Assignment 1

Start Assignment

- Due 26 Feb by 23:59
- Points 10
- Submitting a file upload
- File types pdf
- Attempts 0
- Allowed attempts 2
- Available until 14 Mar at 23:59

- *Before submitting your report please read* ***Guidelines for assignment reports to submit*** *(https://canvas.vu.nl/courses/75847/pages/ga402d5e8b7e83ed13a9a03f2f6799ac8) , the page limit for your report is* ***15****.*
- *Remember that you can use two template files in* ***R-sources and data sets*** *(https://canvas.vu.nl/courses/75847/pages/g0a763d46cba78cc5f8e51b08e07adcb8?wrap=1) for making your report.*
- *Throughout this assignment, tests should be performed using a confidence level* α*=0.05, unless otherwise specified.*
- *Abbreviation CI means* confidence interval*.*
- *Where appropriate, motivate your answers and check the model assumptions by using relevant diagnostic tools.*
- *Beware that the levels of some factors for some data sets may be coded by numbers.*
- Command *read.csv* can be used to upload a csv-file in R, possibly with the option *header=TRUE*.

Exercise 1. Ice cream

The data set **Ice_cream.csv (https://canvas.vu.nl/courses/75847/files/7340354?wrap=1)** ⭳ **(https://canvas.vu.nl/courses/75847/files/7340354/download?download_frd=1)** was collected on 200 high school students and include student's scores on a video game (column *video*) and a puzzle (column *puzzle*), student's favorite flavor of ice cream (column *ice_cream:* 1 - vanilla, 2 - chocolate, 3 - strawberry) and gender (column *female:* 1 - female, 0 - male). (In this exercise we disregard the column *ice_cream*.)

Consider the sample of the video game scores (column *video*), and denote the underlying mean of this sample by μ.

a) Make some relevant plots of this sample, comment on normality. Assuming normality (irrespective of your conclusion about normality), construct a bounded 97%-CI for μ. Evaluate the sample size needed to provide that the length of the 97%-CI is at most 3. Compute a bootstrap 97%-CI for μ and compare it to the above CI.

b) By using a relevant t-test, verify the claim that the mean score on the video game is bigger than μ_0=50. Explain the meaning of the CI in the R-output for this test. Explain what happens with this t-

test and the corresponding CI in case μ_0=51. Comment.

c) Verify the claim that the median score on the video game is bigger than μ_0=50, by using a suitable sign test and a test based on ranks. Compare to b), comment. Design and perform a test to check whether the fraction of the scores less than 42 is at most 25%.

d) Let $X_1, \ldots, , X_{200}$ be the column *video*. By using a bootstrap test with test statistic $T = \min\{X_1, \ldots, X_{200}\}$, determine those $\mu \in [0, 100]$ for which the hypothesis $H_0 : X_1, \ldots, X_{200} \sim N(\mu, 100)$ is not rejected. Can the Kolmogorov-Smirnov test be also applied for this situation? If yes, apply it; if not, explain why not.

Consider now the two samples: the scores on the video game by female and male students (extract these samples from the data frame).

e) The expert claims that the mean score on the video game for the male students is higher than for the female students. Verify this claim by performing three tests: the two samples t-test, the Mann-Whitney test and the Kolmogorov-Smirnov test. Indicate whether these tests are actually applicable in this situation. Is a permutation test applicable? Comment.

Consider now the following two samples: the scores on the video game and the scores on the puzzle (columns *video* and *puzzle*).

f) Investigate whether the columns *video* and *puzzle* are correlated. Apply relevant tests to verify whether the score on the puzzle is higher than the score on the video game. Comment on your findings.

Exercise 2. Hemoglobin in trout

Hemoglobin is measured (g/100 ml.) in the blood of brown trout after 35 days of treatment with four rates of sulfamerazine: the daily rates of 0, 5, 10 and 15 g of sulfamerazine per 100 pounds of fish, denoted as rates 1, 2, 3 and 4, respectively. (*Beware that the levels of the factor rate are coded by numbers.*) Two methods (denoted as A and B) of administering the sulfamerazine were used. The data is collected in data set **hemoglobin.txt (https://canvas.vu.nl/courses/75847/files/7340353?wrap=1)** ↓ **(https://canvas.vu.nl/courses/75847/files/7340353/download?download_frd=1)** .

a) Present an R-code for the randomization process to distribute 80 fishes over all combinations of levels of factors *rate* and *method*.

b) Perform two-way ANOVA to test for effects of factors *rate*, *method* and their interaction on the response variable *hemoglobin*. Comment on your findings.

c) Which of the two factors has the greatest influence? Is this a good question? Consider the additive model. Which combination of rate and method yield the highest hemoglobin? Estimate the mean hemoglobin value for rate 3 by using method A. What rate leads to the highest mean hemoglobin?

d) Test the null hypothesis that the hemoglobin is the same for all rates by a one-way ANOVA test, ignoring the variable *method*. Give the estimated hemoglobin value for each rate. Is it right/wrong or useful/not useful to perform this test on this dataset?

e) Does the Kruskal-Wallis test arrive at the same conclusion about the effect of drug as the test in d)? Explain possible differences between the Kruskal-Wallis and ANOVA tests.

Exercise 3. Sour cream

The file **cream.txt (https://canvas.vu.nl/courses/75847/files/7340352?wrap=1)** **(https://canvas.vu.nl/courses/75847/files/7340352/download?download_frd=1)** contains data on an experiment to produce sour cream. Yogurt was placed in sweet cream, and yogurt bacteria were allowed to develop. Bacteria produce lactic acid, and as a surrogate for the number of yogurt bacteria, the acidity of the cream was measured. Interest was in the effect of the type of yogurt (denoted as starter) on acidity. The mixtures of yogurt and sweet cream were kept at constant temperature in a yogurt maker, in which five different positions could be used. The experiment was carried out with five batches of sweet cream, which were meant to have the same composition. With each batch each of five types of starter was used, with the yogurt placed in one of the five positions. The combinations of levels of three factors form a three-dimensional latin square. (*You may need to install the R-package lme4, which is not included in the standard distribution of R*.)

a) Analyze the data in a three-way experiment without interactions with *acidity* as response and *starter*, *batch* and *position* as factors. By using *summary* command, can you tell whether there is a significant difference between the effects of starter 1 and starter 2 on acidity? Motivate your answer.

b) Recall that the main interest is in the effect of *starter* on *acidity*; factors *position* and *batch* represent the block variables. Remove insignificant block variable(s) if there are such, and perform an ANOVA for the resulting “fixed effects” model. Which starter(s) lead to significantly different acidity? Motivate your answer.

c) For the resulting model from b), can we also apply the Friedman test to test whether there is an effect of starter on acidity?

d) Repeat c) by performing a mixed effects analysis, modeling the block variable(s) (if there are any) as a random effect by using the function *lmer*. (You will need to install the R-package *lme4*, which is not included in the standard distribution of R.) Compare your results to the results found by using the fixed effects model in c). Comment.

EDDA: Assignment 1 (1)

Criteria	Ratings	Pts
Exer.1 a) b) c) d) e) f)		3.4 pts
Exer.2 a) b) c) d) e)		2.8 pts
Exer.3 a) b) c) d)		2.8 pts
General layout		1 pts
		Total points: 10