Empirical Bayes and False Discovery Rate

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

April 7, 2020

Announcements

- HW 7: Was due Tuesday at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Emailed everyone Exam 2 grades and posted solns online
- HW 8: Due April 16 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)

Outline

BH FDR Control and Fdr equivalence

Correlation and Fdp Variability

One and Two Sided p–values

Outline

Review FDR

- $H_{01}, \ldots, H_{0N}$ are hypotheses
- Test procedure results in *false discovery proportion* of

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

  for a particular realization of data.
- $$FDR = \mathbb{E}\left[\frac{a}{R} 1_{R>0}\right] = \mathbb{E}\left[\frac{a}{\max(R,1)}\right]$$
- The BH procedure to reject all $H_{0(i)}$ with $i \leq i_{max}$ where

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

  controls FDR at q.

Review Bayesian Fdr

- $P(H_{0i}\text{ true}) = \pi_0$ (prior probability of null i true)
- $y_i \sim Bernoulli(1 - \pi_0)$ (latent variable indicating null true/false)
- $z_i | y_i \sim f_{y_i}$
- z_i (or p_i) distributed

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$

  where f_0 is the null distribution of the test statistic ($N(0,1)$) or p-value ($Unif[0,1]$)
- Reject all p–values / test statistics in $\mathcal{Z}$

$$\text{Fdr}(\mathcal{Z}) = P(H_0\text{ true}|z \in \mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

$$\overline{\text{Fdr}}(\mathcal{Z}) = \frac{F_0(\mathcal{Z})}{\frac{1}{N}\sum_{i=1}^{N} 1_{z_i \in \mathcal{Z}}}$$

FDR and Fdr Comparison

- FDR and BH: Specify acceptable FDR q and then determine rejection region.
- Fdr: Specify rejection region, estimate Fdr.
- Location of expectations:

$$\text{FDR} = \mathbb{E}\left[\frac{a}{R}1_{R>0}\right]$$

$$\text{Fdr}(\mathcal{Z}) = \frac{n\pi_0 F_0(\mathcal{Z})}{nF(\mathcal{Z})} = \frac{\mathbb{E}[a]}{\mathbb{E}[R]}$$

- No testing procedure can control Fdr
  - If all nulls true $a = R$ so Fdr=1 if $P(R > 0) > 0$.
  - BH sought to control FDR rather than Fdr partially for this reason.

BH Algorithm using Fdr Thresholds

- $p_{(1)}, \ldots, p_{(N)}$
- Let $\mathcal{Z} = [0, p]$
- Recall

$$\overline{\mathsf{Fdr}}(p) = \frac{F_0(p)}{\frac{1}{N}\sum 1_{p_i \le p}}$$

- Further

$$\overline{\mathsf{Fdr}}(p_{(i)}) = \frac{p_{(i)}}{\frac{i}{N}}$$

- Recall

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

- Therefore

$$i_{max} = \max\{i \in \{1, \ldots, N\} : \overline{\mathsf{Fdr}}(p_{(i)}) \le q\}$$

Result: BH algorithm can be expressed in terms of $\overline{\mathsf{Fdr}}$ thresholds.

Interpretation of q

- Original: The expected proportion of false discoveries is bounded by q.
- Using Fdr to Control FDR: The estimated probability the null is true among $z \in \mathcal{Z}$ is bounded by q.
  - For some $z_i \in \mathcal{Z}$, $P(H_{0i} \text{ true}|z_i) < q$
  - For some $z_i \in \mathcal{Z}$, $P(H_{0i} \text{ true}|z_i) > q$
  - On average across set $\mathcal{Z}$ null probability it q across

Outline

BH FDR Control and Fdr equivalence

Correlation and Fdp Variability

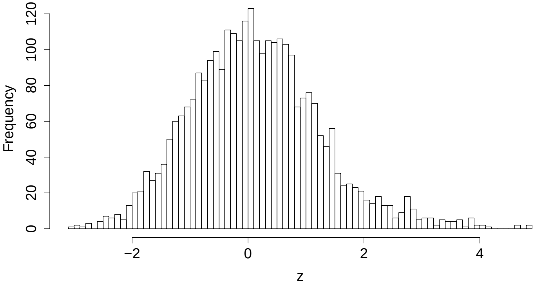
One and Two Sided p–values

Correlation

- $FDR = \mathbb{E}\underbrace{\left[\frac{a}{R}1_{R>0}\right]}_{Fdp}$
- Correlation in test statistics can induce high variability in Fdp
- Even if the reported FDR control remains correct under the correlation, the Fdp for a particular realization of the data can be quite different from the FDR

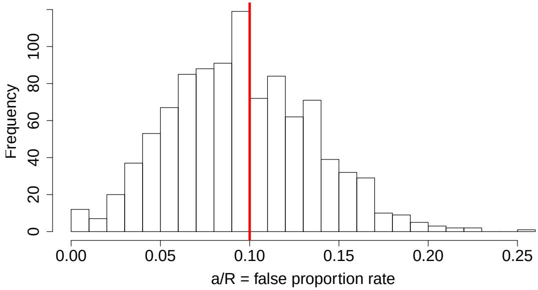
Simulation: Independent Case

- $N = 3000$, $N_0 = 2850$, $\pi_0 = 0.95$
- $f_0 = N(0, 1)$ (null distribution)
- $f_1 = N(2.5, 1)$ (alternative distribution)



- Conduct simulation $M = 1000$ times
- For each run:
  - Use BH to control FDR at $q = 0.1$ each run (right sided p–values)
  - $a =$ number of false rejections
  - $R =$ number of total rejections

False Proportion Rate Distribution

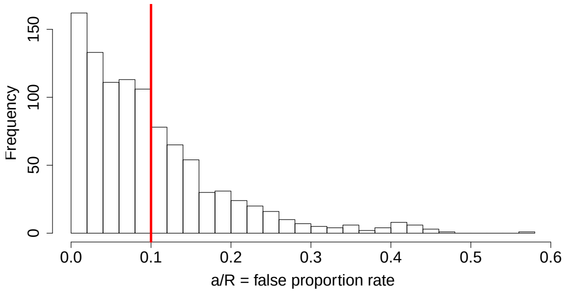


- $0.095 =$ mean of a/R values
  - BH control FDR at $q\pi_0$
- Empirically algorithm is successfully controlling FDR
- 0.903 cases FDP < 0.15

Correlated z-statistics

- $N = 3000$, $N_0 = 2850$, $\pi_0 = 0.95$
- $f_0 = N(0,1)$ (null distribution)
- $f_1 = N(2.5,1)$ (alternative distribution)
- 5 blocks of test statistics
  - Across blocks test statistics independent
  - Within blocks, correlation of 0.2 between pairs of test statistics
  - Alternative hypotheses equally distributed across blocks

Correlated Test Statistics



- $0.097 =$ mean of a/R values
- Empirically algorithm is successfully controlling FDR (despite correlation)
- 0.801 cases FDP $<$ 0.15

Correlation Summary

- Even when correlation does not increase FDR, it can increase variability of Fdp to point where utility of control over FDR is questionable.
    - Also high variability in Fdp whenever N is low.
    - For example with $N = 1$, BH can control FDR at q. But when null true

      $$\frac{a}{R} 1_{R>0} = \begin{cases} 0 & R = 0 \\ 1 & R = 1 \end{cases}$$

      so the Fdp is never near q.
- With correlated z, $\overline{Fdr}$ is generally a more variable an estimate of Fdr than with uncorrelated z. More difficult to assess uncertainty in the estimator

$$\overline{\mathsf{Fdr}}(\mathcal{Z}) = \frac{F_0(\mathcal{Z})}{\frac{1}{N}\sum_{i=1}^{N} 1_{z_i \in \mathcal{Z}}}$$

Outline

BH FDR Control and Fdr equivalence

Correlation and Fdp Variability

One and Two Sided p–values

One and Two Sided p–values

- Let z be a test statistic which is standard normal under H_0
- Three types of p–values:
  - Left sided: $\Phi(z)$
  - Right sided: $1 - \Phi(z)$
  - Two sided: $2(1 - \Phi(|z|))$
- Choice depends on form of null / alternative which is decided by context of problem, e.g. with

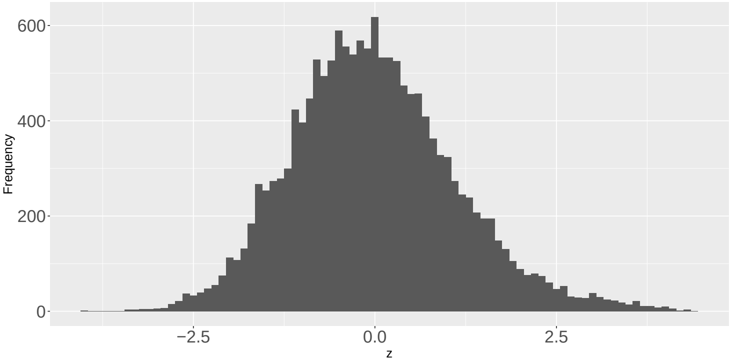
$$H_0 : \mu = 0$$
$$H_a : \mu \neq 0$$

would usually compute two sided p–value.

Discuss Now:

- In multiple testing problems, two-sided p-values often not appropriate.
- Working with test statistics z_i rather than p_i often more straightforward.

DTI Data



- Test statistics z_i from DTI data.
- Distribution center is less than 0
  - Empirical null (histogram if remove small number of true alternatives) does note match theoretical null $N(0, 1)$
  - Will discuss issues for addressing this in Efron Chapter 6
- Right tail heavier than left tail

Why Does Asymmetry Happen: Example

- μ_{0i} is mean gene i expression for healthy tissue
- μ_{1i} is mean gene i expression for cancer tissue
- Test for $i = 1, \ldots, N$:

$$H_{0i} : \mu_{0i} = \mu_{1i}$$
$$H_{1i} : \mu_{0i} \neq \mu_{1i}$$

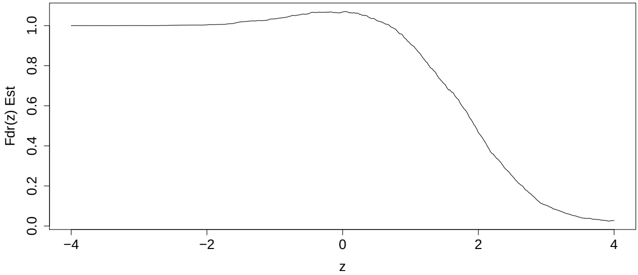
- Test statistic

$$z_i = \frac{\bar{x}_{cancer} - \bar{x}_{control}}{s}$$

- If cancer tends to have no effect (null true) OR increases expression, then z_i corresponding to true alternatives will all be positive

DTI Data

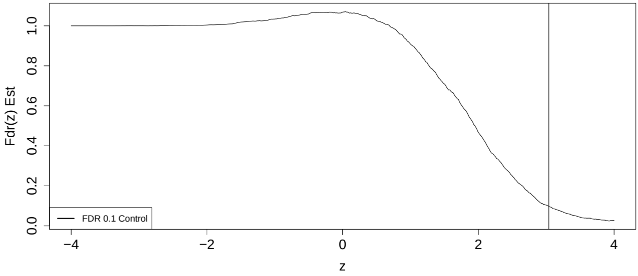
- Consider reject large z, $\mathcal{Z}_R = (z, \infty)$
- Compute
$$\overline{\mathsf{Fdr}}(z) = \frac{1 - \Phi(z)}{\frac{1}{N}\sum_{i=1}^{N} 1_{z_i > z}}$$



FDR

- $p_i = 1 - \Phi(z_i)$
- BH FDR control at q equivalent to reject $z_{(i)}$ for $i \leq i_{max}$ where

$$i_{max} = \max\{i \in \{1, \ldots, N\} : \overline{\mathsf{Fdr}}(z_{(i)}) \leq q\}$$

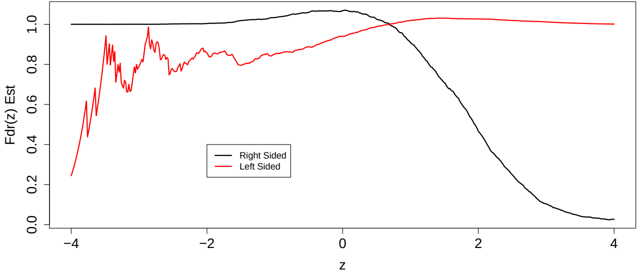


Reject 188 hypothesis at FDR control $q = 0.1$.

DTI Data

- Consider reject small z, $\mathcal{Z}_L = (-\infty, z)$
- Compute

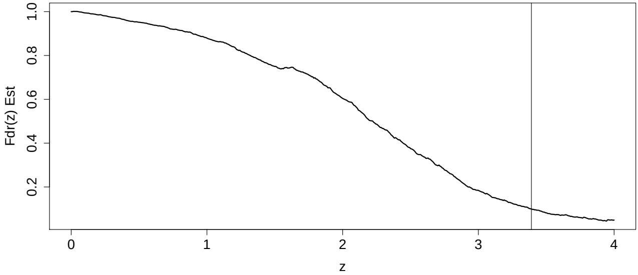
$$\overline{\mathsf{Fdr}}(z) = \frac{\Phi(z)}{\frac{1}{N}\sum_{i=1}^{N} 1_{z_i \leq z}}$$



No rejections for any $q < 0.2$ on left.

Two Sided Tests

- Consider $\mathcal{Z} = (-\infty, -z) \cup (z, \infty)$
- $$\overline{\mathsf{Fdr}}(z) = \frac{\Phi(-z) + 1 - \Phi(z)}{\frac{1}{N}\sum_{i=1}^{N} 1_{z_i < -z} + \frac{1}{N}\sum_{i=1}^{N} 1_{z_i > z}}$$



Reject 108 hypothesis at FDR control $q = 0.1$.

Problems with Two Sided Test

- Hides likely important scientific result that true alternatives nearly all have positive test statistics.
- Rejects fewer hypotheses (108 versus 188) at same FDR control of $q = 0.1$.
- Two sided test control at $q = 0.1$ selects some $\mathcal{Z} = (-\infty, -z) \cup (z, \infty)$ to reject.
  - Region (z, ∞) has Fdr much lower than $q = 0.1$
  - Region $(-\infty, -z)$ has Fdr much higher than $q = 0.1$
  - These average out to $q = 0.1$

$$\overline{\mathrm{Fdr}}(z) = \frac{\Phi(-z) + 1 - \Phi(z)}{\frac{1}{N}\sum_{i=1}^{N} 1_{z_i < -z} + \frac{1}{N}\sum_{i=1}^{N} 1_{z_i > z}}$$

Local Fdr

- Even in one sided test, Fdr varies across rejection region
- Suppose control FDR at $q = 0.1$ and reject in region (z, ∞)
- Fdr in $(z, z+1)$ is higher than Fdr in $(z+1, \infty)$
- But this is not conveyed in standard FDR / Fdr framework, just report q and the set of rejections
- Could select small regions $(z, z+\delta), (z+\delta, z+2\delta), \ldots$ and report Fdr for each
- Taken to the extreme, for each possible z report a test statistic specific Fdr
- This is the idea behind local Fdr
- Cover next week in Chapter 6