Important concepts

Hypothesis testing is a method of statistical inference used to decide whether the formulated hypothesis is true or not for specific data. In statistics, there are two types of hypotheses: **null hypothesis (H0)** and **alternative hypothesis (H1)**. The null hypothesis typically states that there is no relationship between variables and the alternative hypothesis is its opposite. Alternative hypotheses may be **one-sided**, which means that it specifies the direction of the effect or **two-sided**.

P-value- ranges from 0 do 1 and allows us to determine the probability of observing a difference between collected sample means (assuming that the null hypothesis is true).

In hypothesis testing, we compare the calculated p-value with the established **significance level (α)**. It determines the threshold for deciding whether the null hypothesis is true or not. Based on research questions and data type, we need to choose a suitable statistical test and then calculate a test statistic and its associated p-value. Comparing the calculated p-value with specified significance level allows us to decide whether to reject or accept the null hypothesis:

1) If the p-value is smaller than or equal to α, reject the null hypothesis in favor of the alternative hypothesis.
2) If the p-value is greater than α, accept the null hypothesis.

In single hypothesis testing it is possible to make two types of errors:

1) **Type I error (*false positive, FP)***- occurs when the null hypothesis is rejected when it is actually true (incorrect rejection of H0)
2) **Type II error (*false negative, FN*)**- occurs when the null hypothesis is not rejected when it is not true (incorrect rejection of H1)

Basic statistical tests

T-test- compare the means of two groups and determine if there is a statistically significant difference between them. There are two main types of t-tests:

1) **T-test for independent samples**- used to compare the means of two independent samples. In this case, assumptions include the normal distribution of the data and the homogeneity of variances between groups.
2) **T-test for paired samples**- used to compare the means of two variables measured on the same subjects. Assumptions include the normality of the differences between the observations.

Analysis of variance (ANOVA)- assesses whether there are statistically significant differences in means between three or more groups by comparing the variability between groups to the variability within groups. The most popular variations of ANOVA are:

1) **One-way ANOVA**- comparing the means of the groups on a single categorical variable. Assumptions include normal distribution of the data within groups, homogeneity of variances between groups and independence of observations.
2) **Two-way ANOVA**- extends one-way ANOVA by allowing for the comparison of means across two independent variables.

Chi-square test is used to determine if there is a statistical association between two categorical variables by comparing the observed frequencies of certain categories in a contingency table to the expected frequencies under the assumption that the variables are independent.

Multiple hypothesis tests

When conducting multiple hypothesis tests independently, the probability of making at least one Type I error (rejecting H0, when it is actually true) is much higher than the nominal value used for each test, especially when the number of tests is large. This is because the probability of making zero Type I errors in m independent tests is $(1 - \alpha)^m$, where α is the rejection level in each test, which is smaller than the same probability for a single test $(1 - \alpha)$. That's why multiple hypothesis testing procedures aim to make individual tests more conservative so as to minimize the number of Type I errors while controlling an overall **error rate q**. Commonly used Type I error rates are:

1) **Family-wise error rate (FWER)**- denotes the probability of at least one Type I error,
2) **False discovery rate (FDR)**- denotes the expected proportion of false rejections.

When all null hypotheses are true, FDR is equivalent to FWER. However, if the number of true null hypotheses is smaller than the total number of hypotheses, the FDR can be smaller or equal to the FWER. **Modern approaches in multiple hypothesis testing focus on controlling the false discovery rate (FDR) instead of the false-wise error rate (FWER).**

Controlling FDR when p-values are continuous

For all procedures, test m independent null hypotheses $(H_{01}, H_{02},..., H_{0m})$ with corresponding p-values $p_1, p_2,..., p_m$, under the assumption of ordered p-values, which means that $p_1 \leq p_2 \leq ... \leq p_m$.

Benjamini and Hochberg procedure

To control FDR at a specified level q, the BH algorithm rejects all null hypotheses where:

$\{ H_{0(i)}: i \leq max(k: p_{(k)} \leq \frac{i \cdot q}{m}) \}$

It was the first procedure for controlling FDR, but it is still one of the most commonly used methods.

Benjamini and Liu procedure

Unlike the BH algorithm, the BL algorithm is a ***step-down*** procedure, which means that it starts with the least significant p-values and progressively decreases the threshold. The BL algorithm consists of the following steps:

1) For $i = 1,..., m$ calculate the critical values $\delta_i = 1 - [1 - min(1, \frac{m \cdot q}{m - i + 1})]^{1/(m-i+1)}$
2) Reject the null hypotheses $H_{0(1)}, H_{0(2)}, ..., H_{0(k-1)}$, where $k = min\{i: p_{(i)} > \delta_i\}$

Controlling FDR when p-values are discrete

In this case, p-values are no longer uniformly distributed. Moreover, in multiple hypothesis testing with discrete data, the distribution of p-values may vary by test. One of the first ideas for addressing multiple hypothesis testing with discrete data, was based on the use of

midP-values instead of p-values in the BH algorithm. *MidP-values* are calculated as the average of the observed p-value and the next smallest possible p-value and they yield a more uniform distribution under the null hypothesis.

Adaptive procedures

Adaptive procedures involve estimating the number of true null hypotheses, denoted as $m_0 = \pi_0 m$, where π_0 represents the proportion of true null hypotheses among all hypotheses. By replacing m with the estimated m_0 in the Benjamini-Hochberg or Benjamini-Liu algorithms, FDR can be precisely controlled at the specified level q.

Estimating m_0 for continuous tests

Estimating m_0 comes down to estimating π_0 and for this purpose we can use several different methods.

Pounds and Cheng method

The estimator of π_0 is given by:

$$\hat{\pi}_0 = \begin{cases} \min(1, 2\bar{p}) & \text{for two-sided tests} \\ \min(1, 2\bar{t}) & \text{for one-sided tests} \end{cases}$$

The proposed estimator is biased upward, but the bias is negligible when the proportion of true alternative hypotheses is small or when π_0 is close to 1.

Location based estimator

The estimator is given by the following equation:

$$\hat{\pi}_0 = \frac{(1/m)\sum_{i=1}^{m}[-log(1-p_i)]^n}{n!}$$

The LBE is obtained from the expectation of transformed p-values, using the transformation $[-\ log(1\ -\ P)]^n$. The LBE provides a bias-variance balance and, because of its relatively low variance it often performs better than other estimation methods.

Nettleton's method

This method estimates the number of true null hypotheses (m_0) by assessing the proportion of observed p-values conforming to a uniform distribution. The algorithm consists of the following steps:

1) Partition the interval [0, 1] into B beans of equal width
2) Assume all null hypotheses are true, setting the initial estimate of m_0 as m
3) Calculate the expected number of p-values for each bin based on the current estimate of m_0
4) Starting from the leftmost bin, sum the number of p-values in excess of the expected until a bin with no excess is reached
5) Use the sum to update the estimate of m_1, and then use that to update the estimate of $m_0 = m - m_1$
6) Return to *Step 3* and repeat the procedure until convergence is reached.

The recommended number of beans is $B = 20$.

Estimating m_0 for discrete tests

As mentioned before, there is a lack of FDR procedures for discrete data, apart from using *midP-values*. However, the non-uniformity of p-values in discrete data can be addressed in the initial step of estimating the number of true null hypothesis (m_0) followed by utilizing an adaptive FDR method.

Pounds and Cheng method

The proposed estimator is similar to the one proposed for continuous data and is given by:

$$\hat{\pi}_0 = \begin{cases} \min(1, 2\bar{p}) & \text{for two-sided tests} \\ \min(1, 8\bar{t}) & \text{for one-sided tests} \end{cases}$$

This estimator is conservative, but robust for discrete tests.

Multiple hypothesis testing plays a crucial role in genomewide studies, where statistical hypotheses are tested on each of thousands of genes or other genomic features. Approaches based on controlling false discovery rate can be further modified and applied in various biological studies, including the detection of differentially expressed genes or binding sites of transcriptional regulators.