National Technical
University of Athens

HOMEWORK 1

Kreativstorm Data Analysis Course

Full Name: Antonios Mitsis

Email: anmitsis@hotmail.com

Group: B

In this homework every question will be answered with the help of R. Each line of code that will be used will be provided in blue color.

Q1. What is the mean birth weight for babies of non-smoking mothers?

A1. The mean birth weight for babied of non-smoking mothers is 3.5095kg. To get this result first we import the dataset in R using the import dataset option and save it in the environment as df. Then using the function mean() we get the mean of the birthweight for babies and selecting only non-smoking mothers with the brackets []. A mother is considered a smoker if the value of the variable smoker is 1 and non-smoker for a value of 0.

```
df<-Birthweight_reduced_kg_SPSS
mean(df$Birthweight[df$smoker==0])
```

Q2. What is the mean birth weight for babies of smoking mothers?

A2. The mean birth weight for babied of smoking mothers is 3.134091kg.

```
mean(df$Birthweight[df$smoker==1])
```

Q3. What is the mean head circumference for babies of non-smoking mothers?

A3. The mean head circumference for babies of non-smoking mothers is 35.05cm.

```
mean(df$Headcirc[df$smoker==0])
```

Q4. What is the mean gestational age at birth for babies of smoking mothers?

A4. The mean gestational age at birth for babies of smoking mothers is 38.95455 weeks.

```
mean(df$Gestation[df$smoker==1])
```

Q5. What is the maximum head circumference for babies of non-smoking mothers?

A5. The maximum head circumference for babies of non-smoking mothers is 39cm. To find the maximum of a variable we use the function max().

```
max(df$Headcirc[df$smoker==0])
```

Q6. What is the minimum gestational age at birth for babies of smoking mothers?

A6. The minimum gestational age at birth for babies of smoking mothers is 33 weeks.

```
min(df$Gestation[df$smoker==1])
```

Q7. Based on the dataset you have, out of the two, which one would be a better bet:

- Pregnancy period in smoking mothers is shorter
- Pregnancy period in non-smoking mothers is shorter

A7. The better bet would be the first sentence, which is that pregnancy period in smoking mothers is shorter.

Q8. Justify the above choice in a few words.

A8. To find which sentence is a better bet we need to compare the means of the two populations. However we need to assure first that they follow the normal distribution so that our choice is reliable. Using the Shapiro Wilk test we get a p-value of 0.2906 for non-smokers and 0.1283 for smokers, which means that both populations follow the normal distribution. We continue comparing the means and other values using the summary() function and smokers have smaller mean and median. The only issue is that the maximum value is for smokers, but this alone is not enough evidence to change our choice.

Q9. What is the baby birth weight range for babies of smoking mothers?

A9. The baby birth weight range for babies of smoking mothers is [1.92, 4.57].

```
range(df$Birthweight[df$smoker==1])
```

Q10. In your own words describe what the value of the above range for baby's birthweight tells us about smoking versus non-smoking mothers?

A10. After getting the range of non-smoking mothers, which is [2.65, 4.55] we notice that the baby birth weight range for babies of smoking mothers is greater than the range for non-smoking mothers. This means that there is more fluctuation in the baby weight for smoking mothers and therefore we could assume that smoking induces instability in the birth weight of babies. This fluctuation tends to be on the lower side which means that smoking can cause the baby to be born underweight.

Q11. Are head circumference data for babies of smoking mothers normally distributed?

A11. To check whether the head circumference data for babies of smoking mothers are normally distributed we will use the Shapiro Wilk test. The null hypothesis is that the data are normally distributed and the alternative is that they are not.

```
shapiro.test(df$Headcirc[df$smoker==1])
```

```
Shapiro-Wilk normality test

data:  df$Headcirc[df$smoker == 1]
W = 0.95365, p-value = 0.3724
```

Since the p-value is 0.3724 we do not have enough evidence to reject the null hypothesis, therefore our data follows the normal distribution.

Q12. What is the significance value for the above on the Shapiro-Wilk test?

A12. The significance value for the above test is 0.3724 which is greater than 0.05, the most common significance level.

Q13. What is the standard score (Z-score) for head circumference of 35.05 (X=35.05) in non-smoking mothers?

A13. To find the standard score we will subtract from X the mean head circumference in non-smoking mothers and then divide by the standard deviation.

```
(35.05-mean(df$Headcirc[df$smoker==0]))/sd(df$Headcirc[df$smoker==0])
```

The standard score is 0, since the mean head circumference in non-smoking mothers is 35.05, same as X, as we found in Q3.

Q14. How are birth weight data of non-smoking mothers skewed?

A14. To find the skewness of the data we will use the library moments. First we download the package and load it with the function library(). This library contains the function skewness() which gives us the skewness of our data.

```
library(moments)
skewness(df$Birthweight[df$smoker==0])
```

The skewness is positive with a value of 0.3333708.

Q15. Are birth weight data for babies of smoking mothers normally distributed?

A15. The birth weight data for babies of smoking mothers is normally distributed. We perform the Shapiro Wilk test again with the same null hypothesis and alternative hypothesis.

```
shapiro.test(df$Birthweight[df$smoker==1])
```

Q16. What is the significance value for the above on the Shapiro-Wilk test?

A16. The significance value for the above test is 0.9495, which is quite big and the normality of our data is almost certain.

Q17. Based on the dataset you have, how confident can you be in saying that a baby's birth weight will be +/- 1 standard deviation from the mean?

A17. Since the data of the babies birthweights are normally distributed we know that a baby's birthweight will be ± 1 standard deviation with a 68.27% chance. The answer was based on the image in page 12 of the slides.

Q18. Based on the dataset you have, what is the probability that the birth weight for a baby of a smoking mother will be less than 4.2 kg?

A18. To find this probability first we need to find the standard score. Then we will use the function pnrom() in R, or we could look it up in the z-score table.

```
z1<-(4.2-mean(df$Birthweight[df$smoker==1]))/sd(df$Birthweight[df$smoker==1])
pnorm(z1)
```

The probability is 0.9543497.

Q19. Are data for length of baby of non-smoking mothers normally distributed?

A19. We will perform the Shapiro Wilk test to check for normality.

```
shapiro.test(df$Length[df$smoker==0])
```

```
Shapiro-Wilk normality test

data:  df$Length[df$smoker == 0]
W = 0.91225, p-value = 0.07037
```

The p-value is close to 0.05, in a 5% significance level we would accept the normality hypothesis, however in a 10% significance level we would reject it.

Q20. What is the significance value for the above on the Shapiro-Wilk test?

A20. The significance value or p-value of the above test is 0.07037.

Q21. What is the standard score for the length of a baby of 48.5cm for non-smoking mothers?

A21. The standard score for the length of a baby of 48.5cm for non-smoking mothers is -1.014091

```
(48.5-mean(df$Length[df$smoker==0]))/sd(df$Length[df$smoker==0])
```

Q22. Based on the dataset you have, what is the probability that the length of baby for non-smoking mothers will be more than 55 cm?

A22. First we will calculate the standard score and then find the probability.

```
z2<-(55-mean(df$Length[df$smoker==0]))/sd(df$Length[df$smoker==0])
1-pnorm(z2)
```

Since the probability we are looking for is for the length to be more than 55cm, we will calculate the probability of it being less than 55cm and subtract is from 1. The probability we are looking for is 0.162715.