LECTURE 16 NOTES

This lecture is a high-level introduction to global and multiple testing. Most of the material is from Emmanuel Candes' course on modern statistical theory at Stanford University, and we refer to his notes for further details.

1. Global testing. Let $\{H_{0,i}\}_{i\in[n]}$ be a set of n hypotheses. We wish to test the *intersection null*

$$H_0 : \textstyle\bigcap_{i\in[n]} H_{0,i}. \tag{1.1}$$

That is, we wish to test the hypotheses that all the nulls are valid. We assume we are given p-values $\{\mathbf{p}_i\}_{i\in[n]}$ that are stochastically larger than $\mathrm{unif}(0,1)$ when the corresponding null is valid. We wish to combine the p-values to test the global null.

Example 1.1. *Consider an investigation of the genetic causes of a disease. There are n genes under investigation, and we wish to test whether any of them are linked to the disease: the i-th null hypotheses states that the i-th gene has no link to the disease. The global null states that none of the n genes under consideration are linked to the disease. For a concrete example, we refer to Efron (2010), Section 2.1.*

1.1. *Bonferroni's global test.* The simplest method for testing the global null is *Bonferroni's global test.* Given a prescribed level α, Bonferroni's test tests each $H_{0,i}$ at level $\frac{\alpha}{n}$ and rejects H_0 if any of the $H_{0,i}$'s are rejected. Equivalently, Bonferroni's test rejects when $\min_{i\in[n]} \mathbf{p}_i < \frac{\alpha}{n}$. It is possible to show that overall level is at most α:

$$\begin{aligned}\mathbf{P}_0\big(\min_{i\in[n]} \mathbf{p}_i < \tfrac{\alpha}{n}\big) &= \mathbf{P}_0\Big(\textstyle\bigcup_{i\in[n]}\{\mathbf{p}_i < \tfrac{\alpha}{n}\}\Big) \\ &\le \textstyle\sum_{i\in[n]} \mathbf{P}_0\big(\mathbf{p}_i < \tfrac{\alpha}{n}\big) \\ &= \alpha.\end{aligned}$$

Although the union bound seems crude, it is possible to show that if the p-values are independent, the level of the test is

$$\begin{aligned}\mathbf{P}_0\big(\min_{i\in[n]}\mathbf{p}_i \le \tfrac{\alpha}{n}\big) &= 1 - \mathbf{P}_0\Big(\textstyle\bigcap_{i\in[n]}\{\mathbf{p}_i \ge \tfrac{\alpha}{n}\}\Big) \\ &= 1 - \textstyle\prod_{i\in[n]}\mathbf{P}_0\big(\mathbf{p}_i \ge \tfrac{\alpha}{n}\big) \\ &= 1 - \big(1 - \tfrac{\alpha}{n}\big)^n \\ &\to 1 - e^{\alpha},\end{aligned}$$

which is approximately α for any α close to zero. We observe that Bonferroni's test only depends on the smallest p-value. Thus it is suited to situations where we expect at least one of the p-values to be very small. Going back to the heart disease example, Bonferroni's test is a good choice if we expect some the genes to be strongly linked to heart disease.

1.2. *Fisher's combination test.* Fisher's combination test complements Bonferroni's test: it is powerful if there are many small effects. The test rejects if

$$-2\textstyle\sum_{i\in[n]}\log\mathbf{p}_i$$

is large. Under the global null, assuming the p-values are independent, it is possible to derive the exact distribution of the test statistic.

LEMMA 1.2. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{unif}(0,1)$. *We have* $-2\sum_{i\in[n]}\log\mathbf{x}_i \sim \chi^2_{2n}$.

PROOF. We observe that $\mathbf{x}_i \sim \text{unif}(0,1)$, $-2\log\mathbf{x}_i \sim \chi^2_2$. Indeed,

$$\mathbf{P}(-2\log\mathbf{x}_i \le t) = \mathbf{P}\big(\mathbf{x}_i \ge e^{-\frac{t}{2}}\big) = 1 - e^{-\frac{t}{2}},$$

which is the χ^2_2 CDF. We recall the sum of n independent χ^2_2 random variables is a χ^2_{2n} random variable to obtain the stated result. □

We observe that unlike Bonferroni's test, which only depends on the smallest p-value, Fisher's combination test aggregates all the p-values (in log-scale). It is more powerful than Bonferroni's test if the alternative has many weak effects (i.e. small deviations from the null). Bonferroni's test has a hard time detecting many weak effects because it looks for a single strong effect. However, Fisher's combination test considers the aggregate of the p-values, which enables it to detect many weak effects.

2. Multiple hypothesis testing. Thus far, we considered testing the intersection of many hypotheses. We now turn our attention to testing each hypotheses separately. There are only four possible outcomes: $H_{0,i}$ is either true or false; we either accept or reject $H_{0,i}$.

	accept	reject	
$H_{0,i}$ true		$\mathbf{v}$	n_0
$H_{0,i}$ false	$\mathbf{u}$		$n - n_0$
	$n - \mathbf{r}$	$\mathbf{r}$	

2.1. *Controlling the family-wise error rate.* Classical multiple comparison procedures aim to control the family-wise error rate.

Definition 2.1. *The family-wise error rate (FWER) is the probability of making any false rejections,*

$$\text{FWER} := \mathbf{P}(\mathbf{v} > 0).$$

The simplest method that controls the FWER is Bonferroni's method: reject any $H_{0,i}$ whose corresponding p-value is smaller than $\frac{\alpha}{n}$.

Theorem 2.2. *Bonferroni's method controls FWER at level α for any configuration of true and false hypotheses.*

Proof. Without loss of generality, assume $H_{0,1}, \dots, H_{0,n_0}$ are true nulls. By a union bound,

$$\begin{aligned}\mathbf{P}(\mathbf{v} > 0) &= \mathbf{P}\Big(\textstyle\bigcup_{i\in[n_0]}\{\text{falsely rejecting } H_{0,i}\}\Big) \\ &\le \sum_{i\in[n_0]} \mathbf{P}\big(\text{falsely rejecting } H_{0,i}\big) \\ &= \alpha\frac{n_0}{n}\end{aligned}$$

which is at most α. □

There is a close connection between global testing and multiple testing: it is possible to derive a procedure that controls FWER from a global testing procedure. Let $\{H_{0,i}\}_{i\in[n]}$ be a set of n hypotheses. The *closure* of the set of hypotheses is

$$\big\{H_{0,I} := \textstyle\bigcap_{i\in I} H_{0,i} : I \subset [n]\big\}.$$

If $I \subset I'$, we call $H_{0,I}$ a child of $H_{0,I'}$ and $H_{0,I'}$ a parent of $H_{0,I}$. The *closure principle*

1. tests all the hypotheses in the closure at level α. If the hypotheses is an intersection null, test it by a global testing procedure.

2. goes back and accepts all the children of the accepted hypotheses.

We observe that if $H_{0,I}$ is true, its children are necessarily true. The closure principle enforces our observation by going back and accepting all the children of any accepted hypotheses.

Example 2.3. *Consider testing a set of three hypothesis*

$$\big\{H_{0,1}, H_{0,2}, H_{0,3}\big\}$$

The closure of the preceding set of hypotheses consists of 2^3 *hypothesis:*

$$\begin{array}{c} H_{0,\{1,2,3\}} \\ H_{0,\{1,2\}} \quad H_{0,\{1,3\}} \quad H_{0,\{2,3\}} \\ H_{0,1} \quad H_{0,2} \quad H_{0,3} \end{array}$$

Suppose the first step of the closure principle rejected the underlined hypotheses:

$$\begin{array}{c} \underline{H_{0,\{1,2,3\}}} \\ \underline{H_{0,\{1,2\}}} \quad \underline{H_{0,\{1,3\}}} \quad H_{0,\{2,3\}} \\ \underline{H_{0,1}} \quad \underline{H_{0,2}} \quad H_{0,3} \end{array}$$

The second step of the procedure goes back and accepts $H_{0,2}$ *because it is the child of* $H_{0,\{2,3\}}$*, which was not rejected.*

Theorem 2.4. *The closure principle controls FWER on* $\{H_{0,i}\}_{i\in[n]}$.

Proof. Withouth loss of generality, assume the first n_0 hypotheses are true. The closure principle rejects any of these hypotheses only if it rejects $H_{0,[n_0]}$. That is, $\{\mathbf{v} > 0\} \subset \big\{\text{reject } H_{0,[n_0]}\big\}$, which implies

$$\mathbf{P}(\mathbf{v} > 0) \le \mathbf{P}\big(\text{reject } H_{0,[n_0]}\big) \le \alpha.$$

□

The closure principle is a generic recipe for deriving multiple testing procedures that control FWER from global testing procedures. However, it is usually not practical to test all 2^n hypotheses in the closure explicitly. Fortunately, there are ways to avoid testing all the hypotheses in the closure.

For example, consider testing the all intersection nulls in the closure by Bonferroni's test. Let $I_{[i]} \subset [n]$ be the (random) index set that consists of (the indices of) the j largest p-values:

$$I_{[i]} := \{(i), \dots, (n)\}.$$

Since $H_{0,(i)}$ is a child of $H_{0,I_{[1]}}, \dots, H_{0,I_{[i]}}$, the closure principle rejects $H_{0,(i)}$ only if it rejects $H_{0,I_{[1]}}, \dots, H_{0,I_{[i]}}$. Further, if the closure of Bonferroni's test rejects $H_{0,I_{[i]}}$, it also rejects all the children of $H_{0,I_{[i]}}$ that intersects $H_{(i)}$. Indeed,

1. if $H_{0,I}$ is a child of $H_{0,I_{[i]}}$ that intersects $H_{0,(i)}$, $\mathbf{p}_{(i)}$ is the smallest p-value in $\{\mathbf{p}_i : \imath \in I\}$
2. since Bonferroni's test rejects $H_{0,I_{[i]}}$, $\mathbf{p}_{(i)} \le \frac{\alpha}{n-i+1}$, which implies

$$\mathbf{p}_{(i)} \le \tfrac{\alpha}{|I|}.$$

Thus if the closure of Bonferroni's test rejects $H_{0,(i)}$, it also rejects,

$$H_{0,(1)}, \dots, H_{0,(i-1)}.$$

Equivalently, the closure of Bonferroni's test

1. sorts the p-values $\mathbf{p}_{(1)} \le \cdots \le \mathbf{p}_{(n)}$,
2. finds the largest $i_1 \in [n]$ such that $\mathbf{p}_{(i)} \le \frac{\alpha}{n-i+1}$ for all $i \in [i_1]$
3. rejects $H_{0,(1)}, \dots, H_{0,(i_1)}$,

which is also known as *Holm's procedure*.

2.2. *Controlling the false discovery rate.* As the number of hypotheses being tested grows, controlling the FWER becomes a more and more stringent criterion. Intuitively, the chance of a false rejection grows as more hypotheses are tested. In modern applications, practitioners routinely test thousands or tens of thousands of hypotheses, and controlling the FWER is unacceptably conservative. A new approach advanced by Benjamini and Hochberg (1995) aims to control the *false discovery rate*.

Definition 2.5. *The false discovery proportion is the ratio of false rejections to rejections:*

$$\mathrm{Fdp} := \frac{\mathbf{v}}{\max\{\mathbf{r}, 1\}} = \begin{cases} \frac{\mathbf{v}}{\mathbf{r}} & \mathbf{r} > 0, \\ 0 & \mathbf{r} = 0. \end{cases}$$

The false discovery rate is the expected value of Fdp.

We observe that

1. $\mathrm{Fdp} \le \mathbf{1}\{\mathbf{v} > 0\}$. Indeed, if $\mathbf{v} = 0$, $\mathrm{Fdp} = 0$. If $\mathbf{v} > 0$, $\mathrm{Fdp} = \frac{\mathbf{v}}{\mathbf{r}} \le 1$. Thus the FDR is at most the FWER.
2. under the global null, all rejections are false. If $\mathbf{v} > 0$, $\mathrm{Fdp} = 1$. Thus the FDR is equal to the FWER under the global null.

Since FWER is a more stringent criterion than FDR, any procedure that controls the FWER also controls the FDR. However, we wish to avoid controlling the FWER because it is very stringent: too stringent in many modern applications. To control the FDR, Benjamini and Hochberg (1995) proposed the BH(q) procedure:

1. sort the p-values $\mathbf{p}_{(1)} \le \cdots \le \mathbf{p}_{(n)}$,
2. find the largest $i_1 \in [n]$ such that $\mathbf{p}_{(i_1)} \le \frac{i_1}{n} q$.
3. reject $H_{0,(1)}, \dots, H_{0,(i_1)}$.

Theorem 2.6. *If $\{\mathbf{p}_i\}_{i\in[n]}$ are independent, the BH(q) procedure controls the FDR at level $q\frac{n_0}{n}$.*

Proof sketch. Without loss of generality, assume the first n_0 hypotheses are false. Consider rejecting all hypotheses whose p-values are smaller than some threshold t. The Fdp is

$$\mathrm{Fdp}(t) = \frac{\sum_{i\in[n_0]} \mathbf{1}_{[0,t)}(\mathbf{p}_i)}{\sum_{i\in[n]} \mathbf{1}_{[0,t)}(\mathbf{p}_i)} = \frac{\frac{1}{n}\sum_{i\in[n_0]} \mathbf{1}_{[0,t)}(\mathbf{p}_i)}{\widehat{P}_n(t)},$$

where $\widehat{P}_n(t)$ is the empirical CDF of the p-values. Since $\{\mathbf{p}_i\}_{i\in[n_0]}$ are *i.i.d.* $\mathrm{unif}(0,1)$ random variables, we expect

$$\textstyle\sum_{i\in[n_0]} \mathbf{1}_{[0,t)}(\mathbf{p}_i) \approx n_0 t.$$

Thus a (conservative) estimate of the Fdp is

$$\widehat{\mathrm{Fdp}}(t) \approx \frac{n_0 t}{n\widehat{P}_n(t)} \le \frac{t}{\widehat{P}_n(t)}.$$

It remains to choose t. We observe that it is enough to choose among the p-values $\{\mathbf{p}_{(1)}, \dots \mathbf{p}_{(n)}\}$: any threshold in between two p-values leads to the same pattern of rejections as choosing the smaller of the two p-values. The estimate of Fdp if we reject all the p-values smaller than $\mathbf{p}_{(i)}$ is

$$\widehat{\mathrm{Fdp}}(\mathbf{p}_{(i)}) = \frac{\mathbf{p}_{(i)}}{\widehat{P}_n(\mathbf{p}_{(i)})} = \mathbf{p}_{(i)}\big(\tfrac{i}{n}\big)^{-1}.$$

Setting the estimate to be at most α and rearranging, we obtain $p_{(i)} \le q\frac{i}{n}$, which is the i-th BH(q) threshold. Thus the BH(q) procedure chooses the largest t such that $\widehat{\text{Fdp}}(t) \le q$. □

It is possible to rigorize the preceding proof by appealing to the theory of stopping times. The BH(q) procedure starts at $\mathbf{p}_{(n)}$ and steps down to the first sorted p-value such that $\widehat{\text{Fdp}}(\mathbf{p}_{(i)}) \le q$, which is a *stopping time*. It is possible to appeal to theory of stopping times to show that the BH(q) procedure controls the FDR at the nominal level.

The assumption that the p-values are independent is usually violated in practice. However, since there is no better procedure, practitioners use the BH(q) procedure, even if the p-values are dependent. There is some theoretical support for the BH(q) heuristic:

1. the assumption that the p-values are independent can be relaxed to a technical condition called *positive regression dependence on each of a subset (PRDS)*. As long as the p-values are PRDS, the BH(q) procedure controls the FDR at the nominal level.
2. if we allow the p-values to be arbitrarily dependent, it is possible to show that the FDR of the BH(q) procedure is at most $q\frac{n_0}{n}\sum_{i\in[n]} i^{-1}$. Thus, it is possible to maintain FDR control at the nominal level by adjusting the BH(q) thresholds to be $\sum_{i\in[n]} i^{-1}$ times smaller. However, the loss of power is unacceptable in many applications, and the adjusted procedure is rarely used.

Finally, we remark that the BH(q) procedure is conservative by a factor of $\frac{n_0}{n}$. There is a line of research that aims to boost the power of the BH(q) procedure by first estimating the fraction $\frac{n_0}{n}$ and then adjusting the BH(q) thresholds accordingly. Here is such a modified procedure proposed by Storey, Taylor and Siegmund (2004):

1. let $\hat{\pi}_0 = \frac{n-\mathbf{r}(t)}{(1-t)n}$, where $t \in (0,1)$ is a parameter and $\mathbf{r}(t)$ is number of p-values smaller than t.
2. find the largest $i_1 \in [n]$ such that $\mathbf{p}_{(i_1)} \le \frac{i_1\hat{\pi}_0}{n}q$.
3. reject $H_{0,(1)}, \dots, H_{0,(i_1)}$.

The idea is if the non-null p-values are small, all of them are likely smaller than t. Thus $n - \mathbf{r}(t) \approx n_0 - \mathbf{v}(t)$, where $\mathbf{v}(t)$ is the number of null p-value smaller than t. Since the null p-values are unif$(0,1)$, we expect

$$\frac{n_0 - \mathbf{v}(t)}{(1-t)n} \approx \frac{n_0 - n_0 t}{(1-t)n} = \frac{n_0}{n}.$$

To wrap up, we study the BH(q) procedure from a Bayesian perspective. Consider the hierarchical model:

$$\mathbf{h}_i \overset{i.i.d.}{\sim} \mathrm{Ber}(\pi), \quad \mathbf{p}_i \mid \mathbf{h}_i \overset{i.i.d.}{\sim} F_{\mathbf{h}_i},$$

where F_0 and F_1 are the distributions of the p-values under the null and the alternative. The marginal distribution of each $\mathbf{p}_i$ is a mixture:

$$F = \pi_0 F_0 + (1-\pi_0) F_1.$$

The "Bayes false discovery rate" (bFDR) is

$$\begin{aligned} \mathrm{bFDR}(t) &:= \mathbf{P}\big(\mathbf{h}_1 = 0 \mid \mathbf{p}_1 < t\big) \\ &= \frac{\mathbf{P}(\mathbf{p}_1 < t \mid \mathbf{h}_1 = 0)}{\mathbf{P}(\mathbf{p}_1 < t)} \\ &= \frac{\pi_0\, t}{\mathbf{P}(\mathbf{p}_1 < t)}. \end{aligned}$$

In practice, we cannot evaluate $\mathrm{bFDR}(t)$ because the fraction of nulls π_0 and the distribution of the p-values under the alternative F_1 is unknown. However, a conservative estimate of $\mathrm{bFDR}(t)$ is

$$\begin{aligned} \widehat{\mathrm{bFDR}}(t) &:= \frac{nt}{\sum_{i\in[n]} \mathbf{1}_{[0,t)}(\mathbf{p}_i)} \\ &\ge \frac{\pi_0\, t}{\frac{1}{n}\sum_{i\in[n]} \mathbf{1}_{[0,t)}(\mathbf{p}_i)} \\ &\approx \frac{\pi_0\, t}{\mathbf{P}(\mathbf{p}_1 < t)}. \end{aligned}$$

We observe that $\widehat{\mathrm{bFDR}}(\mathbf{p}_{(i)}) = \frac{n\mathbf{p}_{(i)}}{i}$. Rearranging, we deduce

$$\widehat{\mathrm{bFDR}}(\mathbf{p}_{(i)}) \le q \iff \mathbf{p}_{(i)} \le \frac{i}{n} q.$$

Thus the BH(q) procedure picks the largest threshold that keeps $\widehat{\mathrm{bFDR}}(t)$ smaller than the nominal level.

The BH(q) procedure seems mysterious at first blush. However, from a Bayesian perspective, it is quite natural. Here is an example of a Bayesian approach leading to a method that enjoys good frequentist properties. We end by quoting Efron (2010):

> It is always a good sign when a statistical procedure enjoys both a frequentist and Bayesian support, and the BH algorithm passes the test.

References.

BENJAMINI, Y. and 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)* 289–300.

EFRON, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* **1**. Cambridge University Press.

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 187–205.

YUEKAI SUN
BERKELEY, CALIFORNIA
DECEMBER 5, 2015