Lecture 18
36-705

Recall that the FDR (false discovery rate) is

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

where the FDP (false discovery proportion) is

$$FDP = \frac{\text{number of false rejections}}{\text{number of rejections}} = \frac{V}{R}$$

where the ratio is defined to be 0 if there are no rejections. We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.

0.1 The BH procedure

The BH procedure is one that controls the FDR under independence (i.e. the p-values are independent). There is a different version this procedure that works under 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:

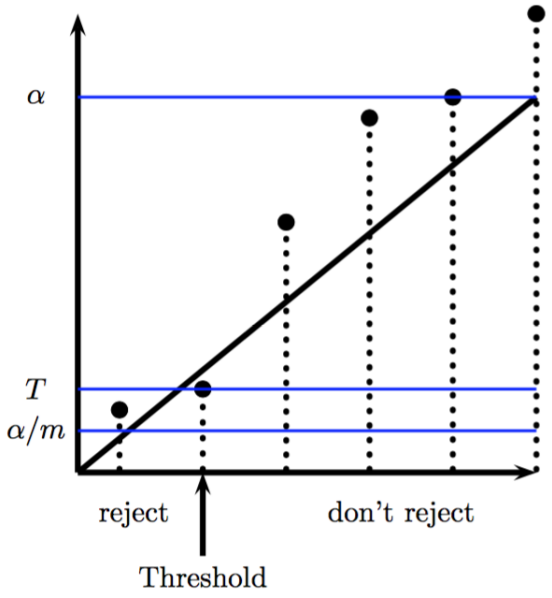
$$t_i = \frac{i\alpha}{d}.$$

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

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

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

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



0.2 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).

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:

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

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).

1 Proving that 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.

Here is a proof from Emmanuel Candes' Stat 300c notes at Stanford. 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 shorthand 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:

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

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:

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

which was the claim we needed to prove. □

Today we will discuss confidence sets and ways to construct them. We have discussed point estimation so far where the goal is to construct an estimate $\widehat{\theta}(X_1, \ldots, X_n)$ of some parameter θ after observing $\{X_1, \ldots, X_n\}$.

The setting here is that we have a statistical model (i.e. a collection of distributions) $\mathcal{P}$. Let $C_n(X_1, \ldots, X_n)$ be a set constructed using the observed data $X_1, \ldots, X_n$. This is a random set. C_n is a $1-\alpha$ confidence set for a parameter θ if:

$$P(\theta \in C_n(X_1, \ldots, X_n)) \geq 1 - \alpha, \quad \text{for all} \;\; P \in \mathcal{P}.$$

This means that no matter which distribution in $\mathcal{P}$ generated the data, the interval guarantees the coverage property described above. Some people would refer to such intervals as *honest* confidence intervals to make explicit the fact that the coverage is uniform over the model.

At a high-level, the confidence interval gives us some idea of how precise our estimate of the unknown parameter is, i.e. a wide interval indicates that our (point) estimate is imprecise.

Example: Suppose that we considered, $X_1, \ldots, X_n \sim U[0, \theta]$. Then we could construct the usual point estimate $\widehat{\theta} = X_{(n)}$. We could perhaps consider two types of confidence intervals:

$$C_1 = [a_1\widehat{\theta}, b_1\widehat{\theta}], \quad 1 \leq a_1 \leq b_1$$
$$C_2 = [\widehat{\theta} + a_2, \widehat{\theta} + b_2], \quad a_2, b_2 \geq 0.$$

Let us try to calculate the coverage probabilities of these two types of intervals. As a preliminary observe that:

$$\mathbb{P}(\widehat{\theta} \leq t) = \left(\frac{t}{\theta}\right)^n, \quad \text{for} \;\; 0 \leq t \leq \theta.$$

1. C_1: We can compute that,

$$\mathbb{P}(\theta \in C_1) = \mathbb{P}(\widehat{\theta} \leq \theta/a_1, \widehat{\theta} \geq \theta/b_1)$$
$$= \mathbb{P}(\widehat{\theta} \leq \theta/a_1) - \mathbb{P}(\widehat{\theta} \leq \theta/b_1)$$
$$= \left(\frac{1}{a_1}\right)^n - \left(\frac{1}{b_1}\right)^n.$$

So we have that for instance choosing $a_1 = 1$, $b_1 = \left(\frac{1}{\alpha}\right)^{1/n}$ guarantees us that this confidence interval has coverage probability exactly $1 - \alpha$.

2. C_2: Similarly we have that,

$$\begin{aligned}\mathbb{P}(\theta \in C_2) &= \mathbb{P}(\widehat{\theta} \leq \theta - a_2, \widehat{\theta} \geq \theta - b_2) \\ &= \mathbb{P}(\widehat{\theta} \leq \theta - a_2) - \mathbb{P}(\widehat{\theta} \leq \theta - b_2) \\ &= \left(\frac{\theta - a_2}{\theta}\right)^n - \left(\frac{\theta - b_2}{\theta}\right)^n.\end{aligned}$$

Notice that now the coverage probability depends on the unknown parameter θ (which is undesirable). Furthermore, if we choose any constants (a_2, b_2) (say depending only on the desired coverage probability α), then as $\theta \to \infty$ we have that the interval has coverage probability that tends to 0, i.e. the interval is not honest for any constants (a_2, b_2).

We will now discuss a few different ways of constructing confidence intervals. Although superficially different most of these methods are roughly the same.

2 Inverting a test

We discussed this method in the last lecture. We suppose that we have a (family of) test(s) for the hypotheses:

$$\begin{aligned}H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0.\end{aligned}$$

These tests have a rejection region and a corresponding acceptance region (where we fail to reject the null). Denote the acceptance region for the test of $\theta = \theta_0$ as $A(\theta_0)$. This is a subset of the sample space.

Given observed data $\{X_1, \ldots, X_n\}$ we consider the random set:

$$C(X_1, \ldots, X_n) = \{\theta_0 : \{X_1, \ldots, X_n\} \in A(\theta_0)\}.$$

Our confidence set is simply the set of parameters θ_0 that we would fail to reject using our family of tests. If our family of tests has level α then the set $C(X_1, \ldots, X_n)$ is a $1 - \alpha$ confidence set.

To see this observe that since our test controls the Type I error we have that for any parameter θ_0,

$$\mathbb{P}_{\theta_0}(\{X_1, \ldots, X_n\} \notin A(\theta_0)) \leq \alpha,$$

so with probability at least $1-\alpha$ we have that, $\{X_1, \ldots, X_n\} \in A(\theta_0)$ and hence that $\theta_0 \in C(X_1, \ldots, X_n)$.

We can also construct tests using confidence intervals, i.e. consider the test that rejects the null hypothesis $\theta = \theta_0$ if $\theta_0 \notin C(X_1, \ldots, X_n)$, then if $C(X_1, \ldots, X_n)$ is a $1-\alpha$ confidence interval this test has level α, i.e.

$$\mathbb{P}_{\theta_0}(\text{reject null } \theta = \theta_0) = \mathbb{P}_{\theta_0}(\theta_0 \notin C(X_1, \ldots, X_n)) \leq \alpha.$$

Let us quickly re-visit the uniform example. Suppose we observe $X_1, \ldots, X_n \sim U[0, \theta]$ and would like to construct a confidence interval, then one method would be to invert the LRT, i.e. we compute the likelihood-ratio for testing $H_0 : \theta = \theta_0$ as if $\theta \geq \max_i X_i$ then:

$$\text{LR} = \frac{\frac{1}{\theta_0^n}}{\frac{1}{(\max_i X_i)^n}} = \frac{(\max_i X_i)^n}{\theta_0^n},$$

and we note that we reject the null for small values of this quantity, i.e. we reject the null if

$$\frac{(\max_i X_i)^n}{\theta_0^n} \leq k_\alpha,$$

for an appropriate choice of k_α. So if we consider the confidence interval obtained by inverting this test, we see that it has the form:

$$C(X_1, \ldots, X_n) = \left\{ \theta : \max_i X_i \leq \theta \leq \frac{\max_i X_i}{k_\alpha^{1/n}} \right\},$$

which is precisely the type of multiplicative interval that we studied in the last lecture (we also calculated a value for k_α that ensures that $C(X_1, \ldots, X_n)$ has coverage $1-\alpha$ in that lecture). This just highlights that in this case, we could have obtained the right kind of interval in a less ad hoc manner (by inverting the LRT).

Example: Suppose we observe $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$, and want to construct a confidence interval for λ.

As our test, suppose we use the LRT, i.e. we define the likelihood ratio:

$$\Lambda = \frac{\lambda_0^n \exp(-\lambda_0 \sum_i X_i)}{(\frac{1}{\overline{X}})^n \exp(-n)}.$$

The acceptance region has the form:

$$A(\lambda_0) = \left\{ \{X_1, \ldots, X_n\} : \left(\lambda_0 \sum X_i\right)^n \exp(-\lambda_0 \sum X_i) \geq k_\alpha(\lambda_0) \right\},$$

where $k_\alpha(\lambda_0)$ needs to be chosen appropriately to control the Type I error. Observe that since $X_i \sim \text{Exp}(\lambda_0)$, $\lambda_0 X_i \sim \text{Exp}(1)$, so $k_\alpha(\lambda_0)$ does not depend on λ_0. Once we determine the cut-off we would obtain the confidence interval by collecting:

$$C(X_1, \ldots, X_n) = \{\lambda : \left(\lambda \sum X_i\right)^n \exp(-\lambda \sum X_i) \geq k_\alpha\},$$

which is an expression that can be solved numerically. Determining k_α and then finding the confidence set can be quite tedious to do exactly (see the Casella and Berger book) and an alternative would be to use large-sample (asymptotic approximations).

3 Inverting Probability Inequalities

In some simple cases, we can use tail bounds to derive confidence intervals. These typically have the advantage of being exact, finite-sample intervals. However, they are rarely used in practice for many reasons including: (1) we do not always have tail bounds for estimators of interest (2) there are usually imprecisely known constants in tails bounds (3) related to (2) they are often very conservative (i.e. the intervals are often too wide to be useful).

Here are a couple of examples:

Example 1 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. By Hoeffding's inequality:*

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(|\widehat{p} - p| > \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$ where $C = (\widehat{p} - \epsilon_n, \widehat{p} + \epsilon_n)$.

Example 2 *Let $X_1, \ldots, X_n \sim F$. Suppose we want a* **confidence band** *for F. We can use VC theory. Remember that*

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence,

$$P_F(L(t) \leq F(t) \leq U(t) \text{ for all } t) \geq 1 - \alpha$$

for all F, where

$$L(t) = \widehat{F}_n(t) - \epsilon_n, \quad U(t) = \widehat{F}_n(t) + \epsilon_n.$$

We can improve this by taking

$$L(t) = \max\left\{\widehat{F}_n(t) - \epsilon_n,\ 0\right\}, \quad U(t) = \min\left\{\widehat{F}_n(t) + \epsilon_n,\ 1\right\}.$$

3.1 Pivots

Another useful way of attempting to construct confidence intervals is to base the intervals on *pivots*. A pivot is a function of the data and the unknown parameter θ – $Q(X_1, \ldots, X_n, \theta)$ – whose distribution does not depend on θ.

Let us consider two examples:

1. Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$ then we can see that $Q(X_1, \ldots, X_n) = \overline{X_n} - \theta \sim N(0, 1/n)$ and so the distribution of Q does not depend on θ.
2. Suppose we consider $X_1, \ldots, X_n \sim U[0, \theta]$ and we consider the function:

   $$Q(X_1, \ldots, X_n, \theta) = \frac{\max_i X_i}{\theta},$$

   has distribution:

   $$P(Q(X_1, \ldots, X_n, \theta) \leq t) = \begin{cases} t^n & 0 \leq t \leq 1 \\ 1 & t \geq 1. \end{cases}$$

   Once again the distribution does not depend on θ.

Given a pivot we can construct confidence intervals in a simple way. Since the distribution of Q does not depend on θ, we can find a, b which do not depend on θ such that:

$$\mathbb{P}_\theta(a \leq Q(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

Now, we construct our confidence interval as:

$$C(X_1, \ldots, X_n) = \{\theta : a \leq Q(X_1, \ldots, X_n, \theta) \leq b\}.$$

By our construction:

$$\mathbb{P}_\theta(\theta \in C(X_1, \ldots, X_n)) = \mathbb{P}_\theta(a \leq Q(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha.$$

Going back to our two examples we find that we will once again obtain the now standard intervals for the two problems (the additive interval for the Gaussian mean, and the multiplicative scale interval for the uniform parameter).

4 Tests Versus Confidence Intervals

Confidence intervals are more informative than tests. Intuitively, p-values are more informative than an accept/reject decision because it summarizes all the significance levels for which we would reject the null hypothesis. Similarly, a confidence interval is more informative that a test because it summarizes all the parameters for which we would (fail to) reject the null hypothesis. More practically, a confidence interval tells us something about the "effect size" as well as something about the uncertainty in our estimate of the "effect size".

Look at Figure 1. Suppose we are testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We see 5 different confidence intervals. The first two cases (top two) correspond to not rejecting H_0. The other three correspond to rejecting H_0. Reporting the confidence intervals is much more informative than simply reporting "reject" or "don't reject."

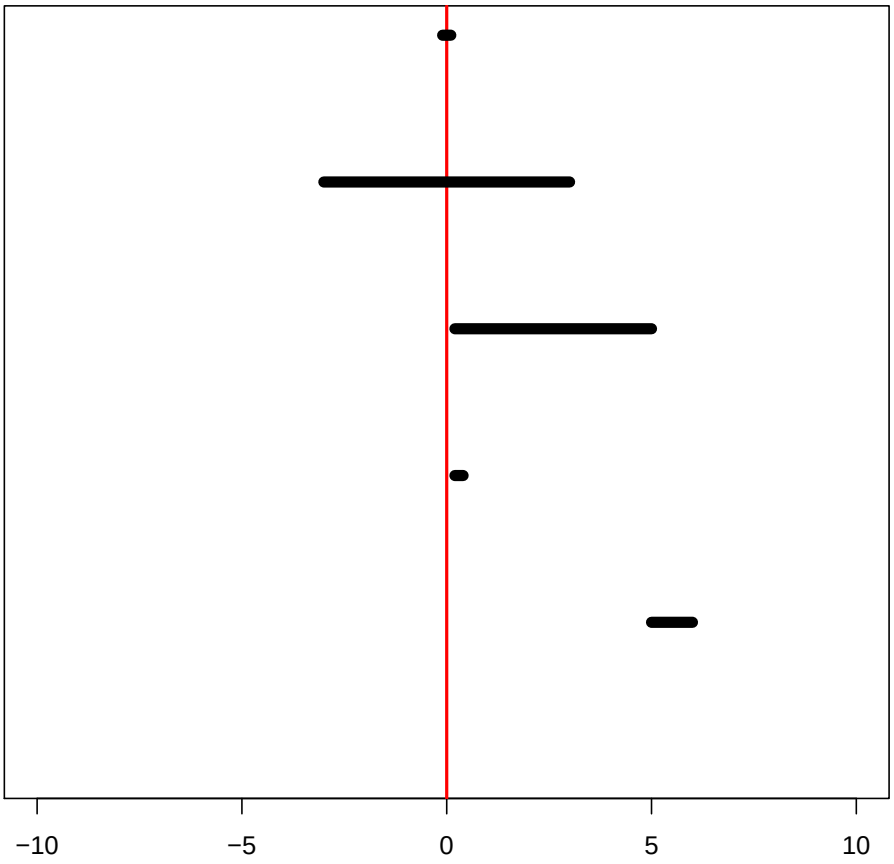


Figure 1: Five examples: 1. Not significant, precise. 2. Not significant, imprecise. 3. Barely significant, imprecise. 4. Barely significant, precise. 5. Significant and precise.