Chapter 15 Large-Scale Hypothesis Testing and False-Discovery Rates   Solution

<u>Notation</u>

n = # of observations

N = # of simultaneous tests being done

α = significance level

FDR = False-discovery rate

$p_{(i)}$ = i^{th} smallest p-value

$H_{0(i)}$ = null corresponding to i^{th} smallest p-value

What does large-scale refer to? N in 1000s or higher

15.1-15.2 We outline 4 procedures related to adjusting for multiple testing covered in these sections.

1. No adjustment. Reject $H_{0(i)}$ if $p_{(i)} \le \alpha$.

2. Bonferroni. Reject $H_{0(i)}$ if $p_{(i)} \le \frac{\alpha}{N}$.

3. Holm's procedure.
   a. Order p-values from smallest to largest
   $$p_{(1)} \le p_{(2)} \le \ldots \le p_{(N)}$$
   b. Let i_0 be the smallest index i such that $p_{(i)} > \frac{\alpha}{(N-i+1)}$.
   c. Reject $H_{0(i)}$ for $i < i_0$.

4. Benjamini-Hochberg FDR procedure.
   a. Order p-values as in Holm's.
   b. Define i_{max} to be the largest index for which $p_{(i)} \le \frac{i}{N} \cdot q$.
   $q = 0.1$ is typical practice.
   c. Reject $H_{0(i)}$ for $i \le i_{max}$. Call results "interesting" rather than "significant."

Be careful with the text's "acceptance" language.

Now we compare / contrast these methods.

FWER stands for... family-wise error rate.

Procedure #1 does NOT control FWER.

Procedures #2 and #3 control FWER @ level $\underline{\alpha}$, but the difference btw them is...

Bonferroni (#2) is more conservative and Holm (#3) is more generous rejecting H_0.

he False discovery proportion is $\dfrac{\#\text{ true null hyp. rejected}}{\#\text{ total rejections}}$. This is Fdp.

For a decision rule D, how is FDR(D) related to Fdp(D)?

$FDR(D) = E(Fdp(D))$ FDR is the <u>expectation</u> of Fdp.

Procedure #4 controls FDR @ level q, ie. $FDR(D) \leq q$.

Example

In a setting where $n = 20$ and $\underline{N = 100}$, with $\alpha = .05$ and $q = 0.1$, the smallest 15 p-values were: Assume these are all 1-sided.

Box are true non-nulls.

$\boxed{0.00005}$ $\boxed{0.00016}$ 0.00196 $\boxed{0.00214}$ $\boxed{0.00694}$ 0.00963 $\boxed{0.01256}$ 0.01657

0.02804 0.04022 0.04024 0.04345 0.05524 0.05822 0.06142

Apply the four procedures. How many times do you reject H_0 for each?

#1 Reject if p-value ≤ 0.05.

Rejects 12 of these 15 H_0's. (so 100 total)

#2 $\dfrac{0.05}{100} = 0.0005 =$ use as α.

Rejects 2 of these 15 H_0's (out of 100 total).

#3 Check indices i. Find cutoff.

Find smallest $i \ni p_{(i)} > \dfrac{\alpha}{101-i}$

$i=1 \Rightarrow .0005$ $p_{(1)} <$ cutoff.

$i=2 \Rightarrow .000505$ $p_{(2)} <$ cutoff. $i=3$

$i=3 \Rightarrow .0005102$ $p_{(3)} >$ cutoff

Reject only 2 H_0's.

#4 Find largest i where $p_{(i)} \leq i \cdot (0.001)$

$i=4$ $0.00214 \leq 0.004$ $i=5$ etc. does not work

Rejects 4 of these. These are the first 4 results (so they are interesting).

Truth: 5 were non-null out of the 100. Of the 4 from fdr, one is a false +.

15.3

The theoretical development in this section uses cdfs and Bayes rule.

Let z_0 be a threshold and z_i be the test statistic for the i^{th} case.

$\text{Fdr}(z_0) = P(\text{case } i \text{ is null} \mid z_i \geq z_0)$

$\text{Fdr}(z_0)$ is the Bayes false-discovery rate, as contrasted with FDR which is Frequentist.

Can obtain an empirical Bayesian estimate of $\text{Fdr}(z_0)$, $\widehat{\text{Fdr}}(z_0)$.

Concludes FDR control relates to Bayes posterior probability of nullness.

$\widehat{\text{Fdr}}(z_0)$ rejects $H_{0(i)}$ when... empirical Bayesian posterior
(general, no formula)
prob. of nullness is too small.

15.4

$\text{Fdr}(z_0)$ is based on tail areas. This is not desirable from a Bayesian perspective. Instead, we can define

$\text{fdr}(z_0) = P(\text{case } i \text{ is null} \mid z_i = z_0)$ as the local false-discovery rate.

We can get reasonable empirical Bayes estimates of fdr.

How are $\text{Fdr}(z_0)$ and $\text{fdr}(z_0)$ related?

The formula is $\text{Fdr}(z_0) = E\{\text{fdr}(z) \mid z \geq z_0\}$

In words, $\text{Fdr}(z_0)$ is the average value of $\text{fdr}(z)$ for $z \geq z_0$.

15.5

Using the empirical null dist. means the significance of an outlying case is judged relative to the dispersion of the majority, not a theoretical ideal.

What are some reasons to doubt the theoretical null in large-scale situations? The txt lists these four.

1. Asymptotics
   (related to Taylor series Approx.)
2. Correlations
3. Unobserved covariates
4. Effect size considerations

Should you always expect to need to adjust the theoretical null?
No. It may need adjustment, or it may not.

15.6

List four "big" take-away messages from the chapter summary. May find others.

1. Large-scale testing is NOT AT ALL like classic Neyman – Pearson theory.

2. N = 5000 (or similar) simultaneous tests gives you your own "long run."

3. fdr's combine frequentist and Bayesian thinking.

4. In large-scale testing, the usual goal is to NOT reject most of the null hypotheses and ID only a few results as "interesting." (and use interesting instead of significant!)