University of Wrocław: Data Science

Theoretical Foundations of Large Data Sets, List 2

1. Global testing for the expected value of the Poisson distribution

   a) Let $X_1, \ldots, X_n$ be the sample from the Poisson distribution. Consider the test for the hypothesis

   $$H_0 : E(X_i) = 5 \ \text{ vs } \ H_1 : E(X_i) > 5 \ ,$$

   which rejects the null hypothesis for large values of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Write the function in R to calculate the p-value for this test.

   b) Consider 1000 of the same hypothesis for $n = 100$ and calculate p-values. Draw histogram of p-values and discuss their distribution.

   c) Consider the meta problem of testing the global hypothesis $H_0 = \bigcap_{j=1}^{1000} H_{0j}$ and use simulations to estimate the probability of the type I error for the Bonferroni and Fisher tests at the significance level $\alpha = 0.05$. Why the probability of the type I error of Fisher's test might be slightly different from α ?

   d) Use simulations to compare the power of the Bonferoni and Fisher tests for two alternatives:

   - *Needle in the haystack*

     $$E(X_1) = 7 \ \text{ and } \ E(X_j) = 5 \ \text{ for } j \in \{2, \ldots, 1000\}.$$

   - *Many small effects*

     $$E(X_j) = 5.2 \ \text{ for } \ j \in \{1, \ldots, 100\} \ \text{ and } \ E(X_j) = 5 \ \text{ for } \ j \in \{101, \ldots, 1000\}.$$

2. Let $X_1, \ldots, X_{100000}$ be iid random variables from *N(0,1)*. For $n \in \{2, \ldots, 100000\}$ plot the graph of the function

   $$R_n = \frac{\max\{X_i, i = 1, \ldots, n\}}{\sqrt{2 \log n}}.$$

   Repeat the above experiment 10 times and plot the respective trajectories of R_n on the same graph. Comment on the results in relation to exercise 4.

3. Let $Y = (Y_1, \ldots, Y_n)$ be the random vector from $N(\mu, I)$ distribution. For the classical needle in haystack problem: $H_0 : \mu = 0$ *vs* H_1 : *one of the elements of* μ *is equal to* γ consider the statistics L of the optimal Neyman-Pearson test

   $$L = \frac{1}{n}\sum_{i=1}^{n} e^{\gamma Y_i - \gamma^2/2},$$

and its approximation

$$\tilde{L} = \frac{1}{n}\sum_{i=1}^{n}\left[e^{\gamma Y_i - \gamma^2/2}\mathbf{I}_{\{Y_i < \sqrt{2\log n}\}}\right] .$$

For $\gamma = (1-\epsilon)\sqrt{2\log n}$ with $\epsilon = 0.1$ and $n \in \{1000, 10000, 100000\}$ use 1000 replicates

a) to draw the histograms of L and $\tilde{L}$;

b) to calculate variances of L and $\tilde{L}$ under the null hypothesis;

c) to estimate $P_{H_0}(L = \tilde{L})$ and compare to the theoretical value.

4. Use simulations to find the critical value of the optimal Neyman Pearson test and compare the power of this test and the Bonferoni test for the needle in haystack problem with $n \in \{500, 5000, 50000\}$ and the needle $\gamma = (1+\epsilon)\sqrt{2\log n}$ with $\epsilon \in \{0.05, 0.2\}$.

   Remark: In case of the Neyman Pearson test work with the log-likelihood ratio $\log L$ rather than with L.

5. Draw

   a) one graph with cdfs of the standard normal distribution and the Student's distribution with degrees of freedom $df \in \{1, 3, 5, 10, 50, 100\}$. Comment on the result.

   b) one graph with cdfs of the standard normal distribution and the standardized chi-square distribution with degrees of freedom $df \in \{1, 3, 5, 10, 50, 100\}$. Please, recall that the standardization is of the form

   $$T = \frac{\chi^2_{df} - df}{\sqrt{2df}}$$

   Comment on the result.