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	12	12	64
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.

Solution:

Notations:

L: Voter supports liberal voter

$\overline{L}$: Voter does not support liberal candidate

A: Voter is from province A

B: Voter is from province B

C: Voter comes from province C

Known (from task):

$$P(L \mid A) = 0.5, \quad P(L \mid B) = 0.6, \quad P(\overline{L} \mid C) = 0.65$$

and

$$P(A) = 0.4, \quad P(B) = 0.25, \quad P(C) = 0.35$$

What we are looking for is $P(B \mid L)$. According to Bayes' formula and the law of total probability:

$$\begin{aligned} P(B \mid L) &= \frac{P(L \mid B)P(B)}{P(L)} \\ &= \frac{P(L \mid B)}{P(L \mid A)P(A) + P(L \mid B)P(B) + P(L \mid C)P(C)} \\ &= \frac{0.6 \cdot 0.25}{0.5 \cdot 0.4 + 0.6 \cdot 0.25 + (1 - 0.65) \cdot 0.35} \\ &= 0.3175 \end{aligned}$$

About 32 % of voters supporting the Liberal candidate are 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.

(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).

Solution:

(a) Read file:

```
nyc <- read.csv("nyc-marathon.csv")

head(nyc)
##    Minutes Gender Age
## 1 200.0667   male  52
## 2 268.4667 female  30
## 3 463.2833 female  43
## 4 286.5500 female  54
## 5 408.1000 female  37
## 6 304.5333   male  51
```

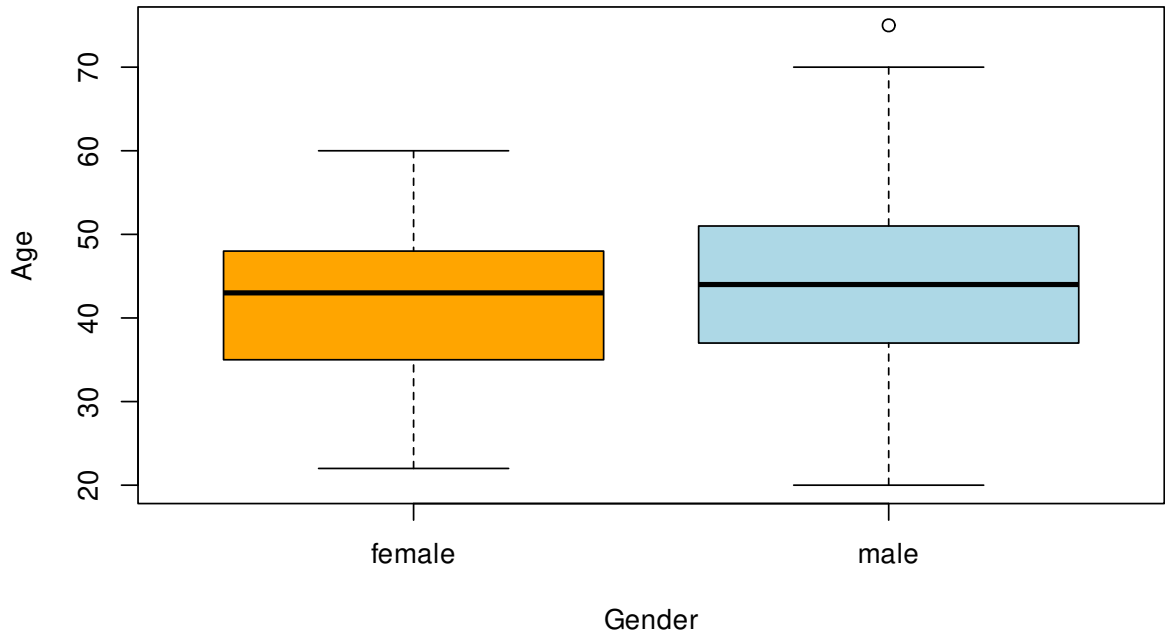
Summary:

```
summary(nyc)
##     Minutes         Gender                Age
##  Min.   :149.9   Length:276         Min.   :20.00
##  1st Qu.:224.3   Class :character   1st Qu.:36.00
##  Median :287.4   Mode  :character   Median :43.00
##  Mean   :291.2                      Mean   :43.39
##  3rd Qu.:339.8                      3rd Qu.:51.00
##  Max.   :509.1                      Max.   :75.00
```

- There are 107 women and 169 men included in this dataset.
- The range of ages is from 20 to 75 years.
- The median and mean are almost the same at 43 years.
- There were 25 % of individuals 36 years old or younger and 75 % 36 years old or older.
- There were 75 % of persons 51 years old or younger and 25 % 51 years old or older.
- Half of the subjects had ages between 36 and 51.

(b) Plot:

```
boxplot(Age ~ Gender, data = nyc, col= c("orange", "lightblue"))
```



- The medians are almost the same and thus the mean age of men and women is the same, 43 years.
- Both boxes are about the same width, about 12 years for women and 13 years for men.
- However, women tend to be younger, as their box is slightly below that of men. Women's mean boxes are in the 36-48 year range; men's mean boxes are in the 38-51 year range.
- The men's box is symmetric, while the women's is not, since the median is not in the middle of the box.

(c) i) Since we do not know a priori in which of the two marathons the average age of women is greater, we do a two-sided test.

ii) We denote the average age of women in the 2005 marathon by μ_{2005}. Correspondingly for μ_{2010}.
Null hypothesis:

$$H_0: \quad \mu_{2005} = \mu_{2010}$$

Alternative hypothesis:

$$H_A: \quad \mu_{2005} \neq \mu_{2010}$$

iii) Output:

```
female <- nyc[nyc[, "Gender"] == "female" ,]
t.test(female$Age, mu = 36.1)
```

```
##
##  One Sample t-test
##
## data:  female$Age
## t = 6.1735, df = 106, p-value = 1.249e-08
## alternative hypothesis: true mean is not equal to 36.1
```

```
## 95 percent confidence interval:
##  39.80704 43.31446
## sample estimates:
## mean of x
##  41.56075
```

The p value is about $1.5 \cdot 10^{-8}$, which is very far below the significance level of 0.05. The null hypothesis is rejected and the averages are statistically significantly different.
From the output, we still find that the average age of women at the 2010 marathon was about 41.6 years. That is, this age is 5.5 years greater than the average of the 2005 Marathon.

iv) We take the confidence interval from the output above:

$$[39.81,\ 43.31]$$

To 95 %, the true mean lies in this interval. Since the value 36.1 *not* lies in this interval, the null hypothesis is rejected. The age difference is statistically significant.

(d) The t test assumes that the data are normally distributed. If it is not known whether the data are normally distributed, the Wilcoxon test is often chosen, which assumes only symmetric distribution.
However, this test is problematic here, since we saw in b) that the distribution is not completely symmetric.
If we apply the Wilcoxon test anyway, the null and alternative hypothesis remain the same as in c).
Output:

```
wilcox.test(female$Age, mu = 36.1)
##
##  Wilcoxon signed rank test with continuity
##  correction
##
## data:  female$Age
## V = 4522, p-value = 3.879e-07
## alternative hypothesis: true location is not equal to 36.1
```

Again, the p value is $3 \cdot 10^{-7}$, well below the significance level of 0.05 and again the null hypothesis is rejected. There is a strong statistically significant difference between the age averages of the women.

(e) i) We can do a one-sided test here, since we suspect from previous data that men's average age is greater than women's.
However, we can also argue for a two-sided test here. The difference may have been random at the 2005 marathon.

ii) Here an unpaired test is made, since no assignment of the men to the women is possible. However, this is already clear since from a) there are more men than women in this data set.

iii) We denote by μ_F the average age of the women and by μ_M the average age of the men.

Null hypothesis (one-sided):

$$H_0: \qquad \mu_F = \mu_M$$

Null hypothesis (one-sided):

$$H_A: \qquad \mu_F < \mu_M$$

iv) Output:

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

t.test(female$Age,male$Age, alternative = "less")
##
##  Welch Two Sample t-test
##
## data:  female$Age and male$Age
## t = -2.4519, df = 252.66, p-value = 0.007443
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.974724
## sample estimates:
## mean of x mean of y
##  41.56075  44.54438
```

The p value is 0.007 below the significance level and the null hypothesis is rejected. Males are statistically significantly older than females.
The confidence interval is

$$[-\infty,\ -0.97]$$

That is, the true difference in averages is to 95 % in the range of $-\infty$ (not very likely) to -1 years. Since 0 years (no difference) is not in this interval, the null hypothesis is also rejected here.

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 %.

Solution:

Read file

```
life <- read.csv("led.csv")
head(life)
```

```
##       Country Year     Status Lifeexpectancy
## 1 Afghanistan 2015 Developing           65.0
## 2 Afghanistan 2014 Developing           59.9
## 3 Afghanistan 2013 Developing           59.9
## 4 Afghanistan 2012 Developing           59.5
## 5 Afghanistan 2011 Developing           59.2
## 6 Afghanistan 2010 Developing           58.8
```

```
##   AdultMortality infantdeaths Alcohol
## 1            263           62    0.01
## 2            271           64    0.01
## 3            268           66    0.01
## 4            272           69    0.01
## 5            275           71    0.01
## 6            279           74    0.01
##   percentageexpenditure HepatitisB Measles  BMI
## 1             71.279624         65    1154 19.1
## 2             73.523582         62     492 18.6
## 3             73.219243         64     430 18.1
## 4             78.184215         67    2787 17.6
## 5              7.097109         68    3013 17.2
## 6             79.679367         66    1989 16.7
##   under.fivedeaths Polio Totalexpenditure Diphtheria
## 1               83     6             8.16         65
## 2               86    58             8.18         62
## 3               89    62             8.13         64
## 4               93    67             8.52         67
## 5               97    68             7.87         68
## 6              102    66             9.20         66
##   HIV.AIDS       GDP Population thinness1.19years
## 1      0.1 584.25921   33736494              17.2
## 2      0.1 612.69651     327582              17.5
## 3      0.1 631.74498   31731688              17.7
## 4      0.1 669.95900    3696958              17.9
## 5      0.1  63.53723    2978599              18.2
## 6      0.1 553.32894    2883167              18.4
##   thinness5.9years Incomecompositionofresources
## 1             17.3                        0.479
## 2             17.5                        0.476
## 3             17.7                        0.470
## 4             18.0                        0.463
## 5             18.2                        0.454
## 6             18.4                        0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

(a) Output

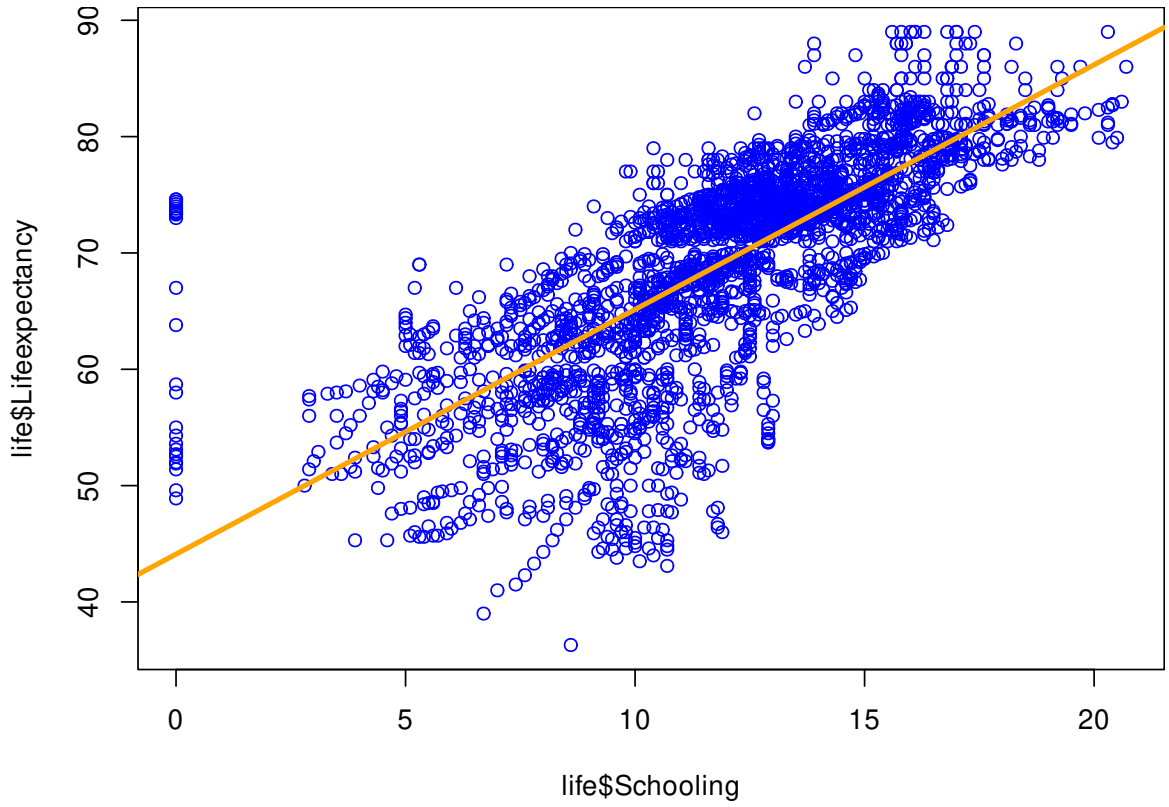
```
mean(life$Lifeexpectancy, na.rm = T)
## [1] 69.22493
```

The mean value of life expectancy is 69.22 years.
The option **`na.rm = T`** *must* be set because the column contains **`NA`**'s:

```
which(is.na(life$Lifeexpectancy == TRUE))
##  [1]  625  770 1651 1716 1813 1910 1959 2168 2217 2714
```

(b) Output:

```
plot(life$Schooling, life$Lifeexpectancy, col="blue")
abline(lm(life$Lifeexpectancy~life$Schooling),col="orange",lwd=3)
```



There is a positive linear relationship between the number of years of schooling and life expectancy: the longer the education, the greater the life expectancy. Of course, this is not a causal relationship. There is a confounder at play here. The richer the country, the longer the schooling and the better the medical care. People in rich countries tend to live healthier lives than those in poor countries.

There is another pattern, the dots on the far left: Having no schooling doesn't seem to hurt life expectancy.

(c) Output:

```
cor(life$Schooling, life$Lifeexpectancy, use = "na.or.complete")
## [1] 0.7519755
```

The correlation coefficient is 0.75 close to 1 and thus the positive linear relationship is also confirmed by the correlation coefficient.

(d) Output:

```
fit <- lm(Lifeexpectancy ~ BMI + Schooling + Status + Population + Year,
          data = life)
summary(fit)
##
## Call:
## lm(formula = Lifeexpectancy ~ BMI + Schooling + Status + Population +
##     Year, data = life)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.0796  -3.4277   0.7045   4.0135  26.1747
##
## Coefficients:
##                    Estimate Std. Error t value
## (Intercept)      -6.895e+01  5.671e+01  -1.216
## BMI               1.037e-01  7.846e-03  13.210
## Schooling         1.710e+00  5.472e-02  31.247
## StatusDeveloping -2.823e+00  3.960e-01  -7.129
## Population        2.806e-09  2.057e-09   1.364
## Year              5.763e-02  2.835e-02   2.033
##                  Pr(>|t|)
## (Intercept)         0.2242
## BMI                < 2e-16 ***
## Schooling          < 2e-16 ***
## StatusDeveloping 1.36e-12 ***
## Population          0.1726
## Year                0.0422 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.977 on 2246 degrees of freedom
##   (686 observations deleted due to missingness)
## Multiple R-squared:  0.6285,Adjusted R-squared:  0.6277
## F-statistic: 760.1 on 5 and 2246 DF,  p-value: < 2.2e-16
```

i) The regression model is:

$$\texttt{lifeexpectancy} = \beta_0 + \beta_1 \cdot \texttt{BMI} + \beta_2 \cdot \texttt{Schooling} + \beta_3 \cdot \texttt{Status} + \beta_4 \cdot \texttt{Population} + \beta_5 \cdot \texttt{Year}$$

ii) We take the corresponding parameters from the **Estimate** column in the output:

$$\texttt{Lifeexpectancy} = -69 + 0.1 \cdot \texttt{BMI} + 1.7 \cdot \texttt{Schooling} - 2.8 \cdot \texttt{Status} + 3 \cdot 10^{-9} \cdot \texttt{Population} + 0.06 \cdot \texttt{Year}$$

Interpretation of parameters:

- $\widehat{\beta}_0 = -69$: If all explanatory variables are 0, the „life expectancy'ïs -69 years. This, of course, makes no sense.
- $\widehat{\beta}_1 = 0.1$: Holding all other explanatory variables constant, a one-unit increase in BMI results in a 0.1 year increase in life expectancy.
  In countries that are very poor, people are also more likely to be underweight and therefore have a low BMI. Life expectancy in these countries is therefore not very high. Improving nutrition will result in a higher BMI and also a greater life expectancy.

- $\widehat{\beta}_2 = 1.7$: Holding all other explanatory variables constant, increasing the number of years of schooling by one year results in an increase in life expectancy of 1.7 years.
  As mentioned earlier in (b), this relationship is not causal, but is due to the confounder of wealth.
- $\widehat{\beta}_3 = -2.8$: Holding all other explanatory variables constant, states in developing countries have life expectancy 2.8 years lower than those in developed countries.
  As mentioned earlier in (b), this relationship is not causal, but is based on the confounder of wealth.
- $\widehat{\beta}_4 = 3 \cdot 10^{-9}$: Holding all other explanatory variables constant, an increase in population by results in an increase in life expectancy by 10^{-9} years.
  Of course, this does not make much sense, since poor population-small countries like Yemen arguably have the same life expectancy as poor large countries like Bangladesh.
  The associated p value is 0.17 is also above the significance level of 0.05 and thus the null hypothesis ($\beta_4 = 0$) is not rejected.
- $\widehat{\beta}_5 = 0.06$: Holding all other explanatory variables constant, increasing the date by one year results in an increase in life expectancy by 0.06 years.
  This has to do with the fact that diseases are better fought and also the food situation improves. There are fewer and fewer major famines.

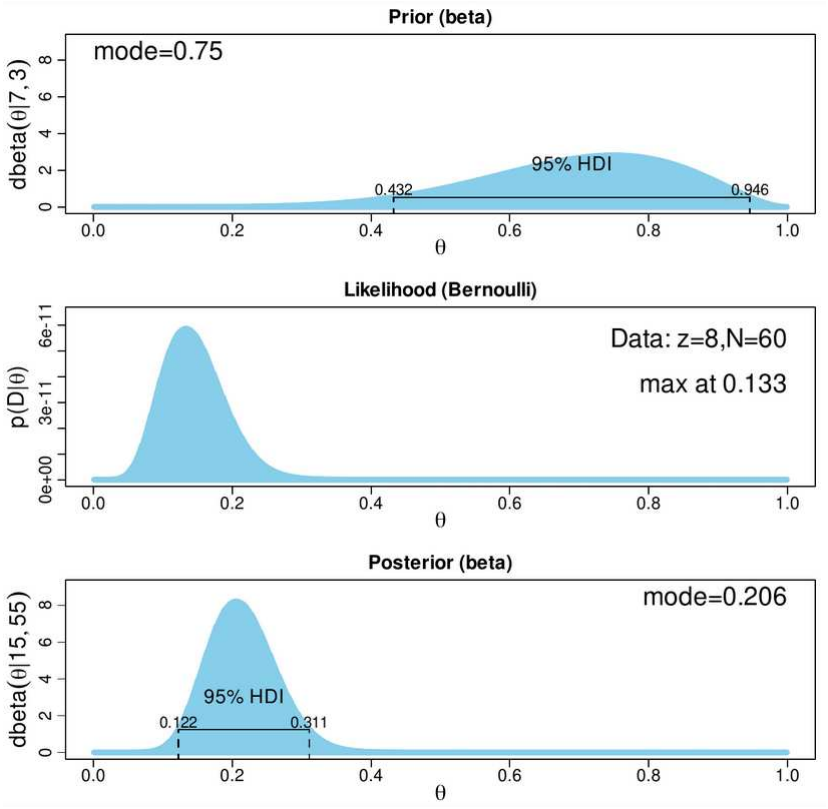
iii) The R^2 value is 0.6285 and thus 62.85 % of the variation is explained by the model.

iv) The p-value of the F-statistic is practically 0 and thus at least one variable affects the target variable **`lifeexpectancy`**.

v) We have already seen in (c) that the explanatory variable **`Population`** probably has no effect on life expectancy and thus it can be omitted.

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 $\theta = 0.75$. *[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 $\theta = 0.75$. *[2]*

Solution:

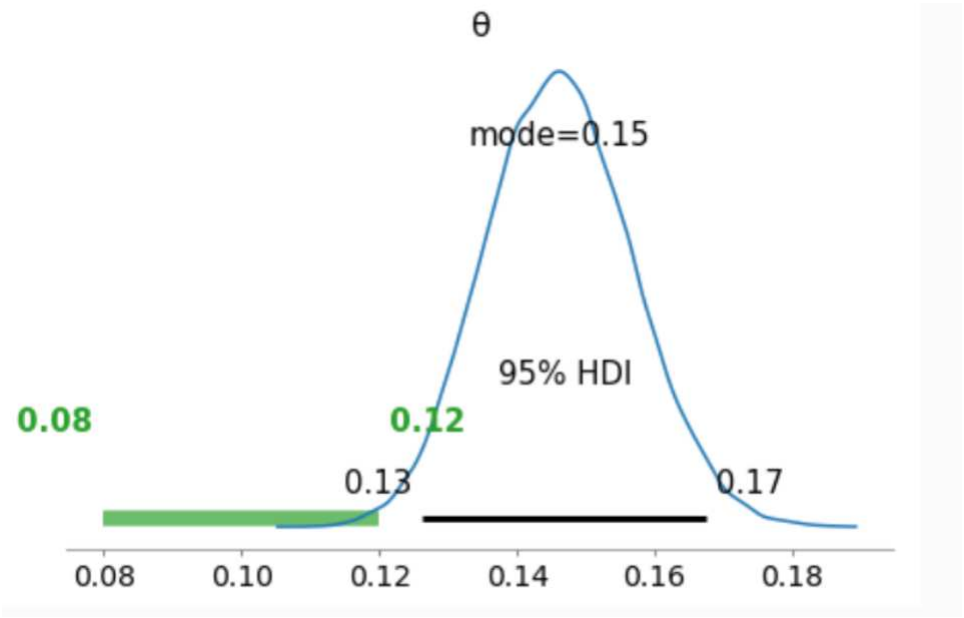
(a) True: For flat prior, we give a lot of θ quite high probabilities. For example, $\theta = 0.6$ has not much the lower value than $\theta = 0.75$ *[2]*

(b) False: For left skewed distributions is the mean to the left of the mode because the values on the far left are also taken into account for the mean. *[2]*

(c) False: Because the prior is so flat (weak conviction about $\theta = 0.75$, we are very likely to belief the data, i.e. the likelihood function. Hence, the likelihood function and the posterior distribution have almost the same shape. *[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 $\theta = 0.1$ is not accepted. *[2]*

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

Solution:

(a) False: The manufacturer is *very convinced* of $\theta = 0.1$, so he chooses a narrow distribution as prior. The distributions are narrow if $a + b$ is large. *[2]*

(b) True: ROPE and HDI do not intersect, so we reject the null value. *[2]*

(c) False: If we choose a ROPE of $[0.05, 0.15]$, the ROPE and HDI would intersect. In this case we have no decision from the data either way. *[2]*