List 3

Kamil Zaborniak

Contents

Introduction **2**

Procedures 3

Multiple testing tasks **3**

$n = 20$ 4

$n = 5000$ 4

Simulation Description 5

Two-step Procedure (Fisher) **5**

Simulation Description 6

Comparison of the Brownian bridge and the empirical process **7**

Simulations 7

Result of simulation 7

Introduction

In the following report we will consider the problem of multiple testing of statistical hypotheses. Let $X_1, ..., X_n$ be samples of random variables such that $X_i \sim N(\mu_i, 1)$, where $i \in \{1, \ldots, n\}$. (in the following tasks, without loss of generality, we will assume that each sample has one element). For each sample X_i we want to perform a hypothesis test:

$$H_{0,i} : \mu_i = 0,$$
$$H_{1,i} : \mu_i > 0.$$

We perform all the above tests at the alpha significance level. If we ran every single test at the alpha significance level, we would make too many false discoveries (rejection of the null hypothesis) and consequently make too many Type I errors, which is what we want to reduce.

In a multiple testing problem, it is difficult to determine what the Type I error is, so we use multiple testing procedures to control for other factors such as FWER and FDR. To explain what these parameters are, consider the table below.

number of $H_{0,i}$ are	Accepted	Rejected	total
True	U	V	n_0
False	T	S	$n - n_0$
total	$n - R$	R	n

Table 1: Null hypothesis contingency table.

Where:

- n — number of tests,
- n_0 — number of true null hypothesis,
- R — number of discoveries (rejected null hypothesis),
- U — number of true non-discoveries,
- V — number of false discoveries,
- T — number of false non-discoveries,
- S — number of true discoveries.

Family-Wise Error Rate is the probability of making at least one false discovery in the process of multiple testing:

$$FWER = P(V \geq 1).$$

A procedure controls the FWER in the **weak sense** if the FWER control at level α is guaranteed only when all null hypotheses are true. A procedure controls the FWER in the **strong sense** if the FWER control at level α is guaranteed for any configuration of true and non-true null hypotheses (whether the global null hypothesis is true or not).

False Discovery Rate is the expected value of False Discovery Proportion—the ratio between the numbers of false discoveries and all discoveries:

$$FDR = \mathbb{E}[FDP] = \mathbb{E}\left[\frac{V}{R}\mathbb{I}_{\{R>0\}}\right].$$

Under the global null hypothesis, FDR and FWEAR are equivalent.

Power of a procedure is the expected value of the proportion of correctly rejected alternatives:

$$power = \mathbb{E}\left[\frac{S}{n-n_0}\right].$$

Procedures

Let's consider the procedures we will use to test the models given in the problem. Let a single test $(H_{0,i}, H_{1,i})$ of sample X_i (mentioned in the introduction) correspond to a p-value:

$$p_i = 1 - \Phi(X_i).$$

Let us set the order $p_{(1)} \leq \ldots \leq p_{(i)} \leq \ldots \leq p_{(n)}$ and let $p_{(i)}$ correspond to $H_{0,(i)}$.

- **The Bonferroni method** reject $H_{0,i}$, when:

  $$p_i \leq \frac{\alpha}{n}.$$

  This method is very conservative. We know from the lecture that Bonferroni's method controls FWER in a strong sense:

  $$FWER \leq \frac{n_0}{n}\alpha.$$

- **The Sidak's procedure** reject $H_{0,i}$, when:

  $$p_i \leq \alpha_n^S,$$

  where $\alpha_n^S = 1 - (1-\alpha)^{1/n}$.

  $$FWER = 1 - (1 - \alpha_n^S)^{n_0}.$$

- **Holm's procedure** reject all $H_{0,(i)}$ for all $(i) < i_0$, where:

  $$i_0 = \min\left\{j : p_{(j)} > \frac{\alpha}{n-j+1}\right\}.$$

  Holm's procedure controls the FWER strongly.

- **Hochberg's procedure** reject all $H_{0,(i)}$ for all $(i) \leq i_0$, where:

  $$i_0 = \max\left\{j : p_{(j)} \leq \frac{\alpha}{n-j+1}\right\}.$$

- **Benjamini-Hochberg procedure** reject all $H_{0,(i)}$ for all $(i) \leq i_0$, where:

  $$i_0 = \max\left\{j : p_{(j)} \leq \frac{j}{n}\alpha\right\}.$$

  This procedure controls FDR at level α:

  $$FDR = \frac{n_0}{n}\alpha.$$

Multiple testing tasks

Let $X_1, ..., X_n$ be single-element samples such that X_i comes from the normal distribution $N(\mu_i, 1)$, where $i \in \{1, \ldots, n\}$ and

$$H_{0,i} : \mu_i = 0,$$
$$H_{1,i} : \mu_i > 0.$$

Let's perform multiple testing at the α significance level for the following cases:

$n = 20$

- a) $\mu_i = 1.2\sqrt{2\log(20)} \cdot \mathbb{I}_{(i=1)}$
- b) $\mu_i = 1.02\sqrt{2\log(2)} \cdot \mathbb{I}_{(i\leq 5)}$
- c) $\mu_i = \sqrt{2\log(\frac{20}{i})} \cdot \mathbb{I}_{(i\leq 10)}$

	Case								
	1.a			1.b			1.c		
Procedure	FWER	FDR	Power	FWER	FDR	Power	FWER	FDR	Power
Bonferroni	0.046	0.0337	0.547	0.034	0.0293	0.0596	0.023	0.0106	0.1487
Holm's	0.048	0.0342	0.548	0.035	0.0298	0.0598	0.027	0.0117	0.1546
Hochberg	0.048	0.0342	0.548	0.035	0.0298	0.0598	0.027	0.0117	0.1546
Sidak's	0.048	0.0347	0.551	0.036	0.0308	0.0600	0.024	0.0107	0.1501
B-H	0.081	0.0469	0.553	0.061	0.0354	0.0716	0.100	0.0251	0.2582

Table 2: The results of simulation for $n = 20$.

We see that for the above case (n=20), all the above-mentioned procedures control the FDR at the given α significance level. We also see that all, except the B–H procedure, control FWER. This supports the theory from the lecture that if a procedure controls FWER in a strong sense, it also controls FDR.

Taking into account the types of cases (1.a - specifies that there is one strong signal in the sample, 1.b - several weak signals, 1.c - many strong and weak signals), we see that the procedures perform best in the first case, and worst in the second one. In each case, the Bonferroni procedure has the worst power and the Benjamini-Hochberg procedure has the best power. An interesting phenomenon is the behavior of the Holm and Hochberg procedures, because they assume equal parameters in each case, and we also know that theoretically the Hochberg procedure should be more powerful than the Holm procedure (it is possible that we are not able to check this using the given cases).

In general, the Bonferroni, Holm, Hochberg and Sidak procedures control FWER and FDR at a similar level relative to each other, and assume similar powers. The Hochberg procedure works perfectly when we only want to control FDR and we are dealing with a case similar to case 1.c.

$n = 5000$

- a) $\mu_i = 1.2\sqrt{2\log(5000)} \cdot \mathbb{I}_{(i=1)}$
- b) $\mu_i = 1.02\sqrt{2\log(25)} \cdot \mathbb{I}_{(i\leq 100)}$
- c) $\mu_i = \sqrt{2\log(25)} \cdot \mathbb{I}_{(i\leq 100)}$
- d) $\mu_i = 1.002\sqrt{2\log(2.5)} \cdot \mathbb{I}_{(i\leq 1000)}$

	Case								
	2.a			2.b			2.c		
Procedure	FWER	FDR	Power	FWER	FDR	Power	FWER	FDR	Power
Bonferroni	0.046	0.0331	0.757	0.041	0.0103	0.0479	0.049	0.0112	0.0428
Holm's	0.046	0.0331	0.757	0.041	0.0103	0.0479	0.049	0.0112	0.0428
Hochberg	0.046	0.0331	0.757	0.041	0.0103	0.0479	0.049	0.0112	0.0428
Sidak's	0.048	0.0346	0.760	0.041	0.0108	0.0485	0.050	0.0114	0.0432

Table 3 continued from previous page

B–H	0.078	0.0491	0.763	0.543	0.0497	0.1661	0.529	0.0477	0.1470

	Case		
	2.d		
Procedure	FWER	FDR	Power
Bonferroni	0.037	0.0170	0.0017
Holm's	0.037	0.0170	0.0017
Hochberg	0.037	0.0170	0.0017
Sidak's	0.038	0.0171	0.0018
B-H	0.369	0.0414	0.0109

Table 4: The results of simulation for $n = 5000$.

For the case when n=5000, we see that the general conclusions regarding the test comparisons are similar to those for n=20. Comparing cases 1.a and 2.a, we see that as the number of hypotheses increased, the power of each test increased significantly. Comparing 2.b and 2.c, we come to the conclusion that the power decreases as the signals decrease, and FWER and FDR increase. Case 2.d shows that the power of the test depends not only on the number of true alternatives, but also on the strength of their signals. We see that for many weak signals, the procedures achieve very low power.

Simulation Description

A single iteration involves generating samples under the corresponding them true alternative hypotheses, and then generating an appropriate number of samples under the corresponding them null hypotheses. The next step is to test the obtained sample using the previously given procedures, and to determine the rejection vector as a result of repeated testing for each procedure. This function returns the V value for each procedure, i.e. a logical value, where 1 means the occurrence of at least one false discovery, FDP—the fraction of false discoveries, and power—the fraction of true discoveries. As a result of repeating this iteration 1000 times, we obtain a data frame whose column averages determine the FWER, FDR, Power coefficients for subsequent procedures.

Two-step Procedure (Fisher)

The Two-step Procedure consists of two steps. First, we test the global null: $H_0 = \bigcap_{i=1}^{n} H_{0,i}$ using the selected global null hypothesis testing procedure. If as a result of the test we reject the null hypothesis, we test each set of hypotheses $(H_{0,i}, H_{1,i})$ at the α significance level, i.e. we check whether $p_i \leq \alpha$. In my opinion, the first step allows us to save time by avoiding repeated testing in some cases. However, by performing the second step, we do not control the probability of a false discovery.

In the simulation we will consider the following null hypotheses for $n \in \{20.5000\}$.

- a) $\mu_i = 1.2\sqrt{2\log(n)} \cdot \mathbb{I}_{(i=1)}$
- b) $\mu_i = 1.02\sqrt{2\log(\frac{n}{10})} \cdot \mathbb{I}_{(i\leq 10)}$
- c) $\mu_i = \sqrt{2\log(\frac{20}{i})} \cdot \mathbb{I}_{(i\leq 10)}$
- d) $\mu_i = 1.002\sqrt{2\log(\frac{n}{40\%n})} \cdot \mathbb{I}_{(i\leq 20\%i)}$

Case	Procedure	strong FWER	weak FWER	FDR	Power
3.a	Bonferroni	0.360	0.039	0.2082	0.572
	χ^2	0.267	0.025	0.1612	0.342
3.b	Bonferroni	0.186	0.046	0.0863	0.1244
	χ^2	0.206	0.040	0.0892	0.136
3.c	Bonferroni	0.318	0.049	0.0641	0.4237
	χ^2	0.391	0.039	0.0791	0.4947
3.d	Bonferroni	0.179	0.048	0.0756	0.154
	χ^2	0.221	0.047	0.0927	0.1635

Table 5: The results of simulation for $n = 20$.

Case	Procedure	strong FWER	weak FWER	FDR	Power
3.a	Bonferroni	0.748	0.044	0.7450	0.748
	χ^2	0.089	0.043	0.0886	0.089
3.b	Bonferroni	0.764	0.040	0.7492	0.7484
	χ^2	0.1450	0.0410	0.1423	0.1420
3.c	Bonferroni	0.121	0.048	0.1185	0.0625
	χ^2	0.0910	0.0410	0.0891	0.0487
3.d	Bonferroni	0.860	0.050	0.2935	0.3322
	χ^2	1	0.0440	0.3411	0.3864

Table 6: The results of simulation for $n = 5000$.

IN the tables above we can see that the theory from the lecture has been confirmed. The Two-step Procedure does not control the FWER in a strong sense, but it does control it in a weak sense. None of the above methods control FDR.

Taking case 3.a for both values of n, and case 3.b for n=5000, we see that the Two-step procedure, using the Bonferroni mote in the first step, works best when there are a very small number of strong signals. In cases where there are many weak and strong signals (cases 3.d for n=5000, 3.c for n=20) we see that both methods assume similar powers, but χ^2 is better and better controls FWER in the weak sense.

Simulation Description

In a single simulation, we generate a sample of n random variables from a normal distribution according to a given case (as in the previous tasks). Based on the obtained sample, we determine p-values (as defined in the introduction). Then we start the Two-step procedure: 1. We test the global null hypothesis based on the generated sample using the Bonferroni correction and the χ^2 test (as it was done in the previous lists). If we accept the global null hypothesis as a result of the test, we determine that there were no discoveries. The result is zero power, zero fraction of false discoveries. 2. If we reject the global null hypothesis, we test each hypothesis at the alpha level, check whether there is at least one false discovery, and determine the fraction of false discoveries and the fraction of true discoveries. As a result of repeating such an iteration 1000 times, we obtain the average number of occurrences of at least one false discovery (FWER estimator in the strong sense), the average value of the fraction of false discoveries (FDR estimator), and the average value of the percentage of true discoveries (power estimator).

In order to determine the FWER estimator in the weak sense, we repeat the experiment 1000 times. It involves generating n random variables from the standard normal distribution, determining their p-values and carrying out the previously described Two-step procedure. As a result of the experiment, we check whether at least one false discovery occurred. As a result of 1000 repetitions, we obtain the fraction of occurrences of at least one false discovery, which is an estimator of FWER in the weak sense.

Comparison of the Brownian bridge and the empirical process

Let us consider two types of trajectories: the Brownian bridge $B(t)$ and the empirical process $U_n(t)$:

$$U_n(t) = \sqrt{n}\,(F_n(t) - t)\,,$$

where $n = 5000$, $t \in [0, 1]$, $n = 5000$, $F_n(t) = \frac{\#\{i:\, p_i \leq t\}}{n}$. To check whether the two processes are equivalent, for each of the 1000 trajectories of each process, let us determine the statistics:

$$K\text{–}S = \sup_{t\in[0,1]} |U_n(t)|,$$

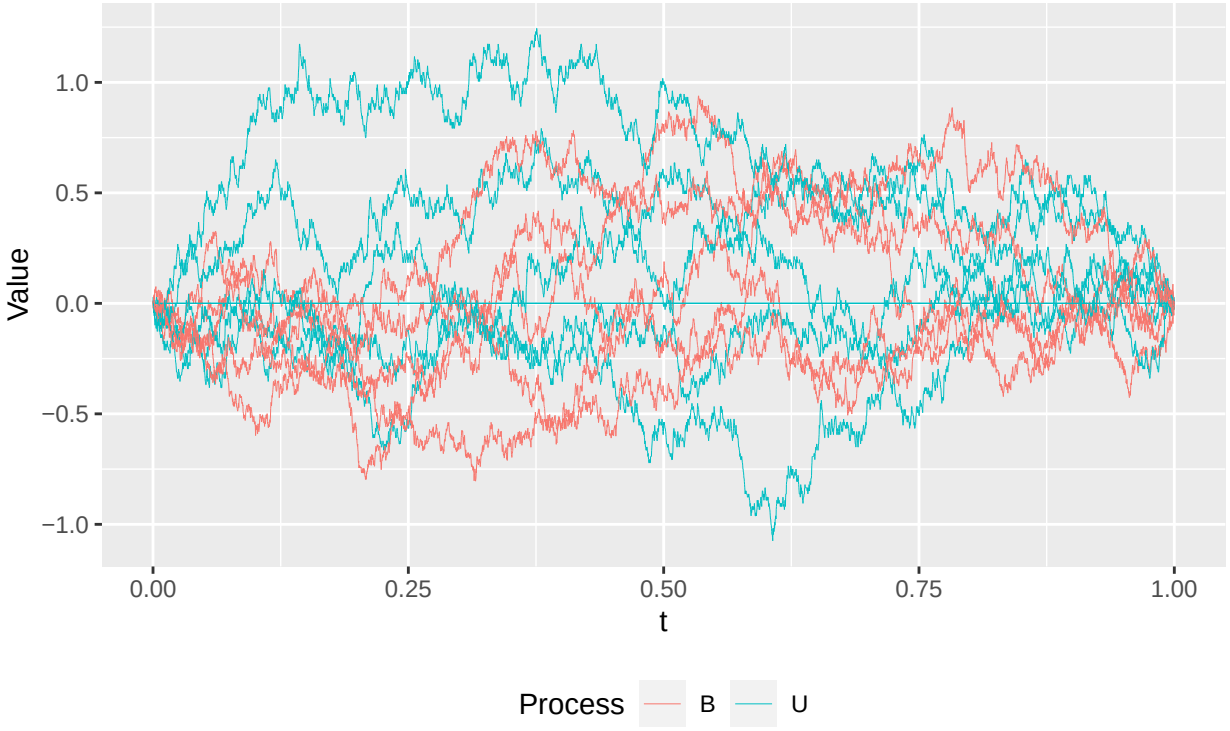
$$T = \sup_{t\in[0,1]} |B(t)|.$$

To compare both processes, run a simulation.

Simulations

The simulation will consist of repeating the iterations 1000 times. A single iteration consists in generating 5,000 single-element samples from the standard normal distribution, determining p-values on their basis, and then determining subsequent process values $U_n(t)$. Based on a single iteration We will calculate the K–S statistic. In addition, during the iteration, we will simulate the $B(t)$ process using an appropriate function in R and determine the T statistic based on it. As a result of 1000 repetitions, we will obtain 1000 trajectories of each process and 1000 element samples of K–S and T statistics. Based on the obtained sample statistics, we will determine their 0.8, 0.9 and 0.95 quantiles.

Result of simulation



In the graph above we can see that the trajectories of both processes are very similar. Process U_n sometimes

takes values further from 0 than B. Based on the graph, we can assume that both processes are equivalent. Let's check the estimated values of the appropriate quantiles of each statistic.

	α		
Statistic	0.8	0.9	0.95
T	1.071357	1.239908	1.372033
K–S	1.060660	1.202082	1.343503

Table 7: α quantiles of K–S statistics (under null hypothesis) and T statistics.

We see that the quantile estimators of each order of both statistics have similar values. The quantiles of the K-S statistic are slightly smaller. The obtained results confirm the theory from the lecture that for large values of n:

$$U_n(t) \longrightarrow B(t).$$