ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

STATISTICS FOR BUSINESS ANALYTICS I

Assignment 1 on Hypothesis Tests

Name: Ilias Dimos

Student Number: f2822102

Professor: Ilias Leriou

Period: 2021-2022

Πίνακας περιεχομένων

Explanation of Assignment's Dataframe

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.

Question 1

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

Output

```
> 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
```

To complete the Question 1, I used "str" function in order to print the types of data to understand the structure of the data frame. First, we can observe that our data frame contains 7 numeric variables (id, salbeg,time,age,salnow,edlevel,work) and 4 factor variables (sex,jobcat,minority,sexrace). Each one of the factor variables is coded.

Analytically,

- "sex" variable is coded 1=female,0=male.
- "Jobcat" variable has 7 levels ("CLERICAL","OFFICE TRAINEE","SECURITY OFFICER","COLLEGE TRAINEE","EXEMPT EMPLOYEE","MBA TRAINEE" "TECHNICAL").
- "Minority" variable has 2 levels ("WHITE","NONWHITE").
- "Sexrace" Variable has 4 levels ("WHITE MALES","MINORITY MALES","WHITE FEMALES","MINORITY FEMALES")

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?

In order to print the summary for the numerical variables we have to create a new variable named “numerical statistics” which contains the variables: “id”, ”salbeg”, “time”, ”age” , ”salnow” , ”edlevel” , ”work”.

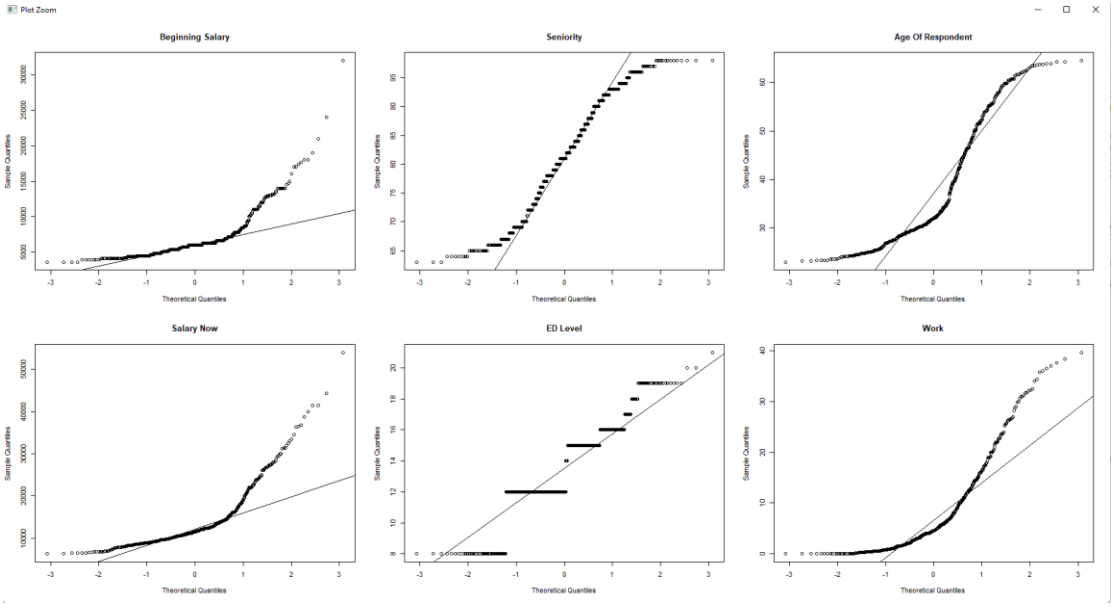
```
> head(salary[,numerical_statistics])
  id salbeg time   age salnow edlevel  work
1  1   8400   81 28.50  16080      16  0.25
2  2  24000   73 40.33  41400      16 12.50
3  3  10200   83 31.08  21960      15  4.08
4  4   8700   93 31.17  19200      16  1.83
5  5  17400   83 41.92  28350      19 13.00
6  6  12996   80 29.50  27250      18  2.42
```

The result of the “summary” function can be seen on the below table

```
> summary(salary[,numerical_statistics])
       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  
```

In this table we can observe all the significant metrics in every numeric variable which are important for our analysis.

Output



To check if the numerical variables are normally distributed, we created “qqnorm” plots for each variable. Also, we draw the qqline in order to help us define if the samples are normally distributed. The definition of the qqline is that the closer the points (of the qqnorm plot) lie to the line the closer the distribution of the sample comes to the normal distribution. As we can see none of the points of the numeric variables are close to the qqline, so none of them is normally distributed.

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.

Solution Structure

Hypothesis Test for a single continuous variable.In order to end up to a result about the test we have to do some other tests first.

First of all we have to check the normallity of the sample (salbeg).Because our sample is large enough (n>50) we check the normallity with two test(Shapiro-Wilk test , Kolmogorov – Smirnov)

```
> lillie.test(beginning_salary)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  beginning_salary
D = 0.25188, p-value < 2.2e-16

> shapiro.test(beginning_salary)

        Shapiro-Wilk normality test

data:  beginning_salary
W = 0.71535, p-value < 2.2e-16
```

As we can see the p-value of the two tests is < a=0.05 so we reject normallity .After that, we proceed checking the symmetry of the sample. This task will be done with the help of the symmetry function.

```
> symmetry.test(beginning_salary)

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

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

Again the p-value < α=0.05 so we reject symmetry in our sample. Finally,we are ready to perfom the **Wilcoxon Test for One Sample.**

The test we want to check is the below :

Ho: beginning salary =1000

H1: beginning salary ≠1000

```
> wilcox.test(beginning_salary,mu=1000)

        Wilcoxon signed rank test with continuity correction

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

The p-value < a=0.05. So we reject the Ho Hypothesis at a significance level of 0.05.

We can assume that the beginning salary of a typical employee can not be considered equal to 1000 dollars.

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.

Solution Structure

For this question we have to build a new variable named “diff” which contains the difference between the Salary of the employee now and the salary of the employee at the beginning of his/her recruitment. The test below are based on this variable.

First, we must check the normality of the sample. Because of the large size of the sample we have to do two normality tests (Shapiro-Wilk test , Kolmogorov – Smirnov).

```
> diff <-salary$salnow-salary$salbeg
> head(diff)
[1]  7680 17400 11760 10500 10950 14254
> lillie.test(diff)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

> shapiro.test(diff)

        Shapiro-Wilk normality test

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

Both tests reject the normality so the variable “diff” isnt normally distributed. Secondly , we check the symmetry of the sample by computing the mean and the median of the variable “diff” . If these two metrics are extremly close from each other then there is symmetry in the “diff” variable .

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

The two metrics are extremely far from each other so there is no symmetry in our sample.

Finally, we proceed implementing the **Wilcoxon Test for Dependent Samples.**

The Hypothesis are:

Ho: diff=0

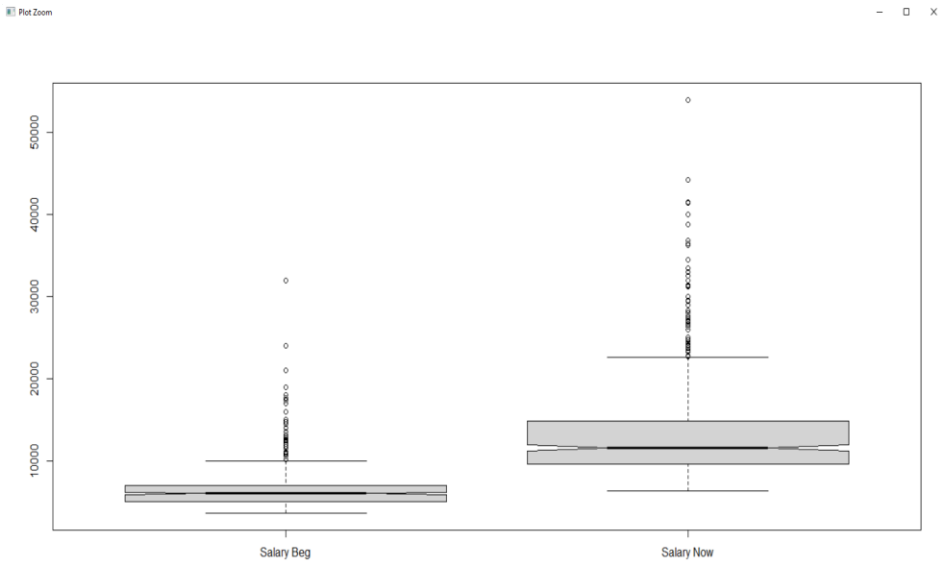
H1: diff≠0

```
> 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 we can see the p-value < α=0.05 so we reject the Ho hypothesis at a significant level a=0.05.

We can assume that there is a significant difference among the beginning and the present salary of the employees.

The above outcome we can confirm it also by looking at the boxplots of the two salaries.



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.

Solution Structure

In this case we have two independent samples. The first one is the Beginning Salary for men and the second one the Beginning Salary of women. In order to compute the hypothesis test for 2 independent samples we have to split the data frame in two samples as we mention above. This task will be executed by using the function “by”.

Firstly, we have to check the normality of the two samples. The size of the two samples separately is large enough so we do two normality tests (Shapiro-Wilk test , Kolmogorov – Smirnov).

```
> by(salary$salbeg,salary$sex,lillie.test) # normality not ok
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) # normality not ok
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
```

The pvalue in these two occasions is < α=0.05 so we reject normality at a significance level α=0.05.

Secondly, we check the symmetry for both samples with the “symmetry.test” function.

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

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

data:  dd[x, ]
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,symmetry.test)
salary$sex: MALES

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

data:  dd[x, ]
Test statistic = 13.829, p-value < 2.2e-16
alternative hypothesis: the distribution is asymmetric.
sample estimates:
bootstrap optimal m
                127

------------------------------------------------------------------------
```

The pvalue is < α=0.05 so we can't assume symmetry in both Females and Males Sample.

Finally, we proceed doing the **Test of Zero Difference between the medians.**

The hypothesis test is:

Ho: M1=M2

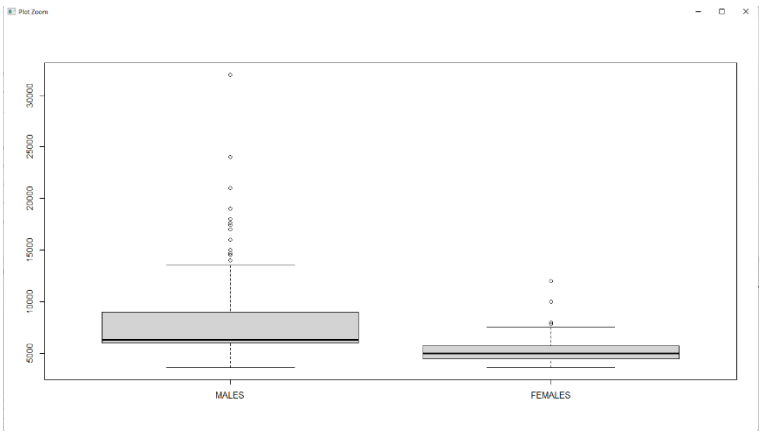
H1: M1≠M2

```
> wilcox.test(dataset1$MALES,dataset1$FEMALES)

        Wilcoxon rank sum test with continuity correction

data:  dataset1$MALES and dataset1$FEMALES
W = 47874, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

The pvalue is < 0.05 so we reject the Ho hypothesis at significance level α = 0.05.

A significance difference is found about the median between the beginning salary of men and the beginning salary of the women.

Also, we have to construct a boxplot for each level in order to verify 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.

Solution Structure

In this Question first we must split the “age” variable into three categories. This can be done very easily with the use of the “cut2” function (library: Hmisc). To implement this task, we have to do analysis of variance (ANOVA).

First, we split the variable and we assign it into the new data frame named “df” to have all the proper information gathered all together. We use the “head” function to print the first 6 rows of our updated data set.

```
   id salbeg   sex time   age salnow edlevel work          jobcat minority     sexrace     age_cut
1   1   8400 MALES   81 28.50  16080      16 0.25 COLLEGE TRAINEE    WHITE WHITE MALES [23.0,29.7)
6   6  12996 MALES   80 29.50  27250      18 2.42 COLLEGE TRAINEE    WHITE WHITE MALES [23.0,29.7)
7   7   6900 MALES   79 28.00  16080      15 3.17        CLERICAL    WHITE WHITE MALES [23.0,29.7)
8   8   5400 MALES   67 28.75  14100      15 0.50        CLERICAL    WHITE WHITE MALES [23.0,29.7)
9   9   5040 MALES   96 27.42  12420      15 1.17        CLERICAL    WHITE WHITE MALES [23.0,29.7)
24 24   8700 MALES   65 28.00  28000      16 1.58 COLLEGE TRAINEE    WHITE WHITE MALES [23.0,29.7)
> 
```

As we can see we have sorted the dataset based on the interval that the age belongs to.

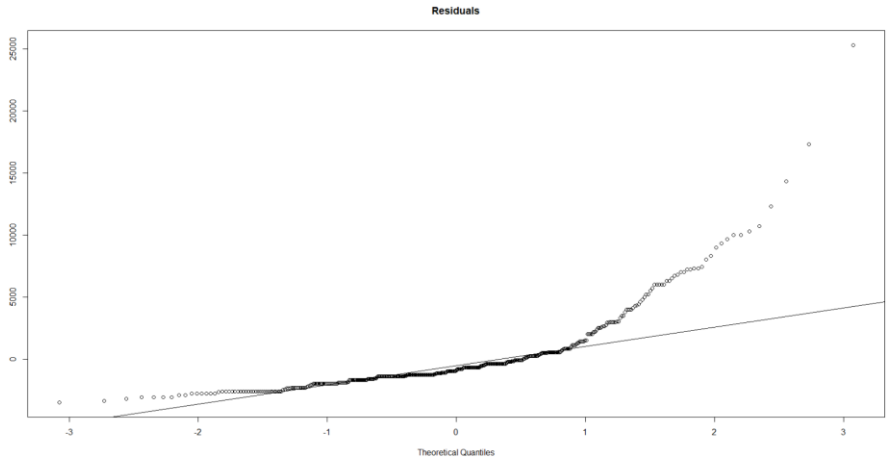
After this step we proceed with the Analysis of Variance

```
> anova1 <- aov(salbeg~age_cut,df)
> 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
> 
```

The anova test is being implementing in the table “df” which contains the age_cut variable.

Second, we must check the normality of the test’s residuals.

In order to see if the sample is normally distributed or not we create the qqnorm diagram for the residuals of the anova test , and we draw the qqline.Doing that we can observe that the sample isnt normally distributed as the points of the qqnorm are not close to the qqline.



As a result we can not assume normality in our sample.Continue, with the test of the sample's symmetry

```
> groups <-aggregate( salbeg~age_cut, df, mean)
> groups
     age_cut   salbeg
1 [23.0,29.7) 5767.788
2 [29.7,39.8) 7997.795
3 [39.8,64.5] 6681.949
```

```
> groups1 <- aggregate(salbeg~age_cut,df,median)
> groups1
     age_cut salbeg
1 [23.0,29.7)   5370
2 [29.7,39.8)   6600
3 [39.8,64.5]   5700
```

There is a significant difference between mean and median in every group so we cant assume symmetry in our sample

Finally, we proceed with the **Test of equality of Medians**.

The Hypothesis Test in this case after all the proper steps above is :

Ho: M1=M2=M3

H1: M1≠M2≠M3

Where M1,M2,M3 are the medians of each age_group.

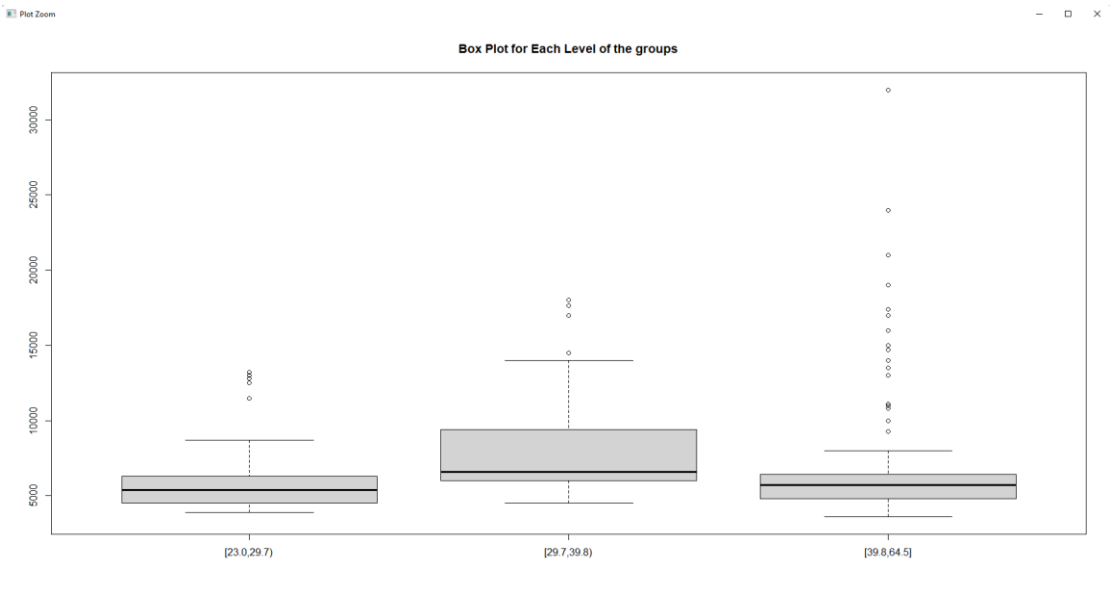
```
> kruskal.test(df$salbeg,df$age_cut) # reject h0

        Kruskal-Wallis rank sum test

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

The pvalue <0.05 so we reject the Ho Hypothesis in a significance level α=0.05.

We assume that on average the begginning salary isnt the same for all the age groups.

In order to identify the groups that differ we have to make pairwise comparisons and build the boxplot for each level of the age_cut group.

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

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  df$salbeg and df$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 group 2 ([29.7,39.8]) differs from the groups 1 ([23.0,29.7]) and 3 ([39.8,64.5]).

Question 8

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.

Solution Structure

First of all, we have to compute the proportion of the **white** males and the **white** females separately.

```
> proportions <- c(white_males_proportion,white_females_proportion)
> names(proportions) <- c('White Males','White Females')
> proportions
  White Males White Females
    0.5243243     0.4756757
```

After that, for our convince we import the two proportions to a new variable in order to have them together in one vector. After that, we compute the size of the **white** males and the white **females,** and we import them too in a vector.

```
> length(sum_of_white_females)
[1] 176
> length(sum_of_white_males)
[1] 194
>
```

```
> proportions
  White Males White Females
    0.5243243     0.4756757
> size <- c(194,176)
> size
[1] 194 176
>
```

In a new variable we compute the total counts by multiplying the size with the proportions of each group (white males and white females).

```
> final_counts <- proportions*size
> final_counts
  White Males White Females
    101.71892      83.71892
```

Finally, we can proceed to the Hypothesis Test which is **Testing for the equality of proportions.**

The test Hypothesis is:

Ho: p1=p2

H1: p1≠p2

Where the p1: proportion of white males and p2: proportion of white females.

```
> prop.test(final_counts,size) #reject h0

        2-sample test for equality of proportions with continuity correction

data:  final_counts out of size
X-squared = 0.68985, df = 1, p-value = 0.4062
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.05866326  0.15596055
sample estimates:
   prop 1    prop 2
0.5243243 0.4756757
```

The pvalue of the test is > 0.05 so we don't reject Ho at a significance level a=0.05.

We don't have enough evidence to assume that the proportion of white male employees is equal to the proportion of white female employees.