STAT 542: Multivariate Analysis **Spring 2021**

Lecture 11: Multiple hypothesis test

Instructor: Yen-Chi Chen

11.1 Introduction

The multiple hypothesis testing is the scenario that we are conducting several hypothesis tests at the same time. Suppose we have n tests, each leads to a p-value. So we can view the 'data' as $P_1, \cdots, P_n \in [0,1]$, where P_i is the p-value of the i-th test. We can think of this problem as conducting hypothesis tests of n nulls: $H_{1,0}, \cdots, H_{n,0}$.

Example. As an illustration example, consider linear regression with a univariate response $Y \in \mathbb{R}$ and a multivariate covariate $X \in \mathbb{R}^d$. We consider a linear model: $\mathbb{E}(Y|X) = \alpha + X^T\beta$. A common scenario in scientific study is to test if every coefficient is 0. Namely, the null hypotheses are

$$H_{1,0}: \beta_1 = 0, H_{2,0}: \beta_2 = 0, \cdots, H_{d,0}: \beta_d = 0,$$

where $\beta = (\beta_1, \cdots, \beta_d)^T$. In this case, $n = d$.

A multiple testing procedure is a map $\Gamma: [0,1]^n \to [0,1]$ the quantity $\Gamma(P_1, \cdots, P_n)$ is the final threshold we will be using. We reject the i-th null hypothesis if

$$P_i < \Gamma(P_1, \cdots, P_n).$$

The case where we do not perform any correct for multiple testing corresponds to the choice $\Gamma_{\mathsf{un}}(P_1, \cdots, P_n) = \alpha$. It is known that such choice could lead to many falsely rejected null hypothesis. For instance, suppose all null hypothesis are correct, we will reject about α proportion of them! The chance that we do not falsely reject any null hypothesis is $1 - (1-\alpha)^n$, which will be close to 1 when n is large.

11.2 Familywise error rate (FWER) control: Bonferroni correction

The *Bonferroni correction* is a simple method that aims at controlling the *Familywise error rate (FWER)*. The FWER is the chance of making any type-1 error when we perform the hypothesis testing for all the n tests. The usual type-1 error rate is $P(\text{reject } H_0; H_0 \text{ is true})$. The FWER is

$$P(\text{there exist } i \text{ such that reject } H_{i,0}; H_{0,i} \text{ is true}).$$

Namely, controlling FWER to be α means that we want to ensure that when we reject any null hypothesis, the chance of falsely reject *any* null is less than α.

As we have argue, if we reject a null when the p-value is less than α, we may not be able to control the FWER to be α.

The Bonferroni correction provides an elegant solution to this problem. Using the Bonferroni correction, we reject null hypothesis i if

$$P_i < \alpha/n.$$

Suppose we have $n = 100$ tests and we want to control FWER at $\alpha = 0.05$ (5%). We will reject any null hypothesis if its p-value is less than $0.05/100 = 0.0005$.

To see why Bonferroni correction leads to a FWER control, we consider the most extreme case that all null hypotheses are correct. The FWER is

$$\begin{aligned} P(\text{there exist } i \text{ such that reject } H_{i,0}; H_{0,i} \text{ is true}) &= P(\text{there exist } i \text{ such that } P_i < \alpha/n) \\ &= P\left(\cup_{i=1}^n \{P_i < \alpha/n\}\right) \\ &\leq \sum_{i=1}^n P(P_i < \alpha/n) \\ &= n \times \alpha/n = \alpha. \end{aligned}$$

Note that Bonferroni correction controls the FWER at α *regardless of the p-values are independent or not.* So it is a conservative approach.

The Bonferroni correction corresponds to the multiple testing procedure

$$\Gamma_{\mathsf{BC}}(P_1, \cdots, P_n) = \frac{\alpha}{n}.$$

11.3 Controlling the FDR: BH approach

While Bonferroni correction is a simple method to control FWER, it tends to reject very few null hypothesis. In some applied research, falsely rejecting a few hypotheses may not be a severe problem as long as the falsely rejection proportion is small. This leads to another concept called *false discovery rate (FDR).*

The FDR is the expected proportion of falsely rejected null hypothesis. In multiple testing, given any threshold $\Gamma(P_1, \cdots, P_n) = \gamma$, the final result can be viewed as in the following table (note that this table is unknown to us but we can think of that there exists such table):

Correct hypothesis	Not Reject	Reject	Total
H_0	U	V	n_0
H_1	T	S	n_1
	W	R	n

Namely, V is the total number of correct nulls but we falsely reject them and S is the total number of incorrect nulls and we successfully reject them. The FWER is the probability $P(V = 0)$.

With the above table, the FDR is the quantity

$$FDR = \mathbb{E}\left(\frac{V}{R \vee 1}\right),$$

where $R \vee 1 = \max\{R, 1\}$. Sometimes people will write it as $FDR = \mathbb{E}\left(\frac{V}{R}\right)$. We modify the denominator to avoid problems when $R = 0$.

Formally, if we use a multiple testing procedure with threshold $\Gamma(P_1, \cdots, P_n) = \gamma$, quantities in Table (11.3) may depends on γ. A proper way to write it is Table 11.3.

So the FWER when we use the procedure $\Gamma(P_1, \cdots, P_n) = \gamma$ is $FWER(\gamma) = P(V_\gamma > 0)$ and the FDR is

$$FDR(\gamma) = \mathbb{E}\left(\frac{V_\gamma}{R_\gamma \vee 1}\right).$$

Correct hypothesis	Not Reject	Reject	Total
H_0	U_γ	V_γ	n_0
H_1	T_γ	S_γ	n_1
	W_γ	R_γ	n

As we have seen in the Bonferroni correction, choosing $\gamma = \alpha/n$ controls FWER to be α. How should we choose γ to control the FDR?

The Benjamini-Hochberg (BH) approach is a very popular method to control the FDR at α. It is based on a simple reference rule from ordered p-values. Let

$$P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$$

be the ordered p-values. The BH procedure first finds the number

$$\widehat{k} = \max\left\{k : p_{(k)} \leq \frac{k}{n}\alpha\right\}. \tag{11.1}$$

And then reject all the null hypotheses with p-values less than the $\widehat{k}$-th smallest p-value. Namely, the threshold is

$$\Gamma_{\mathsf{BH}}(P_1, \cdots, P_n) = p_{(\widehat{k})} = \frac{\widehat{k}}{n}\alpha.$$

It was proved in

> Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

11.4 Controlling the FDR: Storey's approach

In this section, we introduce a famous approach to control the FDR called Storey's approach, which is based on the following paper:

> Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 479-498.

Storey's approach is an asymptotic method for controlling the FDR that has a much better power than the BH approach. Storey's idea is to view the multiple testing as a random procedure as follows.

Suppose we have n hypothesis, each hypothesis can be viewed as IID Bernoulli random variables $A_1, \cdots, A_n \sim \mathsf{Ber}(1-\pi_0)$ such that $A_i = 0$ if the null hypothesis $H_{i,0}$ is correct. The proportion π_0 can be viewed as the proportion of null hypothesis. Let $P_1, \cdots, P_n$ be the p-values of each test. We assume that $P_1, \cdots, P_n$ are independent (but not necessarily identically distributed). Note that the above stochastic model on hypothesis was first appeared in

> Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. Journal of the American statistical association, 96(456), 1151-1160.

If we reject all p-values less than γ, the FDR can be written as the following probability:

$$FDR(\gamma) = P(A = 0|P < \gamma) = \frac{P(A = 0, P < \gamma)}{P(P < \gamma)} = \frac{P(P < \gamma|A = 0)P(A = 0)}{P(P < \gamma)} = \frac{\gamma\pi_0}{P(P < \gamma)}. \tag{11.2}$$

Thus, for any γ, we can estimate its FDR by

$$\widehat{FDR}(\gamma) = \frac{\gamma\widehat{\pi}_0}{\widehat{P}(P < \gamma)}, \tag{11.3}$$

where

$$\widehat{P}(P < \gamma) = \frac{1}{n}\sum_{i=1}^{n} I(P_i < \gamma) = \frac{R}{n}$$

and $\widehat{pi}_0$ is some suitable estimator of π_0, the proportion of null hypothesis.

Storey has a key insight on how to estimate π_0. We know that large p-values are mostly from the null hypothesis, i.e., for $\lambda >> 0$, $P(P > \lambda|A = 1) \approx 0$, and the p-value under H_0 will be from a uniform distribution over $[0, 1]$. This implies that

$$\begin{aligned} P(P > \lambda) &= P(P > \lambda, A = 1) + P(P > \lambda, A = 0) \\ &= \underbrace{P(P > \lambda|A = 1)}_{\approx 0} P(A = 1) + P(P > \lambda|A = 0)P(A = 0) \\ &\approx (1 - \lambda)\pi_0. \end{aligned}$$

The probability $P(P > \lambda)$ can be estimated by empirical proportion $\widehat{P}(P > \lambda) = \frac{1}{n}\sum_{i=1}^{n} I(P_i > \lambda)$, which leads to the estimator

$$\widehat{\pi}_{0,\lambda} = \frac{1}{n(1 - \lambda)}\sum_{i=1}^{n} I(P_i > \lambda). \tag{11.4}$$

Therefore, we obtain an elegant estimator of $FDR(\gamma)$ as

$$\widehat{FDR}^{\dagger}_{\lambda}(\gamma) = \frac{\gamma n\widehat{\pi}_{0,\lambda}}{R} = \frac{\gamma}{(1 - \lambda)R}\sum_{i=1}^{n} I(P_i > \lambda) = \frac{\gamma n\widehat{\pi}_{0,\lambda}}{\frac{1}{n}\sum_{i=1}^{n} I(P_i < \gamma)}. \tag{11.5}$$

Note that the total number of rejected nulls $R = R_\gamma = \sum_{i=1}^{n} I(P_i < \gamma)$. In finite sample case, we may have $R = 0$ so the above estimator is often refined as

$$\widehat{FDR}_{\lambda}(\gamma) = \frac{\gamma}{(1 - \lambda)(R \vee 1)}\sum_{i=1}^{n} I(P_i > \lambda).$$

The quantity λ is a tuning parameter in this procedure.

Note that sometimes we may be interested in the positive FDR (pFDR)

$$pFDR(\gamma) = \mathbb{E}\left(\frac{V}{R}|R > 0\right) = FDR(\gamma)/P(R > 0).$$

In this case, we can estimate $P(R > 0)$ via $1 - (1 - \pi_0)^n$ so an estimator of pFDR is

$$\widehat{pFDR}_{\lambda}(\gamma) = \frac{\widehat{FDR}_{\lambda}(\gamma)}{1 - (1 - \widehat{\pi}_{0,\lambda})^n}.$$

While this idea is elegant, it estimates the FDR *asymptotically*. So in the finite sample case, we may not be able to control FDR exactly.

Finally, a simple threshold to control the FDR to be α is via rejecting all null hypothesis whose p-value is less than $\widehat{\gamma}_\alpha$, where

$$\widehat{\gamma}_{\alpha,\lambda} = \sup\left\{\gamma : \frac{\widehat{\pi}_{0,\lambda}\gamma}{\frac{1}{n}\sum_{i=1}^n I(P_i < \gamma)} \leq \alpha\right\}. \tag{11.6}$$

Using the notation at the beginning, this corresponds to the multiple testing procedure

$$\Gamma_{\mathsf{ST},\lambda}(P_1, \cdots, P_n) = \widehat{\gamma}_{\alpha,\lambda} = \sup\left\{\gamma : \frac{\widehat{\pi}_{0,\lambda}\gamma}{\frac{1}{n}\sum_{i=1}^n I(P_i < \gamma)} \leq \alpha\right\}.$$

The threshold in equation (11.6) corresponds to a population threshold

$$\gamma^*_\alpha = \gamma_\alpha(\pi_0, G) = \sup\left\{\gamma : \frac{\pi_0 \cdot \gamma}{G(\gamma)} \leq \alpha\right\},$$

where $G(t) = P(P < t)$ is the marginal distribution of p-values. If we have any estimator of π_0 and G, we can use it to form a plug-in estimate of the threshold. The threshold $\gamma_\alpha(\pi_0, G)$ is called *oracle threshold* in the following paper:

> [GW2004] Genovese, C., & Wasserman, L. (2004). A stochastic process approach to false discovery control. The annals of statistics, 32(3), 1035-1061.

The Storey's approach corresponds to the threshold $\gamma_\alpha(\widehat{\pi}_{0,\lambda}, \widehat{G})$, where $\widehat{G}$ is the empirical distribution and you can show that the BH approach corresponds to $\gamma_\alpha(1, \widehat{G})$.

11.4.1 Connection to BH approach

Using equation (11.2), we can show that BH approach is a conservative method that controls the asymptotic FDR at $\alpha \cdot \pi_0$, rather than α.

Recall that in BH approach, we reject all null hypothesis whose p-value is less than $\frac{\widehat{k}}{n}\alpha$, where $\widehat{k} = \max\left\{k : p_{(k)} \leq \frac{k}{n}\alpha\right\}$ is from equation (11.1). Using equation (11.2), the choice $\gamma = \frac{\widehat{k}}{n}\alpha$ controls the FDR at

$$\frac{\frac{\widehat{k}}{n}\alpha\pi_0}{\widehat{k}/n} = \alpha\pi_0.$$

Note that we replace $P(P < \gamma)$ by $\widehat{P}(P < \frac{\widehat{k}}{n}\alpha) = \frac{\widehat{k}}{n}$. As a result, choose γ that controls $\widehat{FDR}_\lambda(\gamma) < \alpha$ will lead to a more powerful method (compared to the BH approach) that asymptotically controls the same FDR at α

11.4.2 Identifiability issue

While the stochastic model on multiple testing is appealing, it may not be identifiable. Namely, given the same distribution that we can observed (G: p-value distribution), there could be different pairs (π, F) such that

$$G(t) = \pi \cdot t + (1 - \pi) \cdot F(t),$$

where $\pi = P(A = 0)$ is the chance of null hypothesis is correct and F is the p-value distribution under H_1.

A simple assumption to identification is to assume that the CDF F is *pure*, i.e., the *essential infimum*[1] *of its PDF f is 0*. One way to think of this is that we need the PDF of F to drop to 0 at some point inside $[0, 1]$, so the density at that point is completely from the density of uniform (p-value density under null), which uniquely determines the proportion π. See Section 3.1 of [GW2004].

11.5 False discovery/negative processes

Given a threshold γ, we define the false discovery process (FDP) as

$$FDP(\gamma) = \frac{\sum_{i=1}^n I(P_i < \gamma) A_i}{\sum_{i=1}^n I(P_i < \gamma) + I(\text{all } P_i \geq \gamma)}$$

and the false negative process (FNP) as

$$FNP(\gamma) = \frac{\sum_{i=1}^n I(P_i \geq \gamma)(1 - A_i)}{\sum_{i=1}^n I(P_i \geq \gamma) + I(\text{all } P_i < \gamma)}.$$

They both are stochastic processes (indexed by γ). Moreover,

$$\mathbb{E}(FDP(\gamma)) = FDR(\gamma), \quad \mathbb{E}(FNP(\gamma)) = FNR(\gamma),$$

where FNR is the false negative rate.

[1] see https://en.wikipedia.org/wiki/Essential_supremum_and_essential_infimum