ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
BUSINESS ANALYTICS Master of Science

Statistics for Business Analytics I (Part Time)

Lab Assignment 1: Hypothesis Testing in "**Salary**" Dataset

Despotis Spyridon: p282211

2021

Question 1: Read the dataset "salary.sav" as a data frame and use the function str() to understand its structure.

```
> 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
```

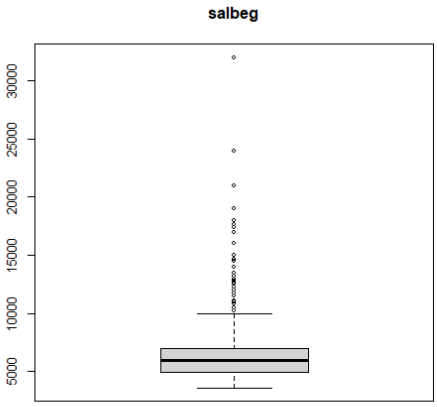
The dataset “Salary” contains **474** raws and **11** columns with 2 different data types: numerical and categorical. The categorical variables are encoded as factors. In more detail, “sex” variable has 2 levels, “jobcat” has 7 levels, “minority” has 2 levels and “sexrace” 4 levels. So we can proceed our statistical analysis.

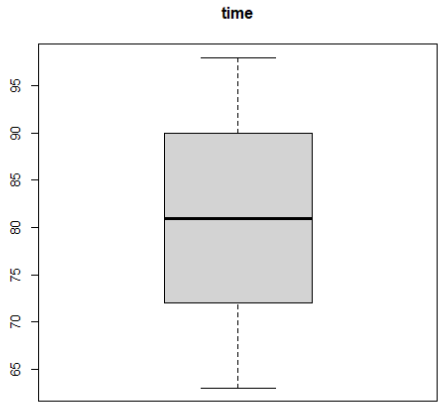
Question 2: Get that summary statistics of the numerical variables in the dataset and visualize their distribution (e.g. use histograms etc). Which variables appear to be normally distributed? Why?

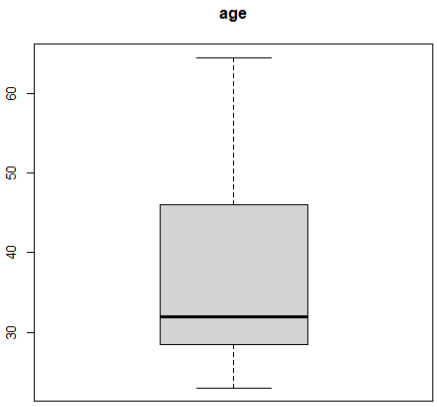
Through the summary table we can see that the medians and means for each variable are different, except for the variable time in which median and mean are more close.

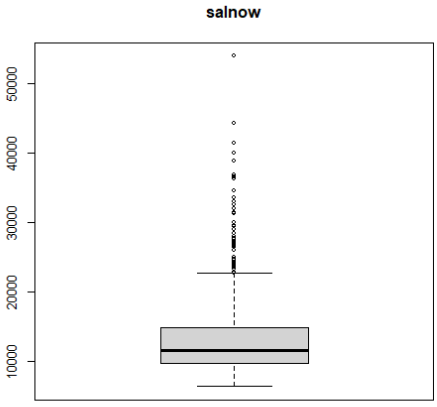
```
> summary (sal_num)
     salbeg          time            age            salnow        edlevel          work       
 Min.   : 3600   Min.   :63.00   Min.   :23.00   Min.   : 6300   Min.   : 8.00   Min.   : 0.000  
 1st Qu.: 4995   1st Qu.:72.00   1st Qu.:28.50   1st Qu.: 9600   1st Qu.:12.00   1st Qu.: 1.603  
 Median : 6000   Median :81.00   Median :32.00   Median :11550   Median :12.00   Median : 4.580  
 Mean   : 6806   Mean   :81.11   Mean   :37.19   Mean   :13768   Mean   :13.49   Mean   : 7.989  
 3rd Qu.: 6996   3rd Qu.:90.00   3rd Qu.:45.98   3rd Qu.:14775   3rd Qu.:15.00   3rd Qu.:11.560  
 Max.   :31992   Max.   :98.00   Max.   :64.50   Max.   :54000   Max.   :21.00   Max.   :39.670  
```

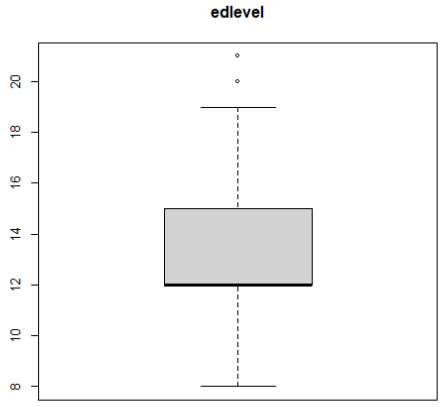
Visualizing our data with QQ-plots, box plots and histograms we can understand more the distribution of values and look for outliers:

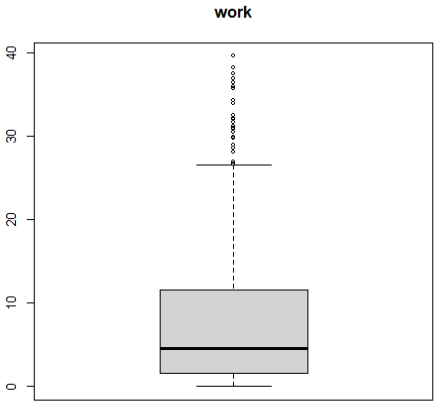

salbeg
5000
10000
15000
20000
25000
30000

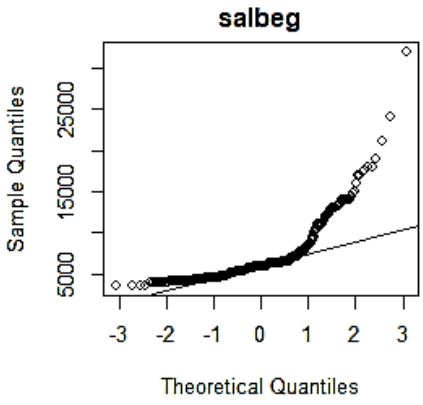

time
65
70
75
80
85
90
95

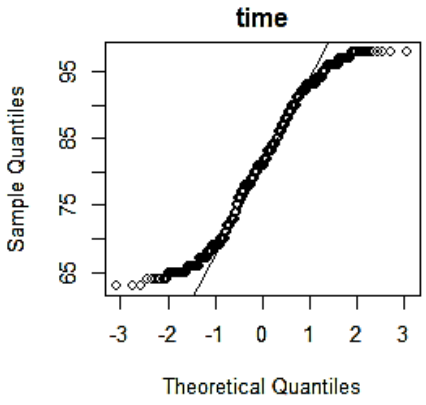

age
30
40
50
60

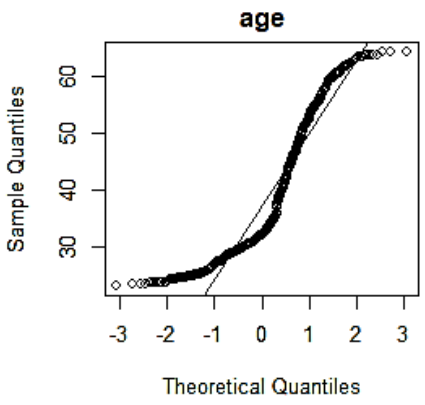

salnow
10000
20000
30000
40000
50000

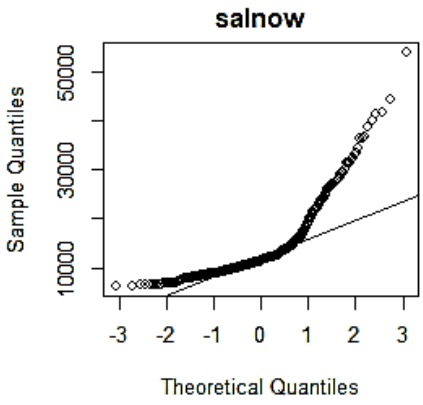

edlevel
8
10
12
14
16
18
20

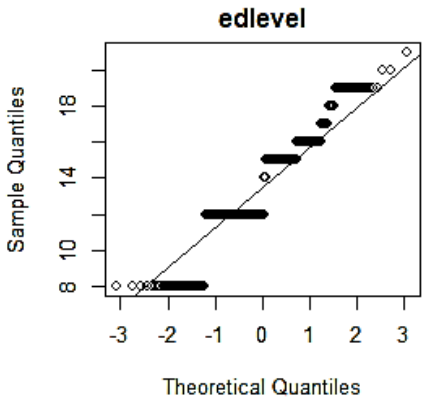

work
0
10
20
30
40

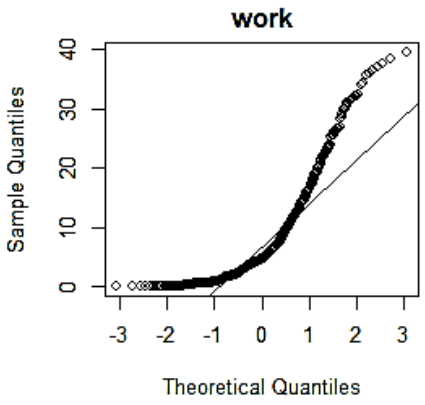

salbeg
Sample Quantiles
5000
15000
25000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

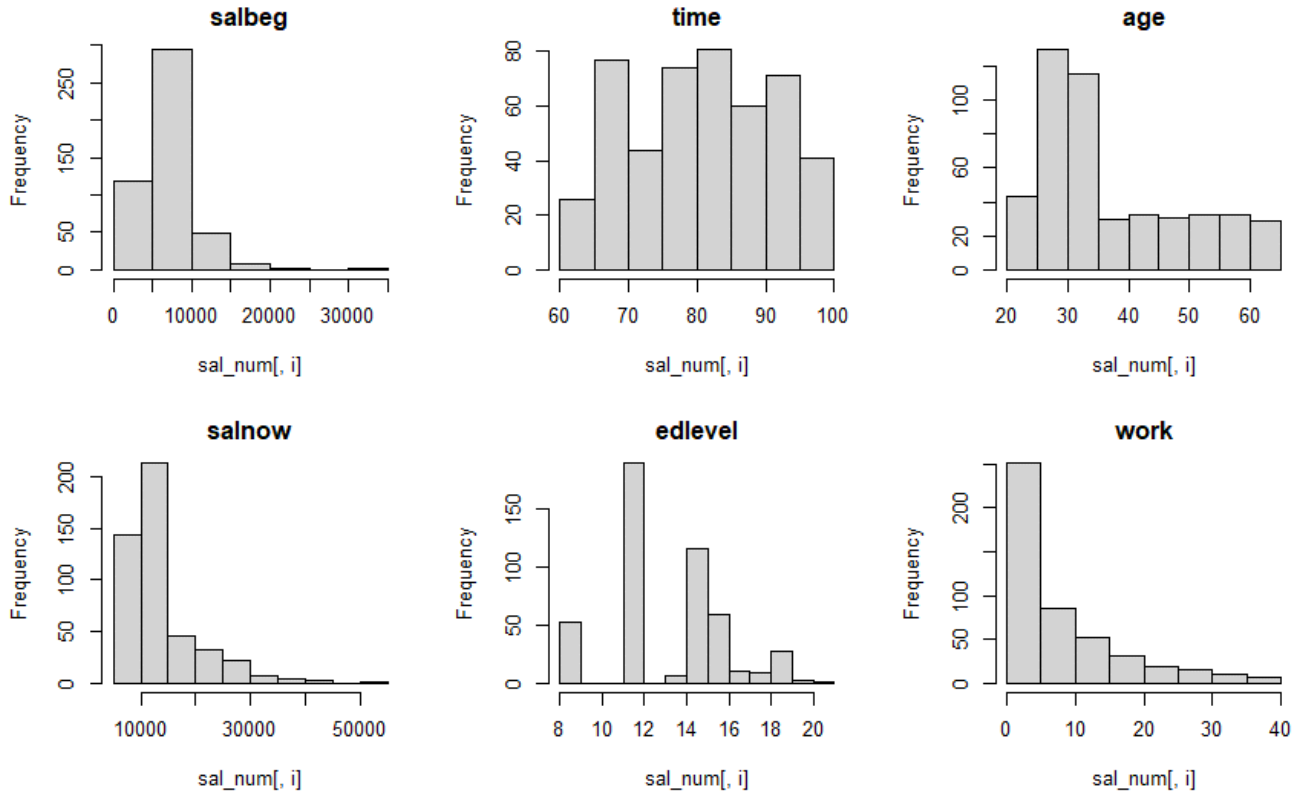

time
Sample Quantiles
65
75
85
95
-3
-2
-1
0
1
2
3
Theoretical Quantiles


age
Sample Quantiles
30
40
50
60
-3
-2
-1
0
1
2
3
Theoretical Quantiles


salnow
Sample Quantiles
10000
30000
50000
-3
-2
-1
0
1
2
3
Theoretical Quantiles


edlevel
Sample Quantiles
8
10
14
18
-3
-2
-1
0
1
2
3
Theoretical Quantiles


work
Sample Quantiles
0
10
20
30
40
-3
-2
-1
0
1
2
3
Theoretical Quantiles



At first look, we can see that none of the variables seem to follow a normal distribution. We would expect the histograms to be approximately bell-shaped and symmetric about the mean, but we have skewness. Also, in QQ-plots, we would expect the data points following the reference line but instead we have heavy tails and the data points curve off. The only variable that seems to be more close to "normal" distribution compared to the others, is "time", since the data points seem to deviate less from the straight line in QQ-plot and we have less skewness but considerable curtosis in the histogram. Last but not least, through the box plots we can see that especially the variables of "salbeg", "work" and "salnow" have many data points located outside the whiskers of the box plots (outliers).

Question 3: 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.

Since we have one sample with one quantitative variable ("beginning salary"), we can use the hypothesis tests for a single continuous variable. Due to the fact that our sample (474 observations) of the population is bigger than 50, we can perform two tests to assume normality: Kolmogorov-Smirnov and Shapiro-Wilkinson.

- *The **Ho** (null) hypothesis states that our sample (beginning salary) follows a normal distribution.*
- *The **H1** (alternative) hypothesis states that our sample (beginning salary) does not follow a normal distribution.*

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

        Shapiro-Wilk normality test

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

> lillie.test(sal_num$salbeg)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

The corresponding p-value is <**2.2*10-16** in both tests, so it is lower than the significance level of alpha = **0.05**. So we reject our null hypothesis that our data follow a normal distribution.

We can see that the raw data do not follow a normal distribution. Therefore we can use log transformed data with the same hypothesis to check if they follow a log-normal distribution.

```
> shapiro.test(logbegs)

        Shapiro-Wilk normality test

data:  logbegs
W = 0.89684, p-value < 2.2e-16

> lillie.test(logbegs)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  logbegs
D = 0.17869, p-value < 2.2e-16
```

The corresponding p-value is < 2.2*10-16 in both tests, lower than the significance level of alpha = 0.05. So we reject our null hypothesis that our log sample of the population does come from a normal distribution. Due to the fact that our sample of the population is large (474 observations) we proceed checking if the mean is a sufficient descriptive measure for central location:

- *The **Ho** (null) hypothesis states that the distribution of our data (beginning salary) is symmetric.*

- *The **H1** (alternative) hypothesis states that the distribution of our data (beginning salary) is asymmetric.*

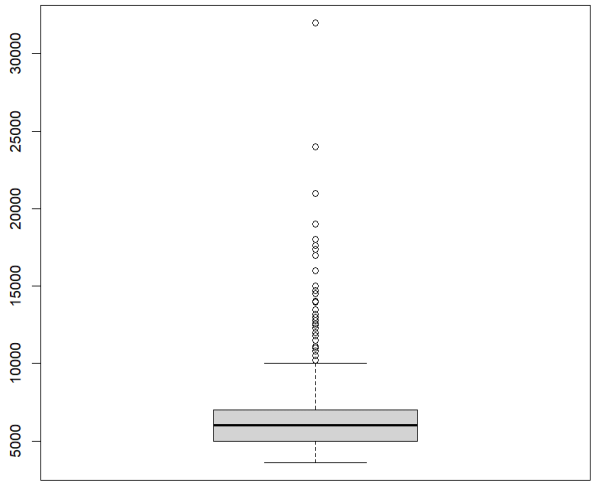
```
        m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)

data:  salary$salbeg
Test statistic = 10.18, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 51
```

From the above output we can see that the corresponding p-value is < 2.2*10-16. Since the p-value is less than the significance level of alpha = **0.05**, we reject the null hypothesis that the population of our sample is symmetric. Therefore, we use the non-parametric statistical hypothesis Wilcoxon signed-rank test, to test for the medians in our sample:

- *The **Ho** (null) hypothesis states that the median beginning salary is equal to 1000.*
- *The **H1** (alternative) hypothesis states that the median beginning salary is equal to 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
```

The output shows that there is significant evidence (p < 0.05) to reject Ho. That means that the beginning salary of a typical employee can not be considered to be equal to 1000 dollars. The data visualizations can confirm our findings:

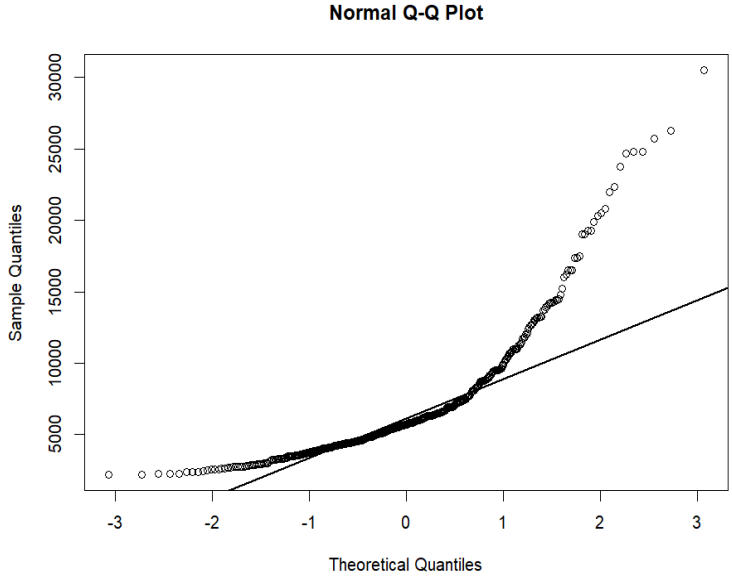


Question 4: Consider the difference between the beginning salary (salbeg) and the current salary (salnow). Test if the there is any significant difference between the beginning salary and current salary. (Hint: Construct a new variable for the difference (salnow – salbeg) and test if, on average, it is equal to zero.). Make sure that the choice of the test is well justified.

In this case, we have two values (i.e., pair of values: beginning and current salary) for the same samples (individuals). The appropriate parametric test for this case is the paired samples t-test, which is used to compare the means between two related groups of samples.

Paired t-test can be used only when the difference between each pair of values is normally distributed.

The QQ-plot for the difference shows that, the points deviate from the reference lines, especially at the tails. Since our sample (474 observations) of the population is bigger than 50, in order to test the normality of the sample we will perform two tests to assume normality: Kolmogorov-Smirnov and Shapiro-Wilk.



- *The **Ho** (null) hypothesis states that our sample follows a normal distribution.*
- *The **H1** (alternative) hypothesis states that our sample does not follow a normal distribution.*

```
> shapiro.test(x)

        Shapiro-Wilk normality test

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

> lillie.test(x)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

The corresponding p-value in both tests, is lower than the significance level of alpha = **0.05**. So we reject our null hypothesis that our sample follows a normal distribution. This conclusion is align with our visual findings.

Due to the fact that our sample is large (>50) we proceed checking if the mean is a sufficient descriptive measure for central location:

- *The **Ho** (null) hypothesis states that the distribution of our sample is symmetric.*
- *The **H1** (alternative) hypothesis states that the distribution of our sample is asymmetric.*

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

data:  x
Test statistic = 10.536, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                263
```

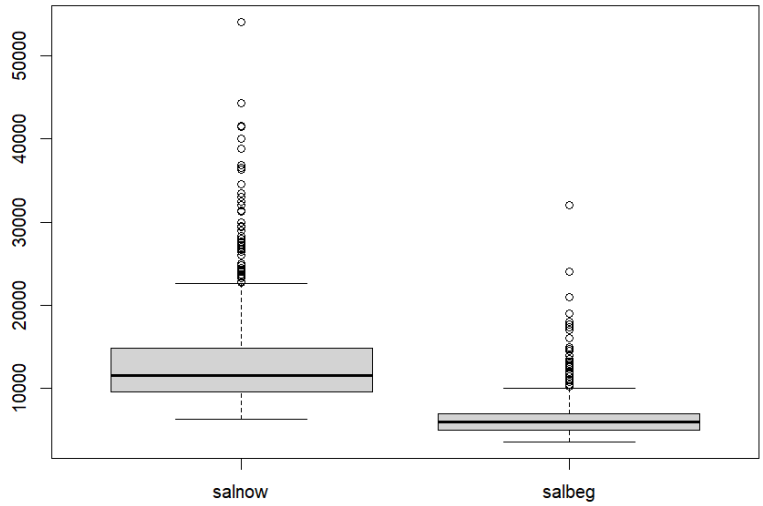
From the output we can see that the corresponding p-value is 2.2*10-16. Since the p-value is less than the significance level of alpha = **0.05**, we reject the null hypothesis which states that the distribution of our sample is symmetric.

Therefore, we use the non-parametric statistical hypothesis Wilcoxon signed-rank test, to test for the medians in our sample:

- *The **Ho** (null) hypothesis states that the median difference in the pair of values is equal to zero. So we do not have difference between the median beginning salary and current salary.*
- *The **H1** (alternative) hypothesis states that the median difference in the pair of values is not equal to zero. So we do have difference between the median beginning salary and current salary.*

```
> wilcox.test(salary$salnow,salary$salbeg, paired=T)

        Wilcoxon signed rank test with continuity correction

data:  salary$salnow and salary$salbeg
V = 112575, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

From the output we can see that the corresponding p-value is < 2.2*10-16. Since the p-value is lower than the significance level of alpha = 0.05, we have strong evidence to reject the null hypothesis. So there is a difference between the median beginning salary and current salary of an employee. The graphs confirm our findings:



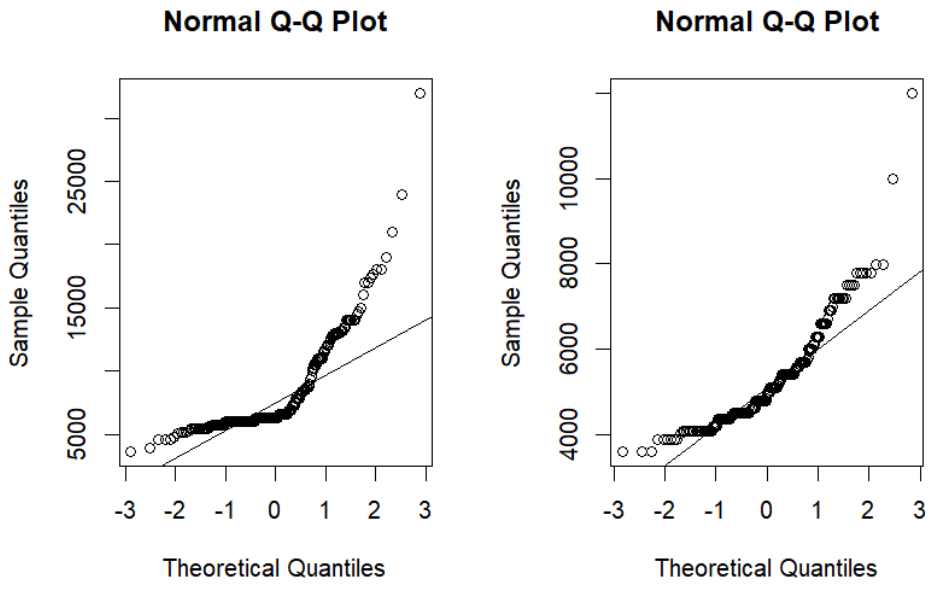
Question 5: Is there any difference on the beginning salary (salbeg) between the two genders? Give a brief justification of the test used to assess this hypothesis and interpret the results.

In this case we have two independent samples (females, males) of one continuous variable (beginning salary). To test for differences using a parametric or a non parametric approach, we first need to test the normality. Due to the fact that the sample is larger (>50) we make 2 tests, Shapiro-Wilk and Kolmogorov-Smirnov:

- *The **Ho** (null) hypothesis states that each sample follows a normal distribution.*
- *The **H1** (alternative) hypothesis states that each sample does not follow a normal distribution.*

```
> library('nortest')
> by(new_dataset$salbeg, new_dataset$method, lillie.test)
new_dataset$method: MALE

        Lilliefors (Kolmogorov-Smirnov) normality test

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

------------------------------------------------------------
new_dataset$method: FEMALE

        Lilliefors (Kolmogorov-Smirnov) normality test

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

> by(new_dataset$salbeg, new_dataset$method, shapiro.test)
new_dataset$method: MALE

        Shapiro-Wilk normality test

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

------------------------------------------------------------
new_dataset$method: FEMALE

        Shapiro-Wilk normality test

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

>
```

From the output we can see that the corresponding p-value in all tests is lower than the significance level of alpha = **0.05** and therefore we reject the null hypothesis that our sample follows the normal distribution. Also we can visualize the distributions using normal qq-plots:



We can see that the data points curve off the reference line, especially at the tails of the distribution. That pattern confirms the results from our tests. Due to the fact that our sample is large 474 (>50), we proceed testing the symmetry to examine if the means can be a sufficient descriptive measure for central location for both groups:

- *The **Ho** (null) hypothesis states that the distribution of each group is symmetric.*
- *The **H1** (alternative) hypothesis states that the distribution of each group is asymmetric.*

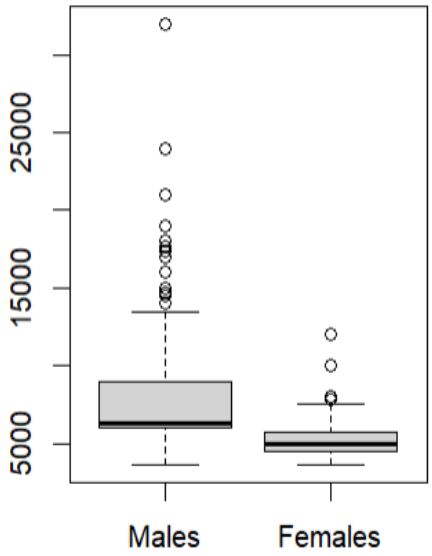
```
> symmetry.test(salmale$salbeg)

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

data:  salmale$salbeg
Test statistic = 13.829, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                 35

> symmetry.test(salfemale$salbeg)

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

data:  salfemale$salbeg
Test statistic = 5.2527, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                135
```

From the output we can see that the corresponding p-value in all tests is lower than the significance level of alpha = 0.05 and therefore we reject the null hypothesis that each group has a symmetrical distribution. We proceed making the non-parametric statistical hypothesis Wilcoxon rank sum test, to test for zero difference between the medians in our groups:

- *The **Ho** (null) hypothesis states that the medians of beginning salary for each gender group are the same.*
- *The **H1** (alternative) hypothesis states that the medians of beginning salary for each gender group are not the same.*

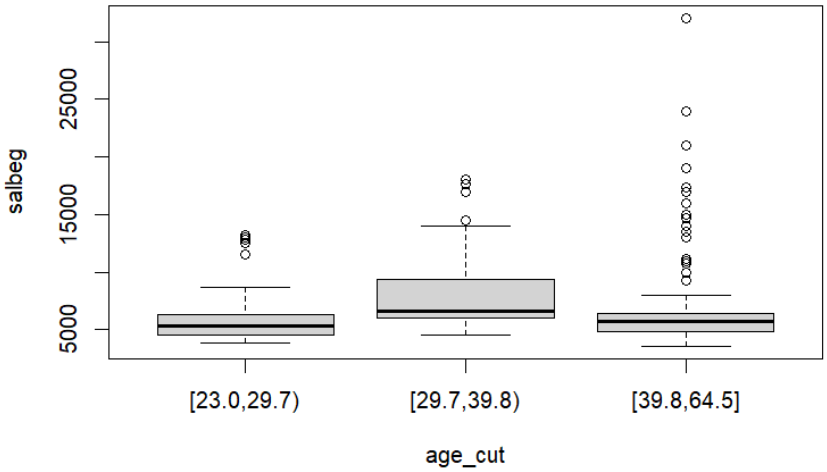
```
> wilcox.test(salmale$salbeg,salfemale$salbeg)

        Wilcoxon rank sum test with continuity correction

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

From the output we can see that the corresponding p-value is < 2.2*10-16. Since the p-value is lower than the significance level of alpha = **0.05**, we have strong evidence to reject the null hypothesis, that the medians of the gender groups (males – females) are equal. So we can conclude that there is a difference in the beginning salary for each group. The following boxplots of beginning salary is each group, are in line with our results:



Question 6: 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.

In this case we have to compare the means or medians of three independent age groups (κ>3).



Plotting our age categories with boxplot we can see that there is a difference in the median salaries of the three groups. The age group [29.7, 39.8) seem to differ from the other two.

For more details, we will proceed with Anova test and test its assumptions (Normality & Homoscedasticity). Firstly we will check the residuals normality. Due to the fact that our sample is large (>50) we conduct two tests: Shapiro-Wilk and Kolmogorov-Smirnov.

- *The **Ho** (null) hypothesis states that the residuals are normally distributed.*
- *The **H1** (alternative) hypothesis states that the residuals are not normally distributed.*

```
> library('nortest')
> shapiro.test(anova1$residuals)

        Shapiro-Wilk normality test

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

> lillie.test(anova1$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

The corresponding p-value in both tests of residuals, is lower than the significance level of alpha = **0.05**. So we reject our null hypothesis that the residuals are normally distributed. Then we proceed testing

Homoscedacity of variances with tests Barlette, Leven and Fligner-Killeen:

- *The **Ho** (null) hypothesis states that the population variances are equal.*
- *The **H1** (alternative) hypothesis states that at least two population variances are not equal.*

```
> bartlett.test(salbeg~age_cut, data=sal_num)

        Bartlett test of homogeneity of variances

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

> fligner.test(salbeg~age_cut, data=sal_num)

        Fligner-Killeen test of homogeneity of variances

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

> library(car)
> leveneTest(salbeg~age_cut, data=sal_num)
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
```

From the output, it can be seen that the p-value of $< 2.2*10^{-16}$ is less than the significance level of alpha = 0.05. This means that there is evidence to reject the null hypothesis and thus the variance between the age groups is heterogenous. Now that we have rejected the normality and homoscedasity ssumptions, we may use the non parametric test of Kruskal-Wallis.

So we proceed testing the medians of the groups.

- *The **Ho** (null) hypothesis states that the medians of beginning salary for the three age groups are equal.*
- *The **H1** (alternative) hypothesis states that the medians of beginning salary for the three age groups are not equal.*

```
> kruskal.test(salbeg~age_cut, data=sal_num)

        Kruskal-Wallis rank sum test

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

From the output we can see that the corresponding p-value is 2.2*10-16. Since the p-value is really small compared to the significance level of alpha = **0.05**, we have strong evidence to reject the null hypothesis that the medians of beginning salary for the three age groups are equal. So we can say that there is a difference in the median beginning salaries for each group. The next step, is to perform pairwise Wilcoxon rank sum test (nonparametric test), comparing all the pairs of groups to identify which groups differ significantly.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  sal_num$salbeg and sal_num$age_cut

           [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
```

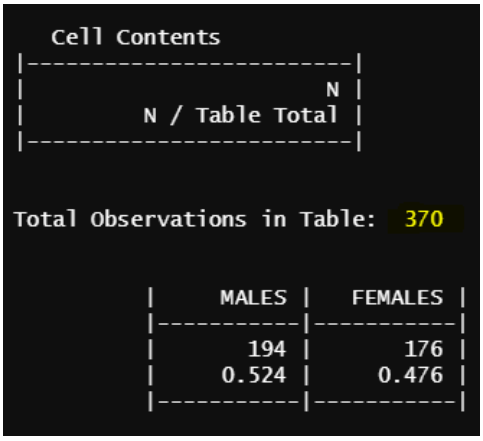
- *The **Ho** (null) hypothesis states that there is not significance difference in the average sum of the ranks (and thus the medians) of the two groups.*
- *The **H1** (alternative) hypothesis states that there is significance difference in the average sum of the ranks (and thus the medians) of the two groups.*

From the output we can see, that in the pair of age groups: [29.7, 39.8) & [23.0,29.7) and [39.8, 69.5) & [29.7,39.8) the p value is less than the significance level of alpha = 0.05. So we reject the null hypothesis that there is not significance difference on medians of salaries between the two groups. In the pair age group [39.8,64.5) with [23.0,29.7), the p value = 0.089, is greater than the significance level of alpha = 0.05, so we do not reject the null hypothesis that there is no significant difference in the medians of salaries between the two groups.

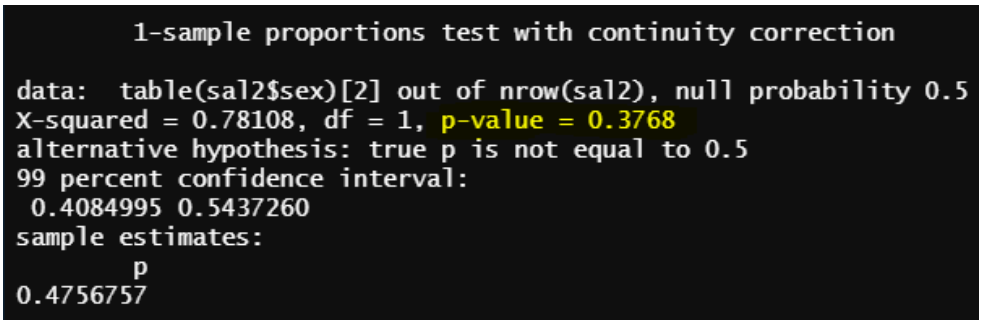
Question 7: By making use of the factor variable minority, investigate if the proportion of white male employees is equal to the proportion of white female employees.

We have one sample of observations (minority= “WHITE”) and we need to test the proportion of the genders for equality. In other words, we need to check if the proportion of either gender is 0.5. The appropriate test is the on*10-sample test for proportions.

The following crosstable shows the number and proportion of the male and female employees:



The prop.test( ) procedure will perform the z-test comparing the proportion of females to the hypothesized value (0.5); input for the prop.test is the number of events (176), the total sample size (370), the hypothesized value of the proportion under the null (p=0.50 for a null value of 50%). The results are the following:



The p-value is 0.38>0.04, thus we fail to reject the null hypothesis that the proportion of females is 50%. Thus, we can infer that both genders are equally represented in the white employees.

R Code Used in Questions

```
require(foreign)
salary <- read.spss("salary.sav", to.data.frame = T)
###custom functions------
#Error Bar code
myerrorbar<-function(x,y, horizontal=F){
  a<-0.05
  sdata <- split(x,y)
  means <- sapply( sdata,mean )
  sds <- sapply( split(x,y), sd )
  ns <- table(y)
  LB <- means + qnorm( a/2 ) * sds /sqrt(ns)
  UB <- means + qnorm( 1-a/2 ) * sds /sqrt(ns)
  nlev <- nlevels(y)
  if (horizontal) { errbar( levels(y), means, UB, LB )
  } else {
    errbar( 1:nlev, means, UB, LB, xlim=c(0,nlev+1), axes=F, xlab=''
)
    axis(2)
    axis(1, at=0:(nlev+1), labels=c('',levels(y),''))
  }}
###exercise 1------
any(is.na(salary))
str(salary)
###exercise 2------
index <- sapply(salary, class) == "numeric"
sal_num <- salary[index]
sal_num <- sal_num[,-1]
summary (sal_num)

par(mfrow=c(2,3))
for (i in 1:ncol(sal_num[,1: ncol(sal_num)])){
  boxplot(sal_num[, i], main = names(sal_num[i]))
}
par(mfrow=c(2,3))
for (i in 1:ncol(sal_num[,1: ncol(sal_num)])){
  qqnorm(sal_num[, i], main = names(sal_num[i]))
  qqline(sal_num[, i])
}
par(mfrow=c(2,3))
for (i in 1:ncol(sal_num[,1: ncol(sal_num)])){
  hist(sal_num[, i], main = names(sal_num[i]))}
###exercise 3-----
library('nortest')
shapiro.test(sal_num$salbeg)
lillie.test(sal_num$salbeg)
logbegs<-log(sal_num$salbeg)
shapiro.test(logbegs)
lillie.test(logbegs)
install.packages('lawstat')
library(lawstat)
symmetry.test(salary$salbeg)
```

```
install.packages('lawstat')
library(lawstat)
symmetry.test(logbegs)
wilcox.test(sal_num$salbeg, mu=1000)
boxplot(
)
qqnorm( y = sal_num$salbeg )
qqline(sal_num$salbeg, col = "black", lwd = 2)
boxplot(sal_num$salbeg)
###exercise 4-----
x<-salary$salnow-salary$salbeg
par(mfrow=c(1,1))
qqnorm( y = x )
qqline(x, col = "black", lwd = 2)
library('nortest')
shapiro.test(x)
lillie.test(x)
library(lawstat)
symmetry.test(x)
wilcox.test(salary$salnow,salary$salbeg, paired=T)
require(Hmisc)
par(mfrow = c(1,1))
library(Hmisc)
myerrorbar(logsalary_diff)
boxplot(salary[, c("salnow", "salbeg")])
###exercise 5-----
salmale<-salary[salary$sex == "MALES", ]
salfemale<-salary[salary$sex == "FEMALES", ]
n_smale<-length(salmale$salbeg)
n_sfemale<-length(salfemale$salbeg)
new_dataset <- data.frame( salbeg=c(salmale$salbeg,
salfemale$salbeg),
                          method=factor( rep(1:2,
c(n_smale,n_sfemale)), labels=c('MALE','FEMALE')))
library('nortest')
by(new_dataset$salbeg, new_dataset$method, lillie.test)
by(new_dataset$salbeg, new_dataset$method, shapiro.test)
par(mfrow=c(1,2))
qqnorm(salmale$salbeg)
qqline(salmale$salbeg)
qqnorm(salfemale$salbeg)
qqline(salfemale$salbeg)
library(lawstat)
symmetry.test(salmale$salbeg)
symmetry.test(salfemale$salbeg)
wilcox.test(salmale$salbeg,salfemale$salbeg)
boxplot(salmale$salbeg, salfemale$salbeg,
names=c("Males","Females"))
###exercise 6-----
library(Hmisc)
sal_num$age_cut <- cut2(salary$age,g = 3)
table(sal_num$age_cut)
anova1 <- aov(salbeg~age_cut, data=sal_num )
```

```r
summary(anova1)
boxplot(salbeg~age_cut, data=sal_num)
library(nortest)
lillie.test(anova1$residuals)
shapiro.test(anova1$residuals)
bartlett.test(salbeg~age_cut, data=sal_num)
fligner.test(salbeg~age_cut, data=sal_num)
library(car)
leveneTest(salbeg~age_cut, data=sal_num)
library(lawstat)
symmetry.test(anova1$residuals)
kruskal.test(salbeg~age_cut, data=sal_num)
pairwise.wilcox.test(sal_num$salbeg, sal_num$age_cut)
qqnorm(anova1$residuals)
qqline(anova1$residuals)
###exercise 7-----
sal2=salary[salary$minority=="WHITE",]
table(sal2$sex)
install.packages(gmodels)
require(gmodels)
CrossTable(sal2$sex)
prop.test(x=table(sal2$sex)[2], n=nrow(sal2), conf.level=0.99)
```