Exercise Set 1

Goals

1) Learn to quantify centre and asymmetry of a dataset.
2) To emphasize that the first step in data analysis is to visualize datasets (boxplot, histogram,...).
3) Familiarize yourself with Python for large datasets.

1 Steel cable resistance

Steel cables for an elevator are tested for tensile resistance. Data (in MPa):

10.1	12.2	9.3	12.4	13.7	10.8	11.6	10.1	11.2	11.3
12.2	12.6	11.5	9.2	14.2	11.1	13.3	11.8	7.1	10.5

a) Compute mean and median by hand and discuss them.

b) Draw a box plot, including outliers if they are present outside the 1.5*IQR limit.

You are the process engineer. You revisit all cables which were measured and find that one of them had not been mounted correctly - it was the one which gave rise to the value of 7.1MPa. You decide to remove it from the dataset.

c) Recompute the first two questions.

d) Naturally, the cable strength is very important for your industry partners. Given the empirical distribution function, what minimal strength can you expect in 90% of your cables? How about in 95%?

2 Grades at a previous statistics exam

2.5	4	4.25	4.75	3.75	4	5	3.5	4.25	4.25	4
4.75	4	5.5	4.5	5.5	4.5	4.5	6	3.5	4	4.75

One goal of the exercise is to practice these calculations and plots by hand. You can afterwards verify your results with Python

a) Compute mean and median.

b) Draw a box plot and a histogram: are there any outliers?

c) Does it make sense to remove them as in the previous exercise?

d) Compute the cumulative distribution: what is the fraction of students that have

   1) a passed?
   2) a received 5.25 or more?

3 Wealth inequalities in the world

One area of current societal debate is the economical injustice on the globe, as well as the role globalization plays in it. To that end, the OECD carries out regular thorough studies of the distribution of global wealth, and its evolution with time.

This is an example of a moderately complex dataset, slightly too big to handle by manual computation. Use Python for this exercise.

a) Compute mean and median of the dataset OECD-World-Wealth-Report.csv available on Moodle. Follow the Python example from the course, import the file, and do this analysis.

b) Draw the box plot, are there any outliers? Is the mean or the median more representative of the center of the distribution?

c) Does it make sense to remove any outliers as in the previous exercise?

d) How many percent of people receive the upper 50% of all income? 10%? 1%?

e) Assuming you are a Swiss student living on 1500CHF/month. Compute your yearly income. Find your quantile: how many percent of people live on less than you?

f) Assume you finished your studies at EPFL in the MX section. Look at the *rapport d'insertion professionnel epfl* on Moodle find the 2016 report. What is the average yearly salary? Again find your quantile.

Data from the OECD - available on Moodle: Cumulative distribution of world net income *per year* corrected for the Swiss purchasing power. As all quantiles, the ECDF(X) quantifies the fraction of the global population that make X or less CHF per year.