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

Lecture 1 — March 29, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Asher Spector*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 to 2022, and scribe notes written by Xavier Gonzalez, Gene Katsevich, Andy Tsao, and Yiguang Zhang.*

Reading: Large-Scale Inference, Section 3.1

In the first lecture, we discuss perhaps the simplest task in multiple testing, called *global testing*, which asks whether at least one of n null hypotheses are false. We introduce two approaches to global testing, *Bonferroni's method* and *Fisher's combination test*, and we begin to discuss the strengths and weaknesses of these techniques. We emphasize that Bonferroni's method often is *not* overly conservative, and we begin to discuss the sense in which Bonferroni's method is optimal for testing against sparse alternatives.

1.1 Course Logistics

Please see the course webpage for logistics.

1.2 Multiple Hypothesis Testing: Motivation

As motivation, suppose we have n genes and data about expression levels for each gene among m_0 healthy patients and m_1 patients with prostate cancer. To get a rough sense of the dimensionality, note humans have roughly $20,000$ genes, so we should expect n to be in the tens of thousands. In a typical study, we might expect m_0 to be in the hundreds, and m_1 in the low hundreds. Let $Y_{ij}^{(0)}$ denote expression level for gene i and healthy patient $j \in \{1, \dots, m_0\}$, and let $Y_{ij}^{(1)}$ denote the expression level for gene i and sick patient $j \in \{1, \dots, m_1\}$.

	Healthy (m_0 patients)	Prostate cancer (m_1 patients)
Expression level of gene i	$Y_{ij}^{(0)}, 1 \le j \le m_0$	$Y_{ij}^{(1)}, 1 \le j \le m_1$

We want to ask the question: for each gene i, do the gene expressions differ between diseased and healthy patients? Formally, we could try to test the null hypothesis

$$H_{0,i} : \mathbb{E}[Y_{ij}^{(0)}] = \mathbb{E}[Y_{ij}^{(1)}]$$

which asserts that the mean expression level for gene i is the same for healthy and diseased patients. Similarly, the following null hypothesis asserts that the *distribution* of the

expression level for gene i is the same for healthy and diseased patients:

$$H_{0,i} : Y_{ij}^{(0)} \stackrel{d}{=} [Y_{ij}^{(1)}].$$

If we were only interested in a single gene i, we could test the first hypothesis easily using, e.g., a two-sample t-test, or we could test the second null hypothesis using a permutation. However, we have to account for the fact that we are interested in testing *many* hypotheses simultaneously, since testing many hypotheses gives us more chances to make false positives.

1.3 Global Testing: Introduction and Two Methods

1.3.1 Setup

Given null hypotheses $H_{0,1}, \ldots, H_{0,n}$, the *global null* is defined as

$$H_0 = \bigcap_{i=1}^{n} H_{0,i}.$$

H_0 is true if and only if every individual null hypothesis is true. *Global testing* is the task of testing the global null.

Throughout this section, we assume have access to p-values $p_1, \ldots, p_n$ for the hypotheses $H_{0,1}, \ldots, H_{0,n}$. Formally, this means that when $H_{0,i}$ is true, p_i is super-uniform, so $\mathbb{P}_{H_{0,i}}(p_i \leq s) \leq s$ for each $s \in [0, 1]$. For simplicity, we will often assume that $p_1, \ldots, p_n$ are exactly uniform under the null, namely that $p_i \sim \text{Unif}(0, 1)$ when $H_{0,i}$ is true.

1.3.2 Bonferroni's method

Procedure

Given a desired level α, *Bonferroni's global test* rejects the global null whenever

$$\min_i p_i \leq \alpha/n.$$

Size

An appealing property of Bonferroni's method is that it controls the Type I error rate even when the p-values $p_1, \ldots, p_n$ are arbitrarily dependent. This fact follows from a simple union

bound, namely that

$$\begin{aligned}
\mathbb{P}_{H_0}(\text{Bonferroni rejects } H_0) &= \mathbb{P}_{H_0}\left(\min_i p_i \le \alpha/n\right) \\
&= \mathbb{P}_{H_0}\left(\bigcup_{i=1}^n p_i \le \frac{\alpha}{n}\right) \\
&\le \sum_{i=1}^n \mathbb{P}_{H_0}\left(p_i \le \frac{\alpha}{n}\right) && \text{by a union bound} \\
&\le \sum_{i=1}^n \frac{\alpha}{n} && \text{since } p_i \text{ is super-uniform} \\
&= \alpha.
\end{aligned}$$

A common misconception is that Bonferroni's method is conservative, meaning that the size for the Bonferroni method is much smaller than α. However, it is easy to see that in the simple case where $p_1, \dots, p_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$, the size of the Bonferroni method is quite close to α. In particular,

$$\begin{aligned}
\mathbb{P}_{H_0}\left(\min_i p_i \le \alpha/n\right) &= 1 - \mathbb{P}_{H_0}\left(\bigcap_{i=1}^n p_i \ge \frac{\alpha}{n}\right) \\
&= 1 - \left(1 - \frac{\alpha}{n}\right)^n && \text{by independence} \\
&\approx 1 - \exp(-\alpha),
\end{aligned}$$

When α is reasonably small, $1 - \exp(-\alpha)$ is not much smaller than α. For example, when $\alpha = 0.05$, $1 - \exp(-\alpha) = 0.0488$.

Discussion

To gain some more intuition about this test, we plot the sorted p-values in Figure 1.1. Bonferroni's test looks only at the smallest p-value (in the bottom left-hand corner of Figure 1.1), and it rejects if and only if this value is below α/n. For example, in Figure 1.1, Bonferroni's test would reject the given the p-values from "Alternative 1", since there are a few extremely small p-values. In contrast, Bonferroni would *not* reject "Alternative 2," because in Alternative 2, no single p-value is smaller than $\frac{\alpha}{n}$, even though most p-values are much smaller than one would expect from a uniform distribution. Thus, Bonferroni's test performs well when we expect to see extremely strong evidence against a few individual null hypotheses, but it cannot combine weak evidence against many null hypotheses.

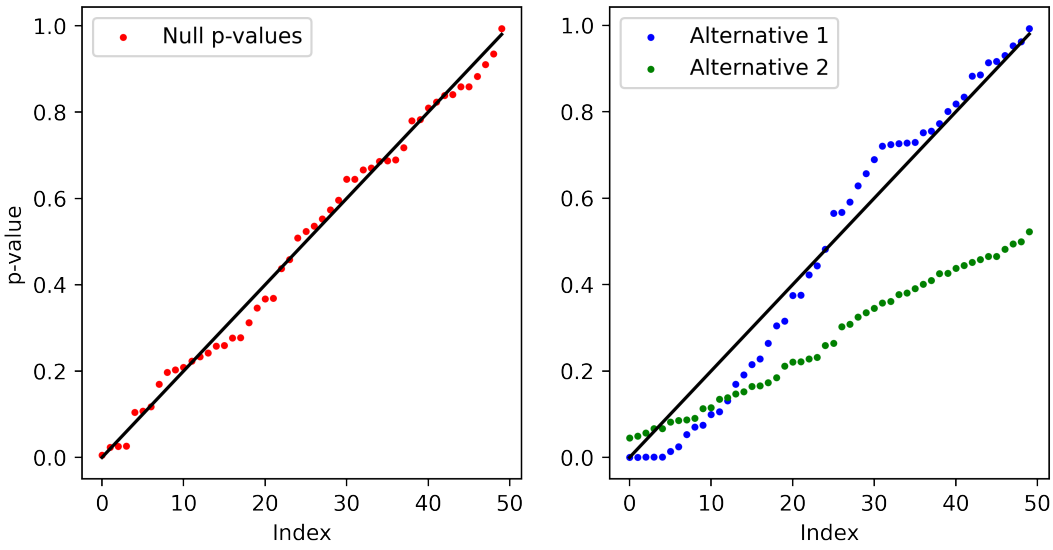

Figure 1.1. Sorted null p-values and 45° line, $n = 50$.

1.3.3 Fisher's Combination Test

Procedure and size

A completely different global testing procedure is *Fisher's combination test*. Given p-values $p_1, \ldots, p_n$, this test rejects for large values of

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

To obtain the finite-sample distribution of T, we need to assume that the p-values $p_1, \ldots, p_n$ are independent.

Proposition 1. Suppose $p_1, \ldots, p_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$ under the global null. Then under the global null, $T \sim \chi^2_{2n}$.

Proof. The proof uses two facts from introductory probability theory. First, if $p_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$, then $-\log(p_i) \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$. Second, note twice an Exp(1) random variable is distributed as a χ^2_2 random variable. Therefore, $-2\log(p_i) \overset{\text{i.i.d.}}{\sim} \chi^2_2$.

Overall, T is a sum of n independent χ^2_2 random variables, so $T \sim \chi^2_{2n}$. □

Therefore, Fisher's test rejects when $T > \chi^2_{2n}(1-\alpha)$.

Discussion of the power of Fisher's test

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 in the individual tests, while Fisher's test works better for detecting many subtle changes. For example, in Figure 1.1, Fisher's test would be able to reject "Alternative 2," whereas the Bonferroni method would not be able to. In contrast, we would expect Bonferroni's method to have more power against "Alternative 1."

Discussion of the independence assumption

A serious drawback of Fisher's combination test is that it assumes $p_1, \dots, p_n$ are independent. In many applications, such as examples in genetics (see Section 1.2), we use the same data to calculate $p_1, \dots, p_n$, so we do not expect the p-values to be independent. In contrast, Bonferroni's test is valid for all dependence structures.

However, Fisher's combination test is often used when we are combining results from different studies. For example, suppose n researchers are testing the same scientific hypothesis, but they each run different experiments using different data and obtain p-values $p_1, \dots, p_n$. In this setting, Fisher's combination test can be quite useful, because we can reasonably expect that $p_1, \dots, p_n$ are independent. Furthermore, this is a setting where we might expect each study to contain some (weak) evidence against the null, and thus Fisher's test is useful because it can combine evidence across studies.

1.4 Preview: Optimality of Bonferroni's Method against Sparse Alternatives

In this section, we will argue that Bonferroni is in some sense optimal against sparse alternatives. Note a "sparse alternative" is an alternative where only a few individual hypotheses are false. For example, in our prostate cancer example, a sparse alternative would mean that only a few genes show significant changes between healthy patients and those with prostate cancer.

1.4.1 Gaussian Sequence Model

To analyze the Bonferroni test, we consider an independent *Gaussian sequence model*:

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1) \text{ for } i = 1, \dots, n,$$

where we are interested in the null hypotheses

$$H_{0,i} : \mu_i = 0.$$

The global null asserts that $\mu_i = 0$ for all i, while under the alternative, there is at least one nonzero μ_i. Bonferroni's method rejects the global null if

- $\max_i Y_i \geq |z(\alpha/n)|$ in the one-sided setting
- $\max |Y_i| \geq |z(\alpha/2n)|$ in the two-sided setting.

This recasts Bonferroni's method in terms of z-scores, whereas we originally presented Bonferroni's method in terms of p-values. (Of course, both formulations are equivalent, since for this Gaussian sequence model, the p-values are monotone functions of the z-scores.)

1.4.2 Magnitude of Bonferroni's Threshold

For simplicity, let us start by considering the one-sided case. To analyze the power of Bonferroni's method, the first step is to find an analytical approximation to $t \coloneqq |z(\alpha/n)|$. To do this, note that under H_0, $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. It follows that under H_0,

$$\frac{\max Y_i}{\sqrt{2\log(n)}} \xrightarrow{p} 1.$$

Thus, the Bonferroni threshold satisfies

$$|z(\alpha/n)| = \sqrt{2\log(n)}(1 + o(1)) \tag{1.1}$$

If we hold α fixed, one can show the stronger asymptotic result that

$$|z(\alpha/n)| = \sqrt{2\log(n)}\left(1 - \frac{\log\log n}{4n}\right).$$

For finite samples, one can use the excellent approximation

$$|z(\alpha/n)| \approx \sqrt{B\left(1 - \frac{\log B}{B}\right)} \text{ where } B \coloneqq 2\log(n/\alpha) - \log(2\pi) \tag{1.2}$$

Of course, since $\log(B)/B \to 0$ as $n \to \infty$, this recovers our original approximation. To evaluate these approximations, Figure 1.2 plots the true normal quantile against $\sqrt{2\log n}$ approximation as well as the approximation from Equation (1.2). It shows that both approximations perform quite well, although the approximation in Equation (1.2) performs better, particularly for small α and n.

1.4.3 Aside: the Gumbel approximation

As an aside, the above analysis (namely Equation (1.1)) is a bit coarse because it ignores the fluctuations of $\max_i Y_i$. A more precise approximation is based on the following result.

Proposition 2. Suppose $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let G denote a standard Gumbel random variable, so $\mathbb{P}(G \le t) = \exp(-\exp(-t))$. Then the distribution of $\max_i Y_i$ is well-approximated by $\mu_n + \sigma_n G$. To be precise, we have the following distributional convergence:

$$\frac{\max_i Y_i - \mu_n}{\sigma_n} \xrightarrow{d} G, \tag{1.3}$$

where

$$u_n = z\left(1 - \frac{1}{n}\right), \quad \sigma_n = z\left(1 - \frac{1}{n}e^{-1}\right) - z\left(1 - \frac{1}{n}\right) \tag{1.4}$$

The same results holds for $\max_i |Y_i|$ if we change n to $2n$ in Equation (1.4).

For a very general proof of this proposition, see David and Nagaraja (2003), Chapter 10.

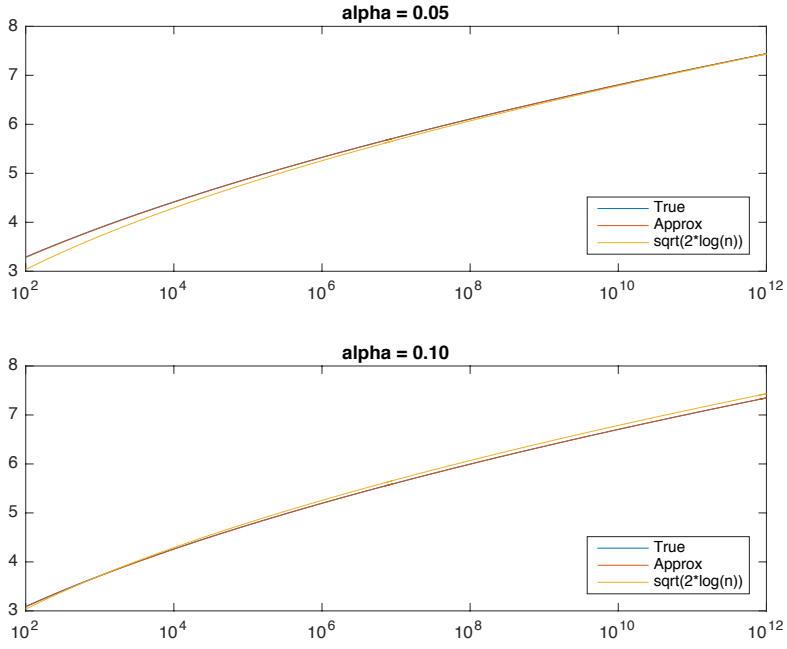

Figure 1.2. $z(\alpha/n)$, $\sqrt{B\left(1-\frac{\log B}{B}\right)}$ ("Approx"), and $\sqrt{2\log n}$. Note that the x-axis denotes the value of n.

1.4.4 Asymptotic Power of Bonferroni's Method

We now find the limiting power of Bonferroni's method as the number of tests $n \to \infty$. To start, we specify the alternative, which is that there is exactly one $i \in [n]$ such that $\mu_i \neq 0$, but we do not know i. We let $\mu^{(n)}$ denote the value of the nonzero μ_i, also known as the "signal size."

This "needle in a haystack" alternative has a sharp detection threshold: if the signal size is slightly above the Bonferroni threshold $\sqrt{2\log n}$, then the Bonferroni method has power 1 in the limit. In contrast, if the signal is slightly below the Bonferroni threshold, then the power of the Bonferroni test goes to α. The proof of this is below.

1. Asymptotic full power above the threshold: Suppose $\mu^{(n)} = (1+\epsilon)\sqrt{2\log n}$ for any $\epsilon > 0$. Also, suppose without loss of generality that $\mu_1 \neq 0$ is the nonzero coefficient (although the analyst does not know this). Then

$$\begin{aligned}
\mathbb{P}_{H_1}\left(\max Y_i > |z(\alpha/n)|\right) &\geq \mathbb{P}\left(Y_1 \geq |z(\alpha/n)|\right) \\
&= \mathbb{P}\left(z_1 > |z(\alpha/n)| - \mu^{(n)}\right) \qquad \text{since } Y_1 \sim \mathcal{N}(\mu^{(n)}, 1) \\
&=\!\to 1.
\end{aligned}$$

Note that above, z_1 represents an arbitrary $\mathcal{N}(0,1)$ random variable, and the last step follows because we know that $|z(\alpha/n)| - \mu^{(n)} \leq -\epsilon\sqrt{2\log n} \to -\infty$.

2. Asymptotic powerlessness below the threshold: Suppose now that $\mu^{(n)} = (1-\epsilon)\sqrt{2\log n}$, with the same notation as before. Then by inclusion-exclusion,

$$\begin{aligned}\mathbb{P}_{H_1}\left(\max Y_i > |z(\alpha/n)|\right) = &\mathbb{P}_{H_1}\left(Y_1 > |z(\alpha/n)|\right) + \mathbb{P}_{H_1}\left(\max_{2\le i\le n} Y_i > |z(\alpha/n)|\right) \\ &- \mathbb{P}_{H_1}\left[\left(Y_1 > |z(\alpha/n)|\right) \cap \left(\max_{2\le i\le n} Y_i > |z(\alpha/n)|\right)\right].\end{aligned}$$

Note that as $n \to \infty$,

$$\mathbb{P}_{H_1}\left(Y_1 > |z(\alpha/n)|\right) = \mathbb{P}\left(z_1 > |z(\alpha/n)| - \mu^{(n)}\right) \to 0$$

since in this setting $|z(\alpha/n)| - \mu^{(n)} \to \infty$. Therefore, we have that

$$\begin{aligned}\mathbb{P}_{H_1}\left(\max Y_i > |z(\alpha/n)|\right) &= \mathbb{P}_{H_0}\left(\max_{2\le i\le n} Y_i > |z(\alpha/n)|\right) + o(1) \\ &\to 1 - \exp(-\alpha),\end{aligned}$$

where the last step follows because we proved in Section 1.3.2 that the Bonferroni has size $1-\exp(-\alpha)$ under independence. (Although technically the previous equation takes the maximum over only $n-1$ random variables, excluding one out of n i.i.d. z-scores is negligible asymptotically.) This proves that if the signal size is beneath the Bonferroni threshold, the test is about as bad as flipping a biased coin that rejects α proportion of the time.

Can we do better than Bonferroni? When the signal is beneath the Bonferroni threshold, we might wish for a better test. However, in the next lecture, we will show that there is *no* test with useful power in this situation.

Bibliography

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, 3rd Edition*. John Wiley & Sons.