Homework 1

Lorenzo Bozzi,
{lorenzo.bozzi}@aalto.fi
Course Name: High-Throughput Bioinformatics
Course Code: CS-E5875
Due date: 11/11/2023

I. Introdution

THE following paper reports the solution to the first assignment.

The experiments for this assigment were conducted on a local machine, specifically a MacBook Air M1 (2020) equipped with 8 GB of RAM and the Apple-designed M1 chip. We utilized Jupyter Notebook (version 6.5.2) as the programming environment and employed R as the primary programming language.

Experiments based on the generation of random numbers have been conducted by setting the seed to 0511.

II. Exercise 1

Let X_k with $k \in \mathbb{N}$ be independent Bernoulli random variables representing the outcome of each toss, and we state that $X_k \in \{0,1\}$, where 1 corresponds to heads. Let $q = \mathbb{P}(X_k = 1)$.

Consider another random variable Y_k for $k \in \mathbb{N}$ defined as:

$$Y_k = \sum_{h=1}^{k} X_h. \tag{1}$$

The new random variable Y, defined in Equation (1), is called binomial and counts the number of ”heads” obtained in k tosses. The probability mass function of the binomial follows a symmetric bell-shaped curve, as illustrated in Figure 1.

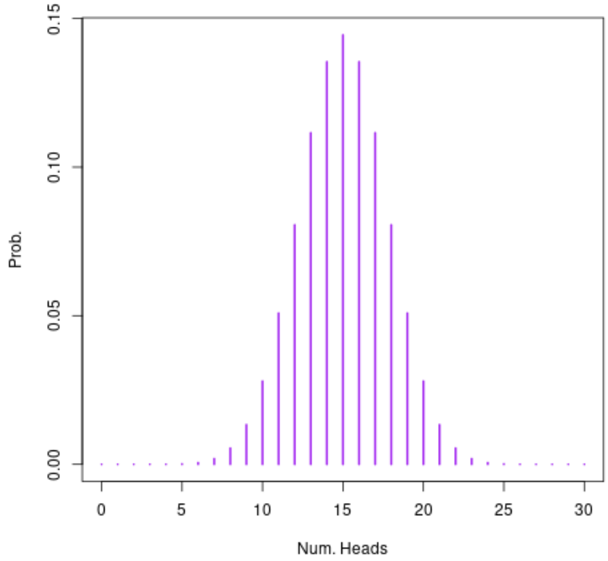

Figure 1: Binomial Distribution.

In the experiment, we are interested in evaluating the probability of obtaining at least k victories in n tosses, that is:

$$\mathbb{P}(Y_n \geq k) = \mathbb{P}\left(\sum_{i=1}^{30} X_i \geq 20\right) = \sum_{i=k}^{30} \binom{30}{i} q^k (1-q)^{30-i}. \tag{2}$$

In particular, for the case where $n = 30$ and $k = 20$, assuming a fair coin (unbiased), evaluating Equation (2) yields a value of 0.049.

However, it is legitimate to question how one can understand the probability p through the obtained victories. Therefore, let's consider two hypotheses to determine if the coin is biased, $p \neq 0.5$: the first assuming that the coin is unbiased, denoted as H_0, and the other assuming that the coin is biased, denoted as H_1. Let's choose $p = 0.05$ as the confidence level; then, we could perform a hypothesis test.

The idea is straightforward since, in our case, we only need to calculate the probability mass function (as shown in Figure 1) and then compute the 95% quantile.

From this, it is observed that 19 tosses are needed to reject the null hypothesis H_0 with this confidence.

III. Exercise 2

In this exercise, the focus is on hypothesis testing, specifically exploring the Bonferroni and Benjamini-Hochberg (BH) correction methods. An example of hypothesis testing is illustrated in the previous exercise with the coin toss game.

In these problems, the concept of the p-value is fundamental. If t is the realisation of a random variable T, then:

$$p = \mathbb{P}\{\text{”test statistic at least as extreme as } t\text{”}\}. \tag{3}$$

In essence, p represents the probability of obtaining a result at least as extreme as the observed one, assuming the null hypothesis is true. It provides a measure of evidence against the null hypothesis. A low p-value (in our case 0.05) suggests sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. A high p-value indicates insufficient evidence to reject the null hypothesis, though it doesn't prove the null hypothesis is true; there might be test power issues or other factors at play.

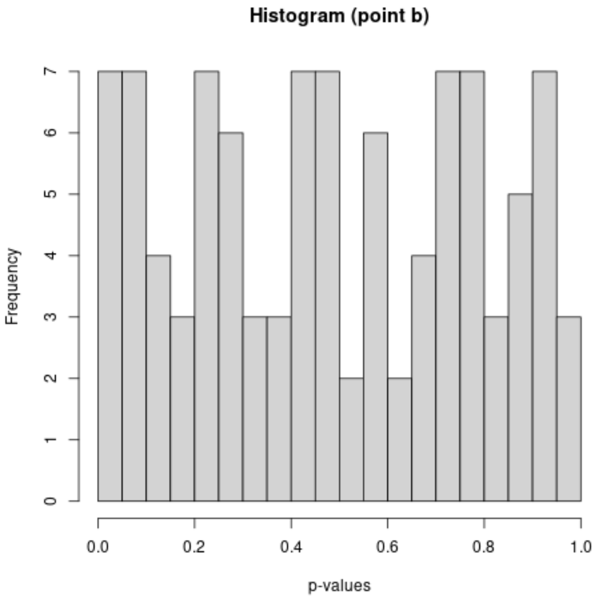

Figure 2: p-values.

In the exercise of interest, data is consistently assumed to follow a normal distribution. Hence, the t-test becomes crucial: the null hypothesis asserts that the means are equal ($\mu_1 = \mu_2$), while the alternative hypothesis suggests they are different ($\mu_1 \neq \mu_2$).

A. *Homogeneous dataset*

Let's assume that we have generated normally distributed expression data for 100 genes, divided into two groups, A and B, each with 8 replicates. The parameters for the data generation are set as follows:

$$\begin{cases} \mu_A = & \mu_B = 0 \\ \sigma_A^2 = \sigma_B^2 = 3 \end{cases} \tag{4}$$

At this point, for each gene, a hypothesis test $H_0\colon \mu_A = \mu_B$ is performed using the two sided t-test when the alternative hypothesis is $H_A\colon \mu_A \neq \mu_B$, resulting in the identification of 7 genes that are statistically significant at level $\alpha = 0.05$(adopted as our significance level), leading to a rejection of the null hypothesis, In Figure 2, the histogram representing the p-values is shown.

Nevertheless, given how the points have been generated, there is no reason to reject the null hypothesis; in fact, the null hypothesis should not be rejected.

In any case, rejecting the null hypothesis does not inherently imply the acceptance of the alternative hypothesis, as this would lead to a false positive.

Hence, it is crucial to perform a correction of the p-value, and the first implemented method is the Bonferroni correction (Bonf. in Figure 3). The second method employed is the Benjamini-Hochberg correction (in Figure 4, new p-values are displayed). The Bonferroni correction is known for its conservative nature, unlike the latter. However, both methods led us to confirm that those seven genes were false positives, as depicted in the aforementioned figures.

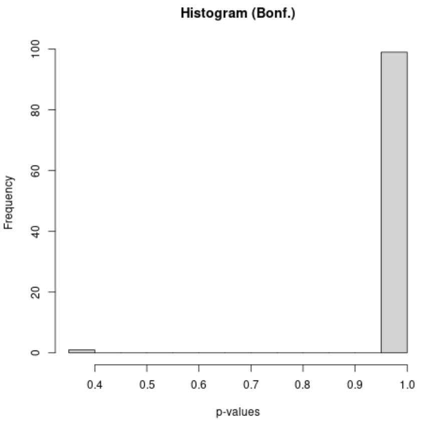

Figure 3: We can see p-values corrected with Bonferroni's method.

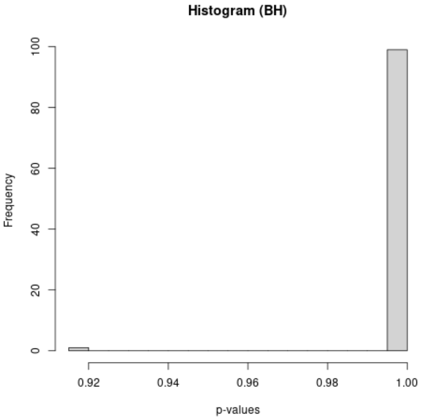

Figure 4: We can see p-values corrected with Benjamini-Hochberg's method.

Table I: Comparisons.

p-values	Rejected(=False Negative)
Original	7
Bonf.	0
BH	0

In summary, as shown in Table I, when applying a correction method, it is observed that the p-values of these 7 genes at a significance level α will be adjusted.

B. *In-Homogeneous dataset*

In this scenario, two groups of 100 genes, denoted as A and B, are constructed. These groups are further partitioned into two subsets, namely $A = A_1 \dot{\bigcup} A_2$ and $B = B_1 \dot{\bigcup} B_2$, with the following parameters:

$$\begin{aligned}\mu_{A_1} =& \mu_{A_2} = \mu_{B_1} = 0\\ \sigma^2_{A_1} =& \sigma^2_{A_2} = \sigma^2_{B_1} = \sigma^2_{B_2}\\ \mu_{B_1} =& 0.\end{aligned}$$

For each gene, a hypothesis test is conducted with $H_0\colon \mu_A = \mu_B$ and alternative $H_A\colon \mu_A \neq \mu_B$, implementing a two-tailed test. It is reasonable to assume that sub-datasets A_1 and B_1 have the same mean by construction. Therefore, if the hypothesis is rejected, false positives are being observed. On the other hand, if H_0 is accepted for sub-datasets A_2 and B_2, these samples will be false negatives. Initially, there is an absence of false negatives and only 6 false positives.

Subsequently, the previously developed correction methods are applied. The Bonferroni method manages to correct all false positives but generates 13 false negatives. On the other hand, the BH method eliminates all false positives but generates false negatives. Regarding FDR, Bonferroni seems to be equivalent to BH, as shown in Table II.

If we consider the alternative hypothesis $H_A\colon \mu_A < \mu_B$, we might expect a behavior similar to the previous one, as the alternative hypothesis remains valid only on the last 2 genes. However, it should be noted that this violates the requirement of a one-tailed test. Again, the BH method seems to have good performance, but Bonferroni consistently has a lower FDR, as highlighted in Table III.

Remember that FDR is defined as:

$$\text{FDR} = \frac{\#\{\text{false positives}\}}{\#\{\text{false positives}\} + \#\{\text{true positives}\}}.$$

Table II: $H_A\colon \mu_A \neq \mu_B$.

	Rejected	**FP**	**FN**	**VP**
Original	26	6	0	20
Bonf.	7	0	13	7
BH	19	0	1	19

Table III: $H_A\colon \mu_A < \mu_B$.

	Rejected	**FP**	**FN**	**VP**
Original	27	7	0	0
Bonf.	12	0	8	8
BH	21	1	0	20

In the end, several tests were performed by varying the dataset; for further details, please refer to the associated notebook.