"Data science" course 2025: 2nd set of exercises

Due: 25.2.2025 22:00

Please submit for each question the answer, including plots, **and** the code you wrote to receive this answer via Moodle. **We strongly recommend writing your own code to benefit from this exercise.** If you use an LLM, like ChatGPT, as assistant, you need to indicate how it helped exactly.

The number of points for a correct answer is indicated for each question. Your grade is one (for submitting a paper with your name on it) plus the sum of your points.

1. Chi-squared Distribution (2.5 point max)

A genome-wide association study using adult height as trait has been used to estimate effect sizes β_i for a larger number of single nucleotide polymorphisms (SNPs). The phenotypes and genotypes have been normalised such that under the null hypothesis of having no genetic contribution to height, the effect size estimates would be distributed normally (with zero mean and unit standard deviation).

Consider a specific gene harbouring $N = 20$ SNPs. We define a score for this gene as $S = \sum_{i=1}^{N} \beta_i^2$, so under the null hypothesis this score should be distributed as a sum of χ^2: $S \sim \chi^2_{20}$.

Task: Plot the corresponding PDF using `curve()`. Specify the range of x-values as [0,40] via the argument `xlim`. What is the value of the lower boundary for the top 10 percentile? What is the probability of observing a score larger than 20 under the null hypothesis?
Hints: `curve()` expects a function and its parameters as arguments (here `dchisq()` and the degrees of freedom `df`). The range of x-values in `xlim` can be passed as a vector of interval bounds.

2. T-test (2.5 point max)

A scientist studying adult height of medical students has gathered the following measurements:

Females heights [cm]: 159, 164, 157, 174, 165, 157, 166, 168, 167
Males heights [cm]: 168, 181, 173, 166, 154, 178, 170, 170

A. Given the data is there any evidence for sexual dimorphism with respect to height?

After getting more funding further measurements were made:

Females heights [cm]: 157, 164, 174, 155, 162, 164, 168, 161, 177, 162
Males heights [cm]: 173, 177, 167, 163, 182, 154, 176, 170, 177, 174

B. Are the estimates for mean height from the first and second sampling consistent?
C. Does the combined data set provide any evidence for sexual dimorphism with respect to height?
D. Please comment on what is the likely reason for your observations!

Hints: In this exercise, you may assume that female height and male height have the same variance.