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

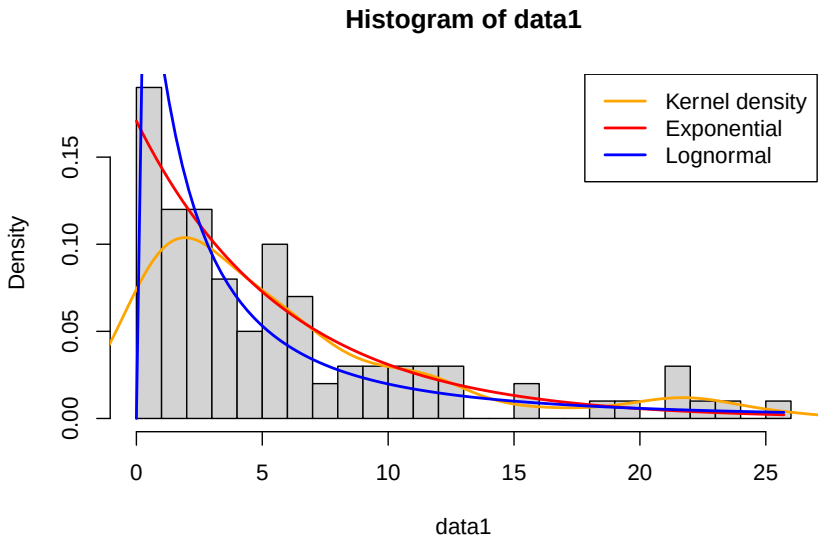
b) Fit lognormal distribution with parameters μ and σ with probability density function

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

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 qqnorm, the lower tail part of the dataset seems to follow the exponential distribution much better than the 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
`0.5741019`	`0.6379937`

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`	`[4.663266, 7.051334]`

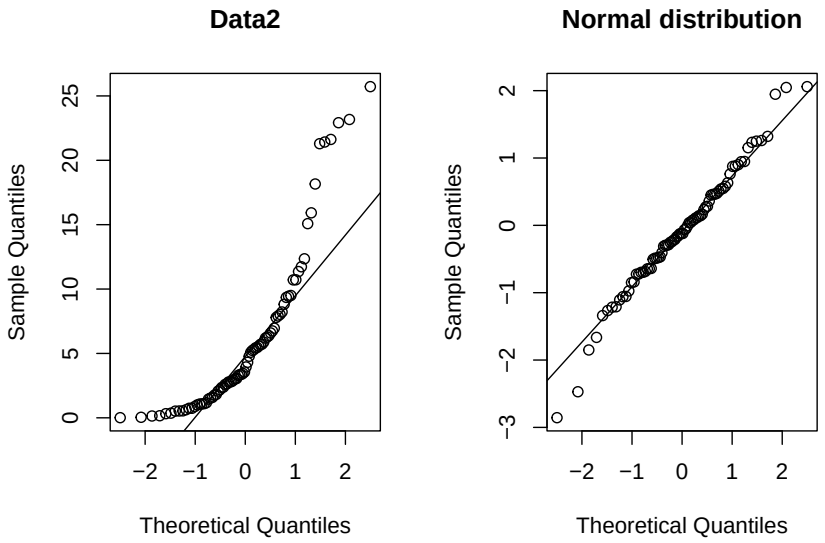
Task 3 - Normality checking (3 points)

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
Shapiro-Wilk Test	8.439703e-09

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 t-test	We are comparing the ages of two different groups of individuals (big spenders and low spenders) to test whether there is a statistically significant difference in their age distributions.

Perform the test.

```r
#load and process data
load("customer_behaviour2.RData")
data$big = as.numeric(data$money_spent > 5000)
big_spenders = data[data$big == 1,]$age
small_spenders = data[data$big == 0,]$age

#two-tailed t-test to test if there is indeed a difference
#H0 - There is not a significant difference between big and small spender
#HA - There is a significant difference
result = t.test(big_spenders, small_spenders)
ifelse(result$p.value <0.05, "H0 rejected", "Failed to reject H0")
```

```
## [1] "H0 rejected"
```

```r
#H0 rejected -> statistically significant evidence that
    #there is a difference in age in both groups
#failed to reject H0 -> not enough statistically significant
    #evidence to tell if there is a difference in age of both groups

#test the direction of the difference
ifelse(mean(big_spenders) < mean(small_spenders),
       "Big spenders are statistically younger",
       "Big spenders are statistically older")
```

```
## [1] "Big spenders are statistically younger"
```

What is you conclusion?

p-value of the test	Formal conclusion	Conclusion with your own words
2.2e-16	Reject H0	Based on the two-tailed t-test result, I can reject the null hypothesis. This means that there is a statistically significant difference in the age of both groups. By comparing their means, I can determine the direction of this difference. This tells me that bigger spenders tend to be younger rather than older.

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
[-22.76265, -17.15859]	I can be 95% confident that the true difference between big and small spenders lies within the confidence interval.	Since I used the two-tailed t-test to test if the age groups differ, the fact that the entire confidence interval is bellow zero suggests that there is a significant difference in the ages of both groups.