Inferring from multiple hypothesis tests

Izabela Fedorczyk

March 2023

1 Single Hypothesis

To make an introduction to the subject of this review which is the multiple hypothesis tests it is good to start with a description of single hypothesis.

Basically in the case of single hypothesis, based on some statistics two hypothesis are tested: the null one - H0 and the alternative one - H1. To decide whether to reject or not the hypothesis H0 in favor of hypothesis H1 some rejection rule is applied. However as it happens in the real world this kind of testing isn't error-free.

There are two types of errors: the one which occurs when the hypothesis 0 is rejected even though it shouldn't be as it is in fact true and this type of error is called the **Type I error** or alternatively **false positive** and the one which occurs in the opposite situation when the null hypothesis H0 isn't rejected although it should - **Type II error/ false negative**. While testing hypothesis there is set a level of significance (usually denoted as α) which is the acceptable chance of a false positive. When the p-value[1] is lower than α the null hypothesis is being rejected.

2 Multiple Hypotheses

In case when we have a lot of hypothesis tests, following the same rejection rule as described in the previous section will cause a very high probability of making at least one Type I error (higher than the nominal level used for each test individually).

We can gain some intuition by comparing this case to multiple dice throws. If one throws the dice once, the chance of getting let's say 3 is 1/6. However if we throw dice n times (let's assume n is a big number) then the probability of getting at least one 3 is significantly higher as it equals $1 - (5/6)^n$.

[1] calculated from a statistical test number describing probability of finding a particular set of observations if the null hypothesis is true

3 Methods for multiple hypothesis adjustment

To overcome the problem with high chance of getting Type I error, focus is on the overall error rate instead of considering individual tests separately. In this regard the most discussed in the literature types of such rates are:

1. Family-wise error rate (FWER) - the probability of at least one Type I error when all null hypotheses are true

2. False discovery rate (FDR) - expected proportion of false rejections (false positives) to all rejected hypothesis (all positive test results).

There are methods of adjustment of multiple hypotheses based on both these rates.

3.1 Controlling FWER

One of the most popular corrections related to the control of FWER is the **Bonferroni correction**. When the FWER rate can be represented by the equation

$$FWER = 1 - (1 - \alpha)^n \tag{1}$$

(where n is the number of tests) in this type of correction the main idea is to set the level of significance for each individual test as α (understood as the cumulative significance level for all tests) divided by number of test (n). This can be summarized with a formula:

$$correctedFWER = 1 - (1 - \frac{\alpha}{n})^n < \alpha \tag{2}$$

.With this correction we obtain the FWER rate always less than or equal the chosen significance level α. However it is based on assumption that all null hypothesis are true. It this condition isn't met then this correction is too strict as nevertheless it can solve the problem with high probability of getting Type I error it can result with many false negatives instead.

3.2 Controlling FDR

Alternative method (which is less strict than the one described before) is the correction related to the control of FDR rate.

The first presented and most common procedure for controlling this rate is called **Benjamini and Hochberg Procedure**. This approach requires certain steps:

1. sorting the p-values of all tests in ascending order

2. Assigning rank to each p-value according to the level of significance (the smaller value gets the rank 1, second smallest gets 2 and so on)

3. calculating critical values for each p-value as:

$$criticalvalue = \frac{i}{n} * q \tag{3}$$

(where i corresponds to the rank, n is the number of tests and q is the overall chosen level of FDR)

4. finding the largest p-value which is less than corresponding to it critical value (lets call it founded p-value)

5. rejecting all null hypothesis related to p-values smaller than founded p-value.

In the result the FDR rate is decreased.

While we focusing on Biology is it also good to mention **Storey Tibshirani method**. If it comes to genomics studies we can't just eliminate possibility of receiving any false positive as it may result in a loss of a lot of important information. Here also the FDR rate is helpful. It tells us how many of features classified as significant are truly null and in context of biology helps us identify false leads.

The **Storey Tibshirani procedure** calculates a specific FDR for each individual test in multiple hypothesis test using following formula:

$$\frac{\pi_0 \cdot n \cdot t}{count\,(p_i \leq t)} \tag{4}$$

(where π_0 is the proportion of hypotheses which are truly null, n is the number of tests, t is the p-value associated with the considered hypothesis, and in the denominator is the count of all p-values from all tests which are less than or equal to considered one (t))(The proportion of null hypotheses (π_0) is derived from the density of values in the uniformly distributed section of p-values).

In this way there is specific measure of significance measured which can help with problem of loosing a lot of significant information from the data.

References

Austin, S. R., Dialsingh, I., & Altman, N. (2014). Multiple hypothesis testing: A review. *J Indian Soc Agric Stat*, *68*(2), 303–14.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445.

[1]The review was mainly based on articles Austin et al. (2014) and Storey & Tibshirani (2003)