Data Analysis Methods 2015
Assignment 1

Aasa Feragen, Christian Igel

In Assignment 1 you will work with probability, statistics and hypothesis testing.

Assignment 1 will be made available 9.2 and your report should be uploaded to Absalon no later than 23.2.

Guidelines for the assigment:

- **The assignments in DAM must be completed and written individually.** This means that your code and report must be unique and written completely by yourself.
- You are, however, encouraged to discuss in small groups while working on the assignments, If you do this, please list your discussion partners in the beginning of your report.
- For this assignment, you are allowed to use numpy functions such as mean, std, sum. You are, however, not allowed to use built-in functions for correlation or hypothesis testing. If you are in doubt, ask!
- In DAM, we will be using an automatic code checking system for some of the exercises. These exercises are written in blue text below, and will be evaluated solely based on your uploaded code. Your solutions to these assignments should therefore be uploaded to the automatic code checking system. To do this, you should:
  - Use the supplied code templates, and do not rename them.
  - Put all your code templates in a zipped folder called handin1.zip
  - Upload your zipped code folder to the website `http://a00508.science.ku.dk/`, using your KU identifier and a password which is sent to you by email.
  - If you have not received a password by email, contact aasa@diku.dk.

  You can do this at any time, as many times as you want, until the deadline. Your auto-generated feedback report will be emailed back to you. There is a queuing system, so if the feedback email takes time, this is due to your fellow students also uploading their solutions.
- Upload your report as a single PDF file (no Word) named `firstname.lastname.pdf`.

Does smoking affect your lung capacity?

It is well known that smoking is not good for your health, but how can we quantify this in a statistical way? In this assignment you will work with a dataset consisting of information on the lung function, smoking status and demographics of 654 youth and children aged 3-19. See the Appendix A below for a detiled description of the data material, and see in particular Appendix B below for a description of the so-called FEV1 measure, which quantifies lung function.

Exercise 1 (Reading and processing data)**.**

a) Read the data from the file `smoking.txt`, and divide the dataset into two groups consisting of smokers and non-smokers. Write a script which computes the average lung function, measured in FEV1, among the smokers and among the non-smokers, using the template `meanFEV1.py` supplied on Absalon. Upload your code in the automatic code checking system.

b) Report your computed average FEV1 scores. Are you surprised?

Deliverables. a) Uploaded code and b) the average lung functions and a one-liner.

Exercise 2 (Boxplots)**.** Make a box plot of the FEV1 in the two groups. What do you see? Are you surprised?

Deliverables. Figure with box plot and a one-liner describing what you find.

Exercise 3 (Hypothesis testing)**.** Next we will perform a *hypothesis test* to investigate the difference between the FEV1 level in the two populations *smokers* and *non-smokers*. We are going to use a two-sample t-test as in the lecture.

The t-statistic is defined as

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the two sample populations; μ_X and μ_Y are the means of their respective underlying distributions, and s_X and s_Y are the standard deviations of the two sample populations. We do not know the underlying distributions and therefore also not the distribution means μ_X abd μ_Y. However, our *null hypothesis* will be $\mu_X = \mu_Y$, which leaves $(\mu_X - \mu_Y) = 0$ in the equation. The rest of the equation for t can be computed from the two sample populations.

As discussed in the lecture, the t statistic approximately follows a t-distribution with parameter

$$\nu = \left\lfloor \frac{(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y})^2}{(\frac{s_X^4}{n_X^2(n_X-1)}) + (\frac{s_Y^4}{n_Y^2(n_Y-1)})} \right\rfloor$$

where $\lfloor a \rfloor$ is a rounded down to the nearest integer. Assume that you observe a t-value t_0 associated to two sample populations X and Y. Under the assumption that the two populations come from distributions with equal means, the probability of observing as extreme a t_0 is therefore given as the integral

$$\int_{-\infty}^{-|t_0|} f_t(x)dx + \int_{|t_0|}^{\infty} f_t(x)dx = 2 \cdot \int_{|t_0|}^{\infty} f_t(x)dx$$

The integral computes the area shown as shaded in Fig. 1, and the equality comes from the symmetry of the t-distribution. The integral can be computed easily using the `t.cdf` function from the `scipy.stats` library, as also discussed in the lecture.

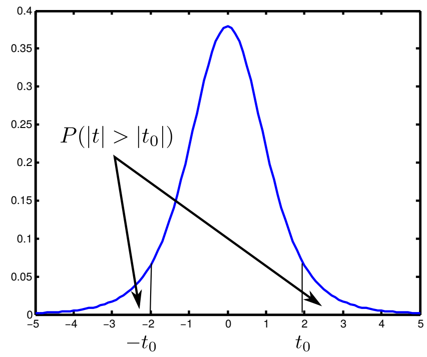

Figure 1: The two-sample t-test measures the probability that two sampled populations indeed stem from the same underlying probability distributions.

a) Based on the supplied template `hyptest.py`, write a script that performs a two-sided t-test whose null hypothesis is that the two populations have the same mean. Use a significance level of $\alpha = 0.05$, and return a binary response indicating acceptance or rejection of the null hypothesis. Please upload your code to the automatic code checker.

b) Report your result and discuss it. Are you surprised?

Deliverables. a) Uploaded code and b) the value of the t-statistic and of the degrees of freedom ν, the returned p-value, whether or not you rejected the hypothesis, and a short discussion of the result.

Confounders

Exercise 4 (Correlation)**.** a) Make a $2D$ plot of age versus FEV1. What do you see?

b) Based on the template `corr.py` provided in Absalon, implement a function which takes two equal-length vectors as input and outputs their correlation, following Chapter 5 in the textbook [1]. Please upload your code to the automatic code checking system.

c) Compute the correlation between age and FEV1, and comment on it.

Deliverables. a) The $2D$ plot and a one-liner; b) uploaded code; c) the correlation and a one-liner.

Exercise 5 (Histograms)**.** Create a histogram over the age of subjects in each of the two groups *smokers* and *non-smokers*. What do you see? Does this explain your results on lung function in the two groups?

Deliverables. The two histograms, and a couple of lines of discussion.

A The data material

The file `smoking.txt`, which can be found in Absalon, contains a 654×6 matrix, where each column corresponds to (in the given order):

- age – a positive integer (years)
- FEV1 – a continuous valued measurement (liter)
- height – a continuous valued measurement (inches)
- gender – binary (female: 0, male: 1)
- smoking status – binary (non-smoker: 0, smoker: 1)
- weight – a continuous valued measurement (kg)

This data is collected from 654 youth and children and each row in the matrix can thus be considered as an observation describing one child/youth.

B Measurement of lung function

Figure 2: Illustration of a spirometry test.

Lung function can be measured using a *spirometry* test, where the person blows in to an apparatus as illustrated in Figure 2, and several parameters are computed based on the result. One of these parameters is the *forced expiratory volume in one second* (FEV1), which measures the volume that a person can exhale in the first second of a forceful expiration after a full inspiration. This measure will be used as an indicator of lung function in this assignment. A decrease in FEV1 generally indicates a decrease in lung function.

References

[1] Joel Grus, *Data Science from Scratch*, 2015.