Statistics and Data Analysis - Lab-01

Barbara Żogała-Siudem

2025/2026

1 Software

Mostly: `python` with packages `numpy`, `pandas`, `scipy.stats`, `matplotlib`

1.1 Where can you learn `python` for data analysis?

There are many tutorials online. There is also a free book:

https://datawranglingpy.gagolewski.com/datawranglingpy.pdf

Check also file `python_numpy_pandas.py` for brief reminder.

1.2 Test basics of `numpy` and `pandas` skills

1.2.1 Excersice

Create a random 1-dimensional vector

```
import numpy as np
np.random.seed(13)
x = np.random.normal(0, 2, 20).round(2)
```

(a) Write to console all numbers form vector `x` that consist in $[-2, -1] \cup [1, 2]$.
(b) Write to console the number and fraction of all non-negative values.
(c) Find the mean of all positive values.
(d) Find the value closest to 0 (remember to keep the sign of this value).
(e) Divide `x` into two vectors `x1` – consisting of all values greater than the mean value of `x` and `x2` – consisting of values smaller or equal to the mean.
(f) Standardize vector `x`, i.e. transform its values so that the mean of a new vector is equal 0 and standard deviation is equal to 1.

1.2.2 Exercise

Create a random matrix `x` with n rows and k columns, for $n = 4$, $k = 5$.

(a) Standardize the columns of `x`.

(b) Standardize the rows of `x`.

(c) Find the maximal value in each row of `x`.

(d) Apply the softmax function to each row of an array, i.e.,

$$x_{i,j} \to \frac{\exp(x_{i,j})}{\sum_{l=i}^{k} \exp(x_{i,j})}.$$

(e) Then one-hot decode the values in each row, i.e., find the column number with the value most close to 1. Return a vector of size n.

1.2.3 Exercise

Assume that an $n \times d$ matrix X represents n points in R^d . Write a function that determines the pairwise distances between all the points in X and a given $y \in R^d$. Return a vector $d \in R^n$ with $d_i = \|x_{i,\cdot} - y\|_2$

1.2.4 Exercise

Load `grades.csv` file into `pandas.DataFrame`.

(a) Select only records for which a column `school` has value `SP1`.
(b) Find the average grade for `math` and `eng` in the whole school.
(c) Find the average grades for each `class` for each school subject.
(d) Sort the frame obtained in (c) by the average `math` grade.
(e) Add a new column to the initial data frame indicating if a student's math grade is above the average grade in school or not.

2 Probability distributions

2.1 Density function and probability mass function

Let us start with basic probability distributions:

- $\mathcal{U}(a, b)$ - uniform distribution,
- $\mathcal{N}(\mu, \sigma^2)$ - normal distribution,
- $\Gamma(k, \theta)$ - gamma distribution
- $\chi^2(n)$ - chi-squared distribution
- $\text{Exp}(\lambda)$ - exponential distribution
- $\text{Bin}(n, p)$ binomial distribution
- $\mathcal{P}(\lambda)$Poisson distribution
- $\text{geom}(p)$ - geometric distribution
-

2.1.1 Exercise:

See how specific distributions are represented in `scipy.stats`. How can you find the density ($f_X(t)$) or the probability mass function ($\mathbb{P}(X = k)$) for specific distribution?

Plot (on one plot) densities of the given probability distributions:

(a) standard normal distribution and standard Cauchy distribution. Can you see a difference. Is it a 'big' difference? Try different parameters for normal distribution, can you pick such ones, so that those densities look alike?

(b) gamma distribution with shape parameter equal to 30 and scale parameter equal to 0.5 along with normal distribution which resembles the given gamma distribution (what parameters would you choose and why?)

2.1.2 Exercise (GS):

Let us assume that a student came totally unprepared for an exam. The exam is in a form of a quiz that consists of 20 questions, and each question is given 4 answers from which exactly one is correct.

- What is the probability of obtaining 0/20, 5/20, 10/20, 15/20, 20/20 points?
- What distribution can be used to model this situation?

Plot the probability mass function for this distribution with appropriate parameters.

2.2 Excersise (KO):

Which probability distributions can appropriately model the following processes:

(a) The number of lorries driving through a state road section between 3 and 4 a.m.?
(b) The number of cars driving out of Warsaw by A2 motorway between 3 and 4 p.m.?
(c) The time of waiting for the first speeding vehicle after speed measurements begin on a given road section?
(d) The time of waiting for the first 3 speeding vehicles?
(e) The time of waiting for the first 100 speeding vehicles?
(f) The number of speeding vehicles in a quarter?
(g) The number of speeding vehicles in a day?
(h) The distance between a certain scalar value and its unbiased estimate?
(i) The distance between a certain n-dimensional vector value and its unbiased estimate?
(j) The number of heads obtained during 5 flips of a loaded coin?