National Technical
University of Athens

HOMEWORK 2

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 and SPSS. Each line of code that will be used in R will be provided in blue color.

Q1. What is the mean father's age?

A1. The mean father's age is 28.90476. To get this result first we import our dataset in R and save it as df. Next we use the function mean() to extract the value that we need. To access a variable in the data frame df we use the $ operator.

```
mean(df$fage)
```

Q2. What is the mean father's age for low birthweight babies?

A2. The mean father's age for low birthweight babies is 24.83333.

```
mean(df$fage[df$lowbwt==1])
```

Q3. Is the father's age normally distributed? Justify your answer.

A3. The father's age is not normally distributed. We perform a Shapiro Wilk test to check for normality and the significance value of the test is 0.0385 which is smaller than 0.05, this means that in a significance level of 5% our data is not normally distributed.

```
shapiro.test(df$fage)
```

Q4. If you apply the log transformation to the father's age, what is the mean score of the transformed variable?

A4. The mean score of the log transformation of the father's age is 1.449259. To apply the log transform first we check for skewness with the skewness() function in the moments library. The skewness is positive so we apply the log with base 10 transformation in the variable.

```
skewness(df$fage)
fage_log<-log(df$fage,base = 10)
mean(fage_log)
```

Q5. Is the above mean score a good representation of the real value? Justify your answer.

A5. The above mean score is not a good representation of the real value by itself. However with just a simple transformation we can extract some useful meaning. We need to raise 10 to the power of the mean score to reverse the changes of the log transformation. After doing this we get the value 28.13578 while the real value is 28.90476 so it is a good representation of the real value.

Q6. Is the new variable (log transform of father's age) normally distributed? Justify your answer.

A6. Yes the new variable is normally distributed as it has a significance value of 0.1287 in the Shapiro Wilk test.

```
shapiro.test(fage_log)
```

Q7. Is the variable "years father was in education" normally distributed?
A7. The variable years father was in education is not normally distributed as the significance value of the Shapiro Wilk test is 4.484e-05, which is extremely small.
shapiro.test(df$fedyrs)

Q8. Mentioning the null and alternative hypotheses, explain the above answer.
A8. The null hypothesis in the Shapiro Wilk test is that our data is normally distributed, while the alternative hypothesis is that our data is not normally distributed. We accept the null hypothesis unless we have sufficient evidence to reject it. To make a decision whether we will accept it we look at the significance value and compare it to 0.05 usually. If it is greater we can accept the null hypothesis but if the significance value is smaller we have enough evidence to reject it. In the above answer the significance value was smaller than 0.05 and thus we reject normality hypothesis.

Q9. What is the mean score for the variable "years father was in education" after you apply the Box-Cox transformation?
A9. The mean score for the variable "years father was in education" after you apply the Box-Cox transformation is 13.7141. To apply the Box-Cox transformation we switch to SPSS and select:

- Transform -> Rank cases (enter variable) -> Rank Type -> Check "Fractional rank"
- Transform -> Compute variable
- Function Group -> Inverse DF
- Function and special variables -> Idf.Normal
- In Target variable -> name of transformed variable
- In numeric expression -> (fractional rank variable, mean, st. deviation)

After we create the new variable we select analyze -> Descriptive Statistics -> explore to get the mean score among other values.

Q10. Is this new variable normally distributed? Explain.
A10. The new variable is not normally distributed. In the explore menu we can select plots and then normality plot with tests to perform a Shapiro wilk test. The significance value of the test is less than 0.01 so we reject the null hypothesis. This is because the new and old variable only have 4 distinct values which are too few to give us evidence that they are normally distributed.

Q11. What is the mean score for this new variable (B-C transformed fathers' years in education) for mothers aged under 35?
A11. The mean score for the new variable is 13.5503 for mothers aged under 35. To get this value we select cases in the data menu and make the condition of mother's age variable to be under 35.

Q12. Which test would you use to investigate the relationship between birth weight and Father's age?

- Pearson product-moment correlation
- Spearman's Rank order correlation
- Point-Biserial correlation
- Phi-Coefficient

A12. Spearman's Rank order correlation

Q13. Justify the above choice in terms of the distribution of data and the nature of the test.

A13. We would use the Spearman's Rank order correlation to investigate the relationship between birth weight and father's age. We cannot use Pearsons test as it requires normally distributed data and father's age is not. Also the last two tests need at least one dichotomous variable and we do not have any. So the Spearman's test is the only one that can provide reliable results.

Q14. What is the direction of that relationship?

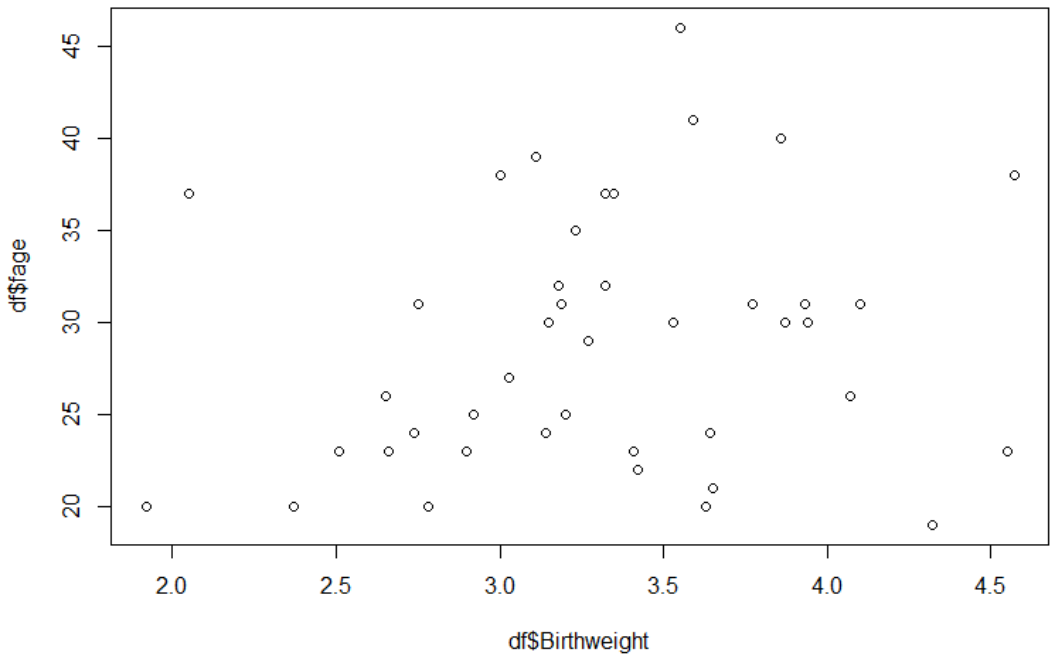
A14. First we perform the spearman's correlation test in R and we get a value of 0.1781063. However the significance value of the test is 0.2591, therefore we accept the null hypothesis that says that the correlation coefficient is 0. The direction of the relationship is very faint but can be defined as positive.

```
cor.test(df$Birthweight,df$fage,method = "spearman")
```

Q15. What is the form of that relationship?

A15. The form of the relationship is non-linear. We can examine the plot of the two variables to tell the form of the relationship.

```
plot(df$Birthweight,df$fage)
```



Q16. What is the degree of that relationship?

A16. The degree of the relationship is weak.

Q17. What test would you use to investigate the relationship between smoking and birth weight?

- Pearson product-moment correlation
- Spearman's Rank order correlation
- Point-Biserial correlation
- Phi-Coefficient

A17. We would use the Point-Biserial correlation as we have one dichotomous variable (smoking) and one continuous (Birthweight).

Q18. Report on the above results including information about direction/form/degree of the relationship.

A18. To perform the Point-Biserial correlation test we will use R and as we already have on dichotomous and one continuous variable we just perform a Pearson's correlation test. The correlation coefficient is -0.3142339 which means that direction is negative, the form is linear(we can fit a regression model to check it) and the degree is weak.

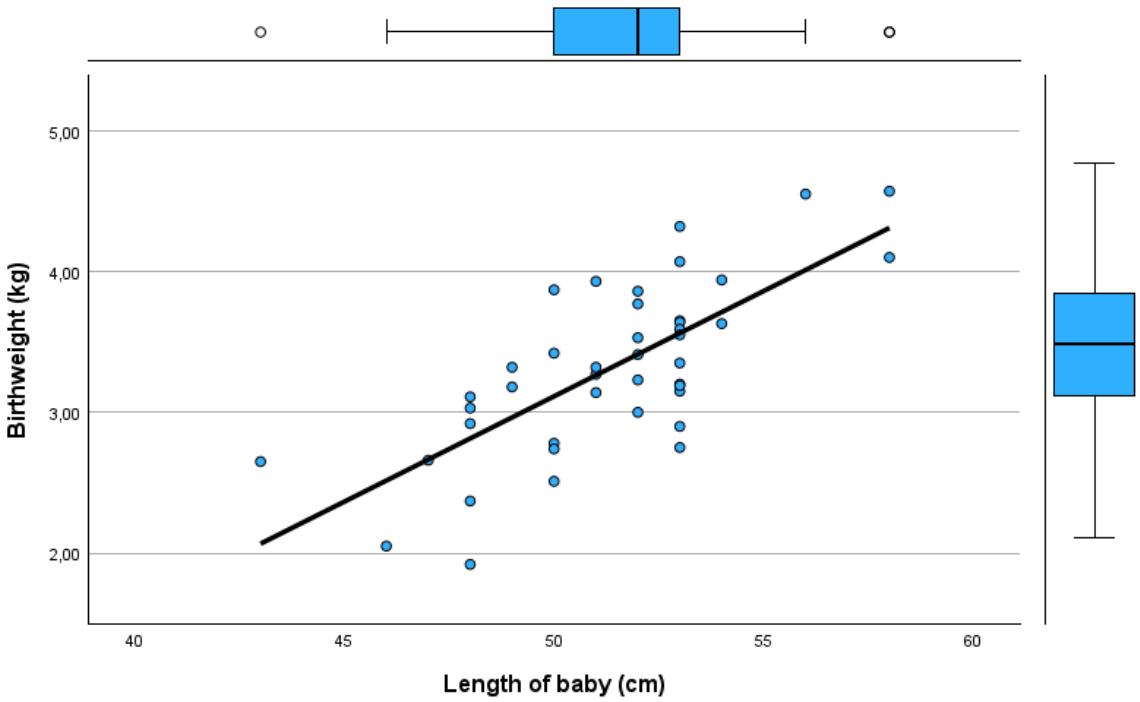
Q19. If you wanted to see the effect of the length of a baby on birthweight, what would your independent variable be?

- Length of baby
- Birthweight

A19. If I wanted to see the effect of the length of the baby on birthweight the independent variable would be the length of the baby.

Q20. In statistics, when creating a scatterplot, it is a common practice to put the independent variable on the x-axis and the dependent variable on the y-axis. With this in mind, create a scatterplot for the above case and provide the regression line. For homework submitted using MS Word, insert a picture of the scatterplot.

A20. We select Graphs -> Regression Variable Plots and for the horizontal axis we select the independent variable and for the vertical axis the dependent. On options we check the linear line and we get the following scatterplot with the regression line. It also contains the bar plots for each variable on the side.



Q21. Is the relationship between the length of baby and birthweight linear?

- Yes
- No

A21. Yes the relationship between the length of the baby and birthweight is linear.

Q22. Justify the above choice.

A22. The scatterplot is one way to check for a linear relationship. A more advanced method is to fit a regression model and perform and F-Test to see if the regression is statistically important. The null hypothesis is that the coefficients of the regression are 0 and the alternative is that the coefficients are not 0. Therefore if we reject the null hypothesis then we can be sure that the relationship between the two variables is linear. To fit a regression model in R we use the following commands.

```
results<-lm(df$Birthweight~df$Length)
summary(results)
```

The significance value of the F-Test is 5.029e-08 which is very small and we reject the null hypothesis.

Q23. Is there any evidence to suggest that the birth weight, length of baby, and head circumference are related? - Q24. Justify the above choice.

A23-A24. To check if these three variables are related we can create the correlation matrix in SPSS by selecting Analyze -> Correlate -> Bivariate and select the three variables.

Correlations

		Birthweight (kg)	Length of baby (cm)	Head circumference (cm)
Birthweight (kg)	Pearson Correlation	1	,727**	,685**
	Sig. (2-tailed)		<,001	<,001
	N	42	42	42
Length of baby (cm)	Pearson Correlation	,727**	1	,563**
	Sig. (2-tailed)	<,001		<,001
	N	42	42	42
Head circumference (cm)	Pearson Correlation	,685**	,563**	1
	Sig. (2-tailed)	<,001	<,001	
	N	42	42	42

**. Correlation is significant at the 0.01 level (2-tailed).

The highlighted values are the significance of the correlation coefficients and we can see that all variables have strong correlation to each other, therefore we can say that they are related.

Q25. Describe the above relationship in your own words and provide evidence for your claims.

A25. The birthweight of a baby has a strong positive correlation with its length, which is totally normal since the larger a baby is the heavier it will be. The same applies to birthweight and head circumference as the larger the head of a baby is the heavier it will be. All in all, these variables are all related because they have to do with the size of the baby. The above correlation table is enough evidence for the claim.