Exercise 2

Do the tasks below and return your answers as a single Word or PDF file. In each task, it is explained which material is intended to be used. The materials can be found either in this file or in the Excel file in Moodle. The file in Moodle has three tabs: people, time series, grocery stores and factory.

1. Use the material on the people tab. Create a frequency table of the variable age so that you classify the age into categories of five years so that the lower limit of the youngest age category is 20 years. Also prepare a histogram describing the frequency distribution. According to this, approximately how old is the typical person in this data set?
2. The last page has a frequency table, which was prepared based on a customer satisfaction survey.

Calculate mean, median and mode for the variable describing satisfaction with customer service when possible. If it is not possible to calculate some of these, explain why not. Justify your answers.

1. Use the material on the time series tab, which shows the monthly prices of a certain raw material

at the beginning of 2000. Prepare four graphs: the first graph shows the values of the time series as they are, the second graph shows the moving average of five consecutive values, the third graph has an exponential smoothing (α=0.4) and the fourth graph shows the percentage changes of the successive values of the original time series. Explain why there are so many positive and negative values in the graph showing percentage changes.

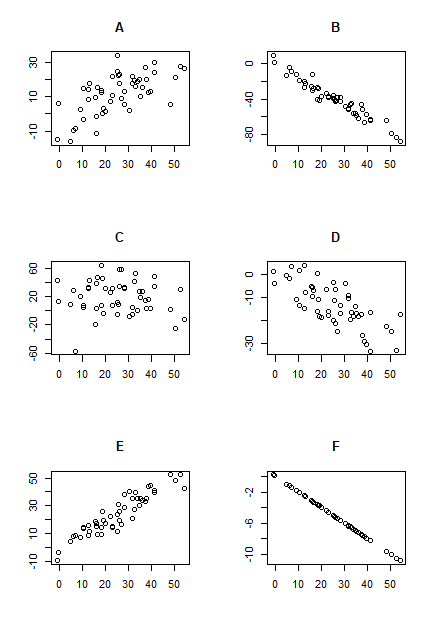
1. The next page shows scatterplots A, B, C, D, E and F. The Pearson correlation coefficients associated with these are

|  |  |
| --- | --- |
| a. | -0.97 |
| b. | -0.74 |
| c. | 0.93 |
| d. | 0.65 |
| e. | -1,00 |
| f. | -0.08 |

in some order. Match the correct correlation coefficient to each figure and briefly justify your answer.

1. Use the material on the grocery store tab. The data set shows grocery stores' one-day customer numbers and one-day sales in euros. Calculate the Pearson correlation coefficient between the variables number of customers and sales. Also prepare a scatterplot between these two variables and add a regression line into it. Attach this figure to your answer. According to the regression model, what is the sales of a grocery store with 350 customers per day? And according to the regression model, how much does one additional customer increase sales in euros? Justify your answer with the equation of the regression line.
2. Use the material on the people tab. Create a cross-tabulation of variables hometown and salary class. Based on this, is there a relationship between variables hometown and salary class? If so, what kind? Use the chi-squared test to find out whether variables hometown and salary class are independent of each other. Explain the steps of testing and the conclusion.
3. Calculate the confidence interval for the mean of height using 99 percent confidence rate. Explain the steps of your calculation. Use the material on the people tab.
4. Battery-operated devices have been serviced at the factory. A random sample of devices has been

selected and battery life (in minutes) has been measured before and after maintenance. The results of the measurements are on the factory tab. Use a suitable test to find out if the maintenance had a statistically significant effect on battery life. In your answer, write the null hypothesis, the alternative hypothesis, the p-value and its verbal interpretation.



*Figure 1: Scatter plots for Exercise 4.*

*Table 1: Table for Exercise 2.*

|  |  |
| --- | --- |
| **Satisfaction with customer**  **service** | **Number of answers** |
| Very dissatisfied | 14 |
| Somewhat dissatisfied | 16 |
| Neither satisfied nor dissatisfied | 15 |
| Somewhat satisfied | 42 |
| Very satisfied | 58 |
| **Total** | **145** |