ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

BUSINESS
ANALYTICS
Master of Science

Statistics for Business Analytics I

Lab Assignment #1

Due Date: December 2nd, 2020

Anastasios Theodorou

(p2822007)

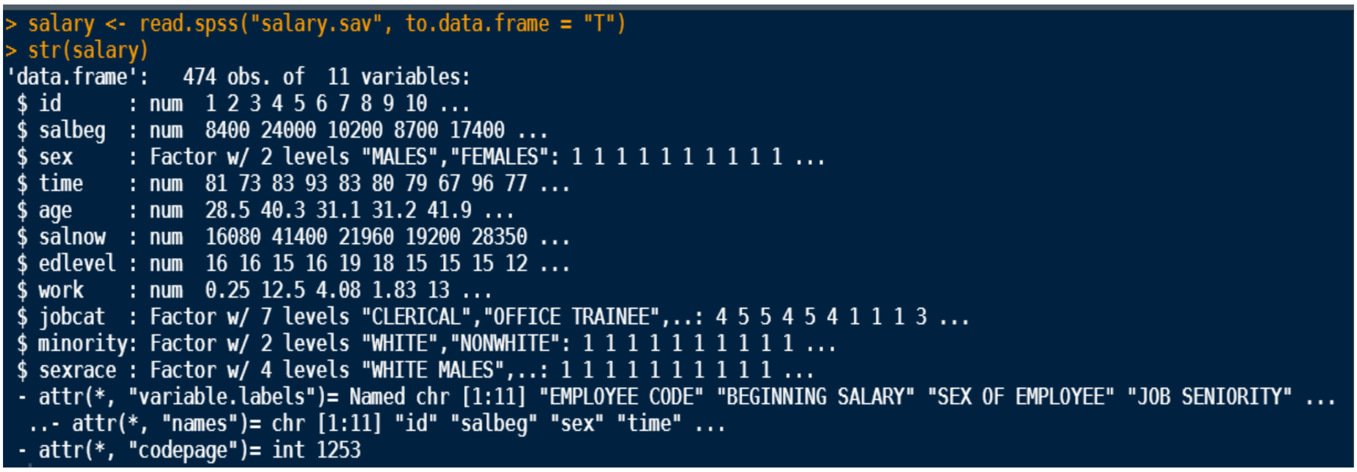
Contents

1st Question

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

1.1 Output

The output of the first question's code in R language is:



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

1.2 Comment

This data frame, which is called "*salary*" contains 474 observations (employees) and 11 variables. These 474 observations form the data frame's rows, while the 11 variables are the columns of it.

By running the code *"str(salary)"* we can conclude that there are 7 numerical vectors ("id", "salbeg", "time", "age", "salnow", "edlevel", "work") and 4 factors ("sex", "jobcat", "minority", "sexrace"). To be more specific, the factor "*sex*" and *"minority"* are consisted of 2 levels ("MALES", "FEMALES" & "WHITE", "NONWHITE" accordingly), the factor *"sexrace"* is consisted of 4 levels ("WHITE MALES" etc) and the *"jobcat"* of 7 levels ("CLERICAL","OFFICE TRAINEE", etc).

Finally, in the last line of this code where is defined as attr we can see the names of the variable labels ("EMPLOYEE CODE", "BEGINNING SALARY" etc), in other words how each variable is called, the name of each column ("id", "salbeg" etc) and the code of this page which is 1253 (Greek ANSI - for writing modern Greek).

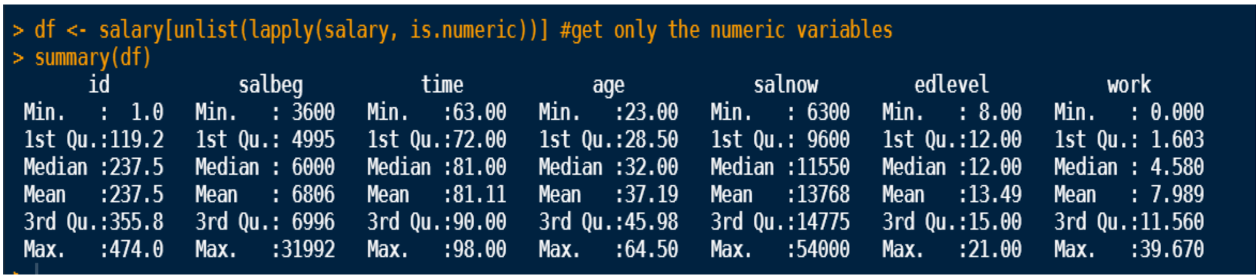
2nd Question

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?

2.1 Output

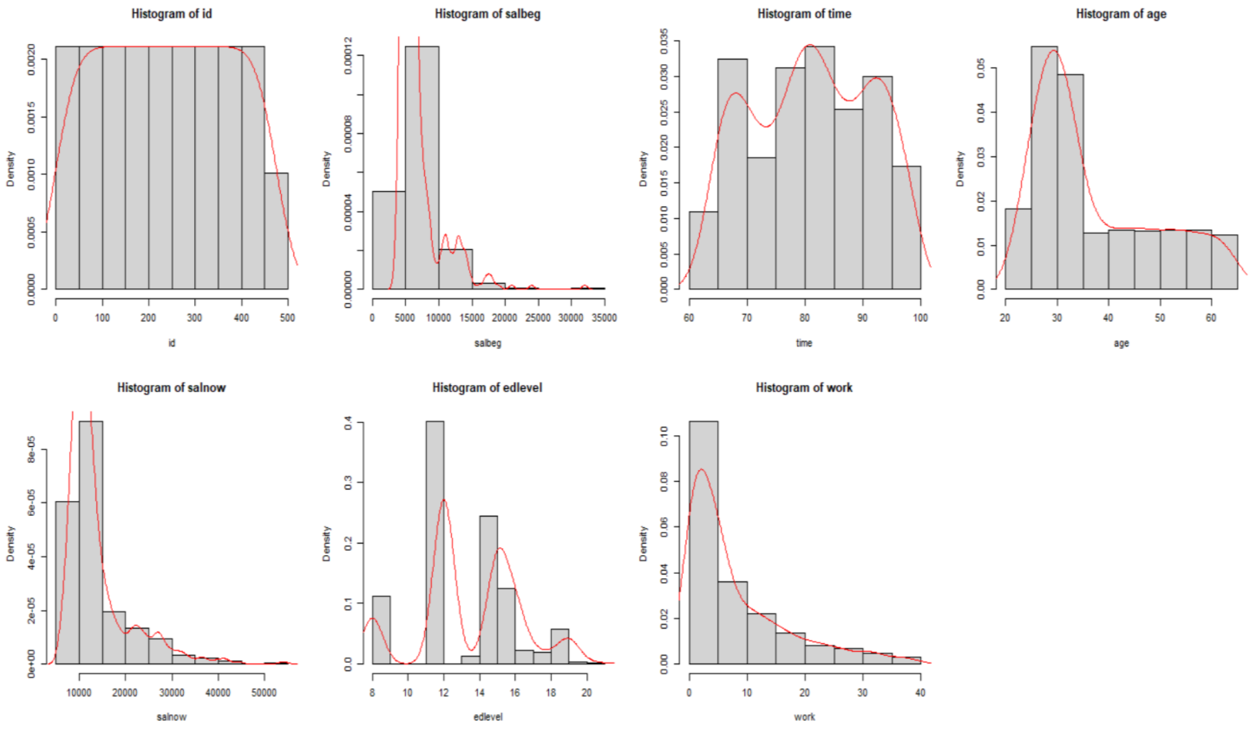
The output of the second question's code in R language is:

- Summary Statistics:

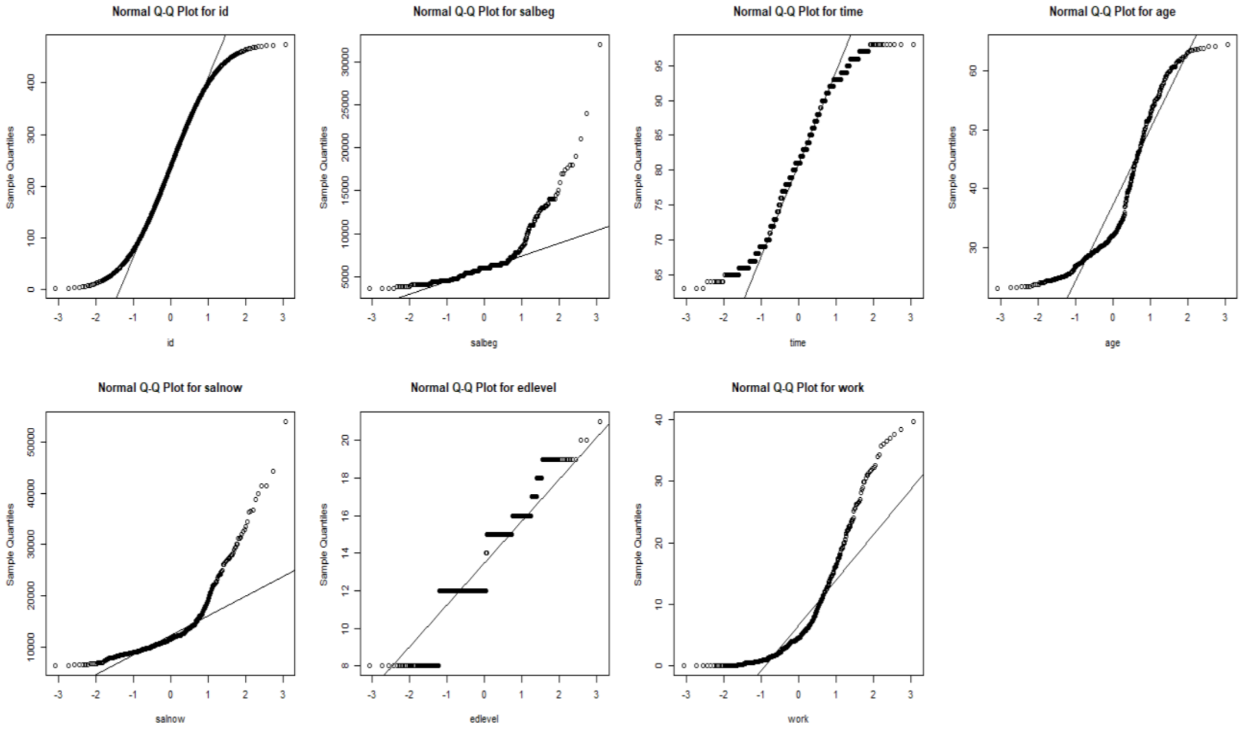


```
> df <- salary[unlist(lapply(salary, is.numeric))] #get only the numeric variables
> summary(df)
       id            salbeg           time            age            salnow          edlevel           work       
 Min.   :  1.0   Min.   : 3600   Min.   :63.00   Min.   :23.00   Min.   : 6300   Min.   : 8.00   Min.   : 0.000  
 1st Qu.:119.2   1st Qu.: 4995   1st Qu.:72.00   1st Qu.:28.50   1st Qu.: 9600   1st Qu.:12.00   1st Qu.: 1.603  
 Median :237.5   Median : 6000   Median :81.00   Median :32.00   Median :11550   Median :12.00   Median : 4.580  
 Mean   :237.5   Mean   : 6806   Mean   :81.11   Mean   :37.19   Mean   :13768   Mean   :13.49   Mean   : 7.989  
 3rd Qu.:355.8   3rd Qu.: 6996   3rd Qu.:90.00   3rd Qu.:45.98   3rd Qu.:14775   3rd Qu.:15.00   3rd Qu.:11.560  
 Max.   :474.0   Max.   :31992   Max.   :98.00   Max.   :64.50   Max.   :54000   Max.   :21.00   Max.   :39.670  
```

- Visualization of distribution:
  1. Histograms:



2. QQPlots:



2.2 Comment

From the above visualization we can conclude the following:

- **id:** it is just an iterator, although on QQ Plot we can conclude that the majority of id values are upon the QQ Line, no normality can be derived from both the plots.

- **salbeg:** mode < median < mean, no normality, despite the fact that it has positive (or right) skew (the 2nd bar is the highest and the rest of them from the 3rd and then are nearly close to 0). In QQ Plot the data appear to follow a curved line as a usual right-skewed distribution does.

- **time:** from the histogram this variable seems to be under-dispersed, relative to a normal distribution, because the bars follow a wavier look. In QQ Plot there is also some kind of normality, although it is a discreet variable.

- **age:** from the two plots we cannot assume any normality because in the histogram the density line doesn’t look like this from the Normal distribution and in QQ Plot the data do not seem to be close to the QQ line as they would be if they were normally distributed.

- **salnow:** mode < median < mean, like the “salbeg” variable no normality can be assumed and it appears to be positive skewed (in the histogram the 2nd bar is the highest and the rest of them from the 3rd and so on are nearly close to 0). Also, the two QQ Plots are very similar.

- **edlevel:** no normality, the bars of this histogram and the data on QQ Plot have not any association between them and in the QQ Plot the dots do not follow the QQ Line.

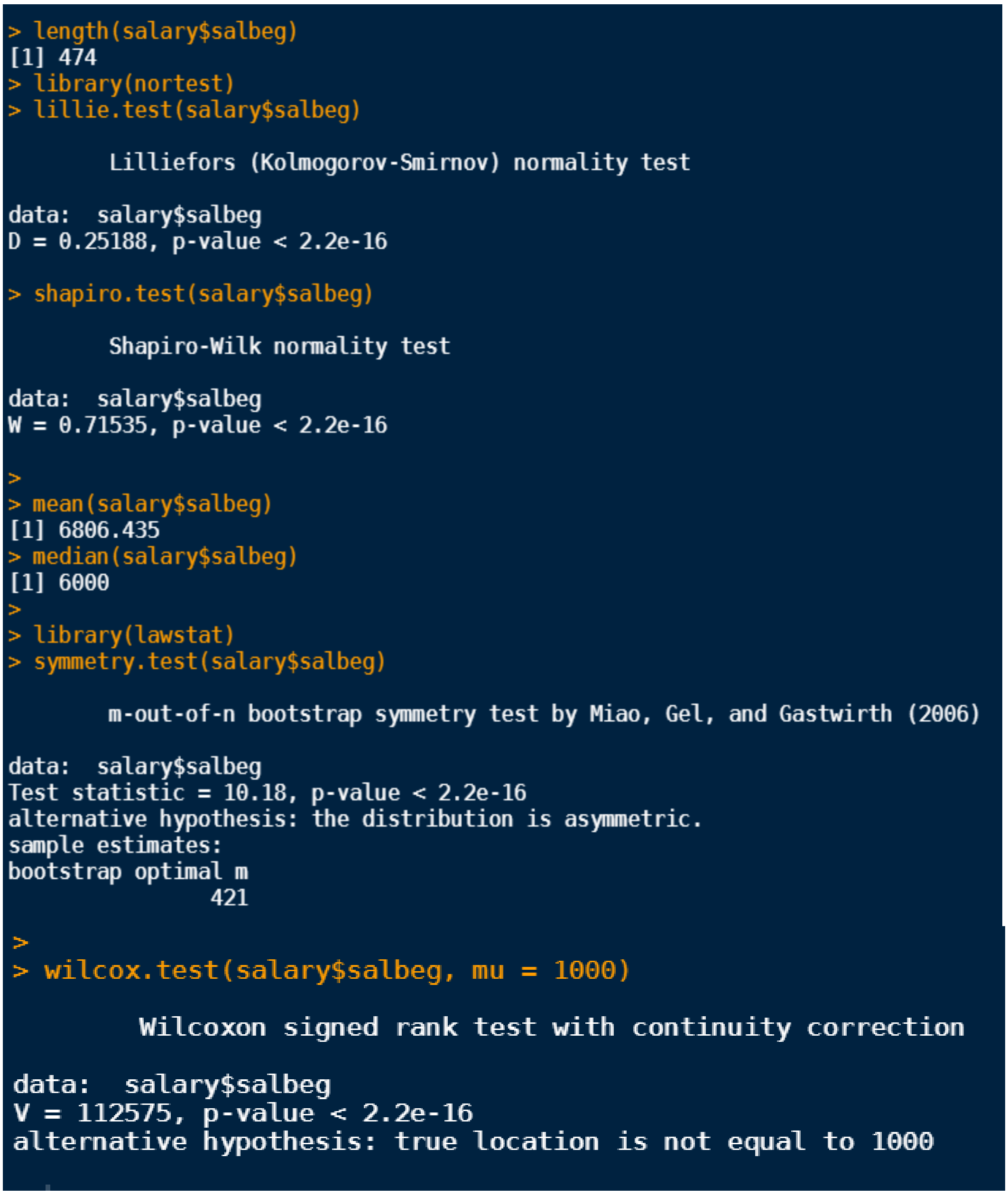
- **work:** mode < median < mean, although in histogram it appears to be not normally distributed, but with positive skewness as the other two variables (“salbeg” and “salnow”) in the QQ Plot the data aren’t near the QQ Line, so it doesn’t seem to be Normal.

3rd Question

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.

3.1 Output

The output of this code in R language is:



```
> length(salary$salbeg)
[1] 474
> library(nortest)
> lillie.test(salary$salbeg)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

> shapiro.test(salary$salbeg)

        Shapiro-Wilk normality test

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

>
> mean(salary$salbeg)
[1] 6806.435
> median(salary$salbeg)
[1] 6000
>
> library(lawstat)
> 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
                421

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

        Wilcoxon signed rank test with continuity correction

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

3.2 Comment

To begin with in order to check if the beginning salary of a typical employee can be considered to be equal to 1000 dollars we should take into consideration the length of the variable, which is in our case pretty large and equal to 474.

Then, we have to check the normality of the beginning salary and we do so with two tests: the Kolmogorov-Smirnov and the Shapiro-Wilk test of normality. Because of the fact that in both two tests the p-value is very small (smaller than the significance level – α = 0,05) we reject the null Hypothesis (Ho) that this variable is normally distributed.

Next, we must check the symmetry of this variable and specifically if the mean is a descriptive measure for central location. For that reason, we check if mean and median is near, which in our case, the first is 6806,435 and the second is 6000. That means that they are too far away one to the other. That's why we call the symmetry test from "lawstat" library. By executing this test we can easily conclude that p-value is again very small, so we reject the null Hypothesis (Ho) that this variable is symmetric.

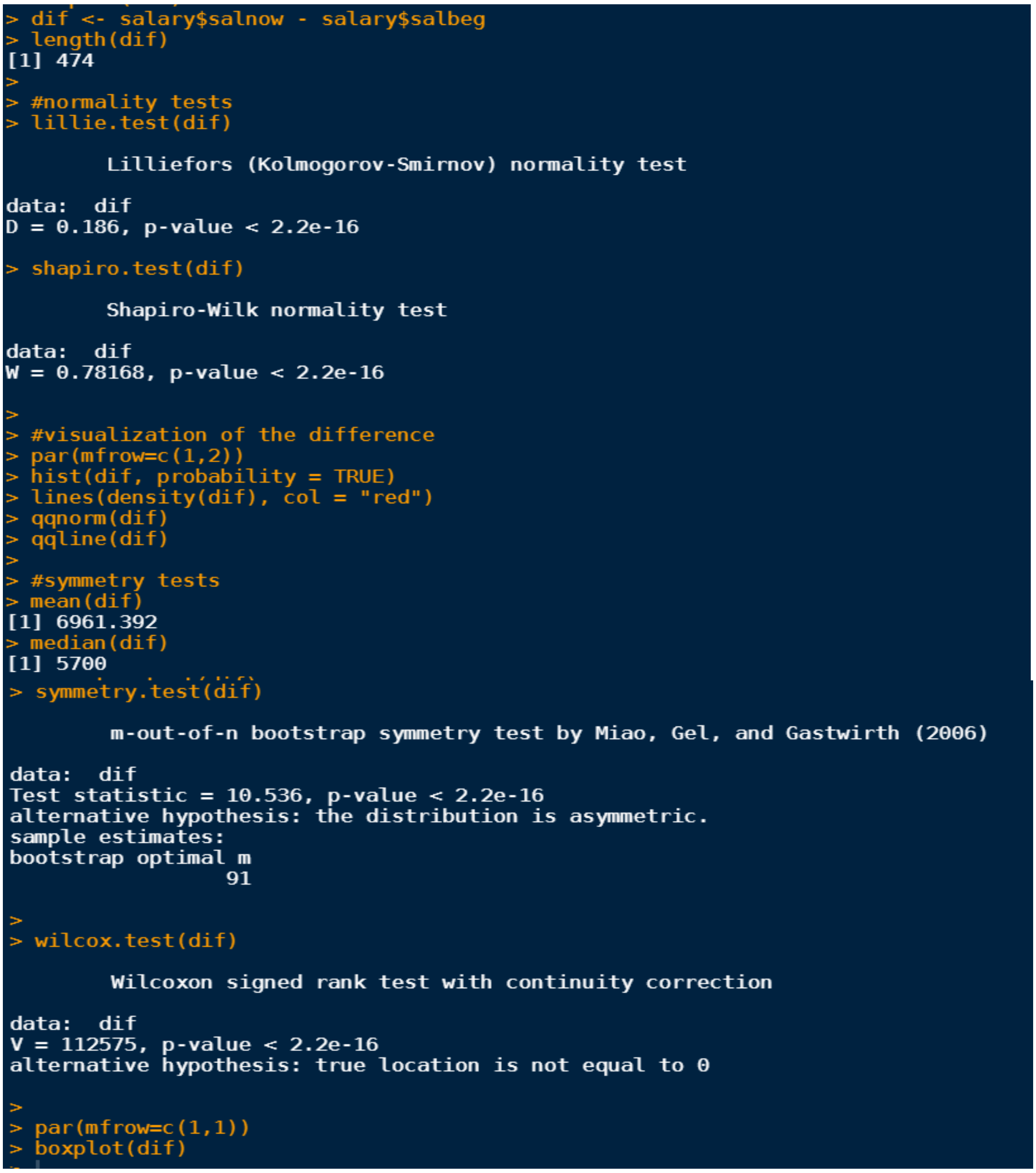
Finally, one last step is to use the non-parametric Wilcoxon test with μ_o = 1000 in order to see if the beginning salary of a typical employee can be considered to be equal to 1000 dollars. From the above output of the R-code we can see one more time that p-value of this test is again very small (smaller than the default critical p-value 0,05) and as a result would be to reject the null Hypothesis (Ho) that $\mu=\mu_o$ (μ is the beginning salary of a typical employee). In this part we should clarify that this test is more appropriate to this case, because it is a non-normal and asymmetric variable. In other case we would use the One sample t-test which is better than this one, because it takes into consideration the confidence intervals.

4th Question

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.

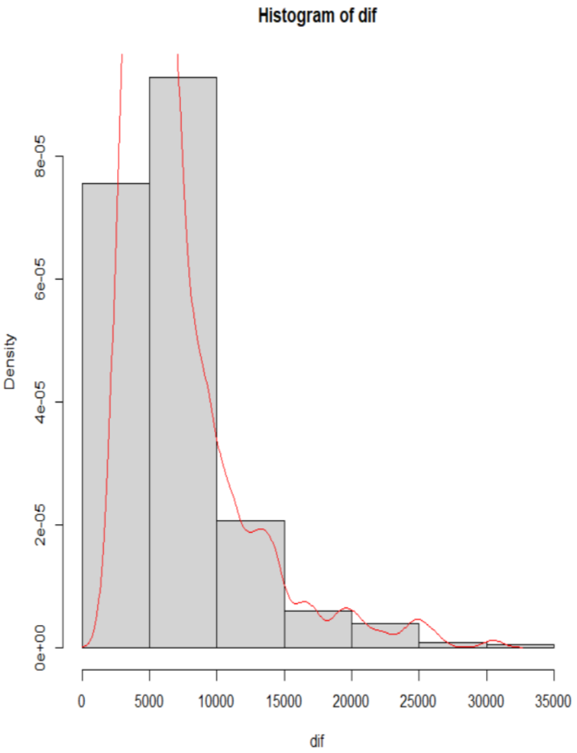
4.1 Output

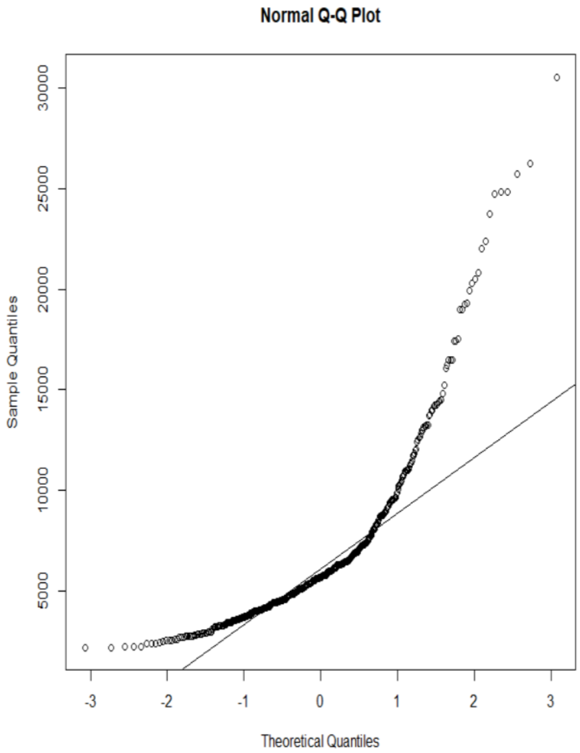
The output of this code in R language is:

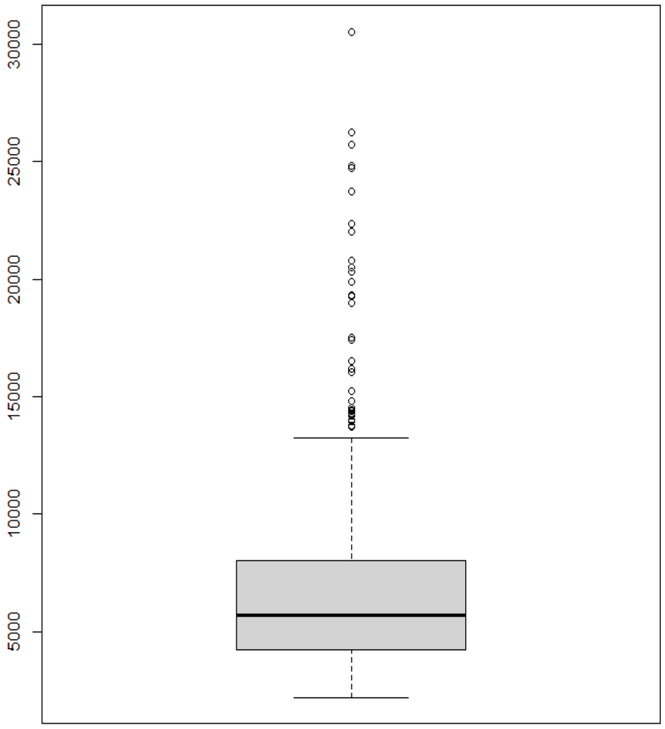


```
> dif <- salary$salnow - salary$salbeg
> length(dif)
[1] 474
>
> #normality tests
> lillie.test(dif)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

> shapiro.test(dif)

        Shapiro-Wilk normality test

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

>
> #visualization of the difference
> par(mfrow=c(1,2))
> hist(dif, probability = TRUE)
> lines(density(dif), col = "red")
> qqnorm(dif)
> qqline(dif)
>
> #symmetry tests
> mean(dif)
[1] 6961.392
> median(dif)
[1] 5700
> symmetry.test(dif)

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

data:  dif
Test statistic = 10.536, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 91

>
> wilcox.test(dif)

        Wilcoxon signed rank test with continuity correction

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

>
> par(mfrow=c(1,1))
> boxplot(dif)
```

The visualization of the above R-code is the following:







4.2 Comment

In order to check if there is any significance difference between two samples (current salary and the beginning salary) that in our case are dependent, we should follow the below steps (in every step we do the computations with the difference of those two variables):

1. **Check Normality:** for large sample, like ours (474 observations) we use first the Kolmogorov-Smirnov test and then the Shapiro-Wilk test of normality for the difference. Because in both two tests the p-value is very small (smaller than the significance level – α = 5%) we reject the null Hypothesis (Ho) that this variable is normally distributed. Also, the histogram and the QQ Plot of the difference are very interesting in checking them, in order to confirm the results of the above two tests.

2. **Check Symmetry:** first, we check if mean and median is near, which in our case, the first is 6961,392 and the second is 5700. That mean that they are far away one to the other. That's why we call the symmetry test from "lawstat" library. By executing this test we can easily conclude that p-value is again very small, so we reject the null Hypothesis (Ho) that this variable is symmetric.

3. **Test for zero median difference:** because of the above two assumptions, we use the non-parametric Wilcoxon test with $\mu_o = 0$ in order to see if the difference between these two variables is equal to zero (in other words we check if $\mu = 0$ → salnow = salbeg). From the R-code we can see one more time that p-value of this test is again very small (smaller than the default critical p-value 0,05) and as a result would be to reject the null Hypothesis (Ho) that $\mu=\mu_o=0$ (μ is the difference). So, there is a significant difference between the current salary and the beginning salary.

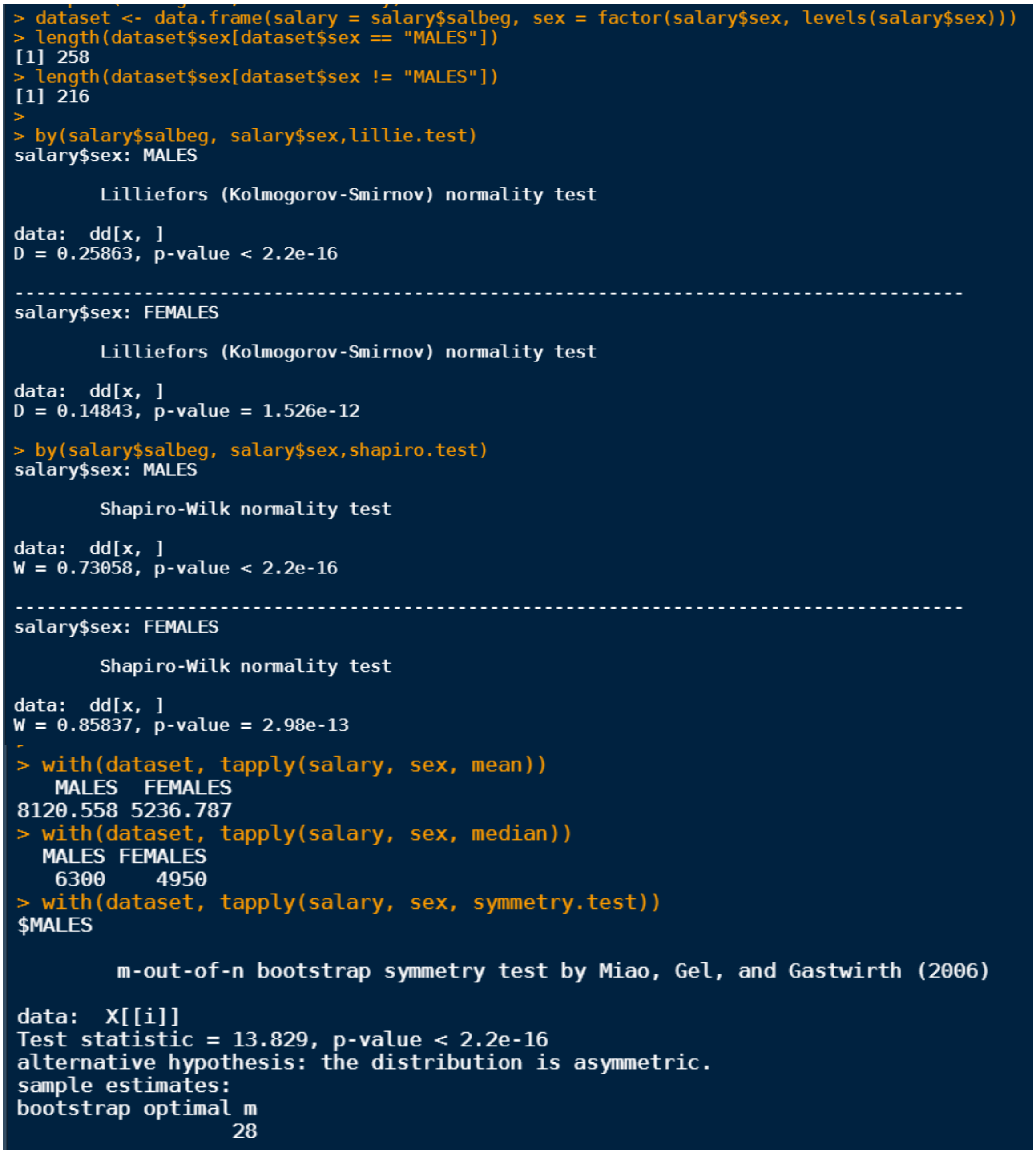
4. **Box-Plot of the difference:** we draw a boxplot to confirm the above results. We see that in the difference of these variables there are a lot of outliers (the dots outside the box). Also, we can see the median (above 5000), the 3rd quartile (the upper border of the box) and the 1st quantile (the lower border of the box). So, because the majority of the data (75%) are above the median that means that the difference could not be zero or near to it.

5th Question

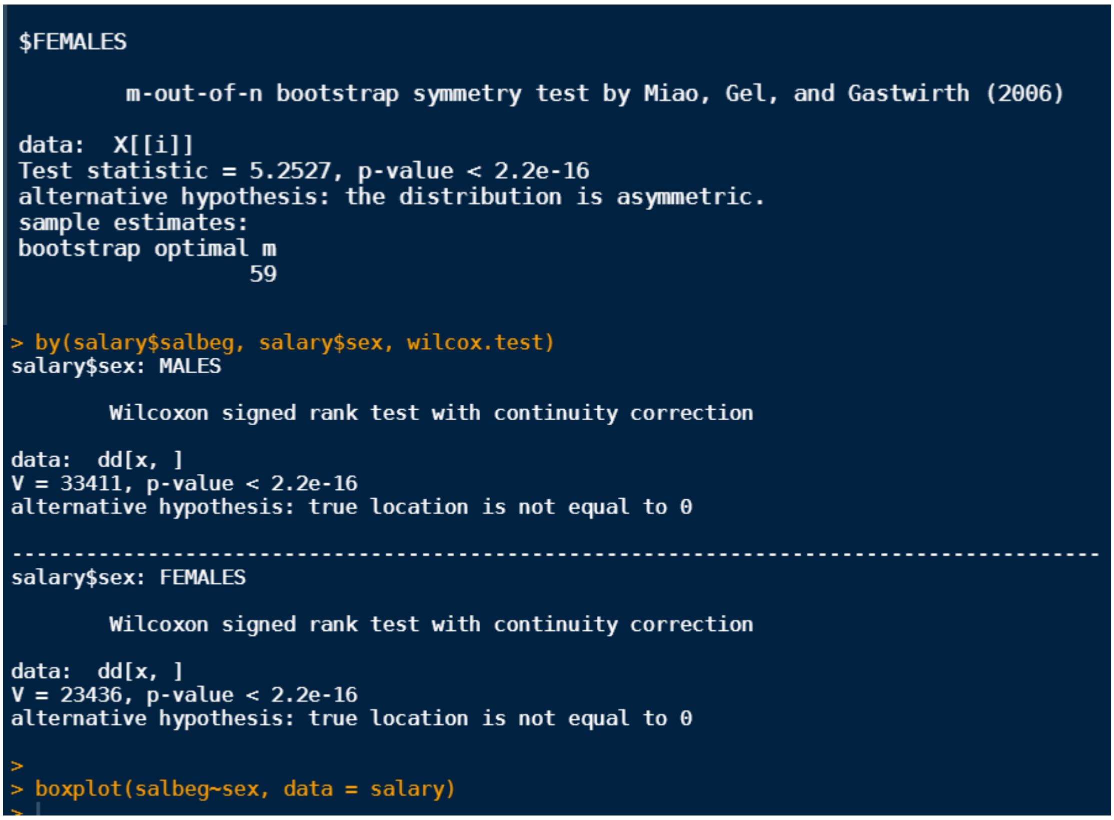
Is there any difference on the beginning salary (**salbeg**) between the two genders? Give a brief justification of the test used to assess this hypothesis and interpret the results.

5.1 Output

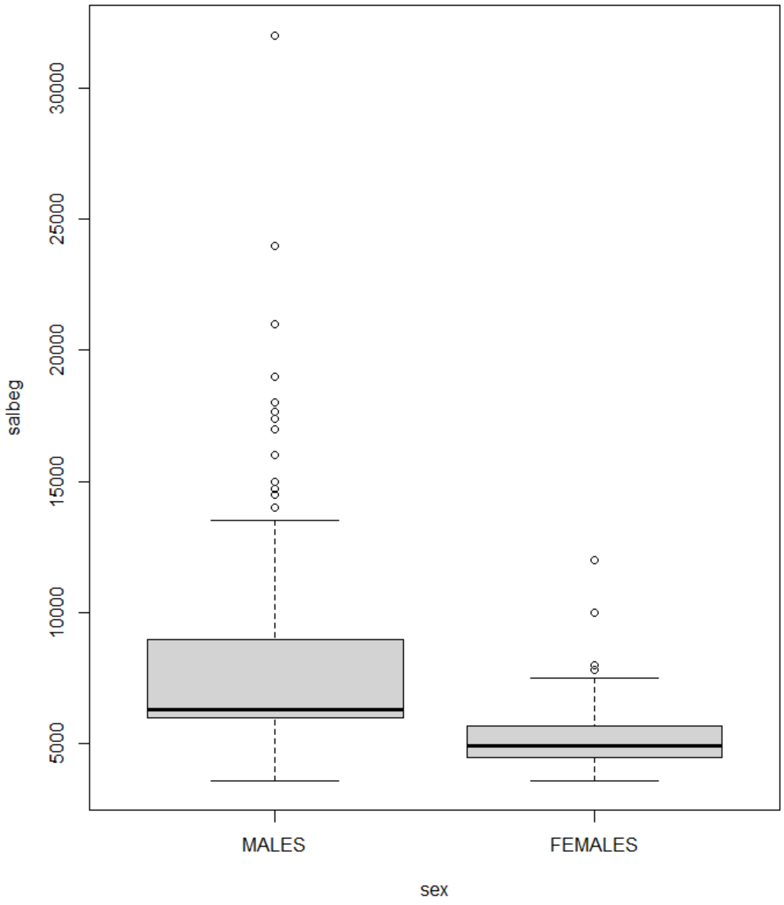
The output of this code in R language and of the plots that are generated from it is:



```
> dataset <- data.frame(salary = salary$salbeg, sex = factor(salary$sex, levels(salary$sex)))
> length(dataset$sex[dataset$sex == "MALES"])
[1] 258
> length(dataset$sex[dataset$sex != "MALES"])
[1] 216
>
> by(salary$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

> 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

> with(dataset, tapply(salary, sex, mean))
   MALES  FEMALES
8120.558 5236.787
> with(dataset, tapply(salary, sex, median))
  MALES FEMALES
   6300    4950
> with(dataset, tapply(salary, sex, symmetry.test))
$MALES

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

data:  X[[i]]
Test statistic = 13.829, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 28
```



```
$FEMALES

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

data:  X[[i]]
Test statistic = 5.2527, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 59

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

        Wilcoxon signed rank test with continuity correction

data:  dd[x, ]
V = 33411, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0

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

        Wilcoxon signed rank test with continuity correction

data:  dd[x, ]
V = 23436, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0

>
> boxplot(salbeg~sex, data = salary)
>
```



5.2 Comment

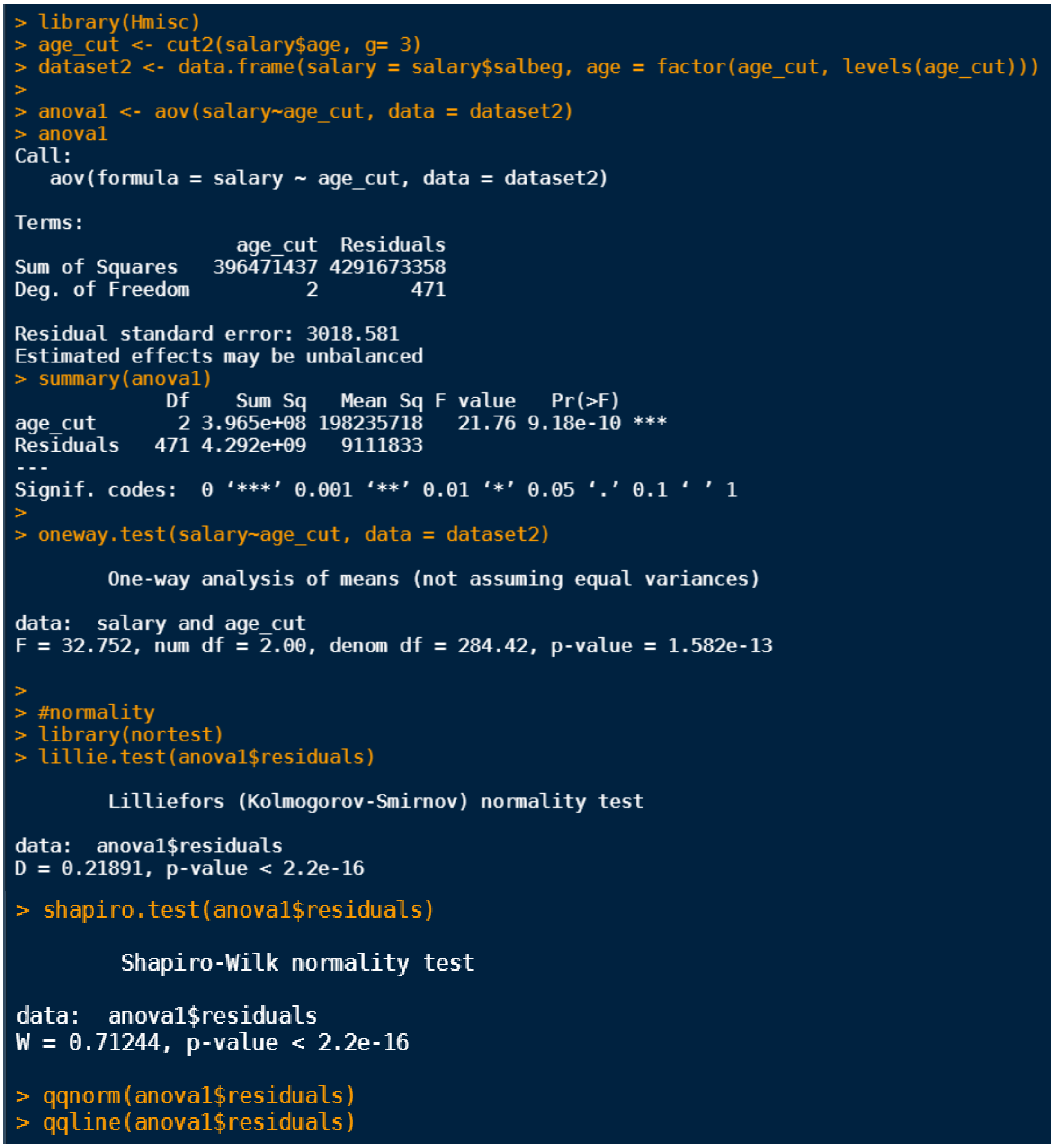
Like the previous two questions we follow the same steps in order to check the difference of the beginning salary between the two genders. Again, we haven't any normality nor symmetry among these two genders (because the length of the sample is larger than 50 we use the Kolmogorov-Smirnov test and then the Shapiro-Wilk test of normality for normality and like the above two questions we use for symmetry and here the comparison between mean and median and the "symmetry.test"). So, we use again the non-parametric Wilcoxon test to see if there is any difference among the medians of these samples. The null Hypothesis is H_0: E(Y|X=1) = E(Y|X=2) vs H1: E(Y|X=1) ≠ E(Y|X=2) (the alternative). In our case the Y is the beginning salary and X is the factor gender ("Male" or "Female"). After executing the R-code we can conclude that p-value is very small (smaller than the significance level 0,05) and as a result would be to reject the null Hypothesis (H_o). That's why, the two genders have significant difference among the beginning salary. This is also very obvious and in the above graph (boxplot). To be more specific, we can easily see in the boxplot that where the females' 75th percentile ends (3rd quartile) approximately the males' 25th percentile starts (1st quartile).

6th Question

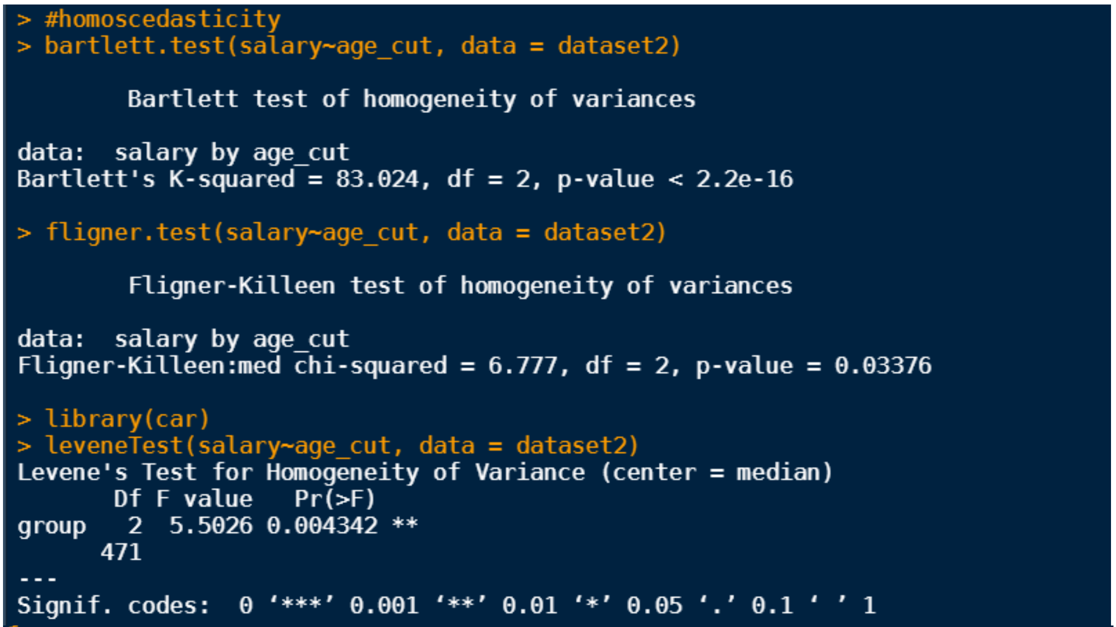
Cut the AGE variable into three categories so that the observations are evenly distributed across categories (Hint: you may find the cut2 function in Hmisc package to be very useful). Assign the cut version of AGE into a new variable called age_cut. Investigate if, on average, the beginning salary (**salbeg**) 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.

6.1 Output

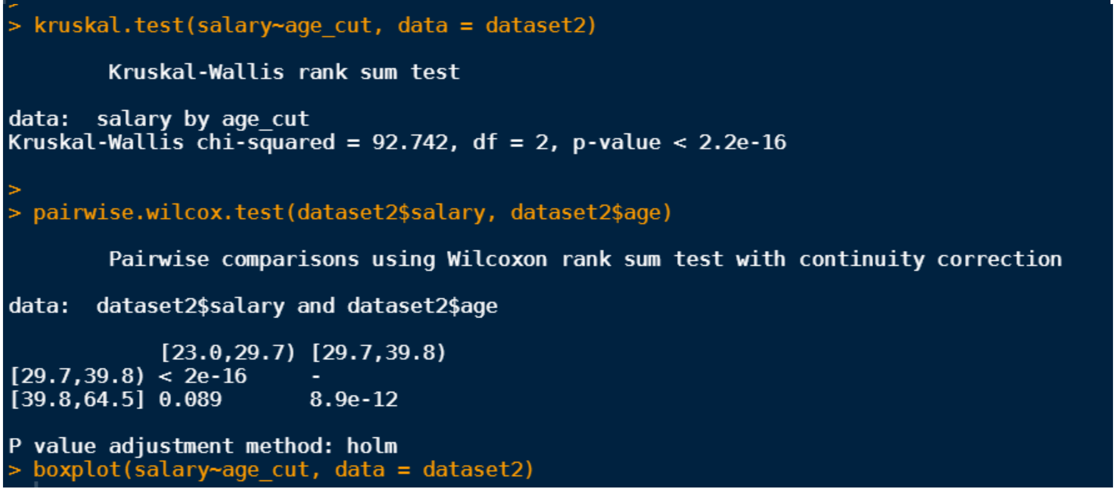
The output of this code in R language and of the plots that are generated from it is:



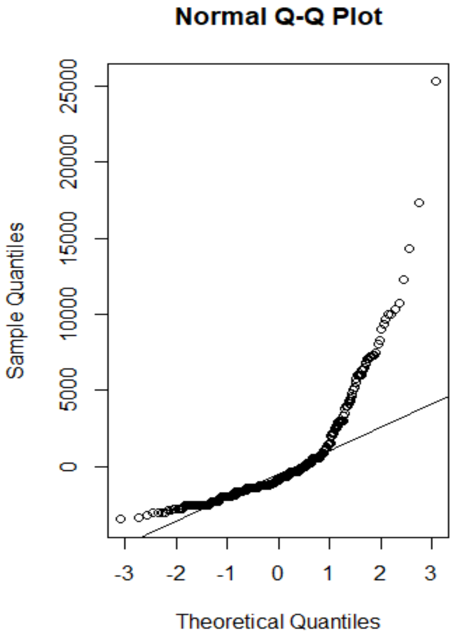
```
> library(Hmisc)
> age_cut <- cut2(salary$age, g= 3)
> dataset2 <- data.frame(salary = salary$salbeg, age = factor(age_cut, levels(age_cut)))
>
> anova1 <- aov(salary~age_cut, data = dataset2)
> anova1
Call:
   aov(formula = salary ~ age_cut, data = dataset2)

Terms:
                   age_cut  Residuals
Sum of Squares   396471437 4291673358
Deg. of Freedom          2        471

Residual standard error: 3018.581
Estimated effects may be unbalanced
> summary(anova1)
             Df    Sum Sq   Mean Sq F value   Pr(>F)    
age_cut       2 3.965e+08 198235718   21.76 9.18e-10 ***
Residuals   471 4.292e+09   9111833                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> oneway.test(salary~age_cut, data = dataset2)

        One-way analysis of means (not assuming equal variances)

data:  salary and age_cut
F = 32.752, num df = 2.00, denom df = 284.42, p-value = 1.582e-13

>
> #normality
> library(nortest)
> lillie.test(anova1$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  anova1$residuals
D = 0.21891, p-value < 2.2e-16

> shapiro.test(anova1$residuals)

        Shapiro-Wilk normality test

data:  anova1$residuals
W = 0.71244, p-value < 2.2e-16

> qqnorm(anova1$residuals)
> qqline(anova1$residuals)
```

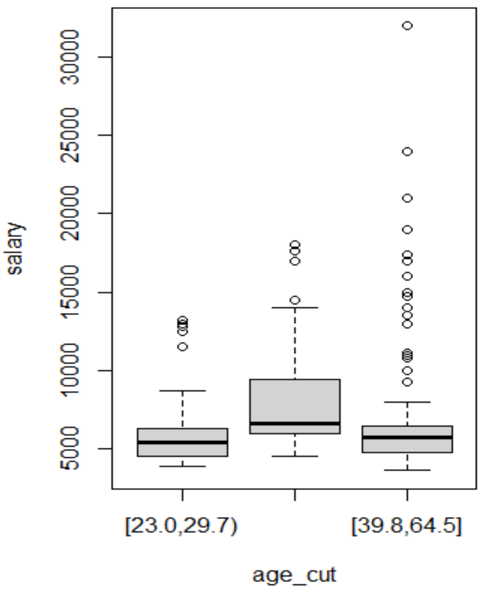


```
> #homoscedasticity
> bartlett.test(salary~age_cut, data = dataset2)

        Bartlett test of homogeneity of variances

data:  salary by age_cut
Bartlett's K-squared = 83.024, df = 2, p-value < 2.2e-16

> fligner.test(salary~age_cut, data = dataset2)

        Fligner-Killeen test of homogeneity of variances

data:  salary by age_cut
Fligner-Killeen:med chi-squared = 6.777, df = 2, p-value = 0.03376

> library(car)
> leveneTest(salary~age_cut, data = dataset2)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)   
group   2  5.5026 0.004342 **
      471                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> kruskal.test(salary~age_cut, data = dataset2)

        Kruskal-Wallis rank sum test

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

>
> pairwise.wilcox.test(dataset2$salary, dataset2$age)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  dataset2$salary and dataset2$age

            [23.0,29.7) [29.7,39.8)
[29.7,39.8) < 2e-16     -          
[39.8,64.5] 0.089       8.9e-12    

P value adjustment method: holm
> boxplot(salary~age_cut, data = dataset2)
```





6.2 Comment

First of all, we need to examine whether the beginning salary on average is the same for all age groups. In other words, we have to check the association of the beginning salary with all the 3 age groups that have been created. That is why, we use the ANOVA test (Analysis of Variance) as Hypothesis test with H_o: $\mu_1 = \mu_2 = \mu_3$ and alternative hypothesis H_1 the other way around (with inequalities). After executing the ANOVA test and from the summary of it, we can conclude that there are significant difference between age groups, because the p-value is very small (Pr(>F): smaller than a = 5%). This assumption is been confirmed and with the one-way analysis of variance (ANOVA), because this test is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups.

After that, we have to examine the normality of the residuals. Each residual is the difference between an entered value and the mean of all values for that group. A residual is positive when the corresponding value is greater than the sample mean and is negative when the value is less than the sample mean. So, we check the normality with Kolmogorov-Smirnov and Shapiro-Wilk tests to see if they follow the normal distribution. Based on the output of the above R-code we can conclude that p-value is again very small, so we reject the null Hypothesis that the residuals are normally distributed. That is also depicted and in the QQ Plot.

Next, we examine the homogeneity of the variances of the variables (salary and age groups) with three (3) tests: Bartlett, Fligner-Killeen and Levene's test. All these tests show us that p-value is very small and specifically smaller than 5%, so we reject the H_o that the variances are homogeneous.

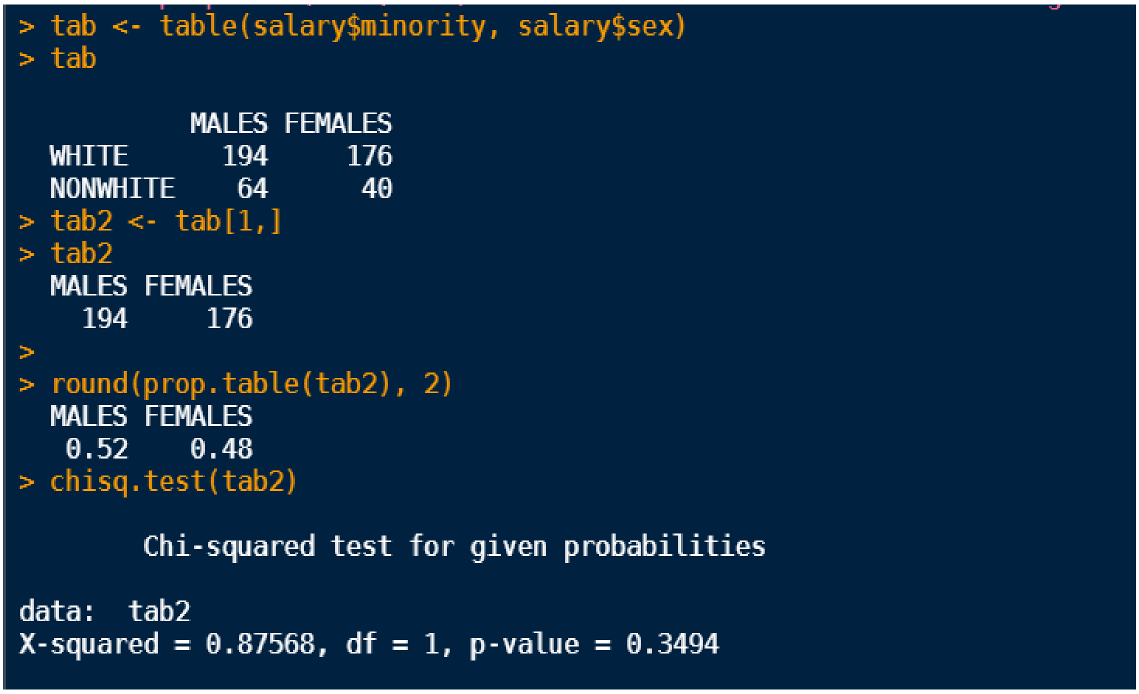
Because of the above last two assumptions, we use the Kruskal-Wallis rank sum test to check the equality of the means of salary, that is divided by 3 age groups. By executing this code, we take as a result very small p-values, so we reject the null Hypothesis that the variables we check do not differ the one from the other. So, we use the pairwise non-parametric Wilcoxon test to see the association of all the groups. From the above R-code we can easily conclude that there isn't any association between the beginning salary of the age groups: [29.7, 39.8) – [23.0, 29.7) & [29.7,39.8) – [39.8, 64.5], but there is an association between the beginning salary of [23.0, 29.7) - [39.8, 64.5] age groups, because the p-value here is 0.089 or 8.9% (larger than the default significant level 0.05). So, in the last case we do not have many evidence to reject the null Hypothesis that the means of beginning salary of these 2 age groups are equal. The above assumption is been represented and in the boxplot of each level of the age groups.

7th Question

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.

7.1 Output

The output of this code in R language is:



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

           MALES FEMALES
  WHITE      194     176
  NONWHITE    64      40
> tab2 <- tab[1,]
> tab2
  MALES FEMALES
    194     176
>
> round(prop.table(tab2), 2)
  MALES FEMALES
   0.52    0.48
> chisq.test(tab2)

        Chi-squared test for given probabilities

data:  tab2
X-squared = 0.87568, df = 1, p-value = 0.3494
```

7.2 Comment

The proportion of white male employees is not equal to the proportion of white female employees, because as we can see from the table we created that the white men are 194 and the white females are 176 (52% and 48% of the total white people accordingly).

The chi-square test is based on a test statistic that measures the divergence of the observed data from the values that would be expected under the null hypothesis of no association. So, after executing this code we can see that p-value is up to 5% (precisely is 35%) and for that reason we conclude that there are not enough evidence to reject the null hypothesis that these two proportions have no association. That is why, we can say that the proportion of white male employees is not equal to the proportion of white female employees.

R Code

The code that has been produced for this Lab Assignment is the following:

```
#Q1
require(foreign)
salary <- read.spss("salary.sav", to.data.frame = "T")
str(salary)

#Q2
library(psych)
df <- salary[unlist(lapply(salary, is.numeric))] #get only the numeric variables
summary(df)

for (j in seq(2)) {
  par(mfrow=c(2,4))
  if(j == 1){
    for (i in colnames(df)) {
      hist(df[,i], probability = TRUE, xlab = i, main = paste("Histogram of", i))
      lines(density(df[,i]), col = "red")
    }
  } else {
    for (i in colnames(df)) {
      qqnorm(df[,i], main = paste("Normal Q-Q Plot for", i), xlab = i)
      qqline(df[,i])
    }
  }
}

#Q3
length(salary$salbeg)

#normality tests
library(nortest)
lillie.test(salary$salbeg)
shapiro.test(salary$salbeg)

#symmetry tests
mean(salary$salbeg)
median(salary$salbeg)
```

```
library(lawstat)
symmetry.test(salary$salbeg)

wilcox.test(salary$salbeg, mu = 1000)

#Q4
dif <- salary$salnow - salary$salbeg
length(dif)

#normality tests
lillie.test(dif)
shapiro.test(dif)

#visualization of the difference
par(mfrow=c(1,2))
hist(dif, probability = TRUE)
lines(density(dif), col = "red")
qqnorm(dif)
qqline(dif)

#symmetry tests
mean(dif)
median(dif)
symmetry.test(dif)

wilcox.test(dif)

par(mfrow=c(1,1))
boxplot(dif)

#Q5
dataset <- data.frame(salary = salary$salbeg, sex = factor(salary$sex, levels(salary$sex)))
length(dataset$sex[dataset$sex == "MALES"])
length(dataset$sex[dataset$sex != "MALES"])

by(salary$salbeg, salary$sex,lillie.test)
by(salary$salbeg, salary$sex,shapiro.test)


with(dataset, tapply(salary, sex, mean))
with(dataset, tapply(salary, sex, median))
with(dataset, tapply(salary, sex, symmetry.test))
```

```
by(salary$salbeg, salary$sex, wilcox.test)

boxplot(salbeg~sex, data = salary)

#Q6
library(Hmisc)
age_cut <- cut2(salary$age, g= 3)
dataset2 <- data.frame(salary = salary$salbeg, age = factor(age_cut, levels(age_cut)))

anova1 <- aov(salary~age_cut, data = dataset2)
anova1
summary(anova1)

oneway.test(salary~age_cut, data = dataset2)

#normality
library(nortest)
lillie.test(anova1$residuals)
shapiro.test(anova1$residuals)
qqnorm(anova1$residuals)
qqline(anova1$residuals)

#homoscedasticity
bartlett.test(salary~age_cut, data = dataset2)
fligner.test(salary~age_cut, data = dataset2)
library(car)
leveneTest(salary~age_cut, data = dataset2)

kruskal.test(salary~age_cut, data = dataset2)

pairwise.wilcox.test(dataset2$salary, dataset2$age)
boxplot(salary~age_cut, data = dataset2)

#Q7
tab <- table(salary$minority, salary$sex)
tab
tab2 <- tab[1,]
tab2

round(prop.table(tab2), 2)
chisq.test(tab2)
```