Warning: These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès’s course from 2018 and 2021, and scribe notes written by Gene Katsevich, Andy Tsao, and Yiguang Zhang.

Reading: Large-Scale Inference, Section 3.1

In the first lecture, we discuss course logistics and begin the discussion of the simplest task in multiple testing, which is global testing. In particular, we discuss two approaches to global testing, Bonferroni’s method and Fisher’s combination test, and the strengths and weaknesses of these techniques. We emphasize the ways in which 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

The logistics for the course are discussed on the course webpage.

1.2 Introduction: Multiple Hypothesis Testing

In multiple testing we wish to consider many hypotheses simultaneously. For example, suppose we have n genes and data about expression levels for each gene among healthy individuals and those with prostate cancer. For reference as to the size of this study, there are approximately 20,000 genes in the human body, and in a such a study of gene expression in prostate cancer, there may be hundreds of healthy patients, and around 50 – 100 patients with prostate cancer.

<table>
<thead>
<tr>
<th>Expression level of gene i</th>
<th>Healthy (m₀ patients)</th>
<th>Prostate cancer (m₁ patients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yᵢⱼ(0), 1 ≤ j ≤ m₀</td>
<td></td>
<td>Yᵢⱼ(1), 1 ≤ j ≤ m₁</td>
</tr>
</tbody>
</table>

The ith null hypothesis, denoted H₀,i, would state that the mean expression level of the ith gene is the same in both groups of patients. Such a null hypothesis might be rendered as

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

or as

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

In this lecture, we will look at the simplest task in multiple testing, which is global testing, and discuss two different global tests: Bonferroni’s test and Fisher’s combination test.

1.3 Global Testing: Introduction and Two Methods

The simplest task in multiple testing is global testing, where we test the global null

\[ H_0 = \bigcap_{i=1}^n H_{0,i}, \]
which holds if and only if all the individual nulls are true.

Suppose that for each individual null hypothesis $H_{0,i}$, we already have a p-value $p_i$. For simplicity, assume that $p_i \sim U[0,1]$. We would like to combine the individual p-values $p_1, \ldots, p_n$ to test the global null $H_0$.

### 1.3.1 Bonferroni’s Method

Bonferroni’s method is a simple and standard way to test the global null. Though it may seem naive at first glance, we will explore its strengths.

**Procedure**

Given a desired level $\alpha$, **Bonferroni’s global test** rejects the global null whenever

$$\min_i p_i \leq \frac{\alpha}{n}.$$ 

**Size**

A desirable feature of Bonferroni’s method is that the dependence structure of the p-values does not matter for level control, which is given by the union bound:

$$\mathbb{P}_{H_0}\left(\text{Type I Error}\right) = \mathbb{P}_{H_0} \left( \bigcup_{i=1}^{n} p_i \leq \frac{\alpha}{n} \right)$$

$$= \sum_{i=1}^{n} \mathbb{P}_{H_0} \left( p_i \leq \frac{\alpha}{n} \right)$$

$$= \sum_{i=1}^{n} \frac{\alpha}{n}$$

$$= \alpha.$$

A common misconception is that Bonferroni’s method is conservative. However, in the case when the hypotheses are independence, the size of Bonferroni’s test is near $\alpha$, as

$$\mathbb{P}_{H_0}\left(\text{Type I Error}\right) = 1 - \mathbb{P}_{H_0} \left( \bigcup_{i=1}^{n} p_i \geq \frac{\alpha}{n} \right)$$

$$= 1 - \left( \frac{1 - \frac{\alpha}{n}}{n} \right)^n$$

$$\xrightarrow{n \to \infty} 1 - \exp (-\alpha) =: q(\alpha).$$

Therefore, if we have many hypothesis, as in our prostate cancer example, then Bonferroni’s test has size approximately $1 - \exp (-\alpha)$, which for small $\alpha$ is approximately $\alpha$. For example, if $\alpha = 0.05$, then $1 - \exp (-\alpha) = 0.0488\ldots$

**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 reject if this value is below $\alpha/n$. Thus, Bonferroni’s method is good when the difference between the groups we are testing is concentrated in a few of the individual hypotheses.
1.3.2 Fisher’s Combination Test

A completely different technique for global testing from Bonferroni’s method is Fisher’s combination test. Fisher’s combination test is a global test that rejects for large value of the statistic

\[ T = -\sum_{i=1}^{n} 2 \log p_i. \]

Thus, while Bonferroni’s method uses only the smallest p-value, Fisher’s test 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.

In order to obtain the finite sample distribution of \( T \), however, we requires the assumption that the hypotheses are independent.

**Proposition 1.** Suppose \( p_1, \ldots, p_n \) are independent. Then under the null hypothesis, \( T \sim \chi^2_{2n}. \)

**Proof.** To prove this proposition, we draw on two facts from introductory probability theory:

- \( p_i \sim U[0, 1] \implies -\log p_i \sim \text{Exp}(1) \)
- \( E \sim \text{Exp}(1) \implies 2E \sim \chi^2_{2} \).

Therefore, if the \( p_i \)’s are independent, \( T \sim \chi^2_{2n}. \)

Therefore, Fisher’s test rejects when \( T > \chi^2_{2n}(1 - \alpha). \)
1.4 Preview: Optimality of Bonferroni’s Method against Sparse Alternatives

Having introduced global testing, we now show that Bonferroni’s method is in some sense optimal for global testing against sparse alternatives. In the case of 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 formalize this intuition, we consider an independent Gaussian sequence model:

\[ Y_i \overset{iid}{\sim} N(\mu, 1), \quad i = 1, \ldots, n, \]

where we are interested in the \( n \) hypotheses

\[ H_{0,i} : \mu_i = 0. \]

The global null asserts that all means \( \mu_i = 0 \), while under the alternative, some means \( \mu_i \neq 0 \).

Bonferroni’s method rejects the global null if

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

In this way, we have recast Bonferroni’s method in terms of z-scores, whereas in our first presentation of Bonferroni’s method in Section 1.3.1, we used p-values.

1.4.2 Magnitude of Bonferroni’s Threshold

What is this rejection threshold \( t := |z(\alpha/n)| \) (in the one-sided case) for z-scores in Bonferroni’s method? Since under \( H_0 \) we have \( Y_i \overset{iid}{\sim} N(0, 1) \), it follows that under \( H_0 \)

\[ \max Y_i \overset{p}{\to} 1, \]

and so intuitively the Bonferroni threshold is

\[ |z(\alpha/n)| = \sqrt{2\log n} \cdot (1 + o(1)), \]

which (remarkably) does not depend on \( \alpha \).

Holding \( \alpha \) fixed, once can show that asymptotically

\[ |z(\alpha/n)| \approx \sqrt{2\log n} \left( 1 - \frac{\log \log n}{4n} \right), \]

which can be understood as meaning that the quantiles grow like \( \sqrt{2\log n} \) with a small correction factor.

For finite samples, one can use the excellent approximation

\[ |z(\alpha/n)| \approx \sqrt{B \left( 1 - \frac{\log B}{B} \right)}, \quad B := 2\log(n/\alpha) - \log(2\pi), \quad (1.1) \]

(where for reference \( \log(2\pi) \approx 1.8379 \)). We can visualize how good an approximation (1.1) is for the Bonferroni threshold in Figure 1.2, where the true value of the Bonferroni threshold as a function of \( n \), plotted in blue, is indistinguishable from the approximation (1.1), plotted in red. The asymptotic approximation \( \sqrt{2\log n} \) in yellow is also quite good, but struggles for smaller values of \( n \) and \( \alpha \).
1.4.3 Asymptotic Power of Bonferroni’s method

We now ask for the limiting power of Bonferroni’s method as the number of tests \( n \to \infty \), i.e. 
\[
\lim_{n \to \infty} \mathbb{P}_{H_1} \left( \max Y_i > |z(\alpha/n)| \right).
\]

“Needle in a Haystack” Problem:

To find this limiting power, we must specify alternative hypotheses. In the needle in a haystack problem, the alternative is that one \( \mu_i =: \mu > 0 \), but we do not know for which mean this difference occurs.

This needle in a haystack problem has a sharp detection threshold: if the signal is slightly above the Bonferroni threshold \( \sqrt{2 \log n} \), then the power of Bonferroni’s method goes to 1 in the limit; but if the signal is slightly below the Bonferroni threshold, then the power of Bonferroni’s method decreases to close to \( \alpha \) in the limit. Formally, writing the observed Gaussians as \( Y_i = \mu_i + z_i \), with \( z_i \sim \mathcal{N}(0,1) \):

1. Asymptotic full power above threshold: Suppose \( \mu(n) > (1 + \epsilon)\sqrt{2 \log n} \). Then

\[
\mathbb{P}_{H_1} \left( \max Y_i > |z(\alpha/n)| \right) \geq \mathbb{P} \left( Y_1 > |z(\alpha/n)| \right)
= \mathbb{P} \left( z_1 > |z(\alpha/n)| - \mu(n) \right)
\to 1.
\]

Figure 1.2. \( z(\alpha/n), \sqrt{B \left( 1 - \frac{\log B}{B} \right)} \) ("Approx"), and \( \sqrt{2 \log n} \) for \( n \in [10^2, 10^{12}] \).
2. Asymptotic powerlessness below threshold: Suppose $\mu(n) < (1 - \epsilon)\sqrt{2\log n}$. Then

$$P_{H_1}(\max Y_i > |z(\alpha/n)|) \rightarrow q(\alpha) = 1 - \exp(-\alpha) \approx \alpha.$$ 

Therefore, if the signal is beneath the Bonferroni threshold, the test is about as bad as flipping a biased coin that rejects $\alpha$ proportion of the time.

**Can we do better than Bonferroni?** When the signal is beneath the Bonferroni threshold, Bonferroni’s method is disappointing, and 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 scenario.