2nd homework assignment

Task 1 - Maximum likelihood method (6 points)

Firstly, generate your own `data1` containing 100 observations. Do not forget to change set.seed() of PRNG with your UČO:

```
uco <- 492875  # insert your UCO
set.seed(uco)
data1 <- round(rgamma(100, 1, 1/5), 2)
data1 <- data1[data1 != 0]
```

If there is a 0 present in your data. Ignore such observation. Consider that the data represent waiting time in *minutes* in a queue at the study department of 100 randomly selected students.

a) Fit exponential distribution with parameter λ to your data (use the same parametrization as in the lecture). Use numerical maximization of the corresponding log-likelihood function to find the maximum likelihood estimate of λ.

$\widehat{\lambda}$
0.170711

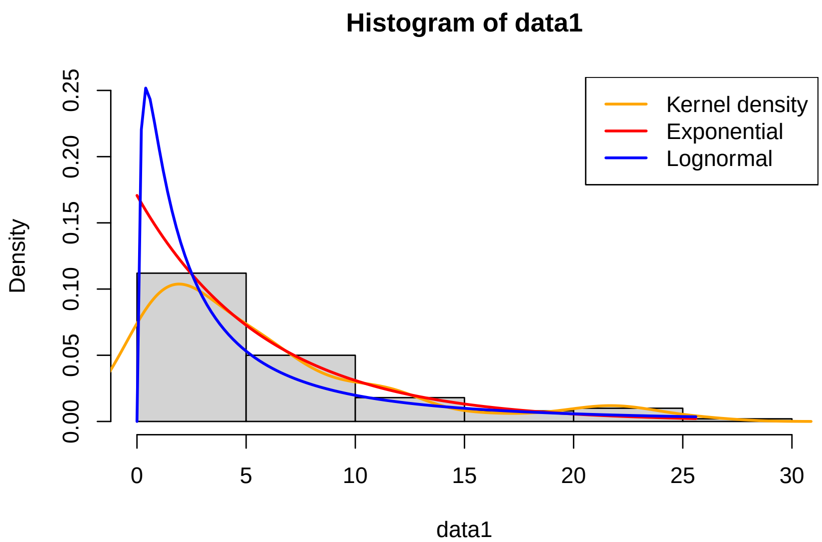
b) Fit lognormal distribution with parameters μ and σ with probability density function

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, & x \geq 0, \\ 0, & \text{otherwise} \end{cases}$$

on your data (see `?dlnorm` in `R` for details). Use numerical maximization of the corresponding log-likelihood function to find the maximum likelihood estimates of μ and σ.

$\widehat{\mu}$	$\widehat{\sigma}$
1.111128	1.411239

c) Plot histogram of your data together with both estimated densities.



d) Which of the two models would you choose? Why? Support your conclusion with a numerical characteristic.

Chosen model	Explanation
`Exponential`	The dataset seems to be closer to exponential distribution rather than the lognormal. Using AIC I get a smaller value for exponential distribution, which suggests that the data is closer to exponential distribution rather than lognormal distribution.

e) Based on both estimated models, what is the probability that you will wait in a queue for more than 5 minutes?

Estimated probability for exponential model	Estimated probability for lognormal model
0.4258981	0.3620063

Task 2 - Statistics I (2 points)

Work with the same data as in the previous task. Use the knowledge that the data represent waiting time in **minutes** between individual events. Using your result from the previous task, estimate the number of students coming to the study department in **one hour**. Construct a corresponding confidence interval.

Estimated number of students	Confidence interval
`10.24266`	`9.048628, 11.436695`

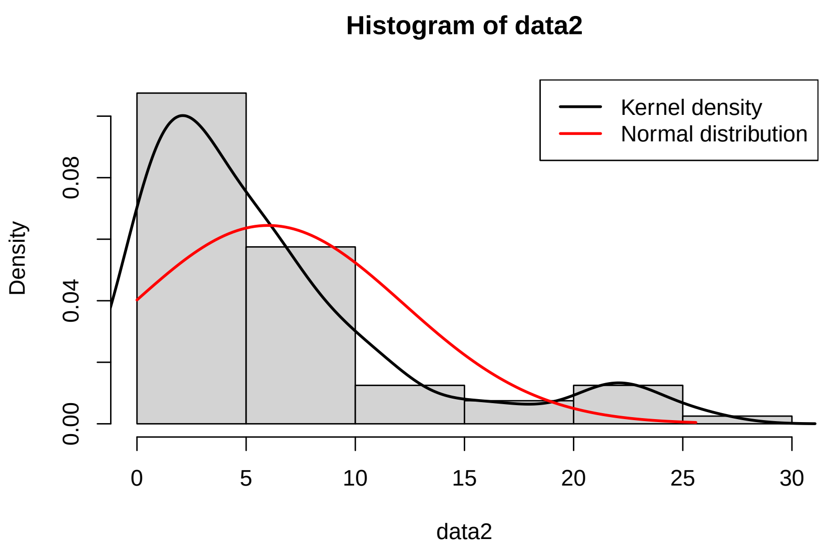
Task 3 - Normality checking (3 points)

Again, start with obtaining your data as a random sample from `data1`.

```
uco <- 492875  # insert your UCO
set.seed(uco)
data2 <- sample(data1, 80)
```

The aim of this task is to decide whether your data might come from a normal distribution. You can use any methods of your choosing, but for the purposes of this task, present only your final decision supported by one graph and the results of one statistical test.

Do your data look normal?
no



Support your claim with ONE suitable plot and a result of ONE statistical test confirming it.

Name of the test	p-value
Lilliefors Test	1.010312e-05

Task 4 - Statistics II (4 points)

In assignment 1, we worked with the **customer_behaviour2** dataset and created a new variable called `big` with values 1 (`big spender`, if the person spent more money than 5000 USD), and 0 (`low spender`, if he spent less or equal). The question is whether the big spenders are rather older people and the low spenders tend to be younger?

a) Use a suitable model and formulate null and alternative hypotheses and choose an appropriate test.

Name of the test you used	Explanation of your choice
Two-sample Wilcox Test	We are testing two independant groups which seem to have non-normal distributions. Specifically I want to use one-tailed version.

Perform the test.

```
#load and process data
load("customer_behaviour2.RData")
data$big = as.numeric(data$money_spent > 5000)
```

```r
big_spenders = data[data$big == 1,]$age
small_spenders = data[data$big == 0,]$age

#H0 - There is no significant difference between the mean age of
#   big spenders and the mean age of small spenders, or the mean age of
#   big spenders is less than or equal to the mean age of small spenders.
#HA - The mean age of big spenders is significantly greater than
#   the mean age of small spenders.
wilcox.test(big_spenders, small_spenders, alternative = "greater",
            conf.int = TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  big_spenders and small_spenders
## W = 11058, p-value = 1
## alternative hypothesis: true location shift is greater than 0
## 95 percent confidence interval:
##  -23.00007       Inf
## sample estimates:
## difference in location
##              -20.99998
```

```r
#p-value = 1 -> fail to reject H0 -> not enough evidence to support HA
#Analysis might even suggest that big spenders might be younger
```

What is you conclusion?

p-value of the test	Formal conclusion	Conclusion with your own words
1	Fail to reject H0	`Based on the result I fail to reject the null hypothesis. This means that I don't have enough evidence to support the claim that big spenders are older than small spenders. The analysis even suggests that big spenders might be younger.`

b) Construct the corresponding confidence interval and use it for testing your hypothesis. Do not forget to interpret your result (what does the confidence interval estimate?).

Confidence interval	Interpretation	Use for hypothesis testing and conclusion
`-23.00007, inf`	`I can be 95% confident that the true difference between big and small spenders lies within the confidence interval.`	`Since the confidence interval for the difference in location includes values below 0, we cannot reject the null hypothesis. And based on the 95% confidence interval, there is not enough evidence to support the claim that big spenders are older than small spenders.`