Advanced Statistics, Summer 2022

Homework 3, 06.07.2022

Data files are provided for you individually in a zip-file numbered with your homework ID. You find your homework ID in the file "Homework ID.pdf". All datasets are in CSV format (comma separated values) with the first line as heading and as Excel sheets.

You can use any programming language that you are familiar with (e.g. MATLAB, python, R, SPSS, …).

Please double-check your solutions. Wrong solutions will be sent back for corrections.

Please, submit the **solutions and the codes** that you used in **one PDF file** labeled "Your Name - Homework 3" to me at: steffen.gais@uni-tuebingen.de.

Please use "ADVSTAT" in the subject line of your emails.

Send your solutions before the next course on **20.07.2022**.

1. Table1.csv / Table1.xlsx contains data of two groups of subjects. Group A has been tested in the morning, Group B has been tested in the evening.
   - Calculate the significance of the difference between the groups once using the **t-test** and once using the **permutation test**. Use a two-tailed test.
     In Matlab, you can use `xr = reshape(x(randperm(2*N)),N,2);` to generate random assignments of measurements to the two columns. Generate the distribution of the group difference and determine the p value from this distribution.
   - Do the same under the assumption of repeated measures, i.e. that the two columns represent two measures of the same subject.
     Permutation has to take into account that only the two measures of the same subject can be permuted randomly.
   - Please do not use an existing "permutation test" function, but implement it yourself.

2. Data in Table2.csv / Table2.xlsx contains two columns. What is the correlation between these two columns? Use the **bootstrap** to find the interval in which the true value of the correlation coefficient lies with 95% confidence? Is the correlation significantly different from zero?
   In Matlab, you can use `randi(N,N,1)` to generate indices for the bootstrap sample.
   Check the result with the **parametric method** (either using the significance provided by your correlation function or calculate yourself using $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with $df = n - 2$).

3. You have a two-factorial dataset (Table3.csv / Table3.xlsx) measuring treatment outcome with two factors. Factor A (Clinic 1, Clinic 2) and Factor B (Treatment 1, Treatment 2). Generate the design matrix for the GLM (A, B, AxB) and calculate **betas**, **F** and **p values** for both factors and the interaction with the code you used in Homework 2.
   By permuting a column, you can test the hypothesis that this predictor contains no relevant information. Permute the column of X corresponding to **Factor A** repeatedly to generate the null distribution of the beta value of Factor A. Calculate the **p value** by comparing the real beta value with the permutation distribution. Do the same for **Factor B** and the **Interaction**.