Exercises 5

1. Statistical tests in literature

Determine what kind of statistical tests are used in the given parts of the following two articles. (NOTE: You may have to browse through the whole article in order to fully understand the study.)

- Article 2 (Figures 1, 2, 4b & 4e): W. He *et al.*, High-salt diet inhibits tumour growth in mice via regulating myeloid-derived suppressor cell differentiation, DOI: https://doi.org/10.1038/s41467-020-15524-1
- Article 4 (Table 1): M.S. Venäläinen *et al.*, Easy-to-use tool for evaluating the elevated acute kidney injury risk against reduced cardiovascular disease risk during intensive blood pressure control, DOI: https://doi.org/10.1097/HJH.0000000000002282
  - NOTE: No tests are mentioned here, so make comments on which tests might have been used.

2. Multivariable tests with toy data

Apply a multivariable test and obtain a P-value for each of the following data sets. What hypotheses do the tests concern? What can you conclude based on the observed p-values?

- Dice data:
  - 2, 3, 5, 4, 4, 3
  - 4, 2, 3, 5, 2, 3
  - 3, 1, 4, 4, 3, 5
- Nordic countries:
  - Fi, Sw, Fi, No, Sw, Fi
  - No, Sw, No, Fi, Fi, Fi
  - Sw, Fi, No, Sw, Sw, No

3. Statistical tests with simulated data

Consider the data in the file `simulated-data.csv`. It has four columns of 100 values each.

- Assume that the data contains four independent samples. Apply a multivariable test and obtain a P-value to compare the following triplets of samples
  - `F`, `G` and `H`
  - `F`, `G` and `I`
- Assume that the data contains 100 observations with four variables. Find a correlation coefficient and its P-value for the following variable pairs
  - `F` and `G`
  - `F` and `H`
  - `F` and `I`

What hypotheses do the tests concern? What can you conclude based on the observed p-values?

4. Gene expression

The file `gene-expression.data` contains simulated gene expression data, in which the expression levels of genes (rows) were measured for subjects (columns). Use the column names as subject identifiers. All subjects belong to either a control group or a treatment group, as indicated by the identifiers in the files `gene-expression-control.ids` and `gene-expression-treatment.ids`, respectively.

- Use a T-test to find differentially expressed genes (i.e. genes for which the means are different between the control and treatment groups). Adjust the p-values with the Benjamini-Hochberg method.
- Create histograms of the unadjusted and adjusted p-values. Why do these two histograms differ?
- How many differentially expressed genes (i.e. statistically significant differences) are there at the false discovery rate of 0.05?

5. Horse Colic data

- Download the file `horse-colic.data` from the Horse Colic data set available at https://archive.ics.uci.edu/dataset/47/horse+colic.
- Load the data and make sure that the missing values (question marks) are handled correctly.
- Does the data provide statistical evidence that the mean `rectal temperature`, `age` or `pulse` are different between colic horses treated without surgery and those treated with surgery?
- Explain how you reached your conclusions and why your design choices are valid.