Empirical Bayes and False Discovery Rate

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

April 2, 2020

Outline

Announcements

- 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)

Outline

Empirical Bayes False Discovery Rates

Estimating the Fdr

Problem Setup

- Null hypotheses: $H_{01}, \ldots, H_{0N}$
- Evaluate hypotheses based on:
  - Test statistics $z_1, \ldots, z_n$
  - p-values: $p_1, \ldots, p_n$
- Distribution of p–values under null assumed $Unif[0, 1]$
  - For some models p_i will be stochastically larger than $Unif[0, 1]$. Most results we discuss will hold in this case
- Mostly assume z_i are standard normal under H_0
  - Often the case naturally
  - If not, can transform original test statistic x_i to $N(0, 1)$:

$$x_i \sim F \text{ (assuming } H_{0i} \text{ true)}$$
$$\implies z_i = \Phi^{-1}(F(x_i)) \sim N(0, 1)$$

- Distribution of z_i and p_i under the alternative is generally unknown

Two Group Model

- $\pi_0 =$ proportion of true nulls
- $\pi_1 = 1 - \pi_0 =$ proportion of true alternatives
- y_i is indicator H_{1i} is true
  - $y_i \sim Bernoulli(\pi_1)$
- z_i (or p_i) drawn from distribution:

$$f_0(z) \text{ if } y_i = 0 \text{ (i.e. } H_{0i} \text{ is true)}$$
$$f_1(z) \text{ if } y_i = 1 \text{ (i.e. } H_{1i} \text{ is true)}$$

- The marginal distribution of z_i is

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

- **Conceptual Shift:** View the sample size as the number of hypotheses N. Later do asymptotics in N. The number of observations is fixed.

Two Group Model

- Let $\mathcal{Z} \subseteq \mathbb{R}$
- The measures for f_0, f_1, f are

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)$$
$$F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)$$
$$F(\mathcal{Z}) = \int_{\mathcal{Z}} f(z)$$

- Can recover the CDFs by letting $\mathcal{Z} = (-\infty, z)$
- The mixture model equation holds with these measures:

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z})$$

Bayes False Discovery Rate

Rejection rule: Suppose report all $z \in \mathcal{Z}$ as non-null.

Resulting False Discovery Rate:

$$\underbrace{\mathsf{Fdr}(\mathcal{Z}) \equiv \phi(\mathcal{Z}) \equiv P(H_0 \text{ true}|z \in \mathcal{Z})}_{\text{notational equivalence}} = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

The last equality follows from Bayes theorem, hence Bayes False Discovery Rate:

$$\begin{aligned}
P(H_0 \text{ true}|z \in \mathcal{Z}) &= P(y = 0|z \in \mathcal{Z}) \\
&= \frac{P(y = 0, z \in \mathcal{Z})}{P(z \in \mathcal{Z})} \\
&= \frac{P(z \in \mathcal{Z}|y = 0)P(y = 0)}{P(z \in \mathcal{Z})} \\
&= \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}
\end{aligned}$$

Local False Discovery Rate

Alternative Strategy: For each hypothesis report probability H_0 is true.

Local False Discovery Rate:

$$\text{fdr}(z) \equiv \phi(z) \equiv P(H_0 \text{ true}|z) = \frac{\pi_0 f_0(z)}{f(z)}$$

- Somewhat analogous to reporting p–values rather than reject / do not reject decisions
- More objective scale which adapts to plausibility of nulls, i.e. value of π_0
- Will discuss more in future lectures, today's discussion is on Fdr.

Outline

Empirical Bayes False Discovery Rates

Estimating the Fdr

Estimating the Fdr

Suppose reject all $z \in \mathcal{Z}$ (e.g. $\mathcal{Z} = (3, \infty)$). Would like to report:

$$\mathsf{Fdr}(\mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

But some quantities in Fdr are unknown, so need to estimate them.

- Fdr depends on π_0, F_0, and F
- $F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)$ where f_0 is density under H_0
- Since f_0 is known, F_0 is known
  - If z are test statistics, then usually $N(0, 1)$ (after transformation)
  - If z p-values, then $Unif[0, 1]$.
  - When null model wrong, null test–statistics/p-values may not follow f_0.
  - Discuss methods to address this in Chapter 6.
- Need estimators for π_0 and F.

Estimating the Fdr

- Since $\pi_0 \approx 1$ (usually) can estimate with 1 and obtain upper bound

$$\mathsf{Fdr}(\mathcal{Z}) \leq \frac{F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

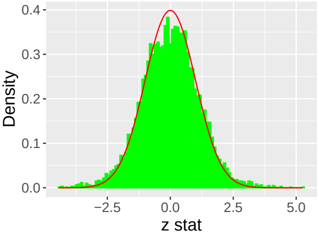
  - Similar to BH FDR which control FDR at $\pi_0 q$ (conservative)
- Since $z \sim f$, the empirical estimator of $F(\mathcal{Z})$ is

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

  - $\overline{F}(\mathcal{Z})$ is unbiased for $F(\mathcal{Z})$ with variance decreasing with N
- **Resulting Estimator:**

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

`prostate data` Application



- Compute z-stat for $N \approx 6033$ hypotheses
- $\mathcal{Z} = (3, \infty)$
- $\overline{F}(\mathcal{Z}) = 49/6033$
- $F_0(\mathcal{Z}) = 1 - \Phi(3) = 0.00135$
- $\overline{\mathsf{Fdr}}(\mathcal{Z}) \approx 0.166$

Quality of Estimator

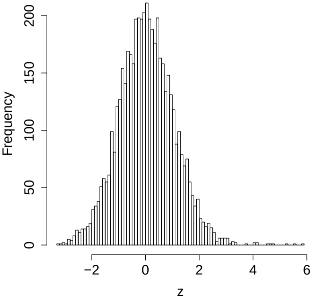
- Choose region $\mathcal{Z}$ and reject $z_i \in \mathcal{Z}$
- Want to know $\mathrm{Fdr}(\mathcal{Z})$
- Report $\overline{\mathrm{Fdr}}(\mathcal{Z})$
- How close is $\overline{\mathrm{Fdr}}(\mathcal{Z})$ to $\mathrm{Fdr}(\mathcal{Z})$?

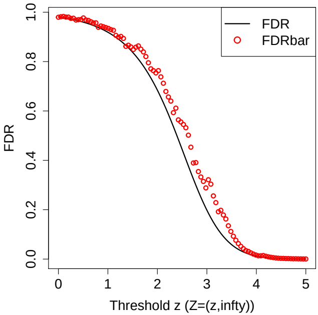
Quality of Estimator: Simulation

- $N = 5000$
- $N_0 = 4900$
- $\pi_0 = 0.98$
- $f_0 = N(0,1)$
- $f_1 = t_{dof=5,ncp=2}$
- Consider rejection regions $\mathcal{Z} = (z, \infty)$

$$\mathsf{Fdr}(\mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

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





Left: Realization of test statistics. **Right:** FDR and $\overline{FDR}$

Quality of Estimator: Mean and Variance

Consider pseudo–estimator:

$$\overline{\mathsf{Fdr}} = \frac{\pi_0 N F_0(\mathcal{Z})}{\sum_{i=1}^{N} 1_{z_i \in \mathcal{Z}}}$$

- Pseudo–estimator because π_0 actually unknown
- But can upper bound with 1 and (usually) induce only small bias (because π_0 near 1)

Define:

$$N_+(\mathcal{Z}) = \sum 1_{z_i \in \mathcal{Z}}$$
$$\underbrace{NF(\mathcal{Z})}_{\equiv e_+(\mathcal{Z})} = \underbrace{N\pi_1 F_1(\mathcal{Z})}_{\equiv e_1(\mathcal{Z})} + \underbrace{N\pi_0 F_0(\mathcal{Z})}_{\equiv e_0(\mathcal{Z})}$$

Note: $e_+(\mathcal{Z}) = \mathbb{E}[N_+(\mathcal{Z})]$

Quality of Estimator

Lemma 2.2 of Efron: Let

$$\gamma(\mathcal{Z}) = \frac{\mathsf{Var}(N_+(\mathcal{Z}))}{e_+(\mathcal{Z})^2}$$

Then

$$\mathbb{E}\left[\frac{\overline{\mathsf{Fdr}}(\mathcal{Z})}{\mathsf{Fdr}(\mathcal{Z})}\right] \approx 1 + \gamma(\mathcal{Z})$$
$$\mathsf{Var}\left(\frac{\overline{\mathsf{Fdr}}(\mathcal{Z})}{\mathsf{Fdr}(\mathcal{Z})}\right) \approx \gamma(\mathcal{Z})$$

Quality of Estimator

Suppressing dependence on $\mathcal{Z}$ and performing a Taylor expansion:

$$
\begin{aligned}
\frac{\overline{\mathsf{Fdr}}}{\mathsf{Fdr}} &= \frac{1}{\mathsf{Fdr}} \frac{e_0}{N_+} \\
&= \underbrace{\frac{1}{\mathsf{Fdr}} \frac{e_0}{e_+}}_{=1} \underbrace{\frac{1}{1 + (N_+ - e_+)/e_+}}_{\text{Taylor expand}} \\
&\approx 1 \underbrace{- \frac{N_+ - e_+}{e_+}}_{\equiv a} + \underbrace{\left(\frac{N_+ - e_+}{e_+}\right)^2}_{\equiv b}
\end{aligned}
$$

a has mean 0 and is higher order than b. So

$$
\mathbb{E}\left[\frac{\overline{\mathsf{Fdr}}}{\mathsf{Fdr}}\right] \approx 1 + \mathbb{E}\left[\left(\frac{N_+ - e_+}{e_+}\right)^2\right] = \frac{\mathsf{Var}(N_+)}{e_+^2}
$$

$$
\mathsf{Var}\left(\frac{\overline{\mathsf{Fdr}}}{\mathsf{Fdr}}\right) \approx \mathsf{Var}\left(-\frac{N_+ - e_+}{e_+}\right) = \frac{\mathsf{Var}(N_+)}{e_+^2}
$$

Quality of Estimator: Independent Case

- Expectation and Variance Depend on $\gamma(\mathcal{Z})$
- Can estimate in straightforward manner if assume independence

$$N_+(\mathcal{Z}) \sim Binomial(N, F(\mathcal{Z}))$$

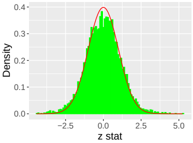
$$\gamma(\mathcal{Z}) = \frac{\mathsf{Var}(N_+(\mathcal{Z}))}{e_+(\mathcal{Z})^2} = \frac{\overbrace{NF(\mathcal{Z})}^{=e_+(\mathcal{Z})}(1 - F(\mathcal{Z}))}{e_+(\mathcal{Z})^2} = \frac{(1 - F(\mathcal{Z}))}{e_+(\mathcal{Z})}$$

- $1 - F(\mathcal{Z}) \approx 1$ and $N_+(\mathcal{Z})/e_+(\mathcal{Z}) \to 1$ so

  $$\widehat{\gamma}(\mathcal{Z}) = \frac{1}{N_+(\mathcal{Z})}$$

  is a reasonable estimator
- Bias is of lower order (in N) than standard deviation
  - Bias $\approx \mathsf{Fdr}(\mathcal{Z})/e_+(\mathcal{Z}) = \mathsf{Fdr}(\mathcal{Z})/(NF(\mathcal{Z})) = O(N^{-1})$
  - s.d. $\approx \mathsf{Fdr}(\mathcal{Z})/\sqrt{e_+(\mathcal{Z})} = \mathsf{Fdr}(\mathcal{Z})/\sqrt{NF(\mathcal{Z})} = O(N^{-1/2})$

Prostate Example



- Compute z-stat for $N \approx 6033$ hypotheses
- $\mathcal{Z} = (3, \infty)$
- $\overline{F}(\mathcal{Z}) = 49/6033$
- $F_0(\mathcal{Z}) = 1 - \Phi(3) = 0.00135$
- $\overline{\mathsf{Fdr}}(\mathcal{Z}) \approx 0.166$
- $\widehat{\gamma}(\mathcal{Z}) = 1/49$
- $\widehat{s.d.}(\overline{\mathsf{Fdr}}) = \overline{\mathsf{Fdr}}\sqrt{\widehat{\gamma}(\mathcal{Z})} = 0.0237$
- 95% CI (assuming asymptotic normality) is $[0.12, 0.21]$

Assumes independence. Discuss more in Chapter 8.

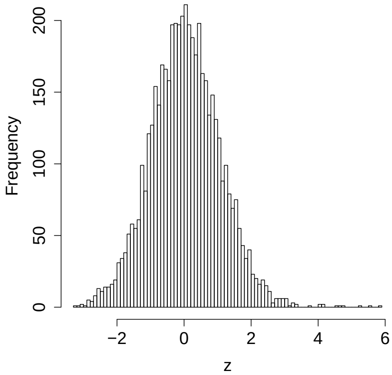
Back to Simulation

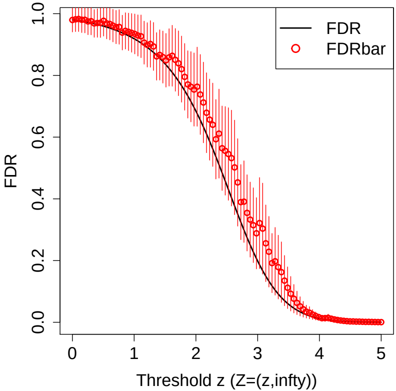
Right Plot:

- Red circle: $\overline{\mathsf{Fdr}}(\mathcal{Z})$
- Red line segments:

$$\overline{\mathsf{Fdr}}(\mathcal{Z}) \pm 2\overline{\mathsf{Fdr}}(\mathcal{Z})\sqrt{\widehat{\gamma}(\mathcal{Z})}$$

- Black line: Fdr





False Discovery Proportion

- Discussed $\overline{\text{Fdr}}$ as estimator for Fdr
- The false discovery proportion is

$$\text{Fdp} = \frac{\#\text{ rejected nulls}}{\#\text{ rejected}} = \frac{N_0(\mathcal{Z})}{N_+(\mathcal{Z})}$$

- Under some assumptions, $\overline{\text{Fdr}}$ is conservatively biased as an estimator of Fdp See Lemma 2.1 in Efron

Summary / Preview

- FDR and the FDR control procedure of Benjamini–Hochberg was developed entirely in a frequentist framework
- Today showed connections with Empirical Bayesian (EB) modeling
- **FDR control of BH and EB presented in different ways:**
  - With BH FDR control, specify acceptable FDR and then determine hypotheses to reject
  - With empirical Bayes FDR, specify hypotheses to reject (i.e. region $\mathcal{Z}$) and then report (estimated) FDR of region
  - Connect these concepts further next class
- BH FDR control constructed for p-values. Empirical Bayes Fdr can be applied to test–statistics or p–values. Mostly discussed test statistics today.
- EB modeling enables definitions and estimators for quantities such as local fdr which are not possible in the strictly frequentist framework
- Discussed estimation of Fdr but not local fdr
  - Estimation of local fdr somewhat more difficult
  - Will discuss in Chapter 5