Assignment 1

Tests should be performed using level $\alpha = 0.05$, unless stated otherwise. Before writing your assignment report, please read carefully the page Guidelines for assignments on Canvas.

Exercise 1.1 Birthweights

The data set `birthweight.txt` contains the birthweights of 188 newborn babies. We are interested in finding the underlying (population) mean μ of birthweights.

a) (0.5 points) Check normality of the data. Compute a point estimate for μ.
b) (0.5) Derive, assuming normality (irrespective of your conclusion in part a)), a bounded 90% confidence interval for μ.
c) (0.8) An expert claims that the mean birthweight is bigger than 2800, verify this claim by using a t-test (hint: the claim of interest should be represented by the alternative).
d) (1.2) In the R-output of the test from c), also a confidence interval is given, but why is it different from the confidence interval found in b) and why is it one-sided?

Exercise 1.2 Kinderopvangtoeslag

We want to investigate the fraction p of working parents that receives childcare benefit (kinderopvangtoeslag). We have taken a (fictive) sample of size 200. In this sample we encountered 140 parents receiving this childcare benefit.

a) (0.2) Give a point estimate for p.
b) (0.6) Derive a 99% confidence interval for p.
c) (1.2) Test the null hypothesis that the fraction is equal to 75%. What is the outcome of this test if you take $\alpha = 0.1$? And other values of α?

Exercise 1.3 Weather

The file `weather.txt` contains the humidity (%) and temperature (°F) of 60 days.

a) (0.5) Make a relevant summary of the data set, both graphically and numerically.
b) (0.5) Investigate the normality of the temperature graphically.
c) (0.8) Assuming normality (irrespective of your conclusion on part b)), give a 90% confidence interval for the mean temperature. Give an interpretation of this confidence interval.
d) (1.0) Derive the minimum sample size for a 95% confidence interval for the mean humidity such that the confidence interval has at most length 2%.

Exercise 1.4 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 admirer 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 admirer, `Sand2`).

a) (0.2) Discuss whether a contingency table test for independence or one for homogeneity is most appropriate here.

b) (1.0) Investigate using these data 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) (1.0) Was the admirer successful in imitating Austen's style? Perform a test including all data. If he was not successful, where are the differences?