Course of Study Bachelor Computer Science	Exercises Statistics WS 2021/22
Sheet V	

Descriptive Statistics - Linear Regression

1. For the X, Y data:

x_i	2	6	3	4	5
y_i	3	7	4	7	6

draw a scatterplot and compute

   (a) covariance

   (b) coefficient of correlation

   (c) regression line: criterion variable Y and predictor variable X

   (d) regression line: criterion variable X and predictor variable Y

2. For a certain class, the relationship between the amount of time spent in exercises (X) and the test score (Y) was examined.

x_i	y_i
10	5
9	5
9	4
11	6
10	7
10	5
6	3
10	4
8	5
12	7
9	4
4	2
12	8

   (a) Draw a scatterplot of the data.

   (b) Is there a positive or a negative association between X and Y?

   (c) Compute the covariance and the coefficient of correlation.

   (d) Compute the regression line $Y = a + bX$.

   (e) Find the predicted test score for someone with 8 units of time spent in exercises.

   (f) Interpret the values of the parameters of the regression line.

   (g) Compute the proportion of variation explained by the simple linear regression.

   (h) Add a the point (20,0) to the data. Inspect how this additional point influences the linear regression.

Descriptive Statistics - Contingency Tables

1. **Exercise 4.1 from Heumann, Schomaker, p. 90**
   A newspaper asks two of its staff to review the coffee quality at different trendy cafes. The coffee can be rated on a scale from 1 (miserable) to 10 (excellent). The results of the two journalists X and Y are:

Cafe	X	Y
1	3	6
2	8	7
3	7	10
4	9	8
5	5	4

   (a) Calculate Spearman's rank correlation coefficient.

   (b) Does the coefficient differ depending on whether the ranks are assigned in a decreasing or increasing order?

   (c) Suppose that the coffee can only be rated as eiter good (> 5) or bad (≤ 5). Do the chances of good rating differ between the journalists?

2. The following 3x2 contingency table categorizes students according to whether or not they pass an introductory statistics course and their level of attendence:

	Course Result		
Attendance	Pass	Fail	Totals
Over 70%	40	10	50
30%-70%	20	10	30
Under 30%	10	10	20
Totals	70	30	100

   (a) Calculate the expected values in case of no association between attendance and course result.

   (b) Calculate χ^2, C and C_{corr}.

3. Exercise 4.4 from Introduction to Statistics and Data Analysis from Heumann, Schomaker, p. 91
   The famous passenger liner Titanic hit an iceberg in 1912 and sank. The data set Titanic of the package titanic contains the survival data of the passengers:

- A total of 325 passengers travelled in first class, 285 in second class, and 706 in third class. In addition, there were 885 staff members on board.
- Not all passengers could be rescued. The following were not rescued: 122 from the first class, 167 from the second class, 528 from the third class and 673 staff.

(a) Determine the contingency table for the variables “travel class” and “rescue status”. This can be done by the following steps

- Create a tibble raw_data with the columns class (possible values: first, second, third, staff) and state (possible value: rescued and not.recued) and fill the rows with the corresponding number of pairs given by the survival data from above.
- Count the number of rescued and non rescued in the classes and the crew by applying table() to the two columns of raw_data.

(b) Use the contingency table to summarize the conditional relative frequency distribution of rescue status given travel class. Could there be an association of the two variables?

(c) What would be the contingency table table from a) look like under the independence assumption? Calculate χ^2, C and C_{corr}. Is there any association between rescue status and travel class?

(d) Combine the categorie “first class” and “second class” as well as “third class” and “staff”. Create a contingency table based on these new categories. Determine and interpret χ^2, C and C_{corr}.