

8. Problem sheet for Statistical Data Analysis

Exercise 1 (8 points)

A mathematician applies for a job at a bank. She/he has to complete an aptitude test consisting of $n = 8$ multiple-choice questions. For each the questions the mathematician has to choose between 4 possible answers and only one is correct.

1. The bank wants to strongly avoid the unfavourable scenario of hiring an incompetent person while they can live with not hiring a competent person. Write up a pair of suitable hypotheses.
2. HR has decided that a mathematician is going to be hired, if she/he answers at least 6 questions correctly. Determine the error of type 1 i.e., H_0 is rejected despite it being true.
3. Construct a test with sensitivity $\alpha = 0.05$. The mathematician answered 4 questions correctly. Is he going to be hired under this set-up?
4. If the mathematician answers questions correctly with a probability of 0.85. How large is the probability of the above test for an error of type 2 (H_0 is not rejected but is not true)?

Exercise 2 (8 Points)

Given the following regularised regression problem:

$$\hat{\beta}^{Ridge} = \arg \min_{\beta \in \mathbb{R}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \quad (1)$$

Show that the solution is

$$\hat{\beta}^{Ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^\top \mathbf{y} \quad (2)$$

Exercise 3 (8 Points)

In order to test customer satisfaction with a given service, we conduct a survey and define a random variable Y_i as follows:

$$Y_i = 1 \quad \text{if customer } i \text{ is satisfied} \quad (3)$$

$$Y_i = 0 \quad \text{if the customer } i \text{ is not satisfied} \quad (4)$$

Given the identical and independent Bernoulli distributed samples y_1, \dots, y_n with

$$P[Y_i = 0] = \theta \tag{5}$$

$$P[Y_i = 1] = 1 - \theta \tag{6}$$

we want to test the hypotheses $H_0 : \theta = \theta_0 = 0.52$ et $H_1 : \theta = \theta_1 = 0.48$

- Construct the likelihood of the observations y_1, \dots, y_n and explain the rejection region of H_0 (i.e., error of type 1) from the test of Neyman-Pearson (for the numerical application, we will choose a risk of the first kind $\alpha = 0.1$).
- Determine $P[H_0 \text{ rejected} | H_1 \text{ true}]$

Exercise 4 (8 Points)

Use the data sets from problem sheet 5 exercise 1 and compute $\hat{\beta}^{Ridge}$ and the corresponding empirical bias and empirical variance. Compare with the solution, variance and bias of the standard regression solution and comment on it.