False Discovery Rate

James Long
jplong@mdanderson.org
Rice STAT 533 / GSBS 1283

March 31, 2020

Outline

Outline

Announcements

- Midterm 2: Due today at 5:00pm, email solutions to me
- HW 6: Due today at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- HW 7: Due April 7 at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Lecture Format
  - Slides (plots / analyses in R)
  - .pdf and .R available on course website
- Lecture Structure
  - Microphones are muted when you enter the class.
  - But please ask questions, remember to unmute / mute
  - Let me know about audio issues (chat window or email if I am not responding)

Review

- Defined FWER and FWER control in multiple testing problems
- Questions / Problems with FWER
  - What set of hypotheses on which to control FWER?
  - FWER is too strict and maybe not what we actually want to control
- Today
  - More on FWER
  - Introduce False Discovery Rate (FDR) and FDR control procedures
- Logic of Course Flow
  - Chapter 3 (FWER), Sections 4.1-4.2 (FDR), Chapter 2 (Empirical Bayes), Section 4.3-4.5 (FDR + Empirical Bayes connections)
  - Order of presentation does not follow ordering of book
  - Instead follow more historical development of multiple testing methodology

Outline

What is the appropriate/correct family?

- The family wise error rate (FWER) is the probability that at least one of the hypotheses $H_{01}, \ldots, H_{0N}$ is falsely rejected.
- How family $H_{01}, \ldots, H_{0N}$ is chosen:
  - <u>All tests on the same model:</u> Linear regression with p predictors. Test whether each predictor coefficient is 0.
  - <u>All tests on a particular data set:</u> Prostate example where a t-test is performed for each gene.
  - <u>All tests in a particular manuscript / paper</u>

Heuristic Behind This

- FWER can prevent / discourage fishing for discoveries
  - Can publish manuscript if obtain p–value < 0.05 for some test
  - Keep adding more predictors to linear model.
  - Even if none are correlated with response, about 1 in 20 will produce p–value < 0.05. (ignoring dependence issues)
- A journal may require some "multiplicity correction" such as controlling FWER to avoid fishing
  - Still requires honesty from authors since could find predictor with p–value < 0.05 and pretend this is the only hypothesis tested
  - Also what if 20 researchers test 20 different true nulls on same data set. They are not going to coordinate together to control FWER. They probably do not even know what others are doing. Publication bias issues arise.

My Opinions

- p–values do not incorporate prior probability that null is true. But this is needed to assess the strength of evidence for the alternative.
  - For single hypothesis tests, prior probability of null may be small (< 0.5) i.e. construct single well thought out hypothesis and test it. Thus a p-value threshold of 0.05 *may be* reasonable.
  - In multiple hypothesis testing problems, nulls more likely to be true. So p–value threshold of 0.05 is too liberal
  - FWER is a (not great) way of addressing this issue. The threshold for concluding alternative becomes stricter as number of hypotheses increases (and hence prior null probability increases)
- FDR and Empirical Bayes move closer to answering questions such as $P(H_{0i}|data)$ which is of most interest but is not computable in frequentist single-testing problems because it requires a prior. This posterior probability is computable/estimable in multiple testing problems.

Outline

FWER Control Procedures

- Bonferroni: No p-value dependence assumptions
- Sidak: Assumes p-value independence, more powerful than Bonferroni
- Holm: No p-value independence assumptions, more powerful than Bonferroni
- Hochberg: Assumes p-value independence, more powerful than Sidak
    - Discuss now, compare to False Discovery Rate control procedures later

Hochberg Procedure

Testing Algorithm:

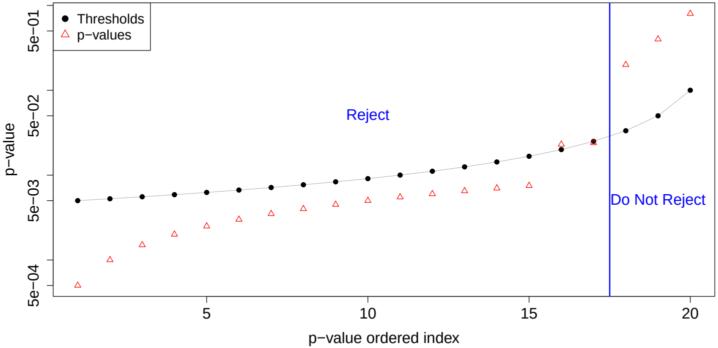
- Let $p_{(1)}, \ldots, p_{(N)}$ be ordered p-values
- Define

$$i = \max\{j \in \{1, \ldots, N\} \text{ s.t. } p_{(j)} \leq \frac{\alpha}{N - j + 1}\}$$

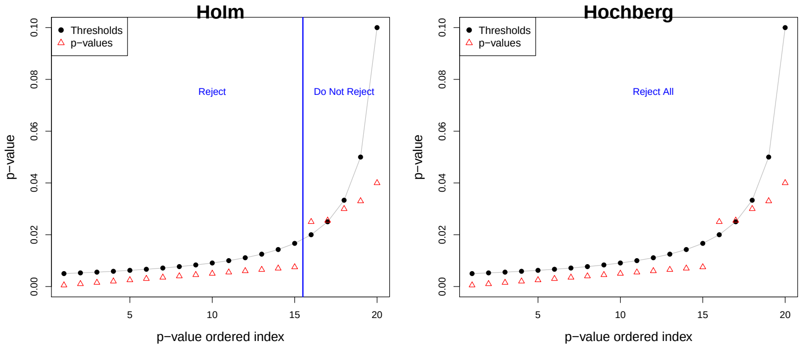
- reject $H_{0(1)}, \ldots, H_{0(i)}$
- If set is empty, fail to reject all hypotheses.

Theorem: If p_i are independent, Hochberg controls FWER at α.

Hochberg Illustration with $\alpha = 0.1$



Holm versus Hochberg



$\alpha = 0.1$. Since the largest p-value is less than α, Hochberg rejects all hypotheses.

More on Hochberg

Notes:

- Hochberg is a "step–up" procedure: Start at largest p-value and work way down. Once some threshold is satisfied reject everything below that threshold.
- Holm is a "step–down" procedure.
- Terminology very confusing because words "step–up" and "step–down" seem to imply opposite

Hochberg Adjusted p-values:

$$\tilde{p}_{(i)} = \min(\min_{j \geq i}\{(N - j + 1)p_{(j)}\}, 1)$$

Hochberg rejects $H_{0(i)}$ with FWER control at α if $\tilde{p}_{(i)} < \alpha$.

Outline

Historical Context

- Mid 1990s: commonly testing 10s of hypotheses
- Many developments on step-up / step-down procedures, e.g. Hochberg and Holm
- Idea: Continue with step-up / step-down algorithms, but control quantity other than FWER

Decisions and Ground Truth

Suppose a test results in the following set of decisions:

		Decision		
		Null	*Non-Null*	
Actual	*Null*	$N_0 - a$	a	N_0
	Non-Null	$N_1 - b$	b	N_1
		$N - R$	R	N

- Known Quantities: N, R
- Unknown Quantities: N_0, N_1, a, b
- A test bounds FWER at α if $P(a > 0) \leq \alpha$

False Discovery Rate

- The *false discovery proportion* (FDP) is

$$\frac{a}{R} = \frac{\text{\# of false rejections}}{\text{\# of rejections}}$$

- FDP control is desirable, but not achievable. For example if all nulls are true FDP equals 0 or 1. No way to have procedure which controls it.
- The *false discovery rate* of a test is

$$FDR = \mathbb{E}\left[\frac{a}{R}1_{R>0}\right] = \mathbb{E}\left[\frac{a}{\max(R,1)}\right]$$

  - $R = a + b$ so $R = 0 \implies a = 0$.

Benjamini–Hochberg FDR Control Algorithm

Algorithm:

- For some $q \in [0, 1]$ let

$$i_{max} = \max\{i \in \{1, \ldots, N\} : p_{(i)} \leq \frac{qi}{N}\}$$

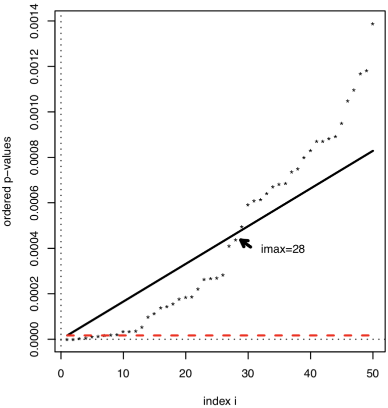
- Rejection set $= BH(q) = \{H_{0(i)}$ s.t. $i \leq i_{max}\}$

Theorem:
If p_i are independent then

$$FDR(BH(q)) \leq \pi_0 q \leq q$$

where $\pi_0 = N_0/N$ is the proportion of true nulls.

BH FDR Control versus Hochberg Illustration



- 50 smallest p-values on `prostate data`
- Black line is $0.1i/N$, BH FDR threshold
- Red dashed line is $0.1/(N-i+1)$ Hochberg FWER threshold.

BH versus Hochberg Threshold Comparison

- BH and Hochberg are both step–up algorithms with different thresholds
- set $\alpha = q$ and

$$\frac{\text{BH threshold}}{\text{Hochberg threshold}} = \frac{iq/N}{q/(N-i+1)} = i\underbrace{\left(1 - \frac{i-1}{N}\right)}_{\approx 1 \text{ for i small}}$$

- If most nulls are true, then ratio when i/N small is most relevant
- For small i/N, BH threshold $\approx i$ times Hochberg
- For example, $N = 6000$ and $i = 50$, ratio is 49.6 (see last slide)

Heuristic that BH Controls FDR

- Consider rejecting all p–values less than t
- $FDR(t)$ is false discovery rate of this procedure
  - $R(t) = \sum_{i=1}^{N} 1_{p_i \leq t} = \#$ rejections
  - $N_0 t =$ expected $\#$ of false rejections
    - Reason: If H_{0i} true then $P(p_i \leq t) = t$
  - $FDR(t) \approx \frac{N_0 t}{R(t)}$
- Select t using BH
  - $i_{max} = \max\{i \in \{1, \ldots, N\} : p_{(i)} \leq \frac{qi}{N}\}$
  - $t_{BH} = \frac{i_{max} q}{N}$

$$FDR(t_{BH}) \approx \frac{N_0 t_{BH}}{R(t_{BH})} = \frac{N_0 \frac{i_{max} q}{N}}{i_{max}} = \frac{N_0}{N} q = \pi q \leq q$$

Argument ignores:

- Cases with 0 rejections where $R(t) = 0$
- Taking expectations of numerators / denominators separately

See 4.2 p 51 of Efron for full proof.

FDR Notes

- For a single hypothesis, FDR control at q is equivalent to controlling type I error at q.
  - Reject H_{01} if $p_1 \leq q/N = q$ for $N = 1$
- FWER control at $q \implies$ FDR control at q (but not converse)

$$FDR = \mathbb{E}\left[\frac{a}{\max(R,1)}\right] \leq \mathbb{E}[1_{a>0}] = P(a > 0) = FWER$$

- FDR controls the expected proportion of false discoveries, not the exact number:
  - Say $N = 6000$ and reject 100 hypotheses with FDR control at $q = 0.1$. So about 10/100 of these hypotheses were falsely rejected. But how close to 10/100 are we likely to be? Can we say it is very unlikely that 50/100 are false discoveries?
  - Empirical Bayes interpretation of FDR will help answer question (Section 4.3-4.5)

FDR Notes

- p-values are often not independent
  - Modify threshold in BH to be more conservative.
  - Empirical Bayes interpretations of FDR are often valid even under dependence (4.3–4.5)
- BH controls FDR at πq so if π not near 1, control is conservative
  - *Adaptive FDR* control estimates π_0 (proportion of true nulls) and achieves more power.
  - Estimate π with $\widehat{\pi}_0$ (discuss methods later)
  - For FDR control at q, use BH with $q^* = q/\widehat{\pi}_0 > q$
  - $FDR \leq \pi_0 q^* = \pi_0 \frac{q}{\widehat{\pi}_0} \approx q$