Biomedical Data Science & AI

# Exercise sheet 2 - Introduction

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

**Due date: May 4th**

**Questions**

**Exercise 1 - Understanding Your Dataset (total: 13 points)**

Load the *processedClevelandData.csv* dataset. The features for the dataset are described in the *featureDescription.csv* file.

1. Perform data cleaning procedures such that your final dataset is usable in the following questions. **(2 points)**
2. For each type of diagnosis of heart disease, find the following for the resting blood pressure: **(2 points)**
    a. Mean
    b. Median
    c. Standard deviation
3. Use Spearman's and Kendall correlation to quantify the correlation between age and the following.
    a. Resting blood pressure
    b. Serum cholesterol level
    c. Maximum heart rate achieved

    Also, which variable(s) are most correlated with age? Illustrate with heatmaps. **(3 points)**
4. From your understanding, which of the features can be labeled as discrete random variables and which features as continuous random variables? **(1 point)**
5. Describe the distribution for the values of the "thalach" feature? Illustrate with a plot. **(1 point)**
6. Plot the frequency of “Sex” variable in the dataset and describe what you observe in the plot. Similarly plot and describe the 'ca' feature for the male participants. **(2 points)**
7. Detect outlier patients for features "trestbps" and "chol". Illustrate with plots. **(2 points)**

**Exercise 2 - Probability (total: 4 points)**

1. Suppose a discrete random variable, MMSE (Mini mental state examination), cognitive test measured for Alzheimer’s disease (AD) has the following probability mass function:

| x | 5 | 8 | 14 | 22 | 24 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|
| pr(X=x) | 0.05 | 0.27 | 0.16 | 0.17 | 0.03 | 0.12 | 0.07 | 0.13 |

Find the probability that MMSE:

a) at least 22 **(1 point)**

b) at least 14 and at most 28 **(1 point)**

2. A company produced antibody testing kits for COVID-19. The false positive rate of the test is known to be 3%. What is the probability to find at least 2 false positive results within 35 tested patients? **(2 points)**

**Exercise 3 - Hypothesis Testing (total: 8 points)**

Using the processed dataset from question 1 answer the following questions.

1. Are all the criteria for carrying out a t-test to identify a significant difference in the age of patients who have heart disease and those who don't, met? **(3 points)**
   - ○ If the criteria is met, carry out a t-test using Python.
   - ○ And if not, point out the unmet conditions for the variables, and mention a possible solution in-order to combat this issue.
2. Identify if women are significantly more likely to get heart disease than men? **(2 points)**
3. Inform yourself about $\chi^2$−test. And using $\chi^2$−test, identify if there is a significant association between exercise induced angina (exang), and the slope of the peak exercise ST segment (slope)? **(3 points)**

**Note:** Please write functions in python wherever possible. Please document and comment your code using this style guide.