Lecture 24: October 28th

Lecturer: Siva Balakrishnan

We first finish up with Holm's procedure and then turn our attention to the False Discovery Rate (FDR) and controlling the FDR.

24.1 Holm's procedure

There are many possible improvements to the Bonferroni procedure. For instance, suppose that I told you that exactly (or at most) d_0 of the null hypotheses are truly nulls. Then you can see that we could have used the cut-off of $\frac{\alpha}{d_0}$ and still maintained control over the FWER.

As a thought experiment consider the following setting. You conduct $d = 5$ experiments and you observe p-values of $(0.7, 0.02, 0, 0, 0)$.

Intuitively, it seems like since we are absolutely sure that the last three experiments are non-nulls we should be able to use the cut-off of $\alpha/2$ for the remaining two tests, and still control the FWER.

At a high-level it seems intuitively clear to us that other p-values for $\{p_j\}_{j\neq i}$ contain information at least about the number of null hypotheses and we can use this to relax the correction for p_i. Holm's procedure translates this intuition into a rigorous procedure.

1. Order the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.
2. If $p_{(1)} < \frac{\alpha}{d}$ then reject $H_{(1)}$ and move on, else stop and accept all H_i.
3. If $p_{(2)} < \frac{\alpha}{d-1}$ then reject $H_{(2)}$ and move on, else stop and accept $H_{(2)}, \ldots, H_{(d)}$.
   ⋮
4. If $p_{(d)} < \alpha$, then reject $H_{(d)}$, else accept $H_{(d)}$.

If you prefer: more succinctly, let

$$i^* = \min\left\{ i : p_{(i)} > \frac{\alpha}{d - i + 1} \right\},$$

and reject all $H_{(i)}$ for $i < i^*$.

The main claim is that Holm's procedure also controls the FWER at level α. Importantly, Holm's procedure does not require independence of the p-values, and strictly dominates the Bonferroni procedure.

Proof: Let I_0 denote the indices of the true nulls. First let us make an observation: if

$$\min_{i \in I_0} p_i > \frac{\alpha}{d_0},$$

then we reject none of the true nulls. This is because the first time we encounter a true null we would compare it to a threshold that is at most α/d_0, and if we fail to reject it we would not reject any of the other true nulls.

So the FWER is:

$$\text{FWER} \leq \mathbb{P}\left(\min_{i \in I_0} p_i \leq \frac{\alpha}{d_0}\right) \leq \alpha,$$

by the union bound.

24.1.1 Something to think about

In the above discussion we assumed that there was a single scientist doing a bunch of tests so he could appropriately correct his procedure for the multiple testing problem.

One thing to ponder is really what error rate should we be controlling, i.e. maybe I am the editor of a journal, and I want to ensure that across all articles in my journal the FWER is $\leq \alpha$. Maybe I want this to be true across the entire field? Should I be adjusting my p-values for people in other disciplines? Sounds absurd but it actually makes sense if you think about each of these procedures and their implications for reproducibility.

24.2 False Discovery Rate

Suppose that we tested $d = 1000$ genes for association with some disease, we got a 1000 p-values, and 100 of them were less than 0.01. We'd expect that roughly $0.01 d_0 \leq 0.01 d = 10$ of these to be falsely rejected nulls, and perhaps this is not a bad tradeoff, i.e. if we rejected the first 100 nulls, we would spend only 10% of our time on falsely rejected nulls, i.e. we would make 90 real discoveries.

The FDR is the expected number of false rejections divided by the number of rejections.

Denote the number of false rejections as V, and the total number of rejections as R. Then the false discovery *proportion* is:

$$\text{FDP} = \begin{cases} \frac{V}{R} & \text{if} \quad R > 0 \\ 0 & \text{if} \quad R = 0. \end{cases}$$

The FDR is then defined as:

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

In this notation we can see that the FWER is:

$$\mathrm{FWER} = \mathbb{P}(V \geq 1).$$

We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.

24.2.1 The BH procedure

The BH procedure is one that controls the FDR under independence (i.e. the p-values are independent). There is a much weaker form of this procedure that works under dependence (see the Wasserman book). It turns out to be very challenging to tightly control FDR under strong dependence.

The procedure is:

1. Suppose we do d tests. Let us take the p-values $p_1, \ldots, p_d$, and sort them, i.e. we create the list: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.

2. Define the thresholds:

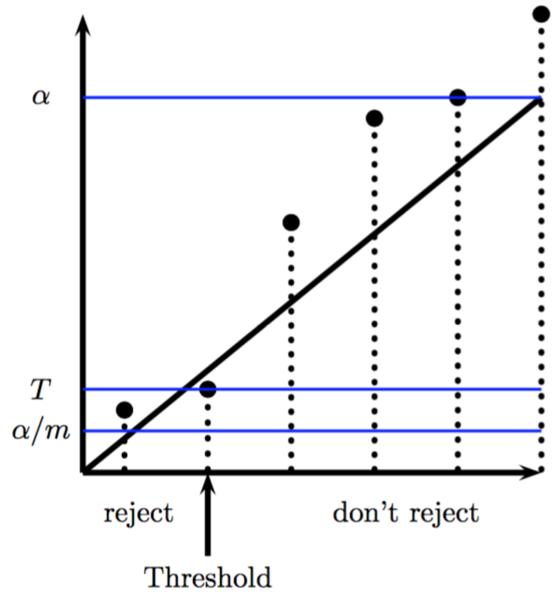
$$t_i = \frac{i\alpha}{d}.$$

3. Find the largest $i_{\max}$ such that

$$i_{\max} = \arg\max_i \{i : p_{(i)} < t_i\}.$$

4. Reject all nulls upto and including $i_{\max}$.

This might seem a bit confusing but here is a simple picture:



24.2.2 Properties of FDR

We have now seen a procedure that controls the FDR under some assumptions. One question of interest is how does FDR control compare to FWER control? Another is just: how do we interpret FDR control?

Interpreting FDR control: The way to think about FDR control is: if we repeat our experiment many times, on average we control the FDP. This is not a statement about the individual experiment we did conduct, and really it does not say much about how likely it is that on a given experiment we have an FDP that is larger than a threshold (think about using Markov's inequality).

FWER on the other hand, does control the error rate for a single experiment. That is, with FWER control, we have managed our false discoveries unless we are very unlucky; with FDR control, on average our test will control FDP, but in our particular experiment we may not have done a very good job. We will see in a second controlling FWER does control the FDR. The way to interpret all of this is that: FDR control is a very weak notion of error control.

Connection to FWER:

1. The first connection is that under the global null (when all the null hypotheses are true) FDR control is equivalent to FWER control.

   Proof: Under the global null, any rejection is a false rejection. There are two possibilities: either we do not reject anything: in this case the FDP $= 0$. If we do reject any null hypothesis then our FDP is 1 (since $V = R$). So we have that:

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{P}(V > 0) * 1 + \mathbb{P}(V = 0) * 0 = \mathbb{P}(V > 0) = \text{FWER}.$$

2. The second connection is that the FWER $\geq$ FDR always. This implies that controlling the FWER implies FDR control.

   Proof: We can see that the following is a simple upper bound on the FDP:

   $$\text{FDP} \leq \mathbb{I}(V \geq 1),$$

   since if $V = 0$, FDP $= 0$, and if $V > 0$ then $V/R \leq 1$. Taking expectations of this expression gives:

   $$\text{FDR} \leq \mathbb{P}(V \geq 1) = \text{FWER}.$$

   The flip-side of this is that FDR control is less stringent so if this is the correct measure for you then you will have *more* power by controlling FDR (rather than controlling FWER).

We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.

24.2.3 The BH procedure

The BH procedure is one that controls the FDR under independence (i.e. the p-values are independent). There is a much weaker form of this procedure that works under dependence (see the Wasserman book). It turns out to be very challenging to tightly control FDR under strong dependence.

The procedure is:

1. Suppose we do d tests. Let us take the p-values $p_1, \ldots, p_d$, and sort them, i.e. we create the list: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.
2. Define the thresholds:

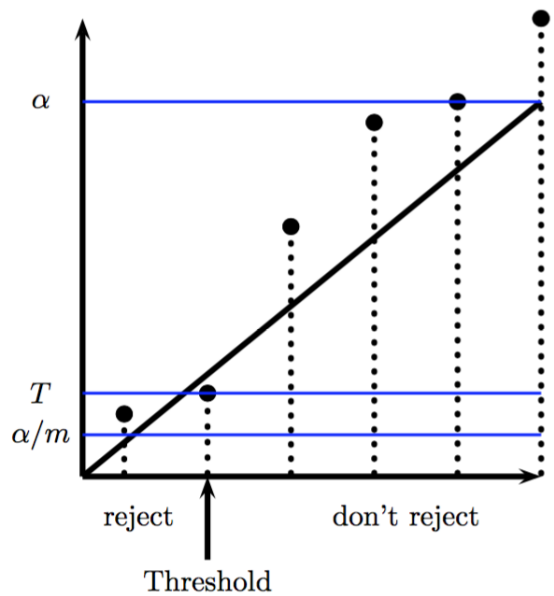
   $$t_i = \frac{i\alpha}{d}.$$

3. Find the largest $i_{\max}$ such that

   $$i_{\max} = \arg\max_i \{i : p_{(i)} < t_i\}.$$

4. Reject all nulls upto and including $i_{\max}$.

This might seem a bit confusing but here is a simple picture:



24.2.4 Properties of FDR

We have now seen a procedure that controls the FDR under some assumptions. One question of interest is how does FDR control compare to FWER control? Another is just: how do we interpret FDR control?

Interpreting FDR control: The way to think about FDR control is: if we repeat our experiment many times, on average we control the FDP. This is not a statement about the individual experiment we did conduct, and really it does not say much about how likely it is that on a given experiment we have an FDP that is larger than a threshold (think about using Markov's inequality).

FWER on the other hand, does control the error rate for a single experiment. That is, with FWER control, we have managed our false discoveries unless we are very unlucky; with FDR control, on average our test will control FDP, but in our particular experiment we may not have done a very good job. We will see in a second controlling FWER does control the FDR. The way to interpret all of this is that: FDR control is a very weak notion of error control.

Connection to FWER:

1. The first connection is that under the global null (when all the null hypotheses are true) FDR control is equivalent to FWER control.

   Proof: Under the global null, any rejection is a false rejection. There are two possibilities: either we do not reject anything: in this case the FDP $= 0$. If we do reject any null hypothesis then our FDP is 1 (since $V = R$). So we have that:

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{P}(V > 0) * 1 + \mathbb{P}(V = 0) * 0 = \mathbb{P}(V > 0) = \text{FWER}.$$

2. The second connection is that the FWER $\geq$ FDR always. This implies that controlling the FWER implies FDR control.

   Proof: We can see that the following is a simple upper bound on the FDP:

$$\text{FDP} \leq \mathbb{I}(V \geq 1),$$

   since if $V = 0$, FDP $= 0$, and if $V > 0$ then $V/R \leq 1$. Taking expectations of this expression gives:

$$\text{FDR} \leq \mathbb{P}(V \geq 1) = \text{FWER}.$$

   The flip-side of this is that FDR control is less stringent so if this is the correct measure for you then you will have *more* power by controlling FDR (rather than controlling FWER).

24.3 Proving BH controls FDR

The main result is the following:

Theorem: Suppose that the p-values are independent, the BH procedure controls the FDR at level α. In fact,

$$\text{FDR} \leq \frac{d_0 \alpha}{d} \leq \alpha.$$

Proof Intuition: Suppose that the BH procedure returned a value $i_{\max}$ then we know that,

$$p_{(i_{\max})} < \frac{i_{\max} \alpha}{d}.$$

We have rejected $i_{\max}$ hypotheses. At the cut-off $\frac{i_{\max}\alpha}{d}$ we expect that $\frac{d_0 i_{\max}\alpha}{d}$ nulls to be rejected. This gives us that the FDR should be roughly:

$$\text{FDR} \approx \frac{d_0 i_{\max} \alpha}{d i_{\max}} = \frac{d_0 \alpha}{d} \leq \alpha.$$

Formalizing this argument is a bit intricate: notice that $i_{\max}$ is a random variable and furthermore the numerator and denominator in the FDP are not independent random variables so we need to be careful while taking the expectation of the ratio. I have included a formal proof that is identical to one from Emmanuel Candes' Stat 300c notes. These notes are in general a great resource that delve much deeper into theoretical aspects of multiple testing.

Proof: When $d_0 = 0$ there are no false discoveries so there is nothing to prove. We will suppose that $d_0 \geq 1$, and denote the set of nulls as I_0. Let us define:

$$V_i = \mathbb{I}(H_i \text{ is rejected}),$$

then we can write the FDP as:

$$\text{FDP} = \sum_{i \in I_0} \frac{V_i}{\max\{R, 1\}},$$

notice that taking the max in the denominator just avoids the 0/0 problem, and is a short-hand way of writing the FDP. Suppose we could prove that:

$$\mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] = \frac{\alpha}{d},$$

then we are done since,

$$\text{FDR} = \sum_{i \in I_0} \mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] = \frac{d_0 \alpha}{d}.$$

To prove the claim we first re-write:

$$\frac{V_i}{\max\{R, 1\}} = \sum_{k=1}^{d} \frac{V_i \mathbb{I}(R = k)}{k},$$

noting that if $R = 0$ both the LHS and RHS are 0. We now need to make some further observations:

1. Suppose that there are k rejections, then we can rewrite:
$$V_i = \mathbb{I}(H_i \text{ is rejected}) = \mathbb{I}(p_i \leq k\alpha/d).$$
2. Suppose that $p_i \leq \alpha k/n$, then we take p_i and set it to 0, and denote the number of rejections as $R(p_i \to 0)$ and note that $R(p_i \to 0)$ is exactly the same as R. On the other hand if $p_i > \alpha k/n$ then $V_i = 0$. So we can write:
$$V_i \mathbb{I}(R = k) = V_i \mathbb{I}(R(p_i \to 0) = k).$$

Now, returning to the main thread suppose we considered the conditional expectation:

$$\mathbb{E}\left[\frac{V_i \mathbb{I}(R = k)}{k} \Big| p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d\right] = \frac{\mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)\mathbb{I}(R(p_i \to 0) = k)|p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d]}{k}$$
$$= \frac{\mathbb{I}(R(p_i \to 0) = k)\alpha}{d},$$

where we use the fact that conditional on the other p-values $\mathbb{I}(R(p_i \to 0) = k)$ is deterministic and that the p-values have uniform distribution under the null, and that the nulls are independent so that:

$$\mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)|p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d] = \mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)] = k\alpha/d.$$

Now, by iterated expectations:

$$\begin{aligned}\mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] &= \sum_{k=1}^{d} \mathbb{E}\left[\mathbb{E}\left[\frac{V_i \mathbb{I}(R = k)}{k}\Big| p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d\right]\right] \\ &= \sum_{k=1}^{d} \frac{\mathbb{I}(R(p_i \to 0) = k)\alpha}{d} = \frac{\alpha}{d},\end{aligned}$$

which was the claim we needed to prove.