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

Lecture 1

1 Motivation

Multiple testing refers to simultaneous testing of more than one hypothesis. Given a set of hypotheses, multiple testing deals with deciding which hypotheses to reject while guaranteeing some notion of control on the number of false rejections.

Microarrays were a game-changer for large-scale data analysis in the biomedical field. These revolutionary devices enabled the assessment of individual gene activity for thousands of genes at once. However, this also posed a significant challenge of carrying out thousands of simultaneous hypothesis tests with the prospect of finding only a handful of interesting genes among a vast number of null cases - akin to searching for a needle in a haystack.

As a specific example, 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. The transformation $z_i := \Phi^{-1}(F_{m_0+m_1-2}(t_i))$ where F_m is the cdf of a t distribution with m degrees of freedom and Φ^{-1} is the inverse function of a standard normal cdf, makes z_i standard normal under the null hypothesis, i.e.,

$$H_{0,i} : z_i \sim N(0, 1).$$

Of course the investigators were hoping to spot some non-null genes, ones for which the patients and controls respond differently. A reasonable model for both null and non-null genes assumes that

$$z_i \sim N(\mu_i, 1)$$

with μ_i being the effect size for gene i. Null genes have $\mu_i = 0$, while the investigators hoped to find genes with large positive or negative μ_i effects.

2 Testing the Global Null

The global null concerns asking whether at least one of n null hypotheses is false. In other words, we want to know if at least one gene has different expression levels for the two groups. Mathematically, the global null is defined as

$$H_0 = \{H_{0,i} \text{ is true for all } 1 \leq i \leq n\} = \cap_{i=1}^{n} H_{0,i}.$$

Global testing is the task of testing the global null. For each hypothesis $H_{0,i}$, we can compute a p-value p_i which follows the uniform distribution under $H_{0,i}$. For instance, we can set $p_i = F_{n_0+n_1-2}(t_i)$ in the prostate cancer example.

We remark that for p_i to be a valid p-value, we only require it to be super-uniform under the null, i.e.,

$$P_{H_{0,i}}(p_i \leq t) \leq t.$$

However, for clarity, we shall assume that p_i is uniform under the null throughout the following discussions.

2.1 Bonferroni correction

Perhaps the simplest approach to test the global null is Bonferroni's method/correction. Let α be the desired Type I error level. The Bonferroni's method rejects H_0 if $p_i \leq \alpha/n$ for some $1 \leq i \leq n$. In other words, it sets a tighter threshold for individual hypotheses so that the Type I error for individual hypotheses is controlled at the level α/n. The Type I error for testing the global null H_0 can be computed as

$$\begin{aligned}
&P_{H_0}(\text{Bonferroni's method rejects } H_0) \\
=&P_{H_0}(p_i \leq \alpha/n \text{ for some } 1 \leq i \leq n) \\
\leq&\sum_{i=1}^{n} P_{H_0}(p_i \leq \alpha/n) \\
=&\sum_{i=1}^{n} \frac{\alpha}{n} = \alpha.
\end{aligned}$$

Bonferroni's method is built upon the union bound, which is robust to the arbitrary dependence within the p-values. When the n p-values are independent, the Šidák correction can be used. Specifically, it rejects H_0 whenever $p_i \leq 1 - (1-\alpha)^{1/n}$ for some $1 \leq i \leq n$. The Šidák correction controls the Type I error:

$$\begin{aligned}
&P_{H_0}(\text{Šidák correction rejects } H_0) \\
=&P_{H_0}(p_i \leq 1 - (1-\alpha)^{1/n} \text{ for some } 1 \leq i \leq n) \\
=&1 - P_{H_0}(p_i > 1 - (1-\alpha)^{1/n} \text{ for all } 1 \leq i \leq n) \\
=&1 - \prod_{i=1}^{n} P_{H_0}(p_i > 1 - (1-\alpha)^{1/n}) \\
=&1 - \prod_{i=1}^{n} (1-\alpha)^{1/n} = \alpha.
\end{aligned}$$

A common misconception is that Bonferroni's method is conservative, meaning that the size of the Bonferroni method is much smaller than α. Consider the case where p_i's are independent. The size of the Bonferroni's method is

$$\begin{aligned}
&P_{H_0}(\text{Bonferroni's method rejects } H_0) \\
=&1 - P_{H_0}(p_i > \alpha/n \text{ for all } 1 \leq i \leq n) \\
=&1 - \prod_{i=1}^{n} P_{H_0}(p_i > \alpha/n) \\
=&1 - \left(1 - \frac{\alpha}{n}\right)^n \to 1 - e^{-\alpha}.
\end{aligned}$$

With $\alpha = 0.1$, $1 - e^{-\alpha} = 0.095$ and when $\alpha = 0.05$, $1 - e^{-\alpha} = 0.0488$. Bonferroni's method becomes more conservative when the p-values have (positive) dependence.

Exercise 1.1: Define $\mathbf{z} = (z_1, \dots, z_n)$. Let $E[\mathbf{z}] = \boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ and $\text{cov}(\mathbf{z}) = \Sigma = (\sigma_{ij})_{i,j=1}^{p}$, where $\Sigma_{ii} = 1$ and $\sigma_{ij} = \rho \in (0,1)$ for $i \neq j$. We are interested in testing the global null that

$$H_0 : \mu_i = 0 \text{ for all } 1 \leq i \leq n.$$

versus the alternative that $\mu_i \neq 0$ for some i. Let $p_i = 2(1 - \Phi(|z_i|))$ be the two-sided p-value. For $\rho \in \{0, 0.1, 0.2, \ldots, 0.9\}$, simulate the z-scores under the global null and perform Bonferroni's test based on the two-sided p-values. Repeat the procedure 1,000 times and calculate the percentage of rejections over the 1,000 simulation runs. Report the percentage of rejections for different values of ρ.

2.2 Fisher's combination test

Fisher's method is a technique for meta-analysis that combines multiple p-values into one test statistic. We state two basic facts before introducing the exact form of the test statistic.

- For $p_i \sim \text{Unif}[0,1]$, we have $-2\log(p_i) \sim \text{Exp}(1/2) =^d \chi^2_2$ (where $\text{Exp}(\lambda)$ denotes the exponential distribution with rate λ). To see this, we note that

$$P(-2\log(p_i) \leq x) = P(\log(p_i) \geq -x/2) = 1 - \exp(-x/2).$$

- Suppose $X_i \sim \chi^2_{k_i}$ are independent χ^2 random variables. We have $\sum_{i=1}^n X_i \sim \chi^2_k$ for $k = \sum_{i=1}^n k_i$.

The Fisher's combination test statistic is defined as

$$T = -2\sum_{i=1}^n \log(p_i).$$

Under H_0, T follows the χ^2_{2n} distribution. Let $q_{2n}(1-\alpha)$ be the $1-\alpha$ quantile of χ^2_{2n}. We reject the global null if $T > q_{2n}(1-\alpha)$.

Unlike Bonferroni's test, which looks only at the smallest p-value, Fisher's test can combine weak evidence against multiple null hypotheses because it is (in some sense) a weighted average over all the p-values. Thus, Bonferroni's method works better for detecting a few larger changes (sparse and strong signals/alternatives) in the individual tests, while Fisher's test works better for detecting many subtle changes (dense and weak signals/alternatives).

A related method is the Pearson's test statistic given by $T_{\text{Pear}} = 2\sum_{i=1}^n \log(1-p_i)$. Another closely related approach is called Stouffer's z-score method. Let Φ be the cdf of $N(0,1)$. Define $z_i = \Phi^{-1}(1-p_i)$ which follows $N(0,1)$ under $H_{0,i}$. Under the alternative $H_{a,i}$, we expect z_i to take large values. Define the test statistic

$$Z = \frac{\sum_{i=1}^n z_i}{\sqrt{n}}$$

which follows $N(0,1)$ under the global null. An advantage of the z-score test is that it allows weights, i.e., one can define

$$Z_w = \frac{\sum_{i=1}^n w_i z_i}{\sqrt{\sum_{i=1}^n w_i^2}}$$

for some weights w_i, which again follows $N(0,1)$ under H_0. Similar to Fisher's test, the z-score test is better suited for detecting dense and weak signals.

2.3 Cauchy combination test

Cauchy combination test [Liu and Xie (2018)] is a recently developed testing method that has the advantage of being robust to the dependence among the z-statistics. Suppose each p_i is computed from a z-score z_i. Define $\mathbf{z} = (z_1, \ldots, z_n)$. Let $E[\mathbf{z}] = \boldsymbol{\mu}$ and $\text{cov}(\mathbf{z}) = \Sigma$. Under the global null, $\boldsymbol{\mu} = \mathbf{0}$.

The Cauchy combination test is defined as

$$W = \sum_{i=1}^n w_i \tan[\{2\Phi(|z_i|) - 1.5\}\pi],$$

where the weights w_is are nonnegative and $\sum_{i=1}^n w_i = 1$. Under $H_{0,i}$, $p_i = 2(1-\Phi(|z_i|))$ is the two-sided p-value which follows Unif$[0,1]$. We have

$$\begin{aligned} &P(\tan[\{2\Phi(|z_i|)-1.5\}\pi] \leq x) \\ =&P(\tan((0.5-p_i)\pi) \leq x) \\ =&P(1-p_i \leq 0.5+\arctan(x)/\pi) \\ =&0.5+\arctan(x)/\pi, \end{aligned}$$

which suggests that $\tan[\{2\Phi(|z_i|)-1.5\}\pi]$ follows the standard Cauchy distribution.

Assume that (z_i, z_j) are jointly normal for all $1 \leq i \neq j \leq n$. It has been shown in Liu and Xie (2018) that

$$\lim_{t\to+\infty} \frac{P(W \geq t)}{P(W_0 \geq t)} = 1,$$

where W_0 is a standard Cauchy random variable.

Remark. The above result can be generalized to transformations of p-values to the other heavy tail distributions, such as levy distribution.

3 Optimality of Bonferroni's Method

We look deeper into Bonferroni's method by showing that it enjoys certain optimality in detecting sparse and strong signals.

The Bonferroni's method considers the test statistic $n\min_{1\leq i\leq n} p_i$ and it rejects the global null if $n\min_{1\leq i\leq n} p_i \leq \alpha$. When $p_i = 2(1-\Phi(|z_i|))$, the statistic is essentially equivalent to the maximum statistic $M = \max_{1\leq i\leq n}|z_i|$.

3.1 Gaussian location models

To analyze the maximum statistic, we consider the Gaussian location model:

$$z_i \sim N(\mu_i, 1), \quad i = 1, 2, \ldots, n,$$

independently over i. The goal here is to test the global null that

$$H_0 : \mu_i = 0 \text{ for all } 1 \leq i \leq n.$$

Note that $n\min_{1\leq i\leq n} p_i \leq \alpha$ is equivalent to

$$\begin{aligned} & n\min_{1\leq i\leq n} 2(1-\Phi(|z_i|)) \leq \alpha \\ \Longleftrightarrow\quad & 1 - \max_{1\leq i\leq n} \Phi(|z_i|) \leq \alpha/(2n) \\ \Longleftrightarrow\quad & \Phi^{-1}(1-\alpha/(2n)) \leq \max_{1\leq i\leq n}|z_i|. \end{aligned}$$

Exercise 1.2: In the one-sided case, we consider the test statistic $\max_i z_i$ and rejects the null if $\max_i z_i \geq \Phi^{-1}(1-\alpha/n)$.

For simplicity, let us consider the one-sided case, where we reject H_0 if $\max_i z_i \geq \Phi^{-1}(1-\alpha/n)$. Under H_0, $z_i \sim N(0,1)$ for all i and

$$\frac{\max_i z_i}{\sqrt{2\log(n)}} \to^p 1. \tag{1}$$

Exercise 1.3: Prove (1). Hint: you may use the Mills ratio, which states that for any $x > 0$,

$$\frac{x}{x^2+1} \leq \frac{1-\Phi(x)}{\phi(x)} \leq \frac{1}{x}$$

where $\phi(x)$ and $\Phi(x)$ are the pdf and cdf of $N(0,1)$ respectively.

Exercise 1.4: Show that $\Phi^{-1}(1-\alpha/n) = \sqrt{2\log(n)}(1+o(1))$ as $n \to +\infty$.

3.2 Asymptotic power analysis

We aim to study the asymptotic power of the maximum test. To begin with, we need to specify the alternative. We assume that there is exactly one μ_i that is nonzero and is equal to μ^*. We shall consider two cases: for some $\epsilon > 0$

- Case 1 (strong signal): $\mu^* = (1+\epsilon)\sqrt{2\log(n)}$;
- Case 2 (weak signal): $\mu^* = (1-\epsilon)\sqrt{2\log(n)}$.

From (1), for the signal to be detectable, we require its magnitude μ^* to be sufficiently larger than $\sqrt{2\log(n)}$. Let us consider the first case where $\mu^* = (1+\epsilon)\sqrt{2\log(n)}$. Suppose μ^* is the mean of the i^*th z-score and write $z_{i^*} = \mu^* + \xi_{i^*}$ for $\xi_i \sim N(0,1)$. We note that

$$\begin{aligned}
&P(\max_i z_i > \Phi^{-1}(1-\alpha/n)) \\
\geq &P(\xi_{i^*} + \mu^* > \sqrt{2\log(n)}(1+o(1))) \\
\geq &P(\xi_{i^*} > \sqrt{2\log(n)}(1+o(1)) - (1+\epsilon)\sqrt{2\log(n)}) \to 1.
\end{aligned}$$

When $\mu^* = (1-\epsilon)\sqrt{2\log(n)}$, we have

$$\begin{aligned}
&P(\max_i z_i > \Phi^{-1}(1-\alpha/n)) \\
=&1 - P(\max_i z_i \leq \Phi^{-1}(1-\alpha/n)) \\
=&1 - P(\xi_{i^*} \leq \Phi^{-1}(1-\alpha/n) - \mu^*) \prod_{i \neq i^*} P(z_i \leq \Phi^{-1}(1-\alpha/n)) \\
=&1 - P(\xi_{i^*} \leq \epsilon\sqrt{2\log(n)}(1+o(1))) \left(1 - \frac{\alpha}{n}\right)^{n-1} \\
\to&1 - \exp(-\alpha).
\end{aligned}$$

3.3 Optimality against sparse alternatives

From the above discussions, we know that Bonferroni's method has trivial power when μ^* is below the detection threshold $\sqrt{2\log(n)}$. A natural question to ask here is whether there exists some test that has non-negligible asymptotic power in this scenario. We show below that this detection threshold cannot be improved using any test of the global null against the sparse alternative.

To prove this, we reduce our composite alternative to a simple hypothesis and show that the optimal test given by the Neyman-Pearson Lemma still does no better than flipping a biased coin. In particular, we assume that

$$H_a : \{\mu_i\} \sim \pi,$$

where π denotes a joint distribution on $\{\mu_i\}$ which randomly select $i^* \in \{1, 2, \ldots, n\}$ and set $\mu_{i^*} = (1-\epsilon)\sqrt{2\log(n)}$ and other μ_is to be zero.

Note that we have a simple null and a simple alternative in this case. Let $\mathbf{z} = (z_1, \ldots, z_n)$. Recall that $\mathbf{z}_i = \mu_i + \xi_i$, where $\xi_i \sim N(0,1)$. We first write down the joint densities of $\mathbf{z}$ under both the null and the alternative:

$$\begin{aligned}f_0(\mathbf{z}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right), \\ f_a(\mathbf{z}) &= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_j-\mu^*)^2}{2}\right) \prod_{i\neq j} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right),\end{aligned}$$

where we have a n-component mixture distribution under H_a. The likelihood ratio is given by

$$l(\mathbf{z}) = \frac{f_a(\mathbf{z})}{f_0(\mathbf{z})} = \frac{1}{n}\sum_{j=1}^{n} \exp\left(z_j\mu^* - \frac{(\mu^*)^2}{2}\right).$$

Exercise 1.5: Show that under H_0, $l(\mathbf{z}) \rightarrow^p 1$.

To construct a α level test, we find $q_{1-\alpha}$ such that

$$P_{H_0}(l(\mathbf{z}) \geq q_{1-\alpha}) = \int_{l(\mathbf{z})\geq q_{1-\alpha}} dP_0(\mathbf{z}) = \alpha,$$

where P_0 and P_a denote the distributions of $\mathbf{z}$ under the null and alternative respectively. We now study the power of the test (denoted by β_n) under H_1. Note that

$$\begin{aligned}\beta_n =& P_{H_a}(l(\mathbf{z}) \geq q_{1-\alpha}) \\ =& \int_{l(\mathbf{z})\geq q_{1-\alpha}} dP_a(\mathbf{z}) \\ =& \int_{l(\mathbf{z})\geq q_{1-\alpha}} \frac{dP_a(\mathbf{z})}{dP_0(\mathbf{z})} dP_0(\mathbf{z}) \\ =& \int_{l(\mathbf{z})\geq q_{1-\alpha}} \{l(\mathbf{z})-1\} dP_0(\mathbf{z}) + \int_{l(\mathbf{z})\geq q_{1-\alpha}} dP_0(\mathbf{z}) \\ =& \int_{l(\mathbf{z})\geq q_{1-\alpha}} \{l(\mathbf{z})-1\} dP_0(\mathbf{z}) + \alpha.\end{aligned}$$

As $l(\mathbf{z}) \rightarrow^p 1$ under the null, using the dominated convergence theorem, we have

$$\int_{l(\mathbf{z})\geq q_{1-\alpha}} \{l(\mathbf{z})-1\} dP_0(\mathbf{z}) \rightarrow 0,$$

suggesting that $\beta_n \rightarrow \alpha$. In other words, the likelihood ratio test has

$$P_{H_0}(\text{Type I error}) + P_{H_1}(\text{Type II error}) \rightarrow 1,$$

where $P_{H_1}(\text{Type II error}) = 1 - \beta_n$.