Theoretical Exercises

Excercise 5.1: (Theoretical) Marginal frequencies

For the retailers from the electronics sector, the following two characteristics were the following two characteristics were determined:

- X: number of employees
- Y: number of televisions sold

The following contingency table is obtained:

X \ Y	3	5	6	8	10	11	12	15
2	3	2	0	0	0	0	0	0
3	1	2	2	0	0	0	0	0
5	1	0	4	4	1	0	0	0
8	0	1	4	5	3	5	2	0
10	0	0	0	1	1	0	3	5

a) Determine the marginal frequencies of X and of Y as well as the conditional relative frequencies of Y under the condition $X = 8$.

b) Among the retailers with eight employees, compute the proportion of those who have sold ten television sets.

c) Compute the proportion of retailers who have eight employees and have sold no more than ten television sets.

d) Compute the proportion of retailers who have sold no more than ten televisions.

Excercise 5.2: (Theoretical) Correlation coefficient

For the companies in the "Chaos" working group of a business association, the following table contains the annual sales (feature Y) and advertising expenditure (feature X) in millions.

Company	1	2	3	4	5	6	7
annual sales	90	110	90	60	120	70	90
advertising expenditure	7	6.5	9	10	9.5	6	8

a) Create a scatter plot of the data.

b) Determine the empirical variance $\tilde{s}_{XY}$ and the correlation coefficient r_{XY} according to the lecture. Interpret the result!

Excercise 5.3: (Theoretical) Steiners theorem

Show that the Identity of Steiners theorem holds: Let $x_1, \ldots, x_n \in \mathbb{R}$, $a \in \mathbb{R}$. Then, there holds

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2.$$

Practical Exercises

Excercise 5.4: (Practical) Anscombe's quartet

Consider the data of the Anscombe's quartet. Load the data with *pandas* and verify that all data pairs have (nearly) the same descriptive statistics (compare mean of x and y, empirical variance of x and y and correlation between x and y)

	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

Excercise 5.5: (Practical) Ames housing

Consider the Ames housing dataset. Write a program that takes two arbitrary columns of the dataset, then computes the contingency table and the corrected Pearson contingency coefficient. What are the resulting correlations?

Hints:

- A contingency table can be computed with *pandas.crosstab*
- The *scipy* package can compute the Pearson contingency coefficient with *scipy.stats.contingency.association*