STATS 300C: Theory of Statistics **Spring 2022**

Lecture 4 — April 7, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Rex Shen*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by Emmanuel Candès, Brad Nelson, and David Ritzwoller.*

4.1 Summary Outline

First, we recap the main material covered in the lectures preceding this one. In particular, we provide concluding remarks on global testing. Then, we introduce multiple testing, in which the modified setting is to test $H_{0,i}$ for $i \in \{1, \ldots, n\}$ separately, as opposed to the global testing settting. We introduce several procedures to control the Family Wise Error Rate (FWER), which is the probability of making at least one Type I Error. Particularly, such methods include Holm's Procedure and Bonferonni. Finally, we preview some of the material covered in Lecture 5, such as the closure principle.

4.2 Concluding Remarks on Global Testing

The material reviewed at the beginning of this lecture is summarized in the scribe notes for Lecture 3, to which we refer the reader for further detail. We enumerate several additional points of emphasis:

1. If Tukey's Higher Criticism Statistic exceeds the threshold

$$\sqrt{(1+\epsilon)2\log\log(n)}$$

for some positive ϵ, we reject the test. Moreover, it asymptotically achieves the optimal detection threshold for the Sparse Mixtures Model of a $N(0,1)$ and $N(\mu,1)$ without knowledge of μ.

2. In practice, the asymptotic approximation to the critical value of the Higher Criticism statistic may not be accurate in finite samples. It is recommended to obtain critical values for Tukey's Higher Criticism statistic with simulation.

3. Moreover, in finite samples, the behavior of the process,

$$W_n(t) = \frac{\sqrt{n}(F_n(t) - t)}{\sqrt{t(1-t)}}$$

may not be well approximated by a Brownian Bridge for small values of t. In particular, for n large and t small, a Poisson(np) distribution provides a more accurate approximation to $F_n(t)$ than a suitably centered and scaled Gaussian; see Pollard (2015) for

further discussion. As a result, the tails of $W_n(t)$ will be much heavier for t close to zero than for t away from zero, and so the behavior of $\sup_{\frac{1}{n} \leq t \leq \alpha_0} W_n(t)$ will highly depend on the behavior of $W_n(t)$ for t close to zero. Therefore, critical values computed with simulation may be conservative and result in a test that performs more similarly to Bonferroni's Method than to a test using Fisher's Statistic.

4.3 Multiple Testing

	Healthy (m_0 patients)	Prostate Cancer (m_1 patients)
Expression Level of Gene i	$Y_{ij}^{(0)}, 1 \leq j \leq m_0$	$Y_{ij}^{(1)}, 1 \leq j \leq m_1$

Table 4.1. Microarray Data

4.3.1 A Motivating Example

Suppose that we have obtained data on the expression levels of $n = 6033$ genes measured on $m_1 = 52$ patients with prostate cancer and $m_0 = 50$ patients who are healthy. We let Y_{ij} indicate the expression level on gene i for patient j, and let the superscripts (1) and (0) indicate whether the observation was collected from patient populations with or without a prostate cancer diagnosis, respectively. We are interested in assessing whether each gene i in $1, \cdots, n$ is differentially expressed in the two patient populations. These data are summarized in Table 4.1.

This problem can be formalized as a collection of tests of the n null hypotheses $H_{0,i}$ against the alternatives $H_{1,i}$, where the null hypotheses $H_{0,i}$ may, for example, indicate either the restrictions that $Y_{ij}^{(0)} \stackrel{d}{=} Y_{ij}^{(1)}$ or that $E[Y_{ij}^{(0)}] = E[Y_{ij}^{(1)}]$. By contrast, tests of the global null $H_0 = \cap_i H_{0,i}$ assess whether there is evidence that prostate cancer has any genetic component. Additionally, suppose that we have computed the p-values p_i for each test i in $1, \ldots, n$ with e.g., a permutation test or a t-test, depending on the form of the hypothesis testing problem. For example, we can compute the approximate t-value

$$T_i = \frac{\bar{Y}_i^{(1)} - \bar{Y}_i^{(0)}}{s_i},$$

where $\bar{Y}_i^{(w)}$ is the sample mean of $Y_{ij}^{(w)}$ over j and s_i^2 is an estimate of the standard error of the numerator, and a corresponding p value is

$$p_i = \mathbb{P}(|t_{100}| > |T_i|),$$

where t_{100} is a Student t Random Variable with 100 degrees of freedom. With this data, we must decide which null hypotheses to reject and which to accept. This example illustrates a Multiple Hypothesis Testing Problem, which we will discuss for the following several lectures. See Chapter 2 of Efron (2012) for further discussion.

	Accepted	Rejected	Total
True	U	V	n_0
False	T	S	$n - n_0$
Total	$n - R$	R	n

Table 4.2. Outcomes in Multiple Testing

4.3.2 Outcomes in Multiple Testing

There are four types of outcomes in multiple testing, illustrated by Table 4.2. 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 and incorrectly accepted hypotheses, *i.e.*, the number of true and false discoveries. The random variables T and S indicate the number of missed detections and correct rejections respectively. The number of true hypotheses is denoted by n_0. The 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, 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.

4.4 Multiple Testing

4.4.1 Familywise Error Rate

Classical multiple testing procedures aim to control the familywise error rate

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

in a strong sense i.e., under all configurations of true and false hypotheses. 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, given by

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

See Lehmann and Romano (2005) for further discussion. We will discuss alternative notions of error control in subsequent lectures.

4.5 Controlling the Familywise Error Rate

4.5.1 Bonferonni's Method

Bonferroni's method for multiple hypothesis testing rejects all hypotheses with p-values below the threshold $\frac{\alpha}{n}$. That is, we reject $H_{i,0}$ if and only if $p_i \leq \frac{\alpha}{n}$ for each i in $1, \ldots, n$. This is a very simple method, which controls FWER.

Theorem 1. Bonferroni's Method controls FWER at level α in a strong sense; that is,

$$\text{FWER} \leq \mathbb{E}[V] \leq \alpha.$$

Proof. Observe that

$$\mathbb{P}(V \geq 1) \leq \mathbb{E}[V]$$

by Markov's Inequality. By the linearity of expectation, we have

$$\mathbb{E}[V] \leq \sum_{i:\ H_{0,i} \text{ is true}} \mathbb{P}\left(p_i \leq \frac{\alpha}{n}\right) = \frac{n_0}{n}\alpha \leq \alpha.$$

This completes the proof. ■

Remark 1. Note, Bonferroni's Method exhibits a stronger notion of error control, in that $E[V] \leq \alpha$. This is referred to as control of the per familywise error rate.

Remark 2. In the proof of Theorem 1, we made no use of the independence or dependence assumptions of the p-values p_i. Thus, Bonferroni's Method is valid for dependent tests.

4.5.2 Sidak's Procedure

Sidak's Procedure is a refinement to the Bonferroni's Method in which hypotheses with p-values below the threshold $1-(1-\alpha)^{\frac{1}{n}}$ are rejected. This procedure is only applicable if the hypotheses are independent. Observe that under independence, if each hypothesis is tested at level α_n, then

$$\mathbb{P}(V \geq 1) = 1 - P(V = 0) = 1 - (1-\alpha_n)^{n_0}.$$

In this case, we would like to control this quantity. Hence, we could solve the equation

$$1 - (1-\alpha_n)^{n_0} \leq 0.05$$

for α_n for which some algebra shows that $\alpha_n \leq 1 - (1-0.05)^{\frac{1}{n_0}}$.

4.5.3 Weak Control

Now, we introduce the following definition.

Definition 1. If a testing procedure controls the FWER under the global null, then it controls the FWER weakly.

To gain some intuition, consider the following two-step multiple testing procedure suggested by Fisher:

1. Implement a global test for $H_0 = \cap_{i=1}^{n} H_{0,i}$
2. If H_0 is rejected, reject $H_{0,i}$ if $p_i \leq \alpha$.

For example, suppose that the first stage consists of Bonferroni's Global Test, rejecting H_0 if $\min_{1\le i\le n} p_i \le \frac{\alpha}{n}$. 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 configurations of null and non-null hypothesis when some hypotheses are non-null. In particular, suppose there is one very strong signal, say with p-value 10^{-8} where $n = 10^4$. Then, if all other hypotheses are null, the number of false discoveries can be very large, in fact, on average, it is $n_0\alpha$. This illustrates how global testing and multiple testing with control of the FWER are very different procedures.

4.5.4 Benjamini-Hochberg procedure

Now, consider the procedure above where the first-stage is implemented with Simes Global Test, which rejects if

$$\min_{1\le i\le n} \left\{ p_i \frac{n}{i} \right\} \le \alpha,$$

and the second step rejects hypotheses $p_{(i)}$ for $i \le j$ with $p_{(j)} \le \frac{j\alpha}{n}$ for ordered p-values $p_{(1)} \le \cdots \le p_{(n)}$. That is, given a rejection of Simes Global Test in the first step, this procedure rejects hypotheses if there are other hypotheses whose p-values that lie below the Simes Critical Line $\frac{i\alpha}{n}$.

This is known as the **Benjamini-Hochberg procedure**. This procedure controls the FWER weakly by the same argument as the preceding example. However, it does not control the FWER strongly. To see this, suppose that the are m strong signals with p-values always very close to zero for some large m. If the remaining hypotheses are all null, their order statistics will be roughly $\frac{1}{n_0}, \frac{2}{n_0}, \ldots$, which will be below the critical thresholds $\frac{(m+1)\alpha}{n}$, $\frac{(m+2)\alpha}{n}, \ldots$, and so will be false rejections. Yet, as we will see, the expected proportion of false discoveries, known as the FDR, is controlled under the Benjamini-Hochberg procedure. See Benjamini, Hochberg (1995) for a rigorous proof. We can write the general algorithm for Benjamini-Hochberg as follows.

Algorithm 1 Benjamini-Hochberg Procedure

$j = n$.
while $p_{(j)} > \frac{\alpha}{n-j+1}$: **do**
  $j = j - 1$
end while
Reject $H_{(1)}, \ldots, H_{(j)}$

Note that this procedure scans backwards and stops at the first p value such that it does not exceed the desired threshold. This is known as a step up procedure. On the other hand, Holm's Procedure scans forward and stops at the first p-value that exceeds the threshold. Contrarily, this is called the step down procedure. It may sound counter-intuitive as to why Benjamini-Hochberg is called a step-up and Holm's Procedure is called a step down, but we

can reason this as follows. In a step-up procedure, we start with the hypothesis associated with the least significant test statistic (i.e. could be a t statistic) and move toward the hypothesis with the most significant, stopping until a significant result is obtained. In other words, the test statistics are increasing, whereas the p values are decreasing. On the other hand, the step down procedure can be reasoned in a similar way. Note, these concepts will be discussed in more detail in Lecture 6.

4.5.5 Holm's Procedure

Consider the ordered p-values

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

and corresponding hypotheses $H_{(1)}, H_{(2)}, ..., H_{(n)}$. Holm's Procedure is a "step down" multiple testing procedure. That is,

Step 1: If $p_{(1)} \leq \frac{\alpha}{n}$, then reject $H_{(1)}$ and go to Step 2. Otherwise, accept $H_{(1)}, H_{(2)}, \ldots, H_{(n)}$ and stop.

Step i: If $p_{(i)} \leq \frac{\alpha}{n-i+1}$, then reject $H_{(i)}$ and go to Step i+1. Otherwise, accept $H_{(i)}, H_{(i+1)}, \ldots, H_{(n)}$ and stop.

Step n: If $p_{(n)} \leq \alpha$, then reject $H_{(n)}$. Otherwise, accept $H_{(n)}$.

In short, the procedure stops the first time $p_{(i)}$ exceeds the critical value $\alpha_i = \frac{\alpha}{n-i+1}$, rejecting all $H_{(j)}$ with $j \leq i$. Note, Holm's Procedure is less conservative than the Bonferroni, as the threshold becomes more "liberal" as larger p-values are considered. Moreover, Holm's Procedure controls the FWER strongly. We write the algorithm for Holm's Procedure below.

Algorithm 2 Holm's Procedure

$j = 0$.
while $p_{(j+1)} \leq \frac{\alpha}{n-j}$: **do**
  $j = j + 1$
end while
Reject $H_{(1)}, \ldots, H_{(j)}$

Theorem 2. Holm's Procedure controls the FWER at level α in a strong sense.

Proof. Let i_0 indicate the rank of the smallest null p-value; that is, Holm's Procedure encounters its first true null at Step i_0. Observe that

$$i_0 \leq n - n_0 + 1$$

as the first true null can be preceded by at most $n - n_0$ other hypotheses. Holm's Procedure commits a false rejection if and only if

$$p_{(1)} \leq \frac{\alpha}{n}, p_{(2)} \leq \frac{\alpha}{(n-1)}, \ldots, p_{i_0} \leq \frac{\alpha}{(n-i_0+1)},$$

which implies

$$p_{(i_0)} \leq \frac{\alpha}{n-i_0+1} \leq \frac{\alpha}{n_0}.$$

Thus, by a union bound, the probability of a false rejection is bounded from above by

$$\mathbb{P}\left(\min_{i:H_{0,i} \text{ is true}} p_i \leq \frac{\alpha}{n_0}\right) = \mathbb{P}\left(\bigcup_{i:H_{0,i} \text{ is true}} p_i \leq \frac{\alpha}{n_0}\right) \leq \sum_{i:H_{0,i} \text{ is true}} \mathbb{P}\left(p_i \leq \frac{\alpha}{n_0}\right) = \alpha.$$

This completes the proof. ■

Remark: Note, Holm's and Benjamini-Hochberg use the same thresholds, but the difference lies in whether they scan forward or backward. Holm's scans forward and stops at the first p-value that exceeds the threshold, whereas Benjamini-Hochberg scans backwards and stops at the first p-value that does not exceed the threshold.

4.6 Preview of Lecture 5

At the end of class, we previewed what Lecture 5 may cover. The fundamental principle addressing this lecture is essentially the closure principle and its applications to global testing procedures. This section introduces the closure principle, illustrating it through concrete examples as well as providing some theory and explanation for why the closure principle controls FWER strongly. See Lecture 5 for more.

4.6.1 The Closure Principle

We begin with a family of hypotheses $\{H_i\}_{i=1}^n$ and the assumption that for each H_i, we can construct a p-value p_i where $p_i \sim U([0,1])$ under H_i. For indices i, j, we define the intersection null as

$$H_{ij} = H_i \cap H_j.$$

More generally, we define the closure of the family as

$$H_I = \bigcap_{i \in I} H_i \text{ for all } I \subseteq \{1, \ldots, n\}$$

Example 1: Consider the case $n = 3$. The closure in this case is simply

$$H_{123}$$

$$H_{12} \quad H_{13} \quad H_{23}$$

$$H_1 \quad H_2 \quad H_3.$$

For each I, consider a valid level α test ψ_I for testing H_I (i.e. reject if $\psi_I = 1$)

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

Note that H_I is the "global null" for the index set I. Thus, the tests ψ_I may be constructed using any global testing procedures (e.g. Bonferroni, Simes, ...).

The Closure Procedure: Reject H_I if and only if for all $J \supseteq I$, H_J is rejected at level α. Mathematically, let $T_I = \min_{I \subseteq J} \psi_J$. Then, we reject H_I if and only if $T_I = 1$.

Thus, in Example 1, we would reject H_1 using the closed test procedure if and only if we reject H_1, H_{12}, H_{13}, and H_{123} using their respective level α tests ψ_I. Note, the closed test will test each H_i using the closure procedure with $I = \{i\}$ to determine whether to reject them. The following is a slightly more complex example of closure.

Example 2: Consider the case $n = 4$ hypotheses. Suppose that the underlined hypotheses are rejected at level α based on ψ_I.

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

Among $\{H_i\}_{i=1}^4$, only H_1 is rejected by the closed test procedure. H_2 is not rejected because not all H_J is rejected for $\{2\} \subseteq J$. That is, H_{23}, H_3, and H_4 are not rejected by the closed test procedure since they are not even rejected by their own test ψ_3, ψ_4.

Next, we show that the closure principle controls the FWER strongly, providing a generic recipe for translating global test procedures into valid FWER controlling procedures.

Theorem 3. The closure principle controls the FWER strongly.

Proof. Let $\mathcal{H}_0 \subseteq \{1, \ldots, n\}$ be the set of indices corresponding to the true nulls. WLOG assume $\mathcal{H}^0$ is non-empty. Otherwise, no rejection can be false. Then, let

$$A = \{\text{Closed Test has at Least One False Rejection}\},$$

$$B = \{\psi_{\mathcal{H}^0} \text{ Rejects } H_{\mathcal{H}^0}\}.$$

For the closure procedure, we claim $A \subseteq B$. Particularly, if the closed test falsely rejects at least one hypothesis H_i, then $i \in \mathcal{H}^0$ and $\psi_{\mathcal{H}^0}$ rejects $H_{\mathcal{H}^0}$ since $\{i\} \subseteq H_0$. It then follows that

$$\mathbb{P}(A) \leq \mathbb{P}(B) \leq \alpha,$$

where the last inequality follows by assumption that $\psi_{\mathcal{H}^0}$ is a level α test. This completes the proof. ■

Lecture 5 will go into detail regarding Closing Bonferonni and Closing Simes, where the former and latter give us precisely Holm's Procedure and Benjamini-Hochberg respectively.

4.7 References

[1] Benjamini, Yoav, and Yosef Hochberg. (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.

[2] Efron, B. (2012). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Volume 1. Cambridge University Press.

[3] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the Familywise Error Rate. The Annals of Statistics, 33(3):1138 – 1154.

[4] Pollard, D. (2015). Chapter 9: Poisson Approximations. http://www.stat.yale.edu/ pollard/Courses/ 241.fall2014/notes2014/Poisson.pdf. Manuscript (Accessed 04-09-2021).