Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Final Exam
SA.01

Tuesday, 25. 6. 2019

Duration: 90 Minutes

Surname, Given Name: _____

IDS Nr.: (On the back of your HSLU-Card) L _____

Signature: _____

Problem	1	2	3	4	5	Total
max. Points	7	17	16	6	6	52
achieved Points						

Important Information

- Allowed aids: Open book
- The use of mobile phones is not allowed. Switch off your phones.
- Place your HSLU-Card in front of you on your table.
- All of your answers should come with explanations! Solutions without understandable justifications obtain no credit.

Good luck!
Peter Büchel

Tuesday, 25. 6. 2019

Problem 1: ..... (7 Points)

The result of a public-opinion poll for a presidential election in three provinces (A, B and C) are as follow: In province A the percentage of voters supporting the Liberal candidate is 50 %. In province B the percentage of voters supporting the Liberal candidate is 60 %. In province C the percentage of voters not supporting the Liberal candidate is 65 %.

The population of the three provinces is distributed as follows: the population of A is 40 % of the total population in A, B and C, 25 % of the total population live in B and the remaining 35 % live in C.

Let us randomly choose a supporter of the Liberal candidate in province A or B or C. Determine the probability that such a voter was chosen from province B.

Problem 2: ..... (17 Points)

The file `nyc-marathon.csv` contains data of 276 persons who participated in the New York marathon in 2010. Included are the gender (`Gender`), the age (`Age`; in years) and the running time (`Minutes`; in minutes).

It was reported that the average age of the participants in the 2005 marathon was 40.5 for men and 36.1 for women respectively.

Hint: The command for loading the dataset is in the `R`-file.

(a) Execute the `summary`-command for this dataset and interpret the output for `Gender` and `Age`. *[2]*

(b) Generate a boxplot for `Gender` as predictor and `Age` as response variable and interpret the plot. *[2]*

(c) We form a new dataset which contains only the data of the women. This can be done with the command (on the stick) *[6]*

```
female <- nyc[nyc[, "Gender"] == "female", ]
```

where `nyc` is the name of the original dataset.

Use a *t*-test to verify the hypothesis that the average age of the women in the 2010 marathon correspond to the average age 36.1 in the 2005 marathon.

i) Do you choose a one- or two-sided test? Justify your answer.

ii) Give the null hypothesis and the alternative hypothesis.

iii) Perform the test and give the test decision on a significance level of 5 %.

iv) Determine the 95 % confidence interval, interpret this interval and give the test decision according to this interval.

(To be continued on the next double page)

(d) Which condition must be met in order to use a t-test in (c)? Do you have an alternative to the t-test? What is the test decision in this case? *[2]*

(e) The data from the 2005 suggest that the participating men were on average older than the women. Use a t-test with the data from 2010 to verify this hypothesis. *[5]*

i) Do you apply a one-sided or two-sided test? Justify your answer.

ii) Do you choose a paired or unpaired test? Justify your answer.

iii) Give the null hypothesis and the alternative hypothesis.

iv) Perform the test and give the test decision on a significance level of 5 %.
Hint: Create a new dataset `male` as in (c).

Problem 3: ..... (16 Points)

The file **led.csv** contains information about the life expectancy of humans depending on 22 factors (country, infant mortality , population of the country, etc). The data was collected for each year between 2000 and 2015. We only consider a few variables.

Hint: The command for loading the dataset is in the **R**-file.

(a) What is the average life expectancy (**Lifeexpectancy**)? *[1]*

(b) Produce a scatter plot for **Schooling** (number of years in school) and **Lifeexpectancy** (in years) with the corresponding regression line. Interpret this plot. Can you give an explanation for the pattern in the scatter plot? *[4]*

(c) Determine the correlation coefficient from (b) and give an interpretation of this value. *[1]*

(d) We use a regression model to investigate whether the life expectancy (**Lifeexpectancy**) is depending on the average body mass index (**BMI**), the number of years in school (**Schooling**), the status (**Status**: country developed=0, developing=1), the population of the country (**Population**) and the year in which the data were collected (**Year**: 2000-2015). *[10]*

i) Write out an equation describing the multiple linear regression model for the variables mentioned above.

ii) Determine the parameters of this model and give an interpretation of these values?

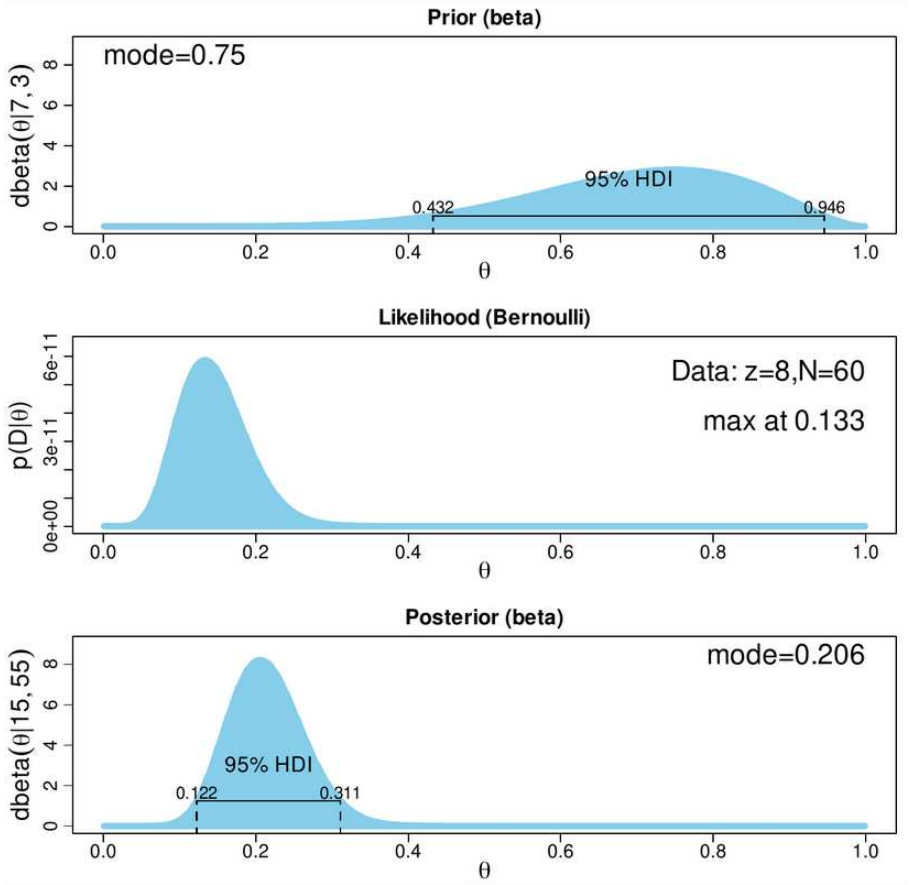
iii) What part of the variance is explained by the regression model?

iv) Interpret the p-value for the corresponding F-value.

v) We consider the individual regression coefficients. Is there any indication that we can remove any variables from the model? Justify your answer with p-values on a significance level of 5 %.

Problem 4: ..... (6 Points)

We flip a coin with the following prior distribution, likelihood function and posterior distribution. For the prior distribution, a and b were chosen so that the mode is 0.75.



Which of the following statements are correct, which are false. *Justify your answer!*

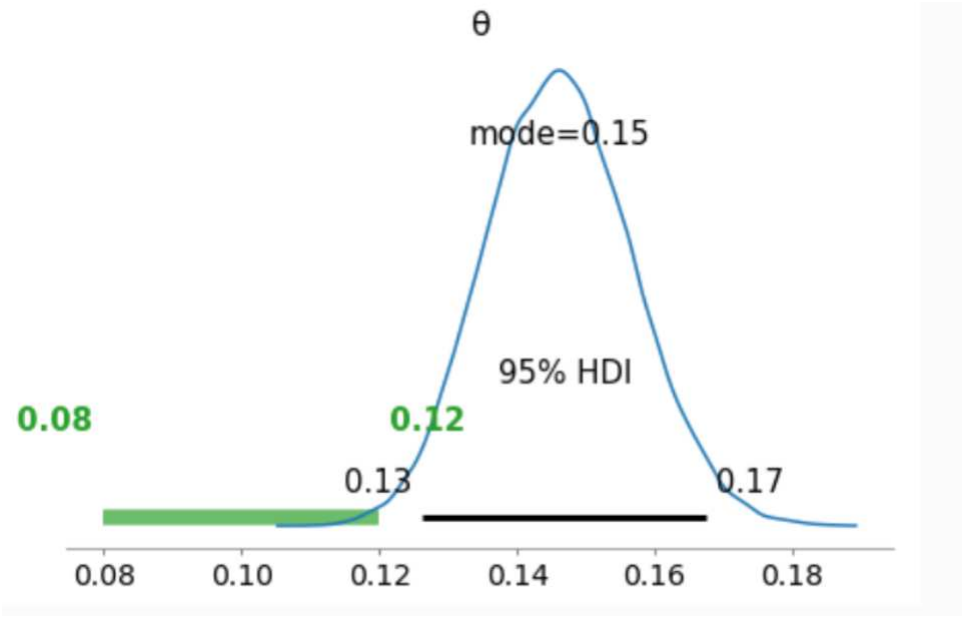
(a) The flat prior distribution shows that we are not very convinced of the value $\vartheta = 0.75$. *[2]*

(b) The mode of the prior distribution here is equal to the expected value. *[2]*

(c) Since the prior and posterior distributions are very different, the data do not affect our belief that $\vartheta = 0.75$. *[2]*

Problem 5: ..... (6 Points)

A manufacturer of computer chips is developing a new chip and would like to know in a first phase how large the proportion of defective chips is. He knows from previous chip developments that about 10 % of the prototypes are defective and is very convinced of this value. 1000 chips are tested and 150 of them are defective. The posterior distribution looks as follows.



Which of the following statements are correct, which are false. *Justify your answer!*

(a) For the prior distribution, a beta distribution with small $a + b$ is chosen. *[2]*

(b) The value of $\vartheta = 0.1$ is not accepted. *[2]*

(c) The choice of the width of the ROPE has no influence on whether $\vartheta = 0.1$ is accepted or not. *[2]*