

# Data Analysis - 2022

## Exercise sheet no 2: Statistical variance and correlation

---

27. September 2022

If possible, use the "skeleton\_sheet2.py" as a starting point for this exercise.

### Exercise 1: Ironman II (16 Points)

Write a script that reads the file `ironman.txt` (see exercise sheet no. 1).

**Remark:** For this exercise, implement the exact function that is asked for, such as mean, standard deviation, histogram etc. yourself as an educational exercise (instead of using `np.mean` and `np.std`) etc. Apart from this exercise sheet, use already existing functions.

- (a) Calculate the mean, uncertainty of the mean, variance, and standard deviation of the
- age distribution
  - total time distribution

directly from the data. (3 points)

*Hint: implement your own algorithm; you can compare your values to `np.mean`, `np.var`, `np.std`*

- (b) Compute the average of the total time for people younger and older than 35 years. Calculate the uncertainty on these two times. Can you conclude that one group is faster than the other? (2 points)
- (c) Histogram and plot the age and total time distributions with error bars according to the statistical uncertainty in each bin. (4 points)
- Hint: implement your own algorithm for the binning, use `plt.bar` for plotting; you can compare your values to `np.histogram`*
- (d) Calculate the mean, variance and standard deviation of the distribution from the histograms of the age and total time. Compare the results with those you got in a). Use different bin widths and comment on what you observe. (2 points)
- Hint: if you did not implement the histogramming flexible enough above, you can also use `np.histogram` to create a histogram here.*
- (e) Calculate the covariance and correlation coefficient between

- the total rank and the total time
- the age (in 2010) and the total time
- the total time and the swimming time
- the cycling time and the running time

Take a look at the scatter plots from exercise sheet no. 1 and compare the results with the results you got. (4 points)

Convert the total time from minutes to seconds and calculate the covariance and correlation coefficient between the age and the total time again. Which of the two changes and which stays the same? (1 point)

## **Exercise 2: Weighted Average (4 Points)**

You have been hired as a medical physicist, and your first task is to report on the level of radiation in the room where the containers of some radioactive sources, used for radiotherapy research, are stored.

You measure the data stored in `radiation.txt` which can be found on the webpage. The measurements are taken with different dosimeters, and under different conditions, and thus have different uncertainties (2nd column). Assume that all measurements are independent from each other. The level of radiation is measured in mSv/h, millisievert per hour<sup>1</sup>.

- (a) Calculate the average radiation level, in mSv per year, and the associated uncertainty. (3 points)
- (b) Based on your result, argue and explain whether the level of radiation in that room is compatible with the natural background radiation, which has been measured to be 2.4 mSv/y. You can assume that the uncertainty on this value is negligibly small compared to the uncertainty on your measurements. (1 point)

**Deadline for submission: Friday, 7th October 2022, 14:00**

**Form: Submission of solutions as a single python script directly to**

**da@physik.uzh.ch. Make sure to adhere to the "exercise rules" and notify us if you did not complete the full setup as described in "pycharm anaconda instructions".**

---

<sup>1</sup>The Sievert is the unit of ionizing radiation dose in the International System of Units, and it represents the effects of radiation on the human body.