3rd homework assignment

Task 1 - Modern statistical methods (4 points)

First, load the data from the file `women.csv`. The data contains information about IQ and weight in *kg* of 30 randomly selected women.

a) Compute the value of Pearson correlation coefficient for `IQ` and `weight`. Compute the 95% confidence interval for the true correlation coefficient. Round all the results to **three** decimal places.

correlation coefficient estimate	lower bound for CI	upper bound for CI
`insert`	`insert`	`insert`

b) Use nonparametric bootstrap (perform 10 000 replications) to estimate 95% confidence interval (percentile) for the true correlation coefficient. Round the results to **three** decimal points.

lower bound for bootstrap CI	upper bound for bootstrap CI
`insert`	`insert`

c) Assume that the data come from a bivariate normal distribution. Is there a connection between IQ and weight? Use the correlation coefficient to test it. Round corresponding p-value to **three** decimal points.

p-value of the test	conclusion
`insert`	`insert`

d) Use Monte Carlo simulations (perform 9 999 replications) to get a simulated p-value of the previous test. Round it to **three** decimal points.

Simulated p-value
`insert`

Hint: Assume normality of the data. Correlation coefficient does not depend on the value of means, nor variances of both variables. Hence they **are not** nuisance parameters and do not need to be taken into account.

Caution: Before every simulation run, do not forget to change `set.seed` of PRNG with your `UCO`:

```r
uco <- 235559  # insert your UCO
set.seed(uco)
```

Task 2 - Testing hypotheses (5 points)

Work with the data samples from `farm1.RData` and `farm2.RData` containing the weights of the total production (in `kg`) in different months at two farms. Your task is to compare their productions.

```
load("farm1.RData")
load("farm2.RData")
```

Firstly answer the question, if the average weight of the production at the farm 1 equals 125 `kg` (as the owner of the farm proclaims), or less. Which statistical tool is appropriate (**explain** Your choice in details)? Why? What is the result (compute p-value and write the conclusion)?

Name of the test	Explanation	p-value	conclusion
`insert`	`insert`	`insert`	REJECTED / NOT REJECTED

Now try to answer, if the average weights of production at the farms are the same. Which statistical test is appropriate (**explain** Your choice in details)? Why? What is the result (compute p-value and write the conclusion)? Support your conclusion by **one** suitable figure, which will visualize the results of your test (the averages of the weights, their difference etc.).

Name of the test	Explanation	p-value	conclusion
`insert`	`insert`	`insert`	REJECTED / NOT REJECTED

In each task **check the assumptions** of the tests you used and **name them all** in the `Explanation` sections.

Task 3 - Linear model (6 points)

Work with the `cholesterol.RData` dataset. It contains information about the cholesterol levels, age, blood pressure, dietary preferences and smoking habits of 100 patients. Your task is to model the cholesterol levels. Try to create a model that best describes the cholesterol levels while being as simple as possible.

Your chosen model formula	adjusted R squared
`insert e.g. cholesterol ~ age + smoker`	`insert`

Hint: Consider different transformations of the explanatory variables at hand.

Warning: Keep in mind the limitations of linear models.

Interpret adjusted R squared. What does it mean?

Value of adjusted R squared	Interpretation
`insert value`	`insert text`

Interpret the coefficients of your chosen model. If your model uses a different number of coefficients as listed here, add or delete table rows.

Variable	Value of coefficient	Interpretation
`insert variable name`	`insert value`	`insert text`
`insert variable name`	`insert value`	`insert text`
`insert variable name`	`insert value`	`insert text`

Interpret the validity of your model (the F statistic in the models summary).

Null hypothesis	p-value	Interpretation
`insert`	`insert`	`insert`