STATISTICS FOR BUSINESS ANALYTICS I

LAB ASSIGNMENT I

Georgia Vlassi – p2822001

Professors: Mr. Ntzoufras, Mr. Leriou

MSc in Business Analytics | Part Time 2020

Department of Management Science & Technology

Athens, Greece
December 2nd 2020

MSc in
Business Analytics

The salary data frame contains information about 474 employees hired by a Midwestern bank between 1969 and 1971. It was created for an Equal Employment Opportunity (EEO) court case involving wage discrimination. The file contains beginning salary (SALBEG), salary now (SALNOW), age of respondent (AGE), seniority (TIME), gender (SEX coded 1 = female, 0 = male) among other variables.

Read the dataset "salary.sav" as a data frame and use the function str() to understand its structure.

To import and read the data frame, set the working directory and use read.spps() function:

```
#import data
require(foreign)
setwd('C:\\Users\\tzina\\OneDrive\\Documents\\LabAssignment')

#read data frame
salary <- read.spss("salary.sav", to.data.frame = T)

#View data frame
View(salary)
```

	id EMPLOYEE CODE	salbeg BEGINNING SALARY	sex SEX OF EMPLOYEE	time JOB SENIORITY	age AGE OF EMPLOYEE	salnow CURRENT SALARY	edlevel EDUCATIONAL LEVEL	work WORK EXPERIENCE	jobcat EMPLOYMENT CATEGORY	minority MINORITY CLASSI
1	1	8400	MALES	81	28.50	16080	16	0.25	COLLEGE TRAINEE	WHITE
2	2	24000	MALES	73	40.33	41400	16	12.50	EXEMPT EMPLOYEE	WHITE
3	3	10200	MALES	83	31.08	21960	15	4.08	EXEMPT EMPLOYEE	WHITE
4	4	8700	MALES	93	31.17	19200	16	1.83	COLLEGE TRAINEE	WHITE
5	5	17400	MALES	83	41.92	28350	19	13.00	EXEMPT EMPLOYEE	WHITE
6	6	12996	MALES	80	29.50	27250	18	2.42	COLLEGE TRAINEE	WHITE
7	7	6900	MALES	79	28.00	16080	15	3.17	CLERICAL	WHITE
8	8	5400	MALES	67	28.75	14100	15	0.50	CLERICAL	WHITE
9	9	5040	MALES	96	27.42	12420	15	1.17	CLERICAL	WHITE
10	10	6300	MALES	77	52.92	12300	12	26.42	SECURITY OFFICER	WHITE

We use the function str() to display the structure of the data frame:

```
> str(salary)
'data.frame':	474 obs. of  11 variables:
 $ id      : num  1 2 3 4 5 6 7 8 9 10 ...
 $ salbeg  : num  8400 24000 10200 8700 17400 ...
 $ sex     : Factor w/ 2 levels "MALES","FEMALES": 1 1 1 1 1 1 1 1 1 1 ...
 $ time    : num  81 73 83 93 83 80 79 67 96 77 ...
 $ age     : num  28.5 40.3 31.1 31.2 41.9 ...
 $ salnow  : num  16080 41400 21960 19200 28350 ...
 $ edlevel : num  16 16 15 16 19 18 15 15 15 12 ...
 $ work    : num  0.25 12.5 4.08 1.83 13 ...
 $ jobcat  : Factor w/ 7 levels "CLERICAL","OFFICE TRAINEE",..: 4 5 5 4 5 4 1 1 1 3 ...
 $ minority: Factor w/ 2 levels "WHITE","NONWHITE": 1 1 1 1 1 1 1 1 1 1 ...
 $ sexrace : Factor w/ 4 levels "WHITE MALES",..: 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "variable.labels")= Named chr [1:11] "EMPLOYEE CODE" "BEGINNING SALARY" "SEX OF EMPLOYEE" "JOB SENIORITY" ...
  ..- attr(*, "names")= chr [1:11] "id" "salbeg" "sex" "time" ...
 - attr(*, "codepage")= int 1253
```

File salary.sav contains 474 observations (objects) with 11 variables each. There are two categories of variables, 7 numeric (int & float) and 5 factor. Factor variables contained by 2, 3, 4 and 7 levels respectively.

Get that summary statistics of the numerical variables in the dataset and visualize their distribution (e.g. use histograms etc.). Which variables appear to be normally distributed? Why?

Assign the numeric variables to a new variable named *sal_num* and remove column id, as it is not necessary for our implementations. Use describe() function to get the summary statistics.

```
> round(t(describe(sal_num)),2)
           salbeg   time    age  salnow edlevel   work
vars         1.00   2.00   3.00    4.00    5.00   6.00
n          474.00 474.00 474.00  474.00  474.00 474.00
mean      6806.43  81.11  37.19 13767.83   13.49   7.99
sd        3148.26  10.06  11.79  6830.26    2.88   8.72
median    6000.00  81.00  32.00 11550.00   12.00   4.58
trimmed   6187.94  81.15  35.84 12479.66   13.54   6.41
mad       1494.46  13.34   9.08  3424.81    4.45   5.43
min       3600.00  63.00  23.00  6300.00    8.00   0.00
max      31992.00  98.00  64.50 54000.00   21.00  39.67
range    28392.00  35.00  41.50 47700.00   13.00  39.67
skew         2.83  -0.05   0.86    2.11   -0.11   1.50
kurtosis    12.18  -1.16  -0.58    5.27   -0.29   1.65
se         144.60   0.46   0.54  313.72    0.13   0.40
```

For each numeric variable we have the mean, the standard deviation (sd), the median, the adjusted mean(trimmed), the mean absolute deviation (mad), the min value, the max value, the range, the skewness, the kurtosis and the standard error.

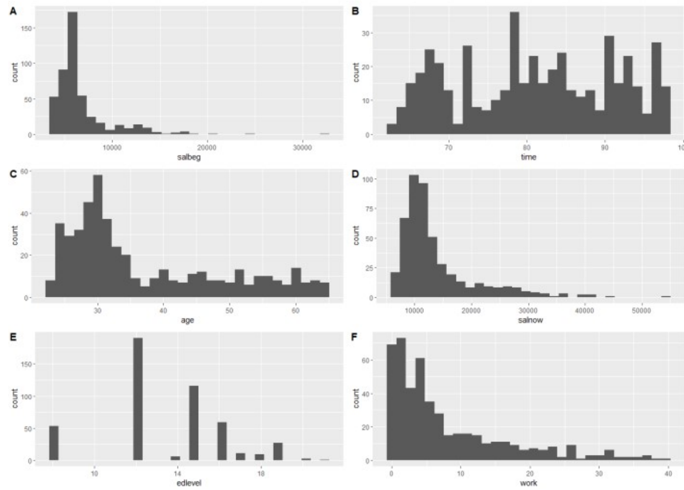
Our null hypothesis (Ho) is that we have normal distribution at each numeric variable of the *sal_num*. To get the normally distribution of each variable, we apply the shapiro.test().

```
> for(i in 1:ncol(sal_num)){
+   print(shapiro.test(sal_num[,i]))
+ }

        Shapiro-Wilk normality test

data:  sal_num[, i]
W = 0.71535, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  sal_num[, i]
W = 0.95425, p-value = 5.954e-11


        Shapiro-Wilk normality test

data:  sal_num[, i]
W = 0.8679, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  sal_num[, i]
W = 0.77061, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  sal_num[, i]
W = 0.90604, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  sal_num[, i]
W = 0.81359, p-value < 2.2e-16
```

We observe that every p-value of each numeric variable is either < 2.2e-16 or 5.954e-11, which is strong evidence against Null hypothesis, that every variable is normally distributed.

We use different visualizations to confirm the results of shapiro.test().

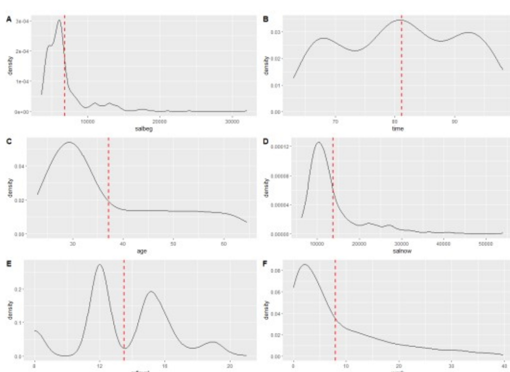
<u>Histograms</u>



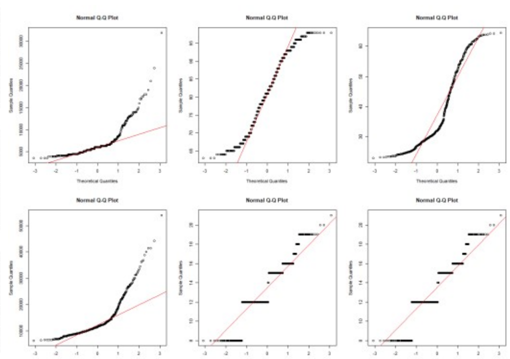
None of the above histograms is bell shaped. <u>Salbeg</u>, <u>salnow</u>, <u>age</u> and <u>work</u> are right-skewed, which indicates that there are several data points, perhaps **outliers,** that are greater than the mode. <u>Time</u> is undefined bimodal, which means that there are intervals equally representing the maximum frequency of the distribution. <u>Edlevel</u> does not contain enough data points to accurately show the distribution of the data.

We use, also Density Plot and QQ plot, which are much more effective ways to view the distribution of a variable. QQ plot can identify from how the values in some section of the plot differ locally from an overall linear trend by seeing whether the values are more or less concentrated than the theoretical distribution would suppose in that section of a plot.

<u>Density Plots</u>



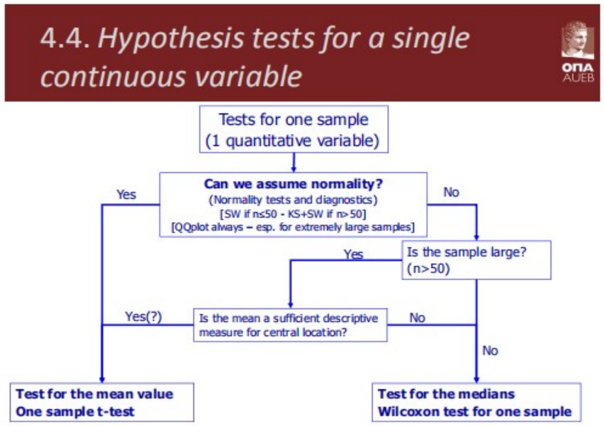
<u>QQ Plots</u>



-

Use the appropriate test to examine whether the beginning salary of a typical employee can be considered to be equal to 1000 dollars. How do you interpret the results? What is the justification for using this particular test instead of some other? Explain.

To examine whether the beginning salary(quantitative) of a typical employee can be considered to be equal to 1000 dollars, we will use the following flowchart of *Hypothesis Testing for one sample*:



The Null hypothesis (Ho) is that the beginning salary is equal to 1000 dollars.

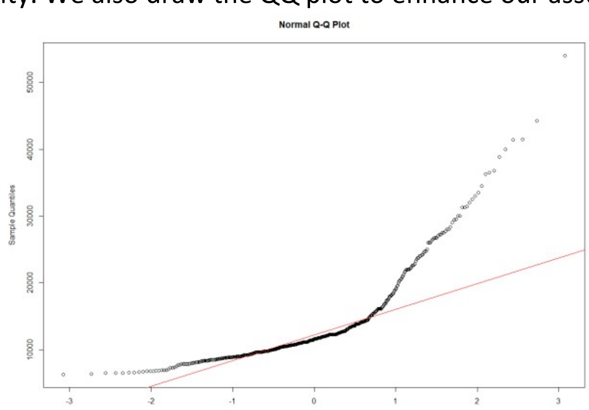
1. We apply lillie.test() and shapiro.test() to test the normality:

```
> #Lillie test
> lillie.test(sal_num$salbeg)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  sal_num$salbeg
D = 0.25188, p-value < 2.2e-16

> #Shapiro test
> shapiro.test(sal_num$salbeg)

        Shapiro-Wilk normality test

data:  sal_num$salbeg
W = 0.71535, p-value < 2.2e-16
```

As the length of our dataset is greater than 50 (n = 474), we should also apply ks.test():

```
> #Kolmogorov test
> ks.test(sal_num$salbeg, 'pnorm')

        One-sample Kolmogorov-Smirnov test

data:  sal_num$salbeg
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

As the p-value of each test is < 2.2e-16, which is strong evidence against Null hypothesis, we cannot assume normality. We also draw the QQ plot to enhance our assumption.



2. We can, also, apply a symmetry.test(). From the result, we assume that there is no symmetry.

```
> symmetry.test(sal_num$salbeg)

        m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  sal_num$salbeg
Test statistic = 10.18, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 72
```

3. As the sample is greater than 50, we have to test the sufficiency of the mean measure.

```
> mean(sal_num$salbeg)
[1] 6806.435
> median(sal_num$salbeg)
[1] 6000
```

   Mean and median are not close enough, so mean is not a sufficient descriptive measure for central location.

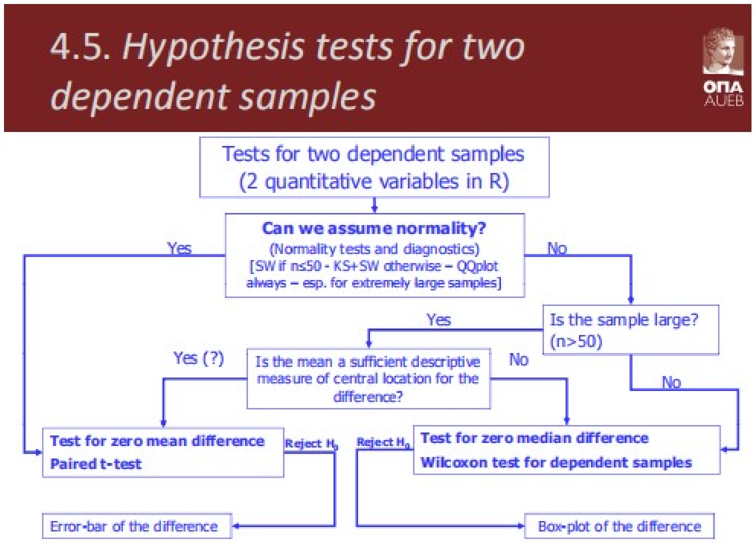
4. We apply wilcox.test() for begging salary equal to 1000 dollars.

```
> wilcox.test(sal_num$salbeg, mu=1000)

        Wilcoxon signed rank test with continuity correction

data:  sal_num$salbeg
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 1000
```

   As the p-value of each test is < 2.2e-16, which is strong evidence against Null hypothesis, we must reject that the begging salary of a typical employee is equal to 1000 dollars.

Consider the difference between the beginning salary (salbeg) and the current salary (salnow). Test if the there is any significant difference between the beginning salary and current salary. (Hint: Construct a new variable for the difference (salnow – salbeg) and test if, on average, it is equal to zero.). Make sure that the choice of the test is well justified.

To examine the difference between the begging salary(quantitative) and the current salary (quantitative), we will use the following flowchart of *Hypothesis Testing for two dependent samples*:



The Null hypothesis (Ho) is that the begging salary is on average equal to the current salary.

All tests will be applied on a new variable diff <- salary$salnow-salary$salbeg

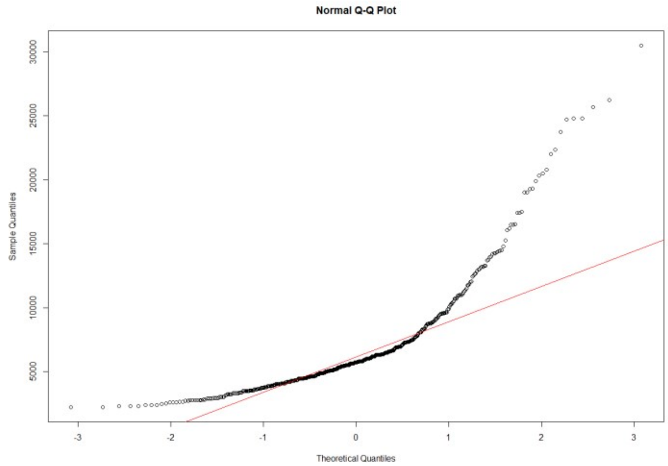
1. We apply lillie.test() and shapiro.test() to test the normality:

```
> #Lillie test
> lillie.test(diff)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  diff
D = 0.186, p-value < 2.2e-16

> #Shapiro test
> shapiro.test(diff)

        Shapiro-Wilk normality test

data:  diff
W = 0.78168, p-value < 2.2e-16
```

As the length of our dataset is greater than 50 (n = 474), we should also apply ks.test():

```
> #Kolmogorov test
> ks.test(diff, 'pnorm')

        One-sample Kolmogorov-Smirnov test

data:  diff
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

As the p-value of each test is < 2.2e-16, which is strong evidence against Null hypothesis, we cannot assume normality. We also draw the QQ plot to enhance our assumption.



2. We can, also, apply a symmetry.test(). From the result, we assume that there is no symmetry.

```
> symmetry.test(diff)

        m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  diff
Test statistic = 10.536, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                296
```

3. As the sample is greater than 50, we have to test the sufficiency of the mean measure.

```
> mean(diff)
[1] 6961.392
> median(diff)
[1] 5700
```

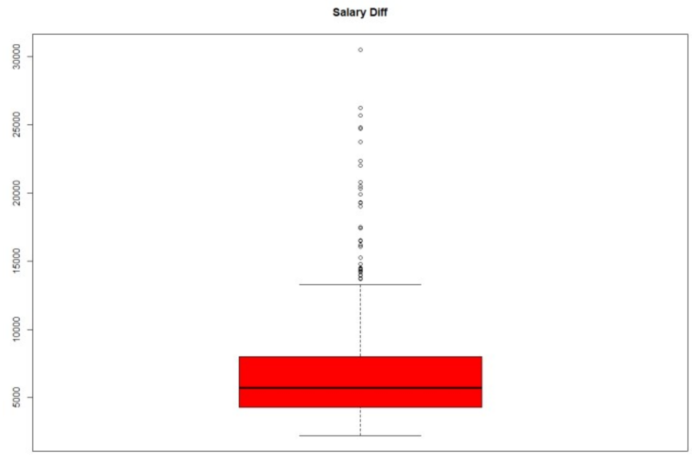
Mean and median are not close enough, so mean is not a sufficient descriptive measure for central location.

4. We apply wilcox.test():

```
> wilcox.test(diff)

        Wilcoxon signed rank test with continuity correction

data:  diff
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
```

As the p-value of each test is < 2.2e-16, which is strong evidence against Null hypothesis, we must reject that the begging salary is on average equal to the current salary.

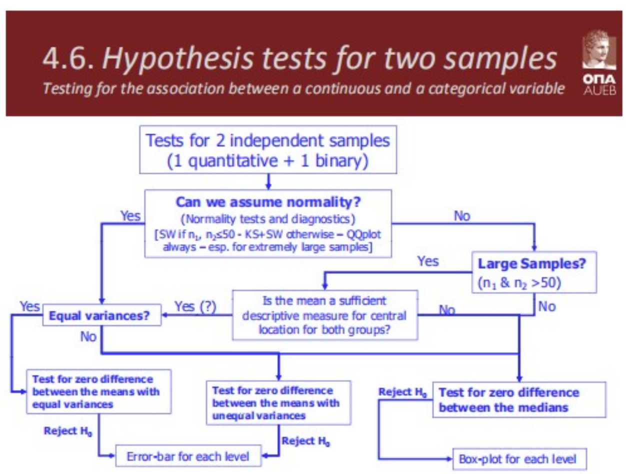
<u>Boxplot</u>



Regarding boxplot, we can conclude that the mean measure is greater than the middle value of the dataset (median). There are, also, several outliers at the maximum, which indicates that the current salary is a lot greater than the beginning salary.

Is there any difference on the beginning salary (salbeg) between the two genders? Give a brief justification of the test used to assess this hypothesis and interpret the results.

To examine the if there is a difference on the beginning salary(quantitative) between the two genders (binary), we will use the following flowchart of Hypothesis Testing for two independent samples:

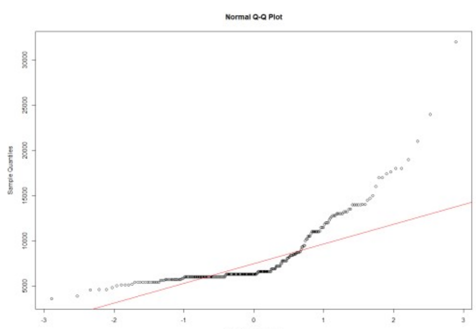


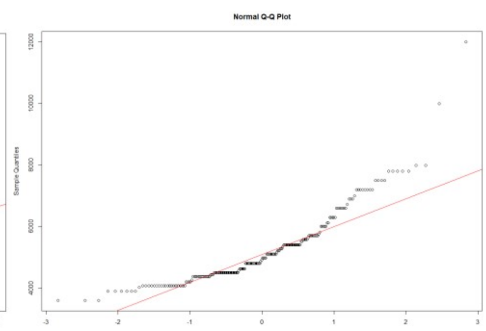
The Null hypothesis (Ho) is that there is zero difference on the beginning salary between the two genders.

1. We apply lillie.test() and shapiro.test() to test the normality:

```
> table(salary$sex)

  MALES FEMALES
    258     216
> by(salary$salbeg, salary$sex, shapiro.test)
salary$sex: MALES

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.73058, p-value < 2.2e-16

---------------------------------------------------------
salary$sex: FEMALES

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.85837, p-value = 2.98e-13

> by(salary$salbeg, salary$sex, lillie.test)
salary$sex: MALES

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  dd[x, ]
D = 0.25863, p-value < 2.2e-16

---------------------------------------------------------
salary$sex: FEMALES

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  dd[x, ]
D = 0.14843, p-value = 1.526e-12
```

As the p-value of each test is < 2.2e-16, which is strong evidence against Null hypothesis, we cannot assume normality. We also draw the QQ plot of males and females respectively to enhance our assumption.





2. We can, also, apply a symmetry.test(). From the result, we assume that there is no symmetry.

```
> symmetry.test(salary$salbeg)

        m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  salary$salbeg
Test statistic = 10.18, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 57
```

3. We compare the mean and median of males and females respectively. Are not close.

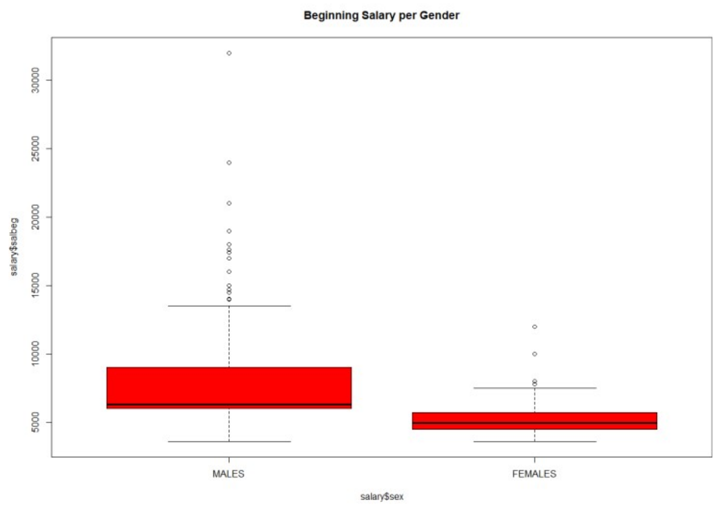
```
> mean(males$salbeg)              > mean(females$salbeg)
[1] 8120.558                      [1] 5236.787
> median(males$salbeg)            > median(females$salbeg)
[1] 6300                          [1] 4950
```

4. We apply wilcox.test(), to test for zero difference between the medians:

```
> wilcox.test(salary$salbeg ~ salary$sex)

	Wilcoxon rank sum test with continuity correction

data:  salary$salbeg by salary$sex
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

As the p-value is < 2.2e-16, which is strong evidence against Null hypothesis, we must reject that there is zero difference about the medians of the beginning salaries between the two genders.

Boxplot



Regarding boxplot, we can conclude that there is a significant difference on the medians between the two genders. The median of the begging salary of males is greater than the median of the begging salary of females. There are, also, more outliers at the maximum of males, which confirms our conclusion.

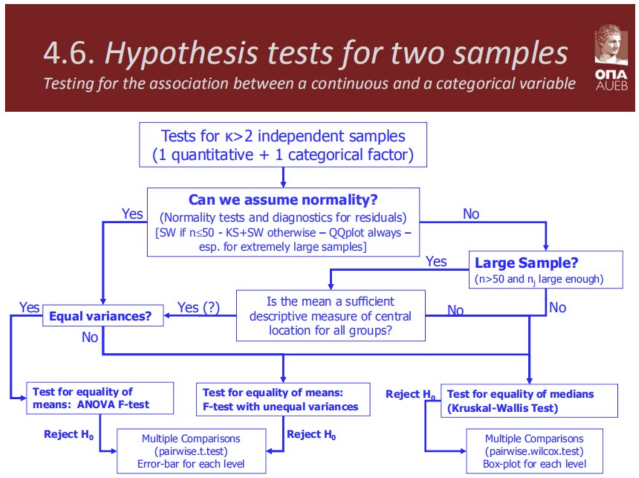
Cut the AGE variable into three categories so that the observations are evenly distributed across categories (Hint: you may find the cut2 function in Hmisc package to be very useful). Assign the cut version of AGE into a new variable called age_cut. Investigate if, on average, the beginning salary (salbeg) is the same for all age groups. If there are significant differences, identify the groups that differ by making pairwise comparisons. Interpret your findings and justify the choice of the test that you used by paying particular attention on the assumptions.

We will construct a new variable age_cut <- cut2(salary$age, g=3), which contains all age observations distributed across three categories. As the age variable is float, we could not distribute it across evenly:

```
Group 1 Group 2 Group 3
    160     156     158
```

All test will be applied on the anova variable, which is the result of aov(sal_num$salbeg~age_cut).

To examine if the beginning salary(quantitative) is the same for all age groups(categorical), we will use the following flowchart of *Hypothesis Testing for two independent samples*:



The Null hypothesis (Ho) is that the beginning salary is the same for all age groups.

1. We apply lillie.test() and shapiro.test() to test the normality:

```
> #Lillie test
> lillie.test(anova$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  anova$residuals
D = 0.21891, p-value < 2.2e-16

> #Shapiro test
> shapiro.test(anova$residuals)

        Shapiro-Wilk normality test

data:  anova$residuals
W = 0.71244, p-value < 2.2e-16
```

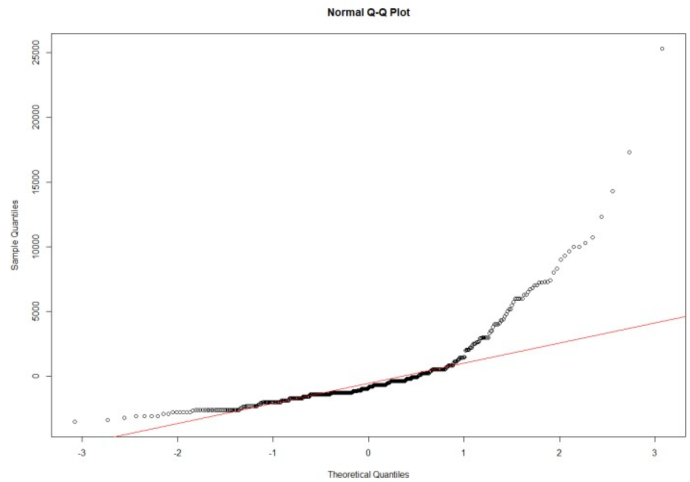
As the size of each group>50,we should also apply ks.test:

```
> ks.test(anova$residuals, 'pnorm')

        One-sample Kolmogorov-Smirnov test

data:  anova$residuals
D = 0.69831, p-value < 2.2e-16
alternative hypothesis: two-sided
```

2. We can, also, apply a symmetry.test(). From the result, we assume that there is no symmetry.

```
> symmetry.test(anova$residuals)

        m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  anova$residuals
Test statistic = 11.477, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 57
```

As the p-value of each test is < 2.2e-16, which is strong evidence against Null hypothesis, we cannot assume normality. We also draw the QQ plot to enhance our assumption.



3. We apply kruskal.test() to test the equality of the medians of each group.

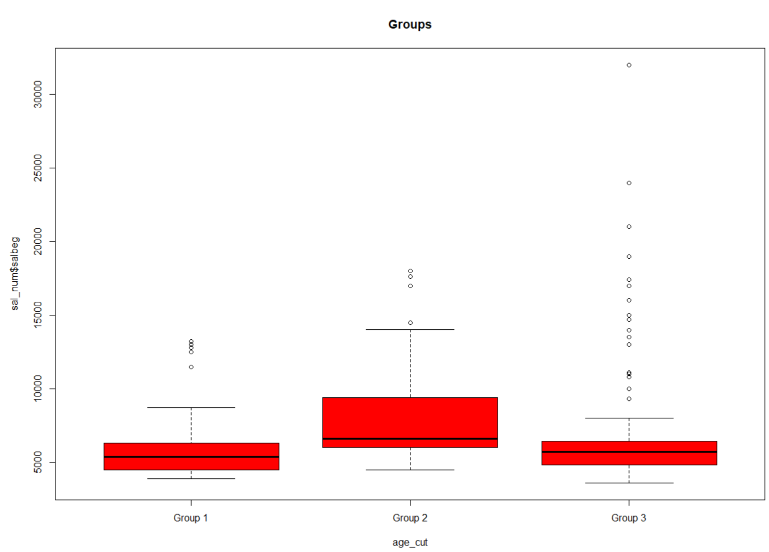
```
> kruskal.test(sal_num$salbeg~age_cut)

        Kruskal-Wallis rank sum test

data:  sal_num$salbeg by age_cut
Kruskal-Wallis chi-squared = 92.742, df = 2, p-value < 2.2e-16
```

As the p-value is < 2.2e-16, which is strong evidence against Null hypothesis, we reject the equality of the medians of groups.

4. We apply pairwise.wilcox.test() to identify the groups that differ:

```
> pairwise.wilcox.test(sal_num$salbeg, age_cut)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  sal_num$salbeg and age_cut

        Group 1 Group 2
Group 2 < 2e-16 -
Group 3 0.089   8.9e-12

P value adjustment method: holm
```

<u>Boxplot</u>



Regarding boxplot, we confirm our conclusion that the medians of groups are not equally. More specific, the median of Group 2 differs from the medians of Group 1 & 3.

By making use of the factor variable minority, investigate if the proportion of white male employees is equal to the proportion of white female employees.

The tests will be applied on a new variable tab <- table(salary$sex, salary$minority).

The Null hypothesis (Ho) is that the proportion of white male employees is equal to the proportion of white female employees.

```
> prop.test(tab)

        2-sample test for equality of proportions with continuity correction

data:  tab
X-squared = 2.3592, df = 1, p-value = 0.1245
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.14102693  0.01527327
sample estimates:
   prop 1    prop 2
0.7519380 0.8148148
```

```
> chisq.test(tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 2.3592, df = 1, p-value = 0.1245
```

```
> fisher.test(tab)

        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.1186
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.429148 1.098149
sample estimates:
odds ratio
 0.6894628
```

In all the above tests the p-values is > 0.05. As a result, we do not reject the Null hypothesis, that the proportion of white male employees is equal to the proportion of white female employees.

Code in R:

To accomplish the assignment, it is necessary to install the following packages and import following libraries:

- foreign, psych, nortest, normtest, moments, lawstat, Hmisc, ggplot2, cowplot, plyr

```
#Question 1

#import data require(foreign)
setwd('C:\\Users\\tzina\\OneDrive\\Documents\\LabAssignment')

#read data frame
salary <- read.spss("salary.sav", to.data.frame = T)

#View data frame
View(salary)

#structure of the data frame
str(salary)

#Question 2

#Keep only the numeric variables of the data frame
index <- sapply(salary, class) == "numeric"
sal_num <- salary[index]

#Delete id column (1st column)
sal_num <- sal_num[,-1]

#Get summary statistics of numerical variables (sal_num)
describe(sal_num)
round(t(describe(sal_num)),2)

#Testing for Normality
for(i in 1:ncol(sal_num)){
  print(shapiro.test(sal_num[,i]))
}
sapply(sal_num, shapiro.test)

#Histograms

salbeg    <- ggplot(sal_num,aes(salbeg))  + geom_histogram()
time      <- ggplot(sal_num,aes(time))    + geom_histogram()
age       <- ggplot(sal_num,aes(age))     + geom_histogram()
salnow    <- ggplot(sal_num,aes(salnow))  + geom_histogram()
edlevel   <- ggplot(sal_num,aes(edlevel)) + geom_histogram()
 work     <- ggplot(sal_num,aes(work))    + geom_histogram()

plot_grid(salbeg, time, age, salnow, edlevel, work,     nrow = 3,     ncol = 2,
     labels = "AUTO")
```

```
#Density plots
den_salbegin <- ggplot(sal_num, aes(x=salbeg))  + geom_density() +
geom_vline(aes(xintercept=mean(salbeg)), color="red", linetype="dashed", size=1)
den_time      <- ggplot(sal_num, aes(x=time))    + geom_density() +
geom_vline(aes(xintercept=mean(time)), color="red", linetype="dashed", size=1)
den_age       <- ggplot(sal_num, aes(x=age))     + geom_density() +
geom_vline(aes(xintercept=mean(age)), color="red", linetype="dashed", size=1)
den_salnow  <- ggplot(sal_num, aes(x=salnow))  + geom_density() +
geom_vline(aes(xintercept=mean(salnow)), color="red", linetype="dashed", size=1)
den_edlevel  <- ggplot(sal_num, aes(x=edlevel)) + geom_density() +
geom_vline(aes(xintercept=mean(edlevel)), color="red", linetype="dashed", size=1)
den_work      <- ggplot(sal_num, aes(x=work))    + geom_density() +
geom_vline(aes(xintercept=mean(work)), color="red", linetype="dashed", size=1)

plot_grid(den_salbegin, den_time, den_age, den_salnow, den_edlevel, den_work,         nrow = 3,
ncol = 2,
        labels = "AUTO")

#QQ Plots library(psych) par(mfrow= c(2,3))
qq_salbeg    <- qqnorm(sal_num$salbeg) + qqline(sal_num$salbeg, col = 'red')
qq_time      <- qqnorm(sal_num$time) + qqline(sal_num$time, col = 'red')
qq_age       <- qqnorm(sal_num$age) + qqline(sal_num$age, col = 'red')
qq_salnow   <- qqnorm(sal_num$salnow) + qqline(sal_num$salnow, col = 'red')
qq_edlevel   <- qqnorm(sal_num$edlevel) + qqline(sal_num$edlevel, col = 'red')
qq_work      <- qqnorm(sal_num$work) + qqline(sal_num$work, col = 'red')

#Question 3

####### Hypothesis Testing for one sample #########

#Lillie test Normality rejected
lillie.test(sal_num$salbeg)

#Shapiro test Normality rejected
shapiro.test(sal_num$salbeg)

#Kolmogorov test Normality rejected
ks.test(sal_num$salbeg, 'pnorm')

#Draw QQ plot of sal_num
qqnorm(sal_num$salnow) + qqline(sal_num$salnow, col = 'red')

#length of salbeg > 50
```

```
length(sal_num$salbeg)

#Symmetry.test
symmetry.test(sal_num$salbeg)

#Check skewness and kurtosis
skewness.norm.test(sal_num$salbeg)
kurtosis.norm.test(sal_num$salbeg)

#Mean and Median are not close
mean(sal_num$salbeg)
median(sal_num$salbeg)

#Apply Wilkcoxon test
wilcox.test(sal_num$salbeg, mu=1000)

#Question 4

####### Hypothesis Testing for two dependent samples #########

diff <- salary$salnow- salary$salbeg

#Lillie test Normality
lillie.test(diff)

#Shapiro test Normality
shapiro.test(diff)

#Length diff >50  length(diff)

#Kolmogorov test
ks.test(diff, 'pnorm')

#Draw QQ plot of diff
qqnorm(diff) + qqline(diff, col = 'red')

#Find length
length(diff)

#As length greater than > 50
mean(diff)
median(diff)

#Test symmetry
```

```
symmetry.test(diff)

#Check skewness and kurtosis
skewness.norm.test(diff)
kurtosis.norm.test(diff)

#Wilcoxon test
wilcox.test(diff)

#Draw the boxplot
boxplot(diff, main = "Salary Diff", col = "red")

#Question 5

####### Hypothesis Testing for two independent samples #########
#1 quantitative & 1 binary

#Length of samples
table(salary$sex)

by(salary$salbeg, salary$sex, shapiro.test)
by(salary$salbeg, salary$sex, lillie.test)
#Create subset of salaries for males and draw QQ plot
males <- subset(salary, salary$sex == 'MALES')
qqnorm(males$salbeg) + qqline(males$salbeg, col = 'red')

#Create subset of salaries for females and draw QQ plot
females <- subset(salary, salary$sex == 'FEMALES')
qqnorm(females$salbeg) + qqline(females$salbeg, col = 'red')

#Test symmetry
symmetry.test(salary$salbeg)

#Compare mean and median for males
mean(males$salbeg)
median(males$salbeg)

#Compare mean and median for females
mean(females$salbeg)
median(females$salbeg)

#test for zero difference between the medians
wilcox.test(salary$salbeg ~ salary$sex)
```

```
#Draw the boxplot
boxplot(salary$salbeg ~ salary$sex, main = "Beginning Salary per Gender", col = "red")

#Question 6
####### Hypothesis Testing for two independent samples #########
#1 quantitative & 1 categorical

#Distribute age in three categories
age_cut <- cut2(sal_num$age, g=3)
age_cut <- factor(age_cut, labels = paste('Group '))

#Show observations of each group table(age_cut)
table(age_cut)

#Anova
anova <- aov(sal_num$salbeg~age_cut)
summary(anova)

#Lillie test
lillie.test(anova$residuals)

#Shapiro test
shapiro.test(anova$residuals)

#Kolmogorov test
ks.test(anova$residuals, 'pnorm')

#Draw QQ plot
qqnorm(anova$residuals) + qqline(anova$residuals, col = 'red')

#Test symmetry
symmetry.test(anova$residuals)

#Kruskal test
kruskal.test(sal_num$salbeg~age_cut)

#multiple comparisons
pairwise.wilcox.test(sal_num$salbeg, age_cut)

#Draw boxplot
boxplot(sal_num$salbeg~age_cut, main = "Groups", col = "red")

#Question 7
```

```
tab <- table(salary$sex, salary$minority)

#Proportion test prop.test(tab)

#Chi squared test
chisq.test(tab)

#fisher test
fisher.test(tab)
```