Exercises 4: Statistics

Introduction to data science and Python programming for medical research (2025)

1 Reading data using Pandas

Download the data files and put them in your working directory or any other directory where you are able to find them from within Python.

Use Pandas to load and summarize data.

- Load the file `dataframe-test-data.csv` into a Pandas DataFrame.
- Use the `describe()` function of the DataFrame to print a summary of the data.

2 Blood pressure medication

Patients at risk of hypertension with high systolic blood pressure were treated with a new blood pressure lowering drug. In the same study the patients were also given a placebo treatment at some other point in time.

Blood pressure was measured and reported in mmHg.

Some of the patients did not show up to do measurements of their blood pressure at both occasions.

- Load the file `blood-pressure-trial.csv` into a Pandas DataFrame.
- Choose a way to handle the missing data. Use the build in functionalities in Pandas.
- Extract the data from the two columns in the DataFrame into two numpy arrays.
- Perform a suitable statistical test to investigate if the treatment had any effect on these patients.
- **(Optional)** Plot the two data distributions using scatter, boxplots or violinplots.

3 Diabetes trial

A new diabetes medicine has been tested in a random clinical trial.
477 patients were enrolled, and completed follow up, in the treatment arm.
512 different patients were enrolled, and completed follow up, in the control group.
Blood glucose levels were measured and recorded in mmol/L.

- Load the data `diabetes-trial-one.csv` into Pandas DataFrame.
- Extract one Series for each column in the DataFrame and remove any missing values. Print the sample size of each Series.
- Extract the data from the two Series into two numpy arrays and perform a suitable statistical test. Report the result of the test.
- **(Optional)** Plot the two data distributions using scatter, boxplots or violinplots.

4 Extended diabetes trial

A second diabetes medicine was added to the study which in the end contained three arms, two with different treatments and one for control.
285 patients were enrolled, and completed follow up, in the new treatment arm.
Blood glucose levels were measured and recorded in mmol/L.

- Load the data `diabetes-trial-two.csv` into Pandas DataFrame.
- Use the `describe()` function of the DataFrame to print a summary of the data.
- Extract one Series for each column in the DataFrame and remove any missing values. Print the sample size of each Series.
- Extract the data from the **three** Series into **three** numpy arrays and perform a suitable statistical test to see if there is any significant difference between any of the groups in the study. Report the result of the test.
- **(Optional)** Plot the **three** data distributions using scatter, boxplots or violinplots.

5 Unbiased variance

Use pure Python to write a function which calculated the **unbiased** variance of the data in a 1D array.
Tip: Look at the examples from today's lecture.

- Use the random number generator in Numpy to generate a 1D array of 100 values sampled from a normal distribution centered on 2 and with standard deviation 3.
- Use your function to calculate the unbiased variance of the array of random numbers.