# 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 |
|---|---|---|
| -0.053 | -0.406 | 0.313 |

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 |
|---|---|
| -0.443 | 0.34 |

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 |
|---|---|
| 0.781 | I fail to reject the null hypothesis. There does not seem to be statistically significant correlation between IQ and weight. |

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 |
|---|
| 0.779 |

**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`:

```
uco <- 492875  # 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 |
|---|---|---|---|
| `one-sample t-test` | `We want to know if the samples come from a population with a specific mean (125kg). H0 - average weight equals 125 kg HA - average weight is less than 125 kg` | `0.18` | I fail to reject null hypothesis. There is not enough evidence to say that the average weight is less than 125kg. |

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 |
|---|---|---|---|
| `tow-sample t-test` | `We have 2 independand dataset and want to test if their average product weight is the same. H0 - Average weight is same HA - Average weight differ signifficantly` | `0.669` | I fail to reject null hypothesis. There is not enough evidence to suggest that the average weight of production between both farms differ significantly. |

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 |
|---|---|
| `log(cholesterol) ~ age * vegetarian` | `0.969` |

**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 |
|---|---|
| 0.969 | It means that about 96.9% of the variation in cholesterol variable can be explained by this model. |

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 |
|---|---|---|
| age | 0.027 | Change in log(cholesterol) for each one-year increase in age. Can be interpreted also as 2.7% increase in log(cholesterol) for each aditional year of age. |
| vegetarian | 0.381 | Difference between expected log(cholesterol) in vegetarians and meat-eaters. We would expect the log(cholesterol) to be on average about 0.381 lower for vegetarians compared to meat-eaters. |
| age:vegetarian meat-eater | -0.005 | Interaction term. It represents the difference in rate of change in log(cholesterol) for a one-year increase in age between meat-eaters and vegetarians. |

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

| Null hypothesis | p-value | Interpretation |
|---|---|---|
| All of the regression coefficients are zero, which would imply that none of the predictors (age, whether they are vegetarian, diastolic blood pressure) have any effect on the response variable log(cholesterol). | 2.2e-16 | fail reject the null hypothesis and can conclude that there is a strong evidence that at least one of the predictors is significantly associated with the response variable log(cholesterol). |