Multiple hypothesis testing for coronavirus disease in Ukraine

I. Kosareva[1] and R. Yamnenko[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: kosarevaivanna777@knu.ua

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: rostyslav.yamnenko@knu.ua

Abstract

In this work, we will consider the data on coronavirus disease in Ukraine by region from the beginning to May 2023[https://index.minfin.com.ua/ua/reference/coronavirus/ukraine/]. The purpose of this study is to find out if the proportion of people who got sick and recovered is equal to 0.5 in each region. The data is arranged in a contingency table and multiple hypothesis testing is planned to be used for its analysis.

Multiple hypothesis testing is a statistical technique used to test multiple hypotheses simultaneously. When it comes to the analysis of contingency tables, multiple hypothesis testing can be used to compare the proportions of different categories across two or more groups.

Multiple hypothesis testing for contingency tables is an important topic in statistics. The need for accurate and efficient analysis of complex data, while avoiding false positives and erroneous conclusions, underscores the relevance of this topic.

The problem with testing multiple hypotheses simultaneously is that the likelihood of making a Type I error (rejecting a true null hypothesis) increases with the number of tests performed. This can lead to spurious or false positive results, which can be misleading and lead to incorrect conclusions. Multiple testing procedures for the contingency table are designed to control the overall error rate while still allowing for the detection of true signals in the data.

In this research, we use the chi-square test to calculate the p-value. The well-known formula for the chi-square statistic used in the chi square test is

$$\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \quad (1)$$

where O_i is the observed value, E_i is the expected value, "i" is the "ith" position in the contingency table and c is the degrees of freedom.

The next step is to perform Monte-Carlo simulation on the data and calculate the p-value using the chi-squared statistic for each shuffle. As we test multiple hypotheses false positive rate has to be controlled with the false discovery rate(FDR) method.

After all these procedures we expect to obtain a lower p-value and get the output of the test about the proportion of people who were infected and recovered.

Keywords: multiple hypothesis testing, contingence table, Monte-Carlo simulation, chi-square test, coronavirus disease in Ukraine.