to 0. Capital Gains may affect working hours per week.

VI. CONCLUSION & SUMMARIZING REMARKS

As indicated by our test results, we can say that there is no much difference among people's level of education, occupations, and respective working hours per week for the generation that was 70 years old in 1994, when the survey was collected. For the first test, we have weak evidence to show there is a difference between adult occupation professional specialty and executive managerial. We came to the conclusion that there is not much difference in distribution of education years between occupation professional specialty and executive managerial jobs, one piece of evidence that people's occupations are less likely dependent on their education levels.

We also find that for participants from private, incorporated self-employment, and unincorporated self-employment working class, the lengths of their working hours do not have a significant difference. Such a result may indicate a nationwide commonly practiced working schedule, and that in order for someone to at least support him or herself, a fixed amount of working hours is necessary.

The above induction about fixed working hours may be reinforced by the next test we performed on the working hours and school degrees. Our test results show that the working hours for people with different degrees are mostly similar, and no increasing or decreasing trends are found. The fact that survey participants with high school graduate education having work hours similar to those with some college education and/or complete college education, along with the working hours and working class relationship we found, may indicate that the society has a set of working schedule that is sufficient to support basic needs of the working population.

We have also performed a test on the correlation between gender and occupations, and our results show that gender is related to people's choice of jobs. This can be seen in our table of gender and jobs. We have more male participants than female participants, and 7.5 times more male working as executive-managerial personnel than female. We can also see that male participants are more involved in labor works such as farming, fishing, craft repairing and cleaning.

The last test we performed determines the effect of education and capital gains on working hours per week. As stated earlier, the questions we wish to solve are first, if people with higher education tend to work shorter; and two, if capital gains stimulate people to be flexible with their working schedules. Our test results show that education level may not affect working hours per week, which agrees with the hypothesis we set in the earlier steps - people need to work for a certain length to support themselves. Capital gains, however, may affect working hours per week. This also can be related to our earlier hypothesis, as the reasoning behind working every week is to earn money to live better, and an extra amount of capital gain leads people directly to the final step of living better, therefore affecting working hours per week.

In conclusion, for the people who were born at 1924, their level of education, occupation, and working hours do not share much relationship or correlation, despite the unique historical periods they grew up with.

VII. SOFTWARE / CODE APPENDIX

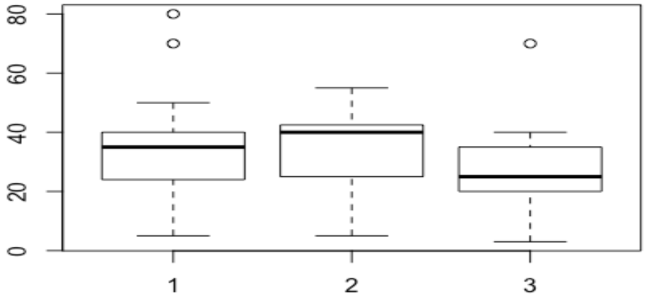
i. Any Difference Between the Two Occupations

```
Directory <- "~/Documents/University of Virginia/2016 SPRING/STAT 3480/
Final Project/Adult.csv"

Adult <- read.csv(Directory)
attach(Adult)

library(jmuOutlier)

## Warning: package 'jmuOutlier' was built under R version 3.2.4

# RMD test
# Test if people who work as professional sepciality have larger educat
ion num than people who work as executive managerial.

index.prof = which(Occupation ==" Prof-specialty")
index.Exec = which(Occupation ==" Exec-managerial")
EducationNum[index.Exec]

##  [1]  9 10 13 13 10 10 10  9 14  9 16 13 10 14 10 10 13

EducationNum[index.prof]

## [1] 15 14 13 13 14  9 15

rmd.test(EducationNum[index.Exec],EducationNum[index.prof], alternative
 = "two.sided")

## [[1]]
## [1] "p-value was estimated based on 20000 simulations."
##
## $alternative
## [1] "two.sided"
##
## $rmd.hat
## [1] 1.326797
##
## $p.value
## [1] 0.6411

wilcox.test(EducationNum[index.prof],EducationNum[index.Exec],alternati
ve = "two.sided")

## Warning in wilcox.test.default(EducationNum[index.prof],
## EducationNum[index.Exec], : cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  EducationNum[index.prof] and EducationNum[index.Exec]
## W = 87.5, p-value = 0.07334
## alternative hypothesis: true location shift is not equal to 0
```

ii. Working Hours and Work Class

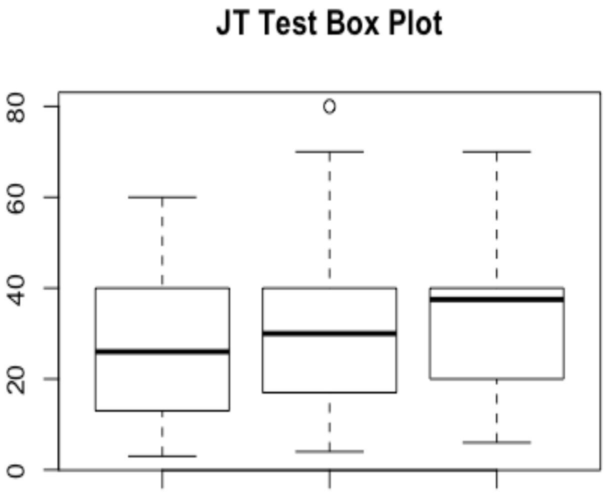
```
index.private = which(Workclass==" Private")
index.Selfempnotinc= which(Workclass==" Self-emp-not-inc")
index.Selfempinc = which(Workclass==" Self-emp-inc")

adult.new = adult[c(index.private,index.Selfempinc,index.Selfempnotin
c),]
boxplot(HoursPerWeek[index.private],HoursPerWeek[index.Selfempinc],Hour
sPerWeek[index.Selfempnotinc])
```



```
kruskal.test(HoursPerWeek~Workclass,data = adult.new)

##
##  Kruskal-Wallis rank sum test
##
## data:  HoursPerWeek by Workclass
## Kruskal-Wallis chi-squared = 2.7133, df = 2, p-value = 0.2575
```

iii. Working Hours and School Degree

```
#Boxplot
HSgrad=HoursPerWeek[which(Education==" HS-grad")]
SomeCollege=HoursPerWeek[which(Education==" Some-college")]
Bachelor=HoursPerWeek[which(Education==" Bachelors")]

boxplot(HSgrad,SomeCollege,Bachelor,main="JT Test Box Plot")
```



```
#Jonckheere-Terpstra Test
library(clinfun)
pieces<-list(HSgrad,SomeCollege,Bachelor)
n<-c(28,21,14)
grp<-as.ordered(factor(rep(1:length(n),n)))
jonckheere.test(unlist(pieces),grp,alternative="increasing")
```

```
## Warning in jonckheere.test(unlist(pieces), grp, alternative = "incre
asing"): Sample size > 100 or data with ties
##  p-value based on normal approximation. Specify nperm for permutatio
n p-value

## 
##  Jonckheere-Terpstra test
## 
## data:
## JT = 748, p-value = 0.07777
## alternative hypothesis: increasing
```

iv. Correlation between Gender and Occupation

```
### CONTINGENCY TABLE & DATA MANIPUTATION

  MaleOcc <- table(Occupation[which(Sex == " Male")])
  FemaleOcc <- table(Occupation[which(Sex == " Female")])
  Row1 <- as.numeric(MaleOcc)     # Male
  Row2 <- as.numeric(FemaleOcc)   # Female
  table(Occupation)

## Occupation
##                  ?       Adm-clerical       Craft-repair
##                 24                  8                  4
##    Exec-managerial    Farming-fishing  Handlers-cleaners
##                 17                  4                  3
##  Machine-op-inspct      Other-service     Prof-specialty
##                  4                  8                  7
##    Protective-serv              Sales       Tech-support
##                  1                  6                  2
##   Transport-moving
##                  1

  sum(Sex == " Male")        # number of males

## [1] 61

  sum(Sex == " Female")      # number of females

## [1] 28

### RUNNING THE PERMUTATION CHI-SQUARE TEST

  ### Make observed contingency table and calculate stat
  Table = rbind(Row1,Row2)
  teststat.obs = chisq.test(Table)$statistic

## Warning in chisq.test(Table): Chi-squared approximation may be incor
rect

  teststat.obs

## X-squared
##  26.84885

  ### create the prefernce data and the gender data
  preference = c(rep("Unknown",24), rep("AdmClerical",8),
                 rep("CraftRepair",4), rep("Exec",17),
                 rep("FarmFish",4), rep("Handler",3),
                 rep("MachineOpIns",4), rep("Other",8),
                 rep("Prof",7), rep("Protect",1),
                 rep("Sales",6), rep("Tech-support",2),
                 rep("Transport",1))
  gender = c( rep("Male",61), rep("Female",28) )
  table(preference); table(gender)

## preference
##   AdmClerical  CraftRepair         Exec     FarmFish      Handler
##             8            4           17            4            3
##  MachineOpIns        Other         Prof      Protect        Sales
##             4            8            7            1            6
##  Tech-support    Transport      Unknown
##             2            1           24

## gender
## Female   Male
##     28     61

  y = preference; x = gender
  teststat = rep(NA, 1000)
  for(i in 1:1000) {
  ### randomly "shuffle" the y data between the x groups
  ySHUFFLE = sample(y)
  ### compute chi-square stat for the shuffled data
  TableSHUFFLE = table(x,ySHUFFLE)
  teststat[i] = chisq.test(TableSHUFFLE)$statistic
  }

### calculate the approximate p-value
  sum(teststat >= teststat.obs)/1000

## [1] 0.004
```

v. The Effect of Education and Capital Gains on the Hours per Week for Working

```
### create our data
oursample = adult
ENthetahat = lm(HoursPerWeek~EducationNum+CapitalGain,data=oursample)$c
oeff[2]
CGthetahat = lm(HoursPerWeek~EducationNum+CapitalGain,data=oursample)$c
oeff[3]
ENthetahat; CGthetahat;

## EducationNum
##   -0.1740954

##  CapitalGain
## 0.0009557792
```

```r
ENthetahat.b = rep(NA,1000); CGthetahat.b = rep(NA,1000);
for (i in 1:1000) {
  ### draw the bootstrap sample and calculate thetahat.b
  index = 1:89
  bootindex = sample(index, 89, replace=T)
  bootsample = oursample[bootindex,]
  ENthetahat.b[i] = lm(HoursPerWeek~EducationNum+CapitalGain,data=bootsample)$coeff[2]
  CGthetahat.b[i] = lm(HoursPerWeek~EducationNum+CapitalGain,data=bootsample)$coeff[3]
}
par(mfrow=c(1,2))
hist(ENthetahat.b); hist(CGthetahat.b);
```

Age	Workclass	fnlwgt	Education	EducationNu	MaritalStatu	Occupation	Relationship	Race	Sex	CapitalGain	CapitalLoss	HoursPerWe	NativeCountr	SalaryType
70	Private	105376	Some-colleg	10	Never-marri	Tech-suppor	Other-relativ	White	Male	0	0	40	United-State	<=50K
70	?	167358	9th	5	Widowed	?	Unmarried	White	Female	1111	0	15	United-State	<=50K
70	Federal-gov	163003	HS-grad	9	Married-civ-	Adm-clerical	Husband	Black	Male	0	0	40	United-State	<=50K
70	Private	30713	HS-grad	9	Married-civ-	Farming-fish	Husband	White	Male	0	0	30	United-State	<=50K
70	Private	131060	7th-8th	4	Married-civ-	Other-servic	Husband	White	Male	0	0	25	United-State	<=50K
70	Private	262345	Some-colleg	10	Never-marri	Adm-clerical	Not-in-famil	White	Female	0	0	6	United-State	<=50K
70	Private	94692	Bachelors	13	Married-civ-	Sales	Husband	White	Male	0	0	70	United-State	>50K
70	Private	35494	HS-grad	9	Married-civ-	Exec-manag	Husband	White	Male	0	0	30	United-State	<=50K
70	Private	121993	HS-grad	9	Married-civ-	Adm-clerical	Wife	White	Female	0	0	5	United-State	<=50K
70	?	173736	Bachelors	13	Married-civ-	?	Husband	White	Male	0	0	6	United-State	<=50K
70	Private	642830	HS-grad	9	Divorced	Protective-s	Not-in-famil	White	Female	0	0	32	United-State	<=50K
70	Self-emp-no	172370	Prof-school	15	Married-civ-	Prof-specialt	Husband	White	Male	0	0	25	United-State	<=50K
70	Private	176285	HS-grad	9	Divorced	Other-servic	Not-in-famil	White	Female	0	0	23	United-State	<=50K
70	?	293076	Some-colleg	10	Married-civ-	?	Husband	White	Male	0	0	30	United-State	<=50K
70	Local-gov	176493	Some-colleg	10	Widowed	Adm-clerical	Not-in-famil	White	Female	0	0	17	United-State	<=50K
70	Self-emp-no	150886	Some-colleg	10	Married-civ-	Craft-repair	Husband	White	Male	0	0	25	United-State	<=50K
70	Private	77219	HS-grad	9	Married-civ-	Craft-repair	Husband	White	Male	0	0	37	United-State	<=50K
70	Private	220589	Some-colleg	10	Widowed	Exec-manag	Not-in-famil	White	Female	0	0	12	United-State	<=50K
70	Local-gov	88638	Masters	14	Never-marri	Prof-specialt	Unmarried	White	Female	7896	0	50	United-State	>50K
70	Private	145419	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	5	United-State	<=50K
70	Private	102610	Some-colleg	10	Divorced	Other-servic	Not-in-famil	White	Male	0	0	80	United-State	<=50K
70	Private	282642	HS-grad	9	Married-civ-	Handlers-cle	Husband	White	Male	0	2174	40	United-State	>50K
70	Self-emp-no	280639	HS-grad	9	Widowed	Other-servic	Other-relativ	White	Female	2329	0	20	United-State	<=50K