Stats 300C: Theory of Statistics Spring 2018

Lecture 9 — April 20, 2018

Prof. Emmanuel Candes

Scribe: Emmanuel Candes, Mona Azadkia

1 Outline

Agenda: FDR theory

1. Dependent p-values
2. BHq under dependence
   (a) Lower bounds on FDR control
   (b) Upper bounds on FDR control

2 BHq under Dependence

Consider the case where we have two hypotheses and assume we are under the global null. In this case the FDR and the FWER are the same. As usual, we shall assume that p-values are uniformly distributed but they may now be dependent.

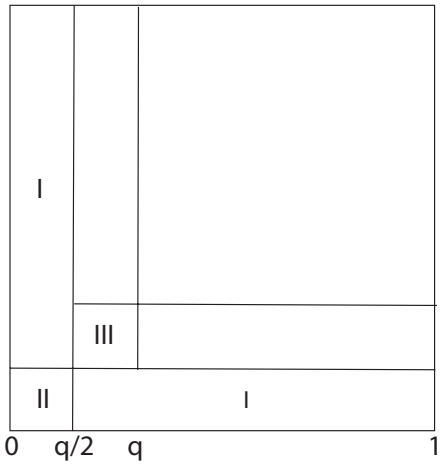


Figure 1: The BHq rejection region.

We refer to Figure 1 and calculate

$$\text{FDR} = \mathbb{P}(I) + \mathbb{P}(II) + \mathbb{P}(III) = q + \mathbb{P}(III) - \mathbb{P}(II),$$

where we have used that $q = \mathbb{P}(I) + 2\mathbb{P}(II)$. Hence,

$$\text{FDR} \leq q + \mathbb{P}(III) \leq q + \mathbb{P}(q/2 < p_1 < q) = 3q/2.$$

So we guaranteed to control at level $3q/2$.

However, there are configurations of p-values for which the FDR is exactly $3q/2$. Consider the joint distribution of p-values as in Figure 2. The distribution is piecewise constant with

- $b = 1/(1-q)$
- $c = 2/q$
- $a = b(1 - bq/2)$
- grey areas have zero probability.

One can check that the marginals are uniform. It easy to see that FDR $= 3q/2$ (under the global null) since FDR $= q + \mathbb{P}(III) - \mathbb{P}(II)$.

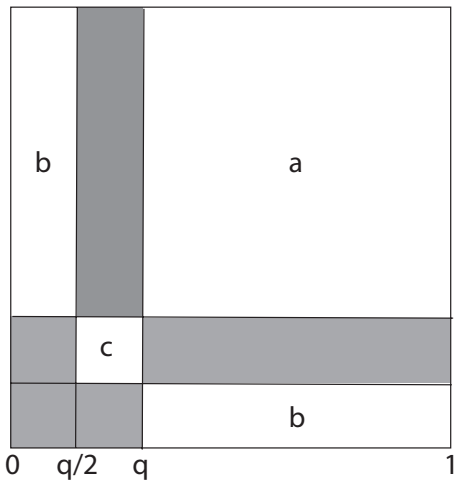


Figure 2: A piecewise constant joint distribution.

There is a result by Guo and Rao (2008), which states that this generalizes to an arbitrary number of hypotheses.

Theorem 1 ([**?**])**:** *There are joint distributions of p-values for which the FDR of the BHq procedure is at least*

$$q \cdot S(n) \wedge 1,$$

where

$$S(n) = 1 + 1/2 + 1/3 + \ldots + 1/n \approx \log n + 0.577.$$

We present an unpublished proof (joint with Rina Foygel Barber) of this phenomenon.

Consider a BH($\boldsymbol{\alpha}$) procedure where the critical values are specified by $0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n \leq 1$. Denote $\hat{k}_{BH} = \max\{k : p(k) \leq \alpha_k\}$ ($\hat{k}_{BH} = 0$ if the set is empty). The procedure then rejects

$H_{(1)}, \ldots, H_{\hat{k}_{BH}}$. This is a general step-up procedure; BHq uses special critical values given by $\alpha_k = qk/n$. No matter the value of the α, we have there is a joint distribution of p-values for which

$$\mathrm{FDR}(\mathrm{BH}(\boldsymbol{\alpha})) \geq \left(\sum_{k=1}^{n} \frac{n(\alpha_k - \alpha_{k-1})}{k} \right) \wedge 1.$$

When $\alpha_k = qk/n$, this gives Theorem 1.

Proof. Here, we shall prove the result only for the case, where $\displaystyle\sum_{k=1}^{n} \frac{n(\alpha_k - \alpha_{k-1})}{k} \leq 1$. (Please do the other case.)

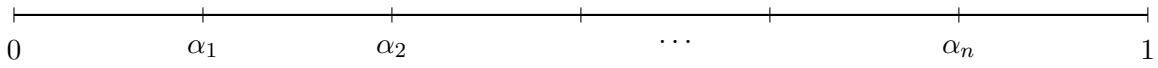
Here, we provide the procedure for generating p-values, for which the lower bound on the FDR is realized. We do this in a hierarchical way. First, sample K from the distribution

$$\mathbb{P}(K = k) = n \cdot \frac{\alpha_k - \alpha_{k-1}}{k}, \qquad k = 1, 2, \ldots, n,$$
$$\mathbb{P}(K = 0) = 1 - \sum_{k=1}^{n} n \cdot \frac{\alpha_k - \alpha_{k-1}}{k}.$$

Second, draw a set of indices $S \subset \{1, \ldots, n\}$ of size K, uniformly at random. To sample the p-values given values of K and S, for each $i = 1, \ldots, n$,

$$p_i \sim \begin{cases} U(\alpha_{K-1}, \alpha_K) & \text{if } i \in S, \\ U(\alpha_n, 1) & \text{if } i \notin S. \end{cases}$$

In case $K = 0$, the set S would be empty, and we assume all $p_i \sim U(\alpha_n, 1)$.



***Claim*:** (i) $p_i \sim U[0, 1]$.

(ii) $\mathrm{FDR}(\mathrm{BH}(\boldsymbol{\alpha})) \geq \displaystyle\sum_{k=1}^{n} \frac{n(\alpha_k - \alpha_{k-1})}{k}$.

Proof: To show (ii): Note that we are under the global null.

If $K = k > 0$, k of the p-values lie in $(\alpha_{k-1}, \alpha_k) \implies \hat{k}_{BH} \geq k$. Now,

$$\mathrm{FDP} \geq \mathbf{1}\{K > 0\} \implies \mathrm{FDR} \geq \mathbb{P}(K > 0) = n \sum_{k=1}^{n} \frac{(\alpha_k - \alpha_{k-1})}{k}.$$

To show (i): conditionally on K and S,

$$p_i \sim \begin{cases} U(0, \alpha_1) & K = 1, \quad i \in S, \\ U(\alpha_1, \alpha_2) & K = 2, \quad i \in S, \\ \vdots & \\ U(\alpha_{n-1}, \alpha_n) & K = n, \quad i \in S, \\ U(\alpha_n, 1) & \text{otherwise.} \end{cases}$$

Now from the distribution of K and the fact that once K is sampled, S is a randomly chosen subset of size K, we have that

$$\mathbb{P}(i \in S \text{ and } K = k) = \alpha_k - \alpha_{k-1}.$$

This gives the marginal distribution of p-values:

$$p_i \sim \begin{cases} U(0, \alpha_1) & \text{wp } \alpha_1, \\ U(\alpha_1, \alpha_2) & \text{wp } \alpha_2 - \alpha_1, \\ \vdots & \\ U(\alpha_{n-1}, \alpha_n) & \text{wp } \alpha_n - \alpha_{n-1}, \\ U(\alpha_n, 1) & \text{wp } 1 - \alpha_n \end{cases}$$

This is precisely the $U[0,1]$ distribution. □

This estimate is surprisingly tight as there is a matching upper bound.

Theorem 2 (Benjamini-Yekutieli (2001))**:** *Under dependence, the BHq procedure controls at level $q \cdot S(n)$. In fact,*

$$\text{FDR} \leq q \cdot S(n) \cdot \frac{n_0}{n}.$$

The proof of this result we give below is from C. and Rina Foygel Barber, which simplfies an argument of Benjamini and Yekutieli.

Proof. First,

$$\text{FDP} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{1 \vee R},$$

Where $V_i = 1$ iff $\mathcal{H}_i$ is rejected. If we show that for any null,

$$\mathbb{E}\frac{V_i}{1 \vee R} \leq \frac{q}{n} S(n), \tag{1}$$

then $\text{FDR} \leq q \cdot (n_0/n) \cdot S(n)$. Setting $\alpha_k = qk/n$, we have

$$\begin{aligned} \frac{V_i}{1 \vee R} &= \sum_{k=1}^{n} \frac{1\{p_i \leq \alpha_k\} 1\{R = k\}}{k} \\ &= \sum_{k=1}^{n} \sum_{\ell=1}^{k} \frac{1\{p_i \in (\alpha_{\ell-1}, \alpha_\ell)\} 1\{R = k\}}{k} \\ &= \sum_{\ell=1}^{n} \sum_{k \geq \ell} \frac{1\{R = k\}}{k} 1\{p_i \in (\alpha_{\ell-1}, \alpha_\ell)\} \\ &= \sum_{\ell=1}^{n} \frac{1\{R \geq \ell\}}{R} 1\{p_i \in (\alpha_{\ell-1}, \alpha_\ell)\} \\ &\leq \sum_{\ell=1}^{n} \frac{1}{\ell} 1\{p_i \in (\alpha_{\ell-1}, \alpha_\ell)\} \end{aligned}$$

Then taking expectation gives (1). □