Exercises for Chapter 5

5.1 Pearson correlation analysis yields to the following findings. Explain these results and suggest ways to improve the analysis.

a) Ice cream sales are highly correlated with accidents in swimming pools.

This may be a spurious correlation: Both features may depend on the season. In summer, ice cream sales are higher than in winter, and so is swimming pool usage and hence also accidents in swimming pools. The analysis may be improved by using partial correlation of ice cream sales and accidents in swimming pools wthout seasonal influence.

b) Average flight delays are highly correlated with the winning numbers in the lottery drawing.

This correlation may be pure coincidence. The analysis may be improved by considering additional data.

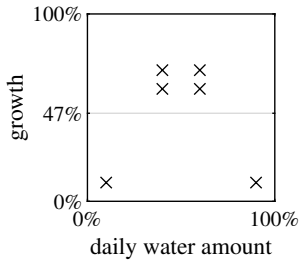
c) The fuel consumption of a car is only weakly correlated with the speed.

The dependecy between fuel consumption and speed may be nonlinear. Pearson correlation is linear and may therefore indicate weak (linear) correlation. The analysis may be improved by using nonlinear correlation methods such as the chi-square test for independence.

5.2 Six flowers receive daily water amounts between 0 and 100%. After one month they have grown according to the following table:

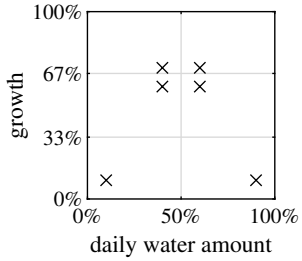
flowers number	1	2	3	4	5	6
daily water amount	10%	40%	40%	60%	60%	90%
growth	10%	60%	70%	60%	70%	10%

a) What is the Pearson correlation between daily water amount and growth?



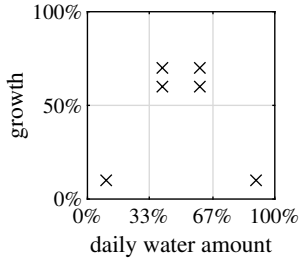
Due to symmetry, the data can be linearly best approximated by a horizontal line at the average growth of about 47%. Since this line is horizontal, the linear correlation will be zero.

b) Would the chi-square test for independence with two bins for daily water amount (0,50,100%) and three bins for growth (0,33,67,100%) indicate a high or low correlation?



This yields the bin counts $H = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$. All bins are equally filled. This indicates a very low (nonlinear) correlation.

c) Would the chi-square test for independence with three bins for daily water amount (0,33,67,100%) and two bins for growth (0,50,100%) indicate a high or low correlation?



This yields the bin counts $H = \begin{pmatrix} 0 & 4 & 0 \\ 1 & 0 & 1 \end{pmatrix}$. The bins have very different counts. This indicates a very high (nonlinear) correlation.

d) How do you interpret the three results?

Pearson correlation is not suitable to detect nonlinear correlations. The number of bins is a crucial parameter for the chi-square test for independence.

5.3 The chi-square formula for two variables is

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n \cdot h_{ij} - h_i^{(1)} \cdot h_j^{(2)}\right)^2}{h_i^{(1)} \cdot h_j^{(2)}}$$

What will be the formula if we consider three variables?

If the three variables are independent, then the probability of a data point falling into the bin combination h_{ijk} is equal to the product of the probability of falling into bin $h_i^{(1)}$, the probability of falling into bin $h_j^{(2)}$, and the probability of falling into bin $h_k^{(3)}$, so $\frac{h_{ijk}}{n} = \frac{h_i^{(1)}}{n} \cdot \frac{h_j^{(2)}}{n} \cdot \frac{h_k^{(3)}}{n} \Rightarrow h_{ijk} = \frac{h_i^{(1)} \cdot h_j^{(2)} \cdot h_k^{(3)}}{n^2} \Rightarrow n^2 \cdot h_{ijk} =$

$h_i^{(1)} \cdot h_j^{(2)} \cdot h_k^{(3)}$, hence the formula for the chi-square test for independence for three variables is $\chi^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{t} \frac{\left(n^2 \cdot h_{ijk} - h_i^{(1)} \cdot h_j^{(2)} \cdot h_k^{(3)}\right)^2}{h_i^{(1)} \cdot h_j^{(2)} \cdot h_k^{(3)}}$