Introduction to Data Science 2023
Assignment 1

Stella Frank, Thomas Hamelryck, Daniel Hershcovich, François Lauze, Lukas Mikelionis

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

Assignment 1 will be made available Tuesday, February 7th, 12.00 (noon), and your report should be uploaded to Absalon no later than Monday, February 20th, 22.00.

Guidelines for the assignment:

- **The assignments in IDS must be completed and written individually.** This means that your code and report must be written by yourself.
- Exercises in blue are coding exercises and will be evaluated as such: do provide a working code! Code files must be handed in in a zip file. If some code templates are provided in Absalon, please use them.
- Upload your report as a single PDF file (no Word) named `firstname.lastname.pdf`.
- Upload your code as `firstname.lastname.zip`. It should consist of one or more Python scripts (text files with "`.py`" extension) or Jupyter/IPython notebooks (with "`.ipynb`" extension). Do not upload the report and source code in a single zip archive. This makes it impossible to use Absalon's SpeedGrader for annotating the reports.
- We will grade using Python 3.10 and the specific package versions specified in `requirements.txt`. Other versions may work, or may break; using these versions is safest. Using Anaconda, create an appropriate `conda env` with this command:
  `conda create --name IDS-A1 python=3.10`
  then activate it with `conda activate IDS-A1`,
  then install the packages with `pip install -r requirements.txt`

Does smoking affect your lung capacity?

It is well known that smoking is not good for your health, but how can we quantify this statistically? 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 Appendix A below for a detailed 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.

While this assignment is not very heavy on implementation, we want nevertheless you to get familiar with Python, `numpy`, `pandas`, `scipy`, `seaborn` and `matplotlib`. You are not allowed to use any other libraries; you do not need to use all of these libraries (e.g. if you want to use pure `matplotlib` instead of `seaborn`, that's fine.)

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

a) Read the data from the file `smoking.csv`, 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.

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*.

a) 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. You should try do implement it by yourself – though not the CDF of the t-distribution, use `scipy`'s `stats.t.cdf`. If you can't, you may use `scipy`'s `stats.ttest_ind`.

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) Compute the correlation between age and FEV1. Make a scatter plot of age versus FEV1 where non-smokers appear in one color and smokers appear in another.

b) What do you see? Comment your results.

Deliverables. The scatter plot, the correlation, and a one-liner.

Exercise 5 (Histograms)**.**

a) Create a histogram over the age of subjects in each of the two groups, *smokers* and *non-smokers*.

b) 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.csv`, 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.

NB. The `smoking.csv` file does not contain headers, so you can use either numpy or pandas to read in the data. If you use pandas, you can add headers by adding arguments to `read_csv(header=None, names=['column1name', 'column2name', ...])`. Also, values in the file are tab-separated, which means you need to specify `sep='\t'` setting in `read_csv` function call.

B Measurement of lung function

Figure 1: 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 1, 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.