Assignment 1

Start Assignment

Due 27 Feb by 23:59 **Points** 10 **Submitting** a file upload **File types** pdf
Attempts 0 **Allowed attempts** 2 **Available** until 5 Mar at 23:59

- *Before submitting your report please read **Guidelines for assignment reports to submit** , the page limit for your report is **20**.*
- *Remember that you can use two template files in **R-sources and data sets** 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.*

Exercise 1. Birthweights

The data set **birthweight.txt** contains the birthweights (in grams) of 188 newborn babies. Denote the underlying mean birthweight by μ.

a) Check normality of the data. Assuming normality (irrespective of your conclusion about normality), construct a bounded 96%-CI for μ. Evaluate the sample size needed to provide that the length of the 96%-CI is at most 100. Compute a bootstrap 96%-CI for μ and compare it to the above CI.

b) An expert claims that the mean birthweight is bigger than 2800 gram. Verify this claim by using a relevant t-test, explain the meaning of the CI in the R-output for this test. Also propose and perform a suitable sign tests for this problem.

c) Propose a way to compute the powers of the t-test and sing test from b) at some μ>2800, comment.

d) Let *p* be the probability that birthweight of a newborn baby is less than 2600 gram. Using asymptotic normality, the expert computed the left end $\hat{p}_l$=0.25 of the confidence interval $[\hat{p}_l, \hat{p}_r]$ for *p*. Recover the whole confidence interval and its confidence level.

e) The expert also reports that there were 34 male and 28 female babies among 62 who weighted less than 2600 gram, and 61 male and 65 female babies among the remaining 126 babies. The

expert claims that the mean weight is different for male and female babies. Verify this claim by an appropriate test.

Exercise 2. Cholesterol

A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat low cholesterol diet. The data set **cholesterol.txt** contains information on 18 people using margarine to reduce cholesterol: columns *Before* and *After8weeks* contain the cholesterol level (mmol/L) respectively before the diet and after 8 weeks on the diet.

a) Make some relevant plots of this data set, comment on normality. Are there any inconsistencies in the data? Investigate whether the columns *Before* and *After8weeks* are correlated.

b) Apply two relevant tests (cf. Lectures 2, 3) to verify whether the diet with low fat margarine has an effect (argue whether the data are paired or not). Is a permutation test applicable?

c) Let $X_1, \ldots, X_{18}$ be the column *After8weeks*. Assume $X_1, \ldots, X_{18} \sim \mathbf{Unif}[3, \theta]$ for $\theta > 3$, then use the central limit theorem to find an estimate $\hat{\theta}$ for θ and construct a 95%-CI for θ. Can you improve this CI?

d) By using a bootstrap test with test statistic $T = \max(X_1, \ldots, X_{18})$, determine those $\theta \in [3, 12]$, for which the hypothesis $H_0 : X_1, \ldots, X_{18} \sim \mathbf{Unif}[3, \theta]$ is not rejected. Can the Kolmogorov-Smirnov test be also applied for this situation?

e) Using an appropriate test, verify whether the median cholesterol level after 8 weeks of low fat diet is less than 6. Next, design and perform a test to check whether the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%.

Exercise 3. Diet

To investigate the effect of 3 types of diet, 78 persons were divided randomly in 3 groups, the first group following diet 1, second group diet 2 and the third group diet 3. Next to some other characteristics, the weight was measured before diet and after 6 weeks of diet for each person in the study. The collected data is summarized in the data frame **diet.txt** with the following columns: *person* – participant number, *gender* – gender (1 = male, 0 = female), *age* – age (years), *height* – height (cm), *preweight* – weight before the diet (kg), *diet* – the type of diet followed, *weight6weeks* – weight after 6 weeks of diet (kg). Compute and add to the data frame the variable *weight.lost* expressing the lost weight, to be used as response variable.

a) Make an informative graphical summary of the data relevant for study of the effect of diet on the wight loss. By using only the columns *preweight* and *weight6weeks*, test the claim that the diet affects

the weight loss. Check the assumptions of the test applied.

b) Apply one-way ANOVA to test whether type of diet has an effect on the lost weight. Do all three types diets lead to weight loss? Which diet was the best for losing weight? Can the Kruskal-Wallis test be applied for this situation?

c) Use two-way ANOVA to investigate effect of the diet and gender (and possible interaction) on the lost weight.

d) Dropped.

e) Which of the two approaches, the one from b) or the one from c), do you prefer? Why? For the preferred model, predict the lost weight for all three types of diet.

Exercise 4. Yield of peas

The dataset *npk* is available in the R package *MASS*. After loading the package *MASS*, type *npk* at the prompt to view this dataset. This dataset gives the *yield* of peas in pounds per plot, based on four factors: in which *block* the plot was located (labeled 1 through 6), and whether nitrogen (*N*), phosphate (*P*) or potassium (*K*) was applied to the soil (1 = applied, 0 = not applied). There are 24 plots, 4 per block. This is incomplete block design but balanced in the sense that within each block each soil additive is received by two plots. Our main question of interest is whether nitrogen *N* has an effect on *yield*.

a) Present an R-code for the randomization process to distribute soil additives over plots in such a way that each soil additive is received exactly by two plots within each block.

b) Make a plot to show the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen, and comment. What is the purpose to take the factor *block* into account?

c) Conduct a full two-way ANOVA with the response variable *yield* and the two factors *block* and *N*. Was it sensible to include factor *block* into this model? Can we also apply the Friedman test for this situation? Comment.

d) Investigate other possible models with all the factors combined, restricting to only one (pair- wise) interaction term of factors *N*, *P* and *K* with *block* in one model (no need to check the model assumptions for all the models). Test for the presence of main effects of *N*, *P* and *K*, possibly taking into account factor *block*. Give your favorite model and motivate your choice.

e) Recall the main question of interest. In this light, repeat c) by performing a mixed effects analysis, modeling the block variable as a random effect by using the function *lmer*. Compare your results to the results found by using the fixed effects model in c). (You will need to install the R-package *lme4*, which is not included in the standard distribution of R.)

EDDA: Assignment 1		
Criteria	**Ratings**	**Pts**
Exer.1 a) 0.5 b) 0.5 c) 0.5 d) 0.4 e) 0.4		2.3 pts
Exer.2 a) 0.4 b) 0.5 c) 0.5 d) 0.6 e) 0.4		2.4 pts
Exer.3 a) 0.5 b) 0.5 c) 0.4 d) e) 0.6		2 pts
Exer.4 a) 0.4 b) 0.4 c) 0.5 d) 0.5 e) 0.5		2.3 pts
General layout		1 pts
		Total points: 10