

This assignment was locked on 19 Mar at 23:59.

- Throughout this assignment, tests should be performed using a confidence level
- $=0.05$, unless otherwise specified.
- Where appropriate, motivate your answers and check the model assumptions by using relevant diagnostic tools.

Exercise 1. Post-operative nausea

The file [nauseatable.txt](#) ☐ [Download nauseatable.txt](#) contains data about post-operative nausea after medication against nausea. Two different medicines were administered to patients that complained about post-operative nausea. One of the medicines, *Pentobarbital*, was administered in two different doses.

- Discuss whether a contingency table test is appropriate here. If yes, perform this test in order to test whether the different medicines work equally well against nausea. Where are the main inconsistencies?
- Perform a permutation test in order to test whether the different medicines work equally well against nausea. Permute the medicine labels for this purpose. Use as test statistic the chisquare test statistic for contingency tables, which can be extracted from the output of the command `chisq.test`. (Hint: make a data frame in R consisting of two columns. One column should contain an indicator whether or not the patient in that row suffered from nausea, and the other column should indicate the medicine.)
- Compare the p-value found by the permutation test with the p-value found from the chisquare test for contingency tables. Explain the difference/equality of the two p-values.

Exercise 2. Airpollution

The data in [airpollution.txt](#) ☐ [Download airpollution.txt](#) were obtained to determine predictors related to air pollution. We want to investigate which explanatory variables need to be included into a linear regression model with *oxidant* as the response variable.

- Make some graphical summaries of the data. Investigate the problem of potential and influence points, and the problem of collinearity.
- Use the added variable plot to depict the relationship between response *oxidant* and predictor *wind*. What is the meaning of the slope of fitted regression for this scatter plot?
- Fit a linear regression model to the data. Use both the step-up and step-down methods to find the best model. If step-up and step-down yield two different models, choose one and motivate your choice.

d) Determine 95% confidence and prediction intervals for *oxidant* using the model you preferred in c) for *wind*=33, *temperature*=54, *humidity*=77 and *insolation*=21.

Exercise 3. Fruit flies

To investigate the effect of sexual activity on longevity of fruit flies, 75 male fruit flies were divided randomly in three groups of 25. The fruit flies in the first group were kept solitary, those in the second were kept together with one virgin female fruit fly per day, and those in the third group were kept together with eight virgin female fruit flies a day. In the data-file [fruitflies.txt](#) ☐ [Download fruitflies.txt](#) the three groups are labelled isolated, low and high. The number of days until death (longevity) was measured for all flies. Later, it was decided to measure also the length of their thorax. Add a column *loglongevity* to the data-frame, containing the logarithm of the number of days until death. Use this as the response variable in the following.

- a) Make an informative plot of the data. Investigate whether sexual activity influences longevity by performing a statistical test, without taking the thorax length into account. What are the estimated longevitys for the three conditions? Comment.
- b) Investigate whether sexual activity influences longevity by performing a statistical test, now including thorax length as an explanatory variable into the analysis. Does sexual activity increase or decrease longevity? What are the estimated longevitys for the three groups, for flies with the minimal and maximal thorax lengths?
- c) How does thorax length influence longevity? Investigate graphically and by using an appropriate test whether this dependence is similar under all three conditions of sexual activity.
- d) Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong?
- e) Perform the ancova analysis with the number of days as the response, rather than its logarithm. Was it wise to use the logarithm as response?


Exercise 4. Personalized system of instruction

The data [psi.txt](#) ☐ [Download psi.txt](#) was collected to study the effect of a new teaching method called “personalized system of instruction” (*psi*), 32 students were randomized to either receive *psi* or to be taught using the existing method. At the end of the teaching period the success of the teaching method was assessed by giving the students a difficult assignment, which they could pass or not. The average grade of the students were also available for analysis: *gpa* on a scale of 0–4, with 4 being the best grade.

- a) Fit a logistic regression model with both explanatory variables, perform relevant tests. Does *psi* work?

- b) Estimate the probability that a student with a *gpa* equal to 3 who receives *psi* passes the assignment. Estimate the same probability for a student who does not receive *psi*. Comment.
- c) Estimate the relative change in odds of passing the assignment rendered by instructing students with *psi* rather than the standard method (for an arbitrary student). What is the interpretation of this number? Is it dependent on *gpa*?
- d) Propose and perform an alternative method of analysis based on contingency tables. Compare its results to the results of the first approach.
- e) Given the way the experiment was conducted, is this second approach wrong? Name both an advantage and a disadvantage of the two approaches, relative to each other.

Exercise 5. School awards

The file [awards.txt](#)  [Download awards.txt](#) contains data on the numbers of awards earned by students at one high school. Predictors of the number of awards earned include the type of program (column *prog*) in which the student was enrolled (1=vocational, 2=general, 3=academic) and the score on their final exam in math (column *math*).

- a) Investigate whether the type of program influences the number of awards by performing a Poisson regression, without taking variable *math* into account. Estimate the numbers of awards for all the three types of program. Which program type is the best for the number of awards for this model?
- b) For the situation in a), can the Kruskal-Wallis test also be used? If yes, apply the test and comment on the results; if no, explain why this test cannot be used.
- c) Now include predictor *math* into analysis and investigate the influence of the explanatory variables *prog* and *math* (and their interaction) on the numbers of awards. Which program type is the best for the number of awards? Comments on your findings. Estimate the number of awards for the vocational program and math score 55.