

EDDA: Final Assignment

- Submit in Canvas a well readable report before the specified deadline, as one pdf-file. Submitting an R-script file is also allowed, it must contain relevant plot commands.
- The submitted report must follow the Guidelines for assignments posted on Canvas.
- Throughout this assignment tests should be performed using a level of 0.05.

Exercise Trees

The ‘Amsterdamsche Bos’ Forestry wishes to estimate the total wood volume of the trees on its domain. To this end the Forestry has cut a sample of 59 trees of their most prevalent types *Beech* and *Oak*. The volume of these trees alongside with their height and trunk diameter have been measured. The tree height and trunk diameter can be measured in the field without sacrificing the tree. The Forestry hypothesizes that these are predictive of the wood volume. The data (in the file `treeVolume.txt`) is to uncover this relationship, taking into account the tree type.

- a) Investigate whether the tree type influences volume by performing ANOVA, without taking the diameter or height into account. What are the estimated volumes for the two tree types?
- b) Investigate whether tree type influences volume, now including diameter and height as explanatory variables into the analysis, which results in an ANCOVA. How does tree type influence volume? What are the estimated volumes for the two tree types with the (overall) average diameter and height?
- c) How does diameter influence volume? Investigate whether this dependence is similar for both tree types.
- d) Propose a transformation of (a subset of) the explanatory variables that possibly yields a better model (verify this). (Hint: think of a natural link between the response and explanatory variables.)

Exercise Jane Austen

Stochastic models for word counts are used in quantitative studies on literary styles. Statistical analysis of the counts can, for example, be used to solve controversies about true authorships. Another example is the analysis of word frequencies in relation to Jane Austen’s novel *Sanditon*. At the time Austen died, this novel was only partly completed. Austen, however, had made a summary for the remaining part. An admiror of Austen’s work finished the novel, imitating Austen’s style as much as possible. The file `austen.txt` contains counts of different words in some of Austen’s novels: chapters 1 and 3 of *Sense and Sensibility* (stored in the **Sense** column), chapters 1, 2 and 3 of *Emma* (column **Emma**), chapters 1 and 6 of *Sanditon* (both written by Austen herself, column **Sand1**) and chapters 12 and 24 of *Sanditon* (both written by the admiror, **Sand2**).

- a) Discuss whether a contingency table test for independence or for homogeneity is most appropriate here.
- b) Using the given data, investigate whether Austen herself was consistent in her different novels. In case you find that Austen was not consistent, find out where the main inconsistencies are.
- c) Was the admiror successful in imitating Austen’s style? Perform a test including all data. If he was not successful, where are the differences?