Department: Department of Management Science and Technology

M.Sc.: Business Analytics (Full-Time)

Course: Statistics for Business Analytics I

Professor: Ioannis Ntzoufras

Professor Assistant: Argyro Damoulaki

Deliverable: R Labs (Second Part) – Assignment #1 (Hypothesis Testing)

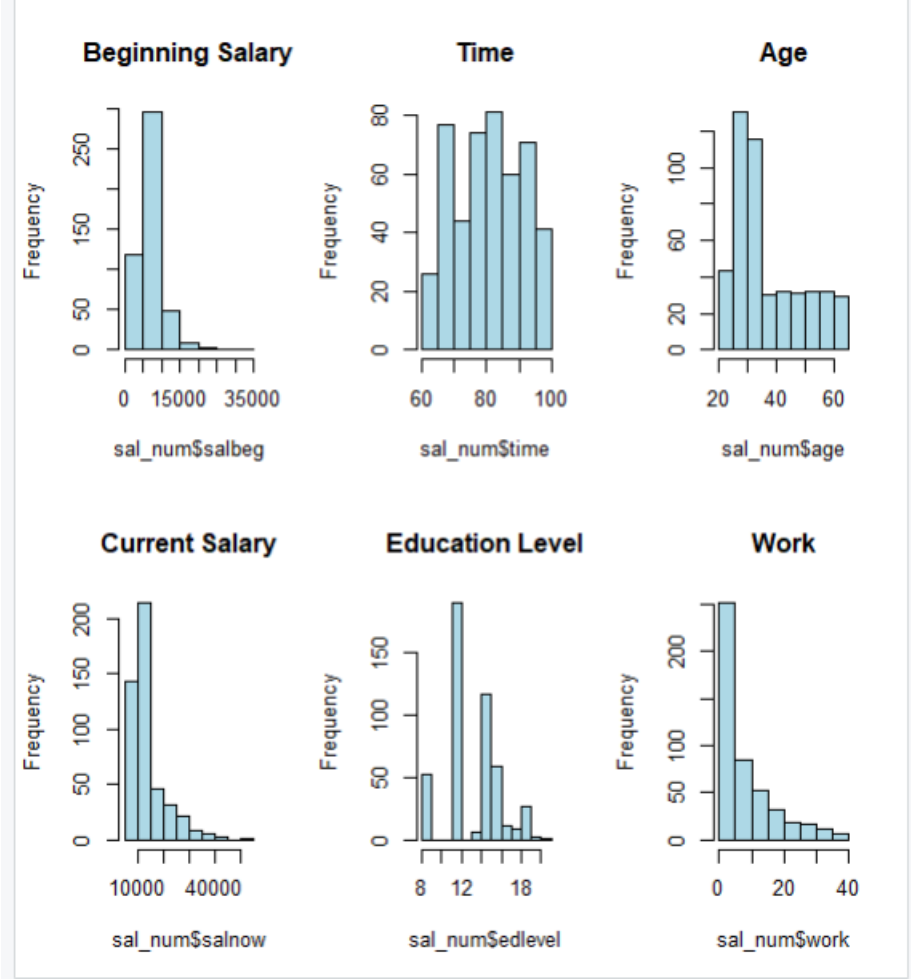
Student Name: Evangelos Lakkas-Pyknis

Registration Number: f2822306

E-mail: eva.lakkaspyknis@aueb.gr

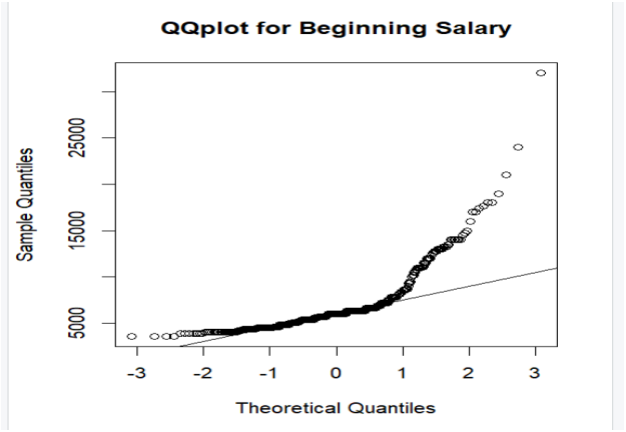
Task 2

None of the numeric variables seem to follow the normal distribution. None of them is symmetric at the mean and has “bell-like” shape, also their skewness and their kurtosis are not similar to those of the normal distribution (this is displayed by the histograms given below but I have also calculated these descriptives in my R code). Finally if we apply the Lillie and Shapiro-Wilks test for these variables in R we get p-values< 0.05 for each of them.



Task 3

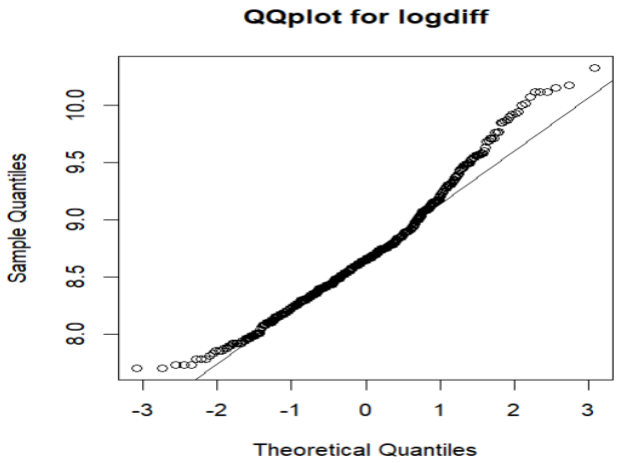
Firstly I examined if the data for beginning salary (SalBeg) come from a normal distribution. By applying the Liliefors and the Shapiro-Wilk tests I see that the p-value is very small(p-value<2.2e-16) so the null hypothesis that beginning salary is normally distributed. Also if we see at the QQplot for beginning salary we see that our data is not close to the QQline.



Given that beginning salary doesn't follow the normal distribution I examine if the sample is large enough, the observations for SalBeg are 474(>50) so I consider the sample big. Then I see if the mean is sufficient measure for the central location which isn't true for this case (mean = 6806, median=6000). So given all of the above, I use the (non-parametric) Wilcoxon test to see if the median of the beginning salary is equal to 1000. This null hypothesis is rejected because p-value is smaller than 2.2e-16 at confidence level of 95%.

Task 4

Given that I want to test a hypothesis for one quintative variable(one sample) I will follow the same process as described in Task 3. The null hypothesis that logdiff variable (the natural logarithm of the difference between salbeg and salnow) comes from a population normally distributed is rejected from the Liliefors and the Shapiro-Wilk tests. Also if we see at the QQplot we observe that is heavy on the tails and it declines from the normal distribution.

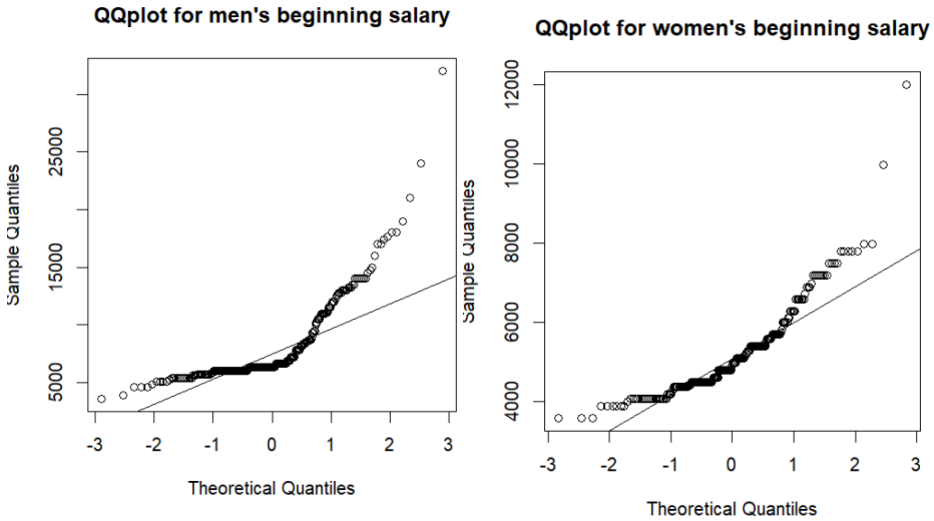


But because my sample is big(>50) and the mean of logdiff is sufficient measure for the central location(median and mean are almost equal), I can use the t.test for the null hypothesis which is that that there is significant difference between the beginning and the current salary(H0: logdiff is on average equal to one). The H0 is rejected(p-value < 2.2e-16).

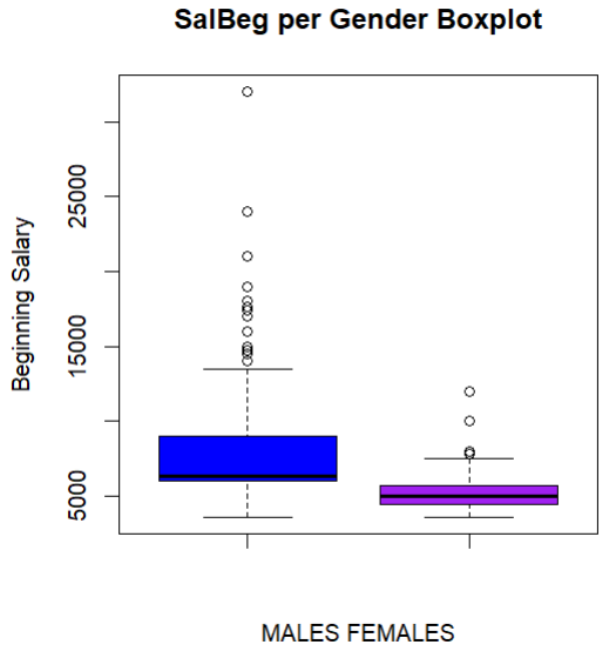
Task 5

I consider that the 2 samples for the women and men employees are independent because the values of the beginning salary variable are not in pairs and the two samples do not share some obvious characteristic. Both samples can be considered big (>50) but they do not come from a population that is normally distributed. Both Lillie and Shapiro-Wilks tests give p-values<a for both samples so the null hypothesis of normality is rejected. And as we see

below the QQnorms are not aligned with the QQline so the samples don't come from a normally distributed population. Then by comparing the values of the median and mean SalBeg for each gender I see that they have a significant difference and thus is not a sufficient measure for the central location.

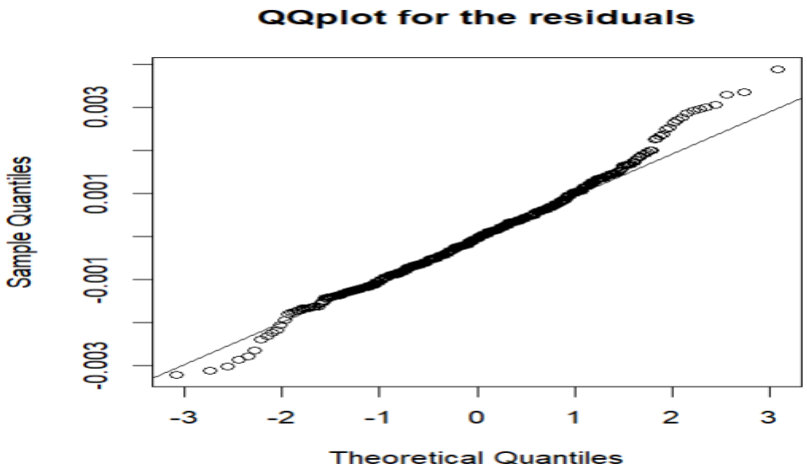


Given all that, I am implementing the Wilcoxon test and reject the null hypothesis (H0: there is not a significant difference for the beginning salary between men and women) because I get a p-value smaller than 0.05. This significant differences are also shown in the following boxplot.



Task 6

Considering that now I have to do a test between one quintative variable and one categorical value which has 3 groups with independent samples, the steps that I will follow in order to test the null hypothesis: the relative salary rise (relSal) is the same for all age groups, will differ from tasks 5 and 6. Firstly I examine the normality for the residuals that I got from anova. Lilie test doesn't reject the hypothesis of normality (p-value=0.79>0.05), Shapiro-Wilks test doesn't reject it. Also if we see the QQplot of the residuals the tails decline from the normal distribution. So I cannot assume normality for the residuals.



All 3 samples have more than 50 observations and mean is a sufficient measure for the central location for all groups because the means of each sample have a reasonable difference from the medians. Then I do the levene test (I used levene because it is more credible than barlett for not normally distributed data) to see if the samples have equal variances (homoscedasticity). The null hypothesis for homoscedasticity is not rejected (p-value= 0.36>0.05). I tested the null hypothesis for the equality of the means of RelSal variable between the different age groups, the H0 is rejected, p-value (from anova F-test)<2e-16<0.05. Finally, I made pairwise comparisons using pairwise t-test to identify the groups that differ.

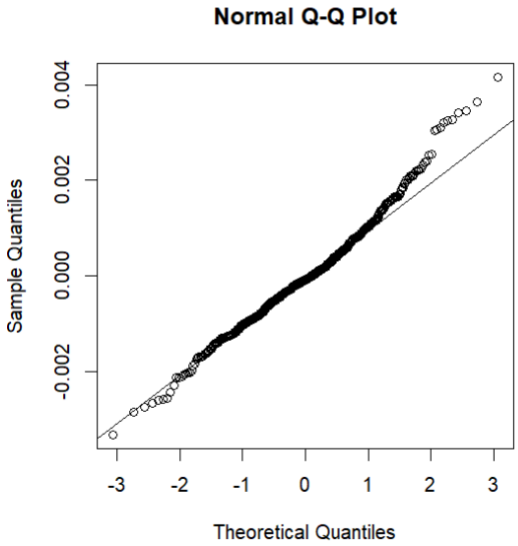
```
> pairwise.t.test(Salary_DF$relSal, Salary_DF$age_cut, p.adjust.method = p.adjust.methods[1])##There are significant differences

        Pairwise comparisons using t tests with pooled SD

data:  Salary_DF$relSal and Salary_DF$age_cut

           [23.0,29.7) [29.7,39.8)
[29.7,39.8) 3.3e-07     -
[39.8,64.5] < 2e-16     1.8e-07

P value adjustment method: holm
```

From the output I got I see that there are significant differences for all groups since all pairs have values<0.05. I can also see that these differences from the following error-bar.

Task 7

I want to test if the null hypothesis that the relative salary rise (relSal) is the same for all job categories. I examine the normality for the residuals for the residuals I got from anova, both Lilie and SW tests reject the null hypothesis of normality because the p-values are smaller

than 0.05, also the sample size for 4 out of 7 groups is small(<50). The QQplot has heavy tails on the right side and it declines from the normal distribution.



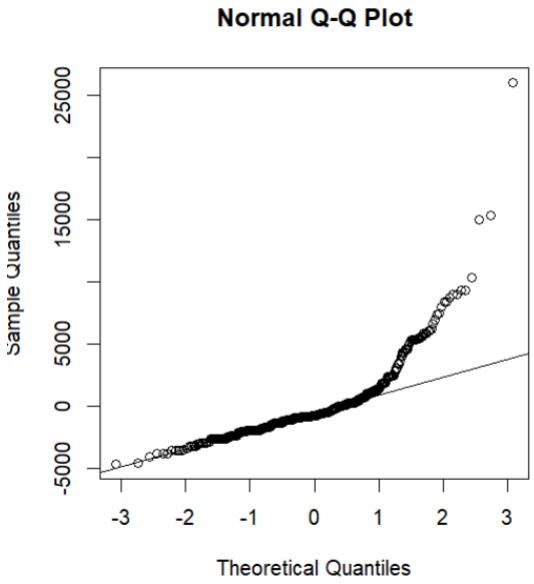
I use the Kruskall-Wallis non-parametric test to examine the H0 which is rejected (p-value=1.968e-11<0.05). Then I use the pairwise wilcox test (because I have small samples for some groups) to make pairwise comparisons. The output I get is the following:

```
> pairwise.wilcox.test(Salary_DF$relSal, Salary_DF$jobcat)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  Salary_DF$relSal and Salary_DF$jobcat

                  CLERICAL OFFICE TRAINEE SECURITY OFFICER COLLEGE TRAINEE EXEMPT EMPLOYEE MBA TRAINEE
OFFICE TRAINEE    1.5e-06  -              -                -               -               -
SECURITY OFFICER  0.022    1.000          -                -               -               -
COLLEGE TRAINEE   4.0e-07  0.110          0.298            -               -               -
EXEMPT EMPLOYEE   1.000    0.276          0.322            0.014           -               -
MBA TRAINEE       0.054    0.322          0.322            1.000           0.260           -
TECHNICAL         1.000    0.470          1.000            0.243           1.000           0.322
```

The pairs with significant differences are those with values <0.05(clerical-off.trainee, clerical-sec.officer, col.Trainee – ex.employee)

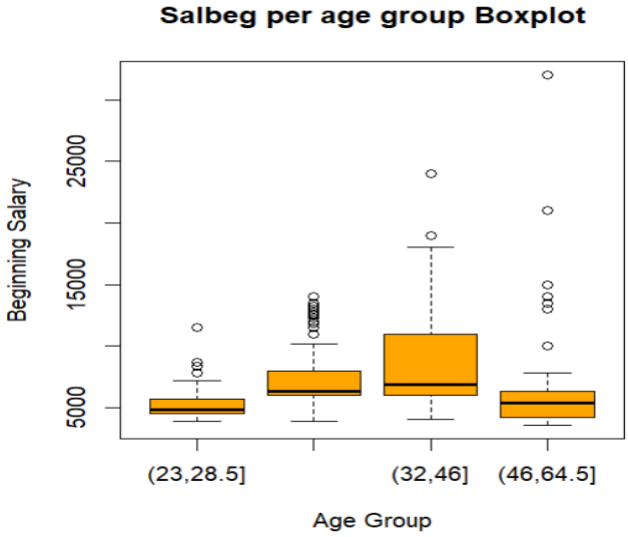
Task 8

I have to test if the null hypothesis that the beginning salary (salbeg) is the same for all 4 age groups. I examine if the residuals I got from anova between the beginning salary and the age groups are normally distributed, Lilie and SW tests reject the null hypothesis. Also the QQplot has heavy tails on the right. All four samples are big (n1,n2,n3 and n4>50).



Then I compare the means and the medians of Salbeg for each group and I see that mean is not a sufficient measure for the central location, because each mean has a significant difference from the median. Considering the above I used the Kruskall-Wallis non-parametric test to examine the H0: the medians of SalBeg for each group are equal on average. Null hypothesis is rejected (p-value < 2.2e-16< a=0.05). I used the pairwise Wilcoxon test to specify the pairs with significant differences. The output of the test shows me that there are significant differences between all pairs (because their values<0.05) except two pairs: (32,46]- (28.5,32] and (46,64.5] - (32,46].

```
> pairwise.wilcox.test(Salary_DF$salbeg,Salary_DF$age_cut2)

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  Salary_DF$salbeg and Salary_DF$age_cut2

          (23,28.5] (28.5,32] (32,46]
(28.5,32] < 2e-16   -         -
(32,46]   < 2e-16   0.11      -
(46,64.5] 0.15      4.3e-12   1.2e-13

P value adjustment method: holm
```



That can be also showed from the following boxplot where there are significant differences between the values of the medians (bold black lines) of all pairs of age groups except those mentioned before.

Task 9

I want to test the null hypothesis: the proportion of white male employees is equal to the proportion of white female employees. Now I have two categorical variables and the expected values are larger than 5, so I am using prop.test and chi-square test (with yates correction), I get p-value=0.12 (>0.5) from both tests so the null hypothesis is not rejected.

Task 10

I want to test the null hypothesis: the proportions of minority among the job categories are equal. Now I have two categorical variables and the expected values are smaller than 5, so I am using Monte Carlo and fisher test and I reject the null hypothesis (because p-value=0.0004<0.05) so there is a significant difference between the proportions of minority among the job categories.