STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 4

1 Multiple testing

1.1 Motivation

Multiple hypothesis testing is concerned with testing several hypotheses simultaneously. As a motivation, we recall the following example from Lecture 1.

Suppose we have m_1 prostate cancer patients and m_0 normal controls from a microarray study. Each man's gene expression levels were measured on a panel of n genes (humans have roughly 20,000 genes), yielding a measurement matrix of the size $n \times (m_0 + m_1)$. Let $x_{ij}^{(1)}$ ($x_{ij}^{(0)}$) be the activity of the ith gene for jth man in the patient (control) group. We ask the question:

- for the ith gene, do the gene expression levels differ between the patient and control groups?

Formally, we can test the null hypotheses:

$$H_{0,i} : \mathbb{E}[x_{ij}^{(1)}] = \mathbb{E}[x_{ij}^{(0)}], \quad i = 1, 2, \ldots, n.$$

For each gene, a two-sample t statistic t_i can be computed comparing gene i's expression levels for the m_1 patients with those for the m_0 controls. Under the Gaussian assumption on the samples and $H_{0,i}$, t_i follows the t distribution with $m_0 + m_1 - 2$ degrees of freedom. Therefore, the p-value can be computed as

$$p_i = P(|T_{m_0+m_1-2}| > |t_i|),$$

where $T_{m_0+m_1-2}$ follows the t distribution with $m_0 + m_1 - 2$ degrees of freedom.

In general, let us assume there are n hypotheses $H_{0,1}, \ldots, H_{0,n}$ to be tested. For the ith hypothesis $H_{0,i}$, we can calculate a p-value p_i. We assume that the p-values are uniformly distributed over $[0, 1]$ when the associated hypothesis is null.

1.2 Potential outcomes

There are four types of outcomes in multiple testing, as illustrated in Table 1.2. The rows indicate the true state of the world with respect to the null hypotheses. On the other hand, the columns indicate acceptance and rejection of the null hypotheses. The random variables U and V indicate the number of correctly accepted and falsely rejected hypotheses, i.e., the number of true negatives and false positives (often called false discoveries). The random variables T and S indicate the number of false negatives and true positives. The number of hypotheses under the null is denoted by n_0. The four random variables U, V, T, and S are unobserved. The random variable R indicates the total number of rejections by a given multiple testing procedure and is observed. Note that the quantities of primary interest are the number of false discoveries V and the number of discoveries R. It is desired to maximize the number of discoveries subject to the constraint that the number of false discoveries remains low.

	H_0 accepted	H_0 rejected	Total
H_0 true	U	V	n_0
H_0 false	T	S	$n - n_0$
	$n - R$	R	n

Table 1: Potential outcomes for testing multiple hypotheses.

1.3 Familywise error rate

Familywise error rate (FWER) is the probability that a multiple hypothesis testing procedure makes at least one false rejection, i.e.,

$$\text{FWER} = \mathbb{P}(V \geq 1).$$

A procedure controls the FWER at level α in the weak sense if the FWER is less than or equal to α under the global null. In contrast, a procedure controls the FWER at level α in a strong sense if the FWER is less than or equal to α under all configurations of true and false hypotheses. One variation of this notion of error control is the k-FWER defined as

$$k\text{-FWER} = \mathbb{P}(V \geq k).$$

2 FWER controlling procedures

In this section, we introduce several classical FWER controlling procedures.

2.1 Bonferron's method and Šidák's procedure

As discussed before, Bonferron's method rejects $H_{0,i}$ if $p_i \leq \alpha/n$. Bonferroni's method controls FWER at level α in a strong sense. To see this, note that

$$\begin{aligned}\text{FWER} =& P(V \geq 1) \leq \mathbb{E}[V] \\ =& \sum_{i:H_{0,i} \text{ is true}} \mathbb{P}(p_i \leq \alpha/n) \\ =& \frac{\alpha n_0}{n} \\ \leq& \alpha,\end{aligned}$$

where the first inequality is due to the Markov inequality.

Šidák's procedure refines Bonferron's method when the p-values are independent. Specifically, it rejects $H_{0,i}$ whenever $p_i \leq 1 - (1-\alpha)^{1/n}$. Notice that under independence among the p-values,

$$\begin{aligned}P(V \geq 1) =& 1 - P(V = 0) = 1 - \prod_{i:H_{0,i} \text{ is true}} P(p_i > 1 - (1-\alpha)^{1/n}) \\ =& 1 - (1-\alpha)^{n_0/n} \\ \leq& 1 - (1-\alpha) = \alpha.\end{aligned}$$

2.2 Weak control

Recall that a procedure controls the FWER at level α in the weak sense if the FWER is less than or equal to α under the global null. To gain some intuition, consider the following two-step multiple testing procedure suggested by Fisher:

- Step 1. Run a procedure to test the global null $H_0 = \cap_{i=1}^n H_{0,i}$.
- Step 2. When the global null is rejected, reject the ith hypothesis if $p_i \leq \alpha$.

This procedure controls the FWER only in a weak sense, i.e., it controls the FWER only when all null hypotheses are null. To see this, observe that if all null hypotheses are true, then the procedure reaches Step 2 with probability at most α, and, therefore, the chance that a single false discovery is made is at most α.

By contrast, it does not control the FWER in a strong sense, i.e., under any configurations of null and non-null hypotheses when some hypotheses are non-null. In particular, suppose there is one very strong signal, say, with an extremely small p-value. Then, if all other hypotheses are null, the procedure will reach the second step with very high probability, and for the second step, there will be, on average, about $n_0\alpha$ false rejections. This illustrates how global testing and multiple testing with control of the FWER are very different procedures.

2.3 Holm's procedure

Holm's procedure provides a strict improvement over Bonferoni's method. Consider the ordered p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ and let $H_{(i)}$ be the hypothesis associated with $p_{(i)}$. Holm's Procedure is a "step down" multiple testing procedure.

- Step 1. If $p_{(1)} \leq \alpha/n$, reject $H_{(1)}$ and go to step 2. Otherwise, accept $H_{(1)}, \ldots, H_{(n)}$.
- Step 2. If $p_{(2)} \leq \alpha/(n-1)$, reject $H_{(2)}$ and go to step 3. Otherwise, accept $H_{(2)}, \ldots, H_{(n)}$.
- Step n. If $p_{(n)} \leq \alpha$, reject $H_{(n)}$ and go to step 3. Otherwise, accept $H_{(n)}$.

In other words, the procedure examines $p_{(i)}$ from $i = 1$ and stops when it encounters the first i such that $p_{(i)} > \alpha/(n-i+1)$, rejecting all $H_{(j)}$ with $j = 1, 2, \ldots, i-1$. It is important to note that Holm's Procedure is less conservative than Bonferroni's method, as the threshold becomes more "liberal" as larger p-values are considered.

Theorem. Holm's procedure controls the FWER at level α in a strong sense.

Proof. Let $\mathcal{N}_0$ be the set of indices corresponding to the hypotheses under the null with n_0 elements.

We first show that if we falsely reject some true hypotheses, then there exists a true hypothesis $H_{(l)}$ such that $p_{(l)} \leq \alpha/n_0$. In this case, $n_0 \geq 1$. Let $H_{(l)}$ be the first true null hypothesis being rejected. Then, $H_{(1)}, \ldots, H_{(l-1)}$ are all correctly rejected and $p_{(i)} \leq \alpha/(n-i+1)$ for $i = 1, 2, \ldots, l$. As $l - 1 \leq n - n_0$, we have $n_0 \leq n - l + 1$, which implies that $p_{(l)} \leq \alpha/(n-l+1) \leq \alpha/n_0$.

Based on the above observation, we get

$$\begin{aligned}
\mathbb{P}(V \geq 1) \leq & \mathbb{P}\left(\bigcup_{H_{(l)} \text{ is true}} \{p_{(l)} \leq \alpha/n_0\} \right) \\
= & \mathbb{P}\left(\bigcup_{i \in \mathcal{N}_0} \{p_i \leq \alpha/n_0\} \right) \\
\leq & \frac{n_0 \alpha}{n_0} = \alpha.
\end{aligned}$$

2.4 Benjamini-Hochberg procedure

The Benjamini-Hochberg (BH) procedure works as follows.

- Given the desired level α, find the largest k such that $p_{(k)} \leq \alpha k/n$.
- Reject the hypotheses $H_{(1)}, \ldots, H_{(k)}$.

Under the global null, the procedure will make one or more than one rejection only when

$$T_n = \min_{1 \leq i \leq n} \frac{n p_{(i)}}{i} \leq \alpha.$$

This is essentially the Simes test, which has been shown to control the Type I error at level α. The BH procedure does not control the FWER strongly (Why?), but we shall revisit the BH procedure later on to show that it controls the so-called false discovery rate (FDR).

3 Closure principle

We introduce the closure principle and its applications to global testing procedures. The section concludes by proving that closed tests control the FWER strongly.

We follow the above setup by assuming that there is a set of hypotheses $\{H_i\}_{i=1}^n$ and for each hypothesise H_i, we observe a p-value p_i which is uniformly distributed over $[0, 1]$ under the null. Let $[n] = \{1, 2, \ldots, n\}$.

For an index set $I \subseteq [n]$, we define

$$H_I = \bigcap_{i \in I} H_i.$$

and the closure of $\{H_i\}_{i=1}^n$ as

$$\mathcal{C} = \{H_I : I \subseteq [n], I \neq \emptyset\}.$$

As an example, consider $n = 4$. The closure is given by

$$\begin{gathered} \underline{H_{1234}} \\ \underline{H_{123}}, \underline{H_{124}}, \underline{H_{134}}, H_{234} \\ \underline{H_{12}}, \underline{H_{13}}, \underline{H_{14}}, H_{23}, H_{24}, H_{34} \\ \underline{H_1}, \underline{H_2}, H_3, H_4 \end{gathered}$$

For each I, consider a valid level α test ψ_I for testing H_I, i.e., we reject H_I when $\psi_I = 1$ and

$$P(\psi_I = 1 | H_I) \leq \alpha.$$

The tests ψ_I may be constructed using any global testing procedures such as Bonferroni, Simes, L_2, etc.

The closure procedure. Reject H_I if and only if for all $J \supseteq I$, H_J is rejected at level α. Let $T_I = \min_{J \supseteq I} \psi_J$. Then, we reject H_I if and only if $T_I = 1$.

Consider the example above with $n = 4$ and assume that the underlined hypotheses are rejected at the level α. According to the closure principle, among $\{H_i\}_{i=1}^4$, only H_1 is rejected.

Theorem. The closure principle controls the FWER strongly.

Proof. Let $\mathcal{N}_0 \subseteq [n]$ be the index set of the true nulls. Suppose $\mathcal{N}_0 \neq \emptyset$. Otherwise, there will be no false rejection. Define $A = \{\text{the closure principle makes at least one false rejection}\}$ and $B = \{H_{\mathcal{N}_0} \text{ is rejected}\}$. Suppose $j \in \mathcal{N}_0$ and H_j is rejected. By the closure principle, $H_{\mathcal{N}_0}$ must be rejected and thus $A \subset B$. We have

$$\text{FWER} = P(A) \leq P(B) \leq \alpha.$$

Therefore, the closure principle controls the FWER at level α regardless of the underlying configuration.

3.1 Closed testing procedure: closing Bonferroni

If we use Bonferroni's procedure to test H_I, then we reject H_I (i.e., $\psi_I = 1$) if

$$\min_{i \in I} p_i \leq \frac{\alpha}{|I|},$$

where p_i is the p-value for testing H_i and $|I|$ denotes the size of I.

3.2 Holm's procedure as a closed testing procedure

Holm's procedure can be viewed as a closed testing procedure with the Bonferroni correction applied locally on each of the intersections of null hypotheses.

Holm's procedure is a shortcut procedure since it makes n or fewer comparisons, while the number of all intersections of null hypotheses to be tested is of order 2^n.

In Holm's procedure, we first test $H_{(1)}$. If it is not rejected (i.e., $p_{(1)} > \alpha/n$), then $\cap_{i=1}^n H_i$ is not rejected using Bonferroni's method. Therefore, the closure principle does not reject any hypothesis. If $p_{(1)} \leq \alpha/n$, then $H_{(1)}$ is rejected in Holm's procedure. On the other hand, any hypothesis H_I containing $H_{(1)}$ has its

smallest p-value being $p_{(1)}$ and hence is rejected using Bonferroni's method. It thus implies that $H_{(1)}$ is also rejected in the closed testing procedure.

If $H_{(1)}$ is rejected, we now consider $H_{(2)}$. If $p_{(2)} > \alpha/(n-1)$, then $H_{(2)}$ is not rejected in Holm's procedure. We note that $H_I = \cap_{i=1}^n H_i \setminus H_{(1)}$ has its smallest p-value being $p_{(2)}$. Thus, H_I is not rejected using Bonferroni's method, which implies that none of the other hypotheses will be rejected in the closed testing procedure. If $p_{(2)} \leq \alpha/(n-1)$, $H_{(2)}$ is rejected in Holm's procedure. Any hypothesis H_I containing $H_{(2)}$ will also be rejected using Bonferroni's method (Why?). Therefore, $H_{(2)}$ is rejected in the closed testing procedure.

The same rationale applies to $H_{(i)}$ for any $1 \leq i \leq n$. The argument relies on the following key observation. Suppose $H_{(1)}, \ldots, H_{(j-1)}$ have already been rejected, i.e., $p_{(j-1)} \leq \alpha/(n-j+2)$. Let

$$H_{I_j} = \cap_{i=1}^n H_i \setminus (\cap_{i=1}^{j-1} H_{(i)}).$$

Then we have

$$\psi_{I_j} = 1 \quad \text{iff} \quad \psi_I = 1 \text{ for any } I \text{ with } (j) \in I$$

and $\psi_{I_j} = 1$ iff $p_{(j)} \leq \alpha/(n-j+1)$ using Bonferroni's method

The above argument provides an alternative way of showing that Holm's procedure controls the FWER in the strong sense.

3.3 Closed testing procedure: closing Simes

Recall that the Simes procedure rejects the global null $\cap_{i=1}^n H_i$ whenever

$$p_{(i)} \leq \frac{\alpha i}{n}$$

for some i. If we define

$$T_n = \min_{1 \leq i \leq n} \left(\frac{n p_{(i)}}{i} \right),$$

Simes test rejects the null if $T_n \leq \alpha$.

We now apply the Simes test statistic in the closed testing procedure. Specifically, to test H_I, we define

$$\psi_I = \mathbf{1} \left\{ \min_{1 \leq i \leq n} \left(\frac{|I| p_{(i,I)}}{i} \right) \leq \alpha \right\}$$

where $p_{(i,I)}$ is the ith smallest p-value among $\{p_i : i \in I\}$. Under independence among the p-values, the Simes procedure controls the type I error at level α. Thus, the closure of the Simes procedure will control FWER at level α in the strong sense. Since Simes is strictly more powerful than Bonferroni, the closure of Simes will be strictly more powerful than Holm's procedure (the closure of Bonferroni). The closure of Simes is called Hommel's procedure; see Hommel (1988, Biometrika) for more details.

3.4 Hochberg's procedure

We discuss Hochberg's procedure, which is more conservative than Hommel's procedure but still more powerful than Holm's procedure. In summary, we have

$$\text{Holm's procedure} \prec \text{Hochberg's procedure} \prec \text{Hommel's procedure},$$

where $A \prec B$ means procedure A is more conservative than procedure B.

Hochberg's procedure. Given the ordered p-values $p_{(1)} \leq p_{(2)} \leq \cdots p_{(n)}$ and the corresponding hypotheses $H_{(1)}, \ldots, H_{(n)}$, we reject $H_{(j)}$ if there exists $j' \geq j$ such that

$$p_{(j')} \leq \frac{\alpha}{n - j' + 1}.$$

We know that the closure of Simes controls FWER at level α under independence. Thus, to conclude that Hochberg's procedure also controls FWER at level α, it suffices to show whenever Hochberg's procedure rejects a hypothesis, then the closure of Simes also rejects.

Proof. Suppose $H_{(j)}$ is rejected in the Hochberg's procedure. Then there exists $j' \geq j$ such that

$$p_{(j')} \leq \frac{\alpha}{n - j' + 1}.$$

Now, we show that for any I with $(j) \in I$, H_I is rejected by the Simes test. Consider a set I with size h. Define the "worse-case" set

$$K_h = \begin{cases} \{(j), (n), (n-1), \ldots, (n-h+2)\} & \text{if } j \leq n-h+1, \\ \{(n), (n-1), \ldots, (n-h+1)\} & \text{if } j > n-h+1. \end{cases}$$

We note that $(j) \in K_h$ and $|K_h| = h = |I|$. If the Simes procedure rejects K_h, it will also reject H_I. This is because the p-values indexed in K_h are larger than the p-values indexed in I. Since the thresholds in the Simes procedure only depend on the size of the set I, this shows that $\psi_{K_h} = 1$ implies $\psi_I = 1$. Therefore, if $\psi_{K_h} = 1$, (j) will be rejected.

Now we show that $\psi_{K_h} = 1$. First, suppose that $j \leq j' \leq n - h + 1$. Then $p_{(j)}$ is the smallest p-value in K_h, and thus

$$p_{(1,K_h)} = p_{(j)} \leq p_{(j')} \leq \frac{\alpha}{n - j' + 1} \leq \frac{\alpha}{h}.$$

It suggests that $\psi_{K_h} = 1$. Now if $j' \geq n - h + 2$, then $(j') \in K_h$. We note that K_h contains h p-values and the $n - j'$ p-values $p_{(n)}, p_{(n-1)}, \ldots, p_{(j'+1)}$ are all in K_h and are all larger than $p_{(j')}$. We must have

$$p_{(h-n+j',K_h)} \leq p_{(j')} \leq \frac{\alpha}{n - j' + 1} \leq \frac{(h - n + j')\alpha}{h}.$$

To see why the last inequality holds, let $a = n - j' + 1$ and $b = h - n + j'$. It suffices to show that $h \leq ab$. But note that $a + b - 1 = h$ and thus $ab - h = (a-1)(b-1)$. Since a and b are both integers greater than or equal to zero, we must have $ab - h = (a-1)(b-1) \geq 0$ and hence $ab \geq h$.

4 Step-down versus step-up procedures

Both Holm's procedure and Hochberg's procedure compare $p_{(j)}$ with the threshold $\alpha/(n - j + 1)$. However,

- Holm's procedure scans forward and stops at the first p-value that is larger than the corresponding threshold. This is called a step-down procedure.
- Hochberg's procedure scans backward and stops at the first p-value that is smaller than the corresponding threshold. Likewise, this is called a step-up procedure.

The names step-up and step-down may seem counter-intuitive and presented in the wrong order. For step-up procedures, we start at the largest p-value and decrease, but for step-down procedures, we start at the smallest p-value and increase. The names make sense if we think about the procedures in terms of z-scores. That is, if our p-values are of the form $p_i = P(Z > z_i)$ where $Z \sim N(0,1)$ and z_i is the z-score for H_i, then the smallest p-values correspond to the largest z-score. In this case, a step-up procedure would start at the smallest z-score and increase. A step-down procedure would start at the largest z-score and decrease.