1 Introduction: Multiple Hypothesis Testing

The multiple hypothesis testing problem is the situation when 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.

	Healthy (k patients)	Prostate Cancer (l patients)
Expression Level of Gene i	$x_{ij}^{(0)}, 1 \le j \le k$	$x_{ij}^{(1)}, 1 \le j \le l$

The ith null hypothesis, denoted $H_{0,i}$, would state that the mean expression level of the ith gene is the same in both groups of patients. Equivalently,

$$H_{0,i} : \mathbb{E}(x_{ij}^{(0)}) = \mathbb{E}(x_{ij}^{(1)}).$$

In this lecture, we define a type of multiple testing called global testing, and then discuss two different global tests: Bonferroni's test and Fisher's Combination Test.

2 Global Testing

One task in multiple testing is called *global testing*, in which we test the *global null*

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

That is, the global null states that all of the individual nulls are true. In the prostate cancer example, the global null would be that $\mathbb{E}(x_{ij}^{(0)}) = \mathbb{E}(x_{ij}^{(1)})$ for all $1 \le i \le n$.

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

3 Bonferroni's Method

A simple method for testing the global null, and one which we will see later is difficult to improve upon when testing against sparse alternatives, is *Bonferroni's Method*.

3.1 Procedure

Given a desired level α, we can test the global null by simply testing each $H_{0,i}$ at level α/n and rejecting H_0 whenever any of the $H_{0,i}$ is rejected. This amounts to rejecting whenever

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

We refer to this as Bonferroni's global test.

3.2 Size

We can prove overall level control via the union bound:

$$\mathbb{P}_{H_0}[\text{Type I Error}] = \mathbb{P}_{H_0}\left[\bigcup_{i=1}^n \{p_i \leq \alpha/n\}\right]$$

$$\leq \sum_{i=1}^n \mathbb{P}_{H_0}[p_i \leq \alpha/n]$$

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

$$= \alpha$$

The union bound might seem crude, but at least in the case when the hypotheses are independent, the size of Bonferroni's test is very near α. Indeed, under independence, the size of the test is

$$\mathbb{P}_{H_0}[\text{Type I Error}] = 1 - \mathbb{P}_{H_0}\left[\bigcap_{i=1}^n \{p_i \geq \alpha/n\}\right]$$

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

$$\overset{n \to \infty}{\longrightarrow} 1 - e^{-\alpha}.$$

This tells us that if we have many hypotheses (e.g. $n = 10,000$ genes in the biological example), then Bonferroni's test has size approximately $1 - e^{-\alpha}$, which for small α is approximately α. For example, if $\alpha = 0.05$, then $1 - e^{-\alpha} = 0.04877\ldots$. So to get a test of size 0.05, we could test each hypothesis at level $0.0512/n$.

3.3 Discussion

To gain some more intuition about this test, we plot the sorted p-values as in Figure 1. Bonferroni's test looks only at the smallest p-value (in the bottom left-hand corner of the figure), and checks if this value is below α/n. Hence, the test is most suited for situations where we expect at least one of the p-values to be very significant. Thus, we expect Bonferroni's test to be powerful against alternatives with this property. In the biological example, we might apply this test if we expect one (or a few) of the genes to be very significantly linked to prostate cancer.
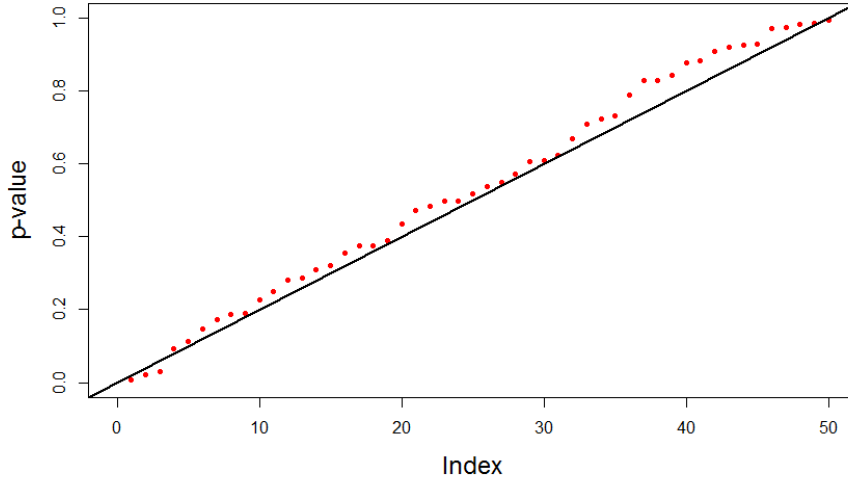
Figure 1: Sorted p-values and $45°$ line, $n = 50$

4   Fisher's Combination Test

Fisher's Combination Test is a global test that rejects for large values of the following statistic:

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

Notice that the function $p \mapsto -\log p$ increases as $p \to 0$, and so it makes sense that smaller p-values will push up the value of T. Assuming the hypotheses are independent, the (finite-sample) distribution of T is known.

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

Proof. First, note that $p_i \sim U[0,1] \Rightarrow -\log p_i \sim \text{Exp}(1)$, since

$$\mathbb{P}(p_i \leq x) = x \Leftrightarrow \mathbb{P}(-\log p_i \geq -\log x) = x.$$

Writing $u = -\log x$, we have $\mathbb{P}(-\log p_i \geq u) = e^{-u}$. Next, we have $-\log p_i \sim \text{Exp}(1) \Leftrightarrow -2\log p_i \sim \chi_2^2$ since

$$\mathbb{P}(-\log p_i \geq u) = e^{-u} \Leftrightarrow \mathbb{P}(-2\log p_i \geq u) = e^{-u/2},$$

which is the cdf of χ_2^2. Indeed, if $X, Y \overset{ind}{\sim} \mathcal{N}(0,1)$, then

$$\mathbb{P}[X^2 + Y^2 \geq u] = \int_{\mathbb{R}^2} I_{\{x^2+y^2 \geq u\}} \frac{1}{2\pi} e^{-(x^2+y^2)/2} dx dy = \int_{\sqrt{u}}^{\infty} re^{-r^2/2} dr = e^{-u/2}.$$

Therefore, if the p_i's are independent, T is a sum of independent χ_2^2, hence a χ_{2n}^2. □

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

3

4.1 Discussion

Note that Fisher's Combination Test aggregates all of the p-values (in log scale), rather than just looking at the minimum p-value. Hence, we expect this test to be powerful when there are many small effects (i.e., when the points in Figure 1 are systematically below the line, although not necessarily by much) and less powerful when there are a few strong effects. In this sense, Fisher's Combination Test is the opposite of Bonferroni's test.

We emphasize that Fisher's Combination Test requires the p_i's to be independent. It is only with this assumption that we were able to conclude that T is distributed as χ^2_{2n}.