

# Lecture Notes 18

## Multiple Testing and Confidence Intervals

Suppose we need to test many null hypotheses

$$\mathcal{H} = \{H_{0,1}, \dots, H_{0,N}\}$$

where  $N$  could be very large. We cannot simply test each hypotheses at level  $\alpha$  because, if  $N$  is large, we are sure to make lots of type I errors just by chance. We need to do some sort of *multiplicity adjustment*.

**Familywise Error Control.** Suppose we get a  $p$ -value  $p_j$  for each null hypothesis. Let  $I = \{i : H_{0,i} \text{ is true}\} \subset \mathcal{H}$ . If we reject  $H_{0,i}$  for any  $i \in I$  then we have made an error. Let  $R = \{j : \text{we reject } H_{0,j}\} \subset \mathcal{H}$  be the set of hypotheses we reject. We say that we have controlled the *familywise error rate* at level  $\alpha$  if

$$\mathbb{P}(R \cap I \neq \emptyset) \leq \alpha.$$

The easiest way to control the familywise error rate is the *Bonferroni method*. The idea is to reject  $H_{0,i}$  if and only if  $p_i < \alpha/N$ . Then

$$\begin{aligned} \mathbb{P}(\text{making a false rejection}) &= \mathbb{P}\left(p_i < \frac{\alpha}{N} \text{ for some } i \in I\right) \\ &\leq \sum_{i \in I} \mathbb{P}\left(p_i < \frac{\alpha}{N}\right) \\ &= \sum_{i \in I} \frac{\alpha}{N} \text{ since } p_i \sim \text{Unif}(0, 1) \text{ for } i \in I \\ &= \frac{\alpha |I|}{N} \leq \alpha. \end{aligned}$$

So we have overall control of the type I error. However, it can have low power.

*The Normal Case.* Suppose that we have  $N$  sample means  $Y_1, \dots, Y_N$  each based on  $n$  Normal observations with variance 1. So  $Y_j \sim N(\mu_j, \sigma^2/n)$ . To test  $H_{0,j} : \mu_j = 0$  we can use the test statistic  $T_j = \sqrt{n}Y_j/\sigma$ . The  $p$ -value is

$$p_j = 2\Phi(-|T_j|).$$

If we did uncorrected testing we reject when  $p_j < \alpha$ , which means,  $|T_j| > z_{\alpha/2}$ . A well known inequality for the tail probability of a Gaussian is

$$\frac{\phi(x)}{x + 1/x} \leq 1 - \Phi(x) \leq \frac{\phi(x)}{x}.$$

From this it can be shown that<sup>1</sup>

$$z_\alpha \approx \sqrt{2 \log(1/\alpha)}.$$

So we reject when

$$|T_j| > \sigma \sqrt{2 \log(2/\alpha)/n}.$$

Under the Bonferroni correction we reject when  $p_j < \alpha/N$  which corresponds to

$$|T_j| > \sigma \sqrt{2 \log(2N/\alpha)/n}.$$

Hence, the familywise rejection threshold grows like  $\sqrt{\log N}$ .

**False Discovery Control.** The Bonferroni adjustment is very strict. A weaker type of control is based on the *false discovery rate*. Suppose we reject a set of hypotheses  $R$ . Define the *false discovery proportion*

$$\text{FDP} = \frac{|R \cap I|}{|R|}$$

where the ratio is defined to be 0 in case both the numerator and denominator are 0. Our goal is to find a method for choosing  $R$  such that

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \alpha.$$

The *Benjamini-Hochberg method* works as follows:

1. Find the ordered p-values  $P_{(1)} < \dots < P_{(N)}$ .
2. Let  $j = \max\{i : P_{(i)} < i\alpha/N\}$ . Let  $T = P_{(j)}$ .
3. Let  $R = \{i : P_i \leq T\}$ .

Let us see why this controls the FDR. Consider, in general, rejecting all hypothesis for which  $P_i < t$ . Let  $W_i = 1$  if  $H_{0,i}$  is true and  $W_i = 0$  otherwise. Let  $\hat{G}$  be the empirical distribution of the p-values and let  $G(t) = \mathbb{E}(\hat{G}(t))$ . In this case,

$$\text{FDP} = \frac{\sum_{i=1}^N W_i I(P_i < t)}{\sum_{i=1}^N I(P_i < t)} = \frac{\frac{1}{N} \sum_{i=1}^N W_i I(P_i < t)}{\frac{1}{N} \sum_{i=1}^N I(P_i < t)}.$$

Hence,

$$\begin{aligned} \mathbb{E}(\text{FDP}) &\approx \frac{\mathbb{E}(\frac{1}{N} \sum_{i=1}^N W_i I(P_i < t))}{\frac{1}{N} \mathbb{E}(\sum_{i=1}^N I(P_i < t))} \\ &= \frac{\frac{1}{N} \sum_{i=1}^N W_i \mathbb{E}(I(P_i < t))}{\frac{1}{N} \sum_{i=1}^N \mathbb{E}(I(P_i < t))} \\ &= \frac{t|I|}{G(t)} \leq \frac{t}{G(t)} \approx \frac{t}{\hat{G}(t)}. \end{aligned}$$

---

<sup>1</sup>In fact, it can be shown that

$$z_\alpha = \sqrt{2 \log(1/\alpha)} - r$$

where  $0 \leq r \leq 1.5$ .

Let  $t = P_{(i)}$  for some  $i$ ; then  $\widehat{G}(t) = i/N$ . Thus,  $\text{FDR} \leq P_{(i)}N/i$ . Setting this equal to  $\alpha$  we get  $P_{(i)} < i\alpha/N$  is the Benjamini-Hochberg rule.

FDR control typically has higher power than familywise control. But they are controlling different things. You have to decide, based on the context, which is appropriate.

**Example 1** *Figure 1 shows an example where  $Y_j \sim N(\mu_j, 1)$  for  $j = 1, \dots, 1,000$ . In this example,  $\mu_j = 3$  for  $1 \leq j \leq 50$  and  $\mu_j = 0$  for  $j > 50$ . The figure shows the test statistics, the  $p$ -values, the sorted log  $p$ -values with the Bonferroni threshold and the sorted log  $p$ -values with the FDR threshold (using  $\alpha = 0.05$ ). Bonferroni rejects 7 hypotheses while FDR rejects 22.*

**Multiple Confidence Intervals.** A similar problem occurs with confidence intervals. If we construct a confidence interval  $C$  for one parameter  $\theta$  then  $\mathbb{P}(\theta \in C) \geq 1 - \alpha$ . But if we construct confidence intervals  $C_1, \dots, C_N$  for  $N$  parameters  $\theta_1, \dots, \theta_N$  then we want to ensure that

$$\mathbb{P}(\theta_j \in C_j, \text{ for all } j = 1, \dots, N) \geq 1 - \alpha.$$

To do this, we construct each confidence interval  $C_j$  at level  $1 - \alpha/N$ . Then

$$\mathbb{P}(\theta_j \notin C_j \text{ for some } j) \leq \sum_j \mathbb{P}(\theta_j \notin C_j) \leq \sum_j \frac{\alpha}{N} = \alpha.$$

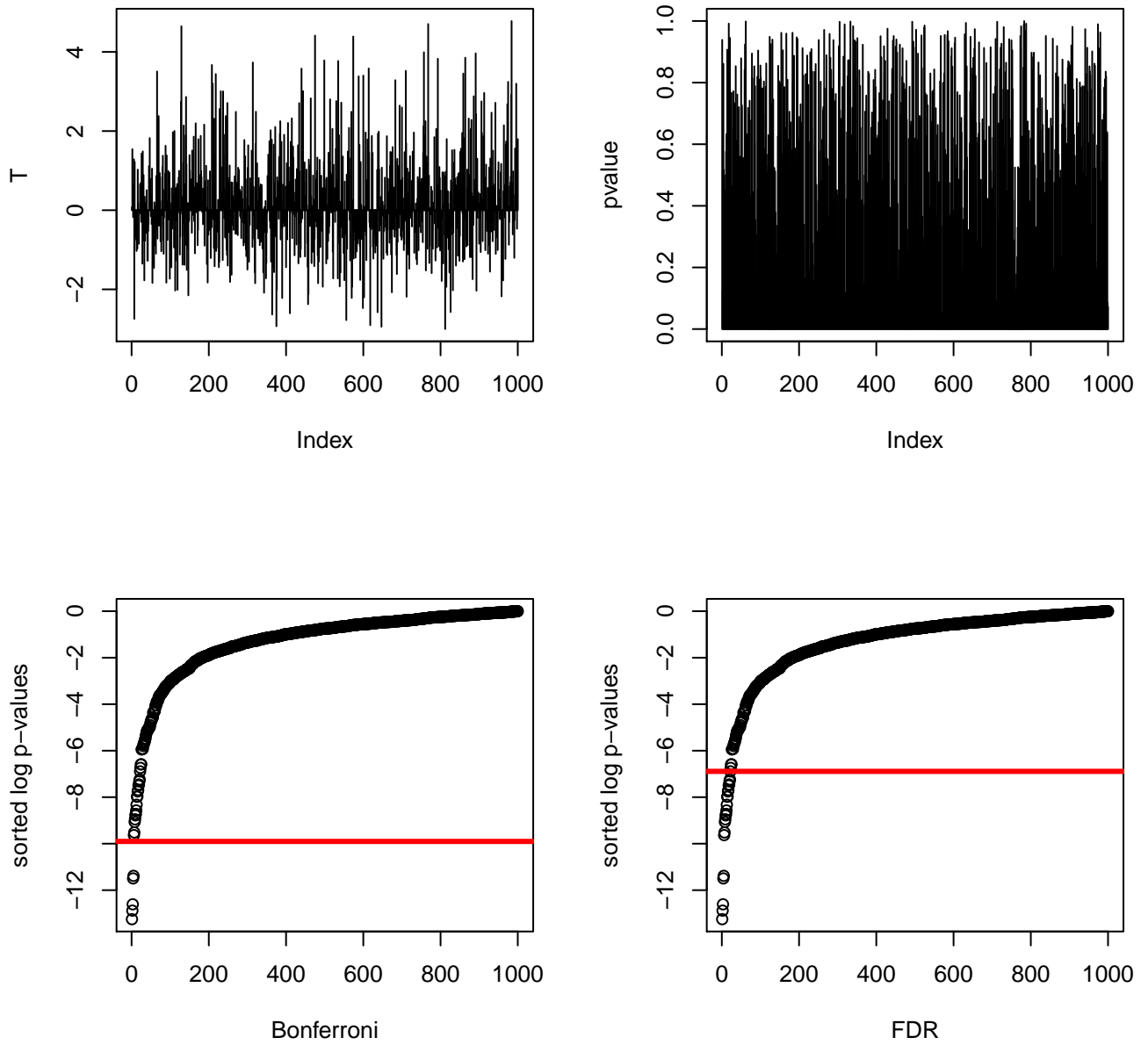


Figure 1: Top left: 1,000 test statistics. Top right: the p-values. Bottom left: sorted log p-values and Bonferroni threshold. Bottom right: sorted log p-values and FDR threshold.