

Exercise sheet 3

SoSe2025

Prof. Dr. Holger Fröhlich, Jonas Botz, Diego Valderrama, Anna Weller

Due date: April 29th

Questions

Exercise 1 – Vampire Cup (total: 7 points)

The Vampire Cup is an annual blood donation marathon that takes place around Halloween and is a competition to collect as much blood and encourage as many people as possible to donate in order to raise awareness of blood donation and counteract the shortage of blood reserves. Students of the Bonn University pharmacy department participate in the cup in cooperation with the blood donation center at the university hospital.

The two most important characteristics of a person's blood are the so-called rhesus factor (positive or negative), and the blood group. The following shows which blood groups can be subordinated under which rhesus factor, and how the blood groups are distributed in Germany:

Rh+		Rh-	
A+	37 %	AB-	6 %
O+	35 %	B-	6 %
B+	9 %	O-	2 %
AB+	4 %	A-	1 %

1. On the first day of the cup, 42 people came in to donate blood. Considering the donors as being sampled at random from the population, what is the probability of gathering
 - a. 21 blood units of blood group O (Rh+ and Rh-)? **(1 point)**
 - b. 6 blood units typed as Rh-? **(1 point)**
2. Of the 42 donated blood units, blood groups are observed with the following frequencies: A+: 16, O+: 14, B+: 5, AB+: 1, AB-: 2, B-: 3, O-: 1, A-: 0. What is the probability of gathering 42 blood units with exactly this composition of blood groups? **(2 points)**
3. One hour before the blood donation center closes on the final day of the cup, the organizers calculate that 7 blood units are still missing to reach the self-imposed goal of 300l total blood donations. Over the course of the cup, it has been observed that on average, 5 people came in per hour. Assuming a suitable statistical distribution, how likely is it that at least 7 people donate blood in the one remaining hour? **(3 points)**

Exercise 2 - Vaccine Adverse Event Analysis (total: 6 points)

A new malaria vaccine is tested on $n = 1200$ participants in a clinical trial. Historical data suggest the probability of severe allergic reaction (anaphylaxis) is $p = 0.003$ per individual.

1. Assuming a binomial distribution, determine the probability of a) exactly 3 severe allergic reactions and b) more than 5 severe allergic reactions in the clinical trial. **(1 point)**
2. Choose a suitable value for λ to use the Poisson distribution for the computation of a) exactly 3 severe allergic reactions and b) more than 5 severe allergic reactions in the clinical trial. Compare the results to the results obtained in 1. **(2 points)**
3. Assuming again a binomial distribution, consider different pairs of n and p :
 - $n = 25, p = 0.3$
 - $n = 50, p = 0.15$
 - $n = 100, p = 0.075$
 - $n = 1000, p = 0.0075$

In each case, choose a suitable λ for the Poisson distribution and compare the probability mass functions of the binomial and the Poisson distribution on the interval $[0, 16]$. Under which conditions does the Poisson distribution closely approximate the binomial distribution? **(3 points)**

Exercise 3 - Hypothesis Testing (total: 5 points)

Use the dataset `processedClevelandData.csv` for the following questions:

1. Is it feasible to carry out a t-test to identify a significant difference in the age of patients who have heart disease and those who don't? (Hint: Shapiro-Wilk test) **(3 points)**
2. Inform yourself about the χ^2 test and use it to identify if there is a significant association between exercise induced angina (`exang`) and the slope of the peak exercise ST segment (`slope`). **(2 points)**

Exercise 4 - Hypothesis Testing (total: 7 points)

This exercise illustrates a gene expression data set with its normally distributed values. Consider the gene expression data of the *Golub* dataset. Load the file `"golub.csv"`. It contains gene expression data of 3051 genes from 38 tumor mRNA samples. The expression data is organized in a matrix where rows correspond to genes and columns to samples. The tumor class of the columns is given in the file `"golub.cl"`. The names of the genes (rows) are given in `"golub.gnames"`.

1. Calculate the sample mean $\hat{\beta}$ of all genes in the pooled expression matrix. Use these means to determine the overall mean β_0 by just taking the average. **(1 point)**

2. Based on the t-statistic defined as follows:

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\text{s.e.}(\hat{\beta})}$$

obtain the 50 most significant genes. [Hint: $\hat{\beta}$ is the sample mean of a particular gene] **(1 point)**

3. Perform two-sampled student t-tests for all genes comparing the distributions for ALL and AML. **(1 point)**
4. Based on the p-values obtained in 3., obtain the top 10 genes with the lowest p-values. **(1 point)**
5. Inform yourself about the multiple testing problem. Apply one appropriate method to deal with it to the results from 3. and explain how it works. **(3 points)**