Description

The purpose of the project work is to practice how to solve statistical problems by using R and RStudio. The project work consists of ***eight*** separate tasks. In addition to just doing the tasks, you need to prepare a report documenting what you have done. The document is done with some text editor (like *Word*) and stored in PDF form. You have to submit you work to Moodle before the deadline. You are allowed to do only one submission.

The form of the document is free, but the first page needs to contain the name of the word "Statistical mathematics project work" and your name.

The tasks need to be documented in the same order as the appear in the description document. All the scripts you have used need to be included into document and also the plots (figures).

We have already considered several R scripts in:

1. the first exercise session,
2. lectures,
3. assignment sessions 4 and 5.

You may also use RStudio's help and internet to find information.

The **deadline** for the report is **16th March at 23:59**.

The project work is **graded**: ***rejected, 1, 2, 3, 4, or 5.*** The grade is based on correctness, logical structure of presentation, and clarity of writing.

In the case your project work gets rejected, you need to edit and **resubmit** it in a **week** after rejection.

To pass the course, you need to pass the exam and the project work. The final grade of the course is an ***weighted average*** of the following:

- Exam 70%
- Project work 30%

Task 1

1. Describe shortly what is **geometric distribution** and what are its parameters.
2. Plot geometric distribution for the probability 0.45 and the number of failures in {0,1,2,...,12}. Add a suitable title to your plot.

Note that you can change the type of your plot by parameter `type` and you can add e.g. points to an existing plot by the function `points()`.

Task 2

1. Describe by few words what is **binomial distribution** and what are its parameters.
2. Plot binomial distribution for p = 0.6 and size = 30. Add a suitable title to your plot.

Task 3

1. Describe briefly **Poisson distribution** and its parameter(s).
2. Plot binomial distribution for lambda = 20 and k = 0,..40. Add a suitable title to your plot.

Task 4

Import the dataset "`data_set1.csv`" to a dataframe. Then answer the following questions:

1. The dataset contains only one column: what is its name?
2. How many rows there are?
3. What is the minimal value of the dataset?
4. What is the maximal value?
5. What is the mean?
6. What is median?
7. What is the sample variance?
8. What is the sample standard deviation?

Task 5

The data used in this task is the same as in Task 4.

1. First plot an "approximation" of the density function of the **normal distribution** so that x-axis values are between 0, 1, 2, ..., 100. Using the mean and the standard deviation of **Task 4**. Plot the curve using plot-type "`l`" (line). The plot should have the title "Data set vs normal distribution" and note that the plot is not finished yet, but you add more things to it.

2. Use the function `density()` to compute the density of the dataset "`data_set1.csv`". Store the density to a variable called `d`. Note that the function "`density`" has a parameter bandwidth (`bw`) which tells how much the density curve is “smoothed”. R estimates `bw` automatically but you may also set it by yourself. You may try different values: when you plot the density curve, you can see what looks best.

3. Add the density curve (in variable `d`) to the plot using the function `points()`. The color of the curve should be "red" so that can be distinguished from the density plot of the dataset.

4. Add a straight green vertical line that shows the mean using the function `abline()`.

Task 6

R contains the dataset `mtcars`, whose the columns are explained in lectures. Using Pearson correlation coefficient, find **four variables** which are most correlated with horsepower (hp). In other words, compute the correlation coefficient of the other variables with respect to hp. Select those which have the biggest negative or positive correlation.

Task 7

1. Build a linear model which models hp using the four variables you found in Task 6.
2. Using the function `predict()` to add a variable `hp_hat` to the the dataset.
3. Compute the residuals (`hp - hp_hat`)
4. Make the residual plot using `density()` and `plot()` functions. The title of the plot is "Density of residuals"
5. Does your residual plot has a bell-shape?
6. What is the R-squared value of the model? Is your model good?

Task 8

A company produces wooden sticks whose length is supposed to be 30 centimeters. The owner of the company is not very satisfied with the results and expects that the produced sticks are too long. She orders a quality inspector to do some testing.

The inspector collects a sample of 1200 sticks and carefully measures the length of the sticks. They are stored in the file "data_set2.csv".

The **zero hypothesis** is $\mu = 30$.

Is it acceptable based on the "data_set2.csv"? Explain how you reached this conclusion.