Data Analytics – Exercises

(Week 07)

In these exercises, you will learn:

- to statistically analyze contingency tables using the Chi-squared test
- to perform a Pearson correlation including a significance test

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.

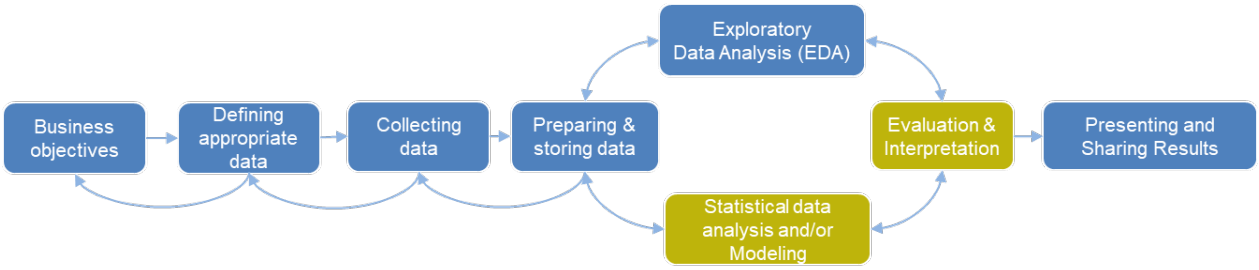

Figure 1: Data analytics process model (see slides of week 01)

Task 1

In this exercise, you will learn to statistically analyze contingency tables using the Chi-squared test. The tasks are:

a) Run the Jupyter notebook 'analysis_of_contingency_tables.ipynb' and try to find out, what the Python code does.

b) The example Chi-squared test in the Jupyter notebook uses only part of the full contingency table (see the 2x2 table below). In detail it tests the null hypothesis that there is no difference in the number of apartments in the specified area- and price categories. This type of analysis can be used, for example, to investigate whether the quantities of specific apartments differ on the market.

	apartment_id	
price_cat	25 - 32	>= 33
area_cat		
50 - 64	54	41
65 - 84	87	55

c) Extent the latter analysis with a Chi-squared test based on the full contingency table. Note that the full contingency table is already included in the Jupyter notebook, i.e.:

	apartment_id			
price_cat	0 - 19	20 - 24	25 - 32	>= 33
area_cat				
0 - 49	0	4	22	66
50 - 64	5	30	54	41
65 - 84	14	102	87	55
85 - 99	37	76	42	16
100 - 125	31	57	37	14
> 125	38	20	13	7

d) Follow this link to a research study in which a Chi-squared has been used to study animal behavior. In detail, the research is based on the social interactions of two spotted hyenas (*Crocuta crocuta*) which were observed over 40 hours in a U.S. zoo. The 4x2 contingency table below has been taken from this research study.

	Female	Male	*Total*
Fear	0	4	*4*
Greeting	7	6	*13*
Aggression	9	0	*9*
No Response	15	10	*25*
Total	*31*	*20*	*51*

e) In the Jupyter notebook, you will find a section 'Applying the Chi-squared test to animal behavior'. Go to this section and perform a Chi-squared test based on the table above (the table is already included in the Jupyter notebook) to see if the calculations presented in the research study are correct.

f) Calculate the expected frequencies of the contingency table "by hand". You can do this in Python based on the Total values of columns and rows of the contingency table by typing:

```
31*4 / 51 = 2.4314
31*13 / 51 = 7.902
31*9 / 51 = 5.4706
…
```

Compare your expected frequencies with those from the chi2_contingency() function in Python.

g) Calculate the Chi-square test statistic of the contingency table "by hand". You can do this in Python based on the observed and expected frequencies by typing:

```
((0-2.4314)**2/2.4314) + ((7-7.902)**2/7.902) + … + … =
```

Compare your Chi-square value with the Chi-square value from the chi2_contingency() function in Python.

h) In the Jupyter notebook, state, whether the results of the original study are correct or not. The link to the research study can be found above.

To be submitted on Moodle:

- The Jupyter notebook as html-file 'analysis_of_contingency_tables.html' extended according to c), d), e), f), g) and h).

Task 2

In this exercise, you will learn to perform correlation analysis. The tasks are:

a) Run the Jupyter notebook 'correlation_analysis.ipynb' and try to find out, what the Python code does.

b) For this one time, let's study the car market. On Moodle you will find a dataset 'autoscout24_data.csv'. In the Jupyter notebook section 'Correlation analysis based on car data', part of the car data is already imported to a data frame.

c) Explore the numerical car data (i.e.: Price, Kilometer, PS) exploratively using a paired scatterplot (see previous exercises).

d) Perform correlation analyses inclusive significance tests for the variables:
   - Price versus Kilometer
   - Price versus PS
   - Kilometer versus PS

To be submitted on Moodle:

- The Jupyter notebook as html-file 'correlation_analysis.html' extended according to c) and d).