Data Analytics – Exercises

(Week 06)

In these exercises, you will learn:

- to simulate data
- to perform random sampling and derive measures from samples
- to perform an unpaired two-sample t-test
- to perform a one-way Analysis of Variance (ANOVA)

In the data analytics process model, these exercises cover part of the steps "Statistical data analysis and/or Modeling" and "Evaluation & Interpretation" (see figure 1). Results of the exercises must be uploaded as separate files (**no .zip files!**) by each student on Moodle. Details on how to submit the results can be found in the tasks below.

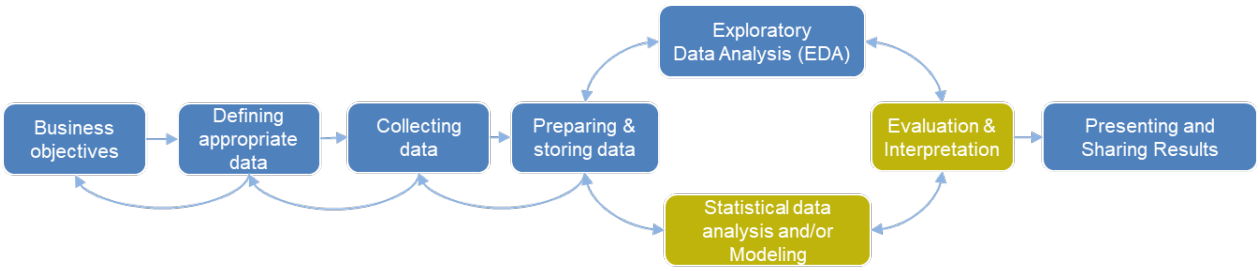

Figure 1: Data analytics process model (see slides of week 01)

Task 1

In this exercise, you will learn what's the difference between a population and a sample. For this, you will use random sampling and bootstrapping. The tasks are:

a) Run the Jupter notebook 'population_and_samples.ipynb' step by step and try to understand what the code does.
b) Find an own practical example with realistic mean and standard deviation. This could be, for example, diameters of trees in a forest, fuel consumptions of cars in a country, weights of fish in a pond, prices of mobile phones, etc.
c) Run the Jupyter notebook with your example.
d) In the single sample and bootstrap section, vary the sample size and/or number of samples and see how the shape of the histograms changes.

To be submitted on Moodle:

- The Jupyter notebook as html-file 'population_and_samples.html' with the changes made according to b), c) and d)

Task 2

In this exercise, you will learn to perform an unpaired two-sample t-test. The tasks are:

a) Run the Jupter notebook 'unpaired_two_sample_t-test.ipynb' step by step and try to understand what the code does.
b) Add a section 'Compare rental apartment prices in the cities of Winterthur and Zuerich' to the Jupyter notebook.
c) Create two subsets of the rental apartments data frame. The first subset must contain the prices per m2 of rental apartments in the city of Winterhur. The second subset must contain the prices per m2 of rental apartments in the city of Zürich. Use the following code to create the subsets:

```
df_winti = df.loc[df['bfs_name'] == 'Winterthur']['price_per_m2']
df_zueri = df.loc[df['bfs_name'] == 'Zürich']['price_per_m2']
```

d) Perform an unpaired two-sample t-test based on the two samples to investigate, whether the prices per m2 of rental apartments differ between the cities of Winterthur and Zuerich. Use a significance level of 5%.
e) Interpret the results of the statistical test in the notebook.

To be submitted on Moodle:

- The Jupyter notebook as html-file 'unpaired_two_sample_t-test.html' with the changes made according to b), c), d) and e).

Task 3

In this exercise, you will learn to perform a one-way ANOVA. The tasks are:

a) Run the Jupter notebook 'analysis_of_variance_ANOVA.ipynb' step by step and try to understand what the code does.
b) Create three subsets (groups) with prices per m2 of the apartments data frame analog to task 2. The subsets must include municipalities with:

   - `low population density (pop_dens <= 600)`
   - `moderate population density (pop_dens > 600 & pop dens <=1500)`
   - `high population density (pop dens >1500)`

c) Perform a one-way ANOVA based on the three subsets to investigate, whether the **`prices per m2`** of rental apartments differ between the three groups of municipalities. Define a null and an alternative hypothesis. Choose a significance level to confirm or reject the null hypothesis. Interpret the results of the statistical test in the notebook.

d) Perform a one-way ANOVA based on the three subsets to investigate, whether the **areas** of rental apartments differ between the three groups of municipalities. Define a null and an alternative hypothesis. Choose a significance level to confirm or reject the null hypothesis. Interpret the results of the statistical test in the notebook.
e) Perform a one-way ANOVA based on the three subsets to investigate, whether the **number of rooms** of rental apartments differ between the three groups of municipalities. Define a null and an alternative hypothesis. Choose a significance level to confirm or reject the null hypothesis. Interpret the results of the statistical test in the notebook.

To be submitted on Moodle:

- The Jupyter notebook as html-file 'analysis_of_variance_ANOVA.html' with the changes made according to b), c), d) and e).