Math stats

Efron (2010) exercises

2.5

Within Lemma 2.1, we have two expectations of functions

$$E\{\overline{Fdr}(\mathcal{Z})|N_1(\mathcal{Z})\} = E\{\frac{e_0(\mathcal{Z})}{N_0(\mathcal{Z})+N_1(\mathcal{Z})}|N_1(\mathcal{Z})\} = E\{\eta(N_0(\mathcal{Z}))|N_1(\mathcal{Z})\}$$
$$E\{Fdp(\mathcal{Z})|N_1(\mathcal{Z})\} = E\{\frac{N_0(\mathcal{Z})}{N_0(\mathcal{Z})+N_1(\mathcal{Z})}|N_1(\mathcal{Z})\} = E\{\zeta(N_0(\mathcal{Z}))|N_1(\mathcal{Z})\}$$

where $\eta(N_0(\mathcal{Z}))$ is convex and $\zeta(N_0(\mathcal{Z}))$ is concave.

Meanwhile, we have one function of expectations

$$\phi_1(\mathcal{Z}) = \frac{e_0(\mathcal{Z})}{e_0(\mathcal{Z})+N_1(\mathcal{Z})}$$

here, with null case independence,

$$e_0(\mathcal{Z}) = E\{N_0(\mathcal{Z})\} = E\{N_0(\mathcal{Z})|N_1(\mathcal{Z})\}$$

Then, we have

$$\phi_1(\mathcal{Z}) = \frac{e_0(\mathcal{Z})}{E\{N_0(\mathcal{Z})\}+N_1(\mathcal{Z})} = \eta(E\{N_0(\mathcal{Z})|N_1(\mathcal{Z})\})$$
$$\phi_1(\mathcal{Z}) = \frac{E\{N_0(\mathcal{Z})\}}{E\{N_0(\mathcal{Z})\}+N_1(\mathcal{Z})} = \zeta(E\{N_0(\mathcal{Z})|N_1(\mathcal{Z})\})$$

Finally, with different function conditions for Jensen's inequality, we have:
for convex $\eta(\cdot)$,

$$\eta(E\{N_0(\mathcal{Z})|N_1(\mathcal{Z})\}) \le E\{\eta(N_0(\mathcal{Z}))|N_1(\mathcal{Z})\}$$
$$\leftrightarrow \phi_1(\mathcal{Z}) \le E\{\overline{Fdr}(\mathcal{Z})|N_1(\mathcal{Z})\}$$

for concave $\zeta(\cdot)$,

$$\zeta(E\{N_0(\mathcal{Z})|N_1(\mathcal{Z})\}) \ge E\{\zeta(N_0(\mathcal{Z}))|N_1(\mathcal{Z})\}$$
$$\leftrightarrow \phi_1(\mathcal{Z}) \ge E\{Fdp(\mathcal{Z})|N_1(\mathcal{Z})\}$$

That completes our proof for (2.30).

4.1

In this frequentist multiple testing situation, we condition on fixed N_0 null cases and N_1 non-null cases $(N_0, N_1 > 0)$. Here, we define test i size α_i $(i \in I_0, |I_0| = N_0)$ and test j power β_j $(j \in I_1, |I_1| = N_1)$, with $\{1, ..., N\} = I = I_0 \bigcup I_1$, as,

$$\alpha_i = Pr\{a_i = 1|H_{0i}\ true\} = E\{a_i|i \in I_0\}$$
$$\beta_j = Pr\{b_j = 1|H_{1j}\ true\} = E\{b_j|j \in I_1\}$$

where

$$a = \sum_{i\in I_0} a_i,\ \ b = \sum_{j\in I_1} b_j$$

So, now we have

$$E\{\frac{a}{N_0}\} = \frac{1}{N_0}\sum_{i\in I_0} E\{a_i|i\in I_0\} = \frac{1}{N_0}\sum_{i\in I_0}\alpha_i = \bar{\alpha}$$
$$E\{\frac{b}{N_1}\} = \frac{1}{N_1}\sum_{j\in I_1} E\{b_j|j\in I_1\} = \frac{1}{N_1}\sum_{j\in I_1}\beta_i = \bar{\beta}$$

4.2

Conditioned on R, $R>0$, we get

$$a = \sum_{i\in I_R} a_i$$

where

$$I_R \subset I = \{1,...,N\}, |I_R| = R$$

Also, from the setting, we have the two-groups model, *i.i.d.* $z_i, i = 1,...,N$

$$\begin{aligned} Pr\{H_{0i}\} = \pi_0;\ \ z_i|H_{0i} &\sim f_0, F_0 \\ Pr\{H_{1i}\} = \pi_1;\ \ z_i|H_{1i} &\sim f_1, F_1 \\ \rightarrow z_i &\sim f, F \\ f = \pi_0 f_0 + \pi_1 f_1; F &= \pi_0 F_0 + \pi_1 F_1 \end{aligned}$$

Meanwhile, the decision rule rejects H_{0i} for $z_i \in \mathcal{Z}$, i.e.,

$$\{i\in I_R\} \leftrightarrow \{z_i \in \mathcal{Z}\}$$

So, for each a_i, it is a Bernulli variable with the same probability p,

$$\begin{aligned} p &= Pr\{a_i = 1|i\in I_R\} \\ &= Pr\{H_{0i}|z_i\in\mathcal{Z}\} \\ &= \frac{Pr\{H_{0i}\}\cdot Pr\{z_i\in\mathcal{Z}|H_{0i}\}}{Pr\{z_i\in\mathcal{Z}\}} \\ &= \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})} \\ &= Fdr(\mathcal{Z}) = \phi(\mathcal{Z}) \end{aligned}$$

Naturally, we get

$$\begin{aligned} a_i &\sim^{i.i.d.} Bern(\phi(\mathcal{Z})) \\ \rightarrow a = \sum_{i\in I_R} a_i &\sim Bi(R, \phi(\mathcal{Z})) \end{aligned}$$

Finally, we have the scaled version

$$a/R \sim Bi(R,\phi(\mathcal{Z}))/R$$

4.4

Theorem 4.1:
For independent p-value p_i of each testing case i, rule BH(q) rejects

$$H_{0(1)},...,H_{0(i_{max})}$$

and accept others, where (i) are the ordered indices

$$p_{(1)} \le p_{(2)} \le ... \le p_{(i)} \le ... \le p_{(N)}$$

and for a fixed value of q,

$$i_{max} \equiv max\{i|p_{(i)} \leq \frac{i}{N}q\}$$

This leads to that, $\pi_0 = N_0/N$,

$$E\{Fdp_{BH(q)}\} = \pi_0 q$$

Corollary 4.2:
For independent real-valued test statistics z_i, without of loss of generality, corresponding to left-tailed p-value p_i, $(i = 1, ..., N)$

$$p_i = F_0(z_i) \quad \leftrightarrow \quad z_i = F_0^{-1}(p_i)$$

where, for all i, $z_i|H_{(0i)} \sim f_0, F_0$ and $z_i \sim f, F$, empirical CDF $\bar{F}(z) = \#\{z_i \leq z\}/N$.
Rule EB(q) rejects

$$H_{0(1)}, ..., H_{0(i_{max})}$$

and accpet others, where (i) are the ordered indices equivalent to those $p_{(i)}$ indices

$$z_{(1)} \leq z_{(2)} \leq ... \leq z_{(i)} \leq ... \leq z_{(N)}$$

and for a fixed q,

$$i_{max} \equiv max\{i|\overline{Fdr}(z_{(i)}) \leq q\}$$

The condition here is equivalent to BH(q) in that

$$\begin{aligned} \overline{Fdr}(z_{(i)}) &\leq q \\ \frac{\pi_0 F_0(z_{(i)})}{\bar{F}(z_{(i)})} &\leq q \\ \frac{\pi_0 p_{(i)}}{i/N} &\leq q \\ p_{(i)} &\leq \frac{i}{N}\frac{q}{\pi_0} \end{aligned}$$

So, even π_0 is unknown, with the equivalency to Theorem 4.1, we can see that

$$E\{Fdp_{EB(q)}\} = \pi_0 \frac{q}{\pi_0} = q$$

4.5

In Figure 4.3, we are plotting $\overline{Fdr}(z)$, $\overline{Fdr}^{(c)}(z)$ vs. z-values,

$$\begin{aligned} \overline{Fdr}(z) &= \pi_0 F_0(z)/\bar{F}(z) \\ \overline{Fdr}^{(c)}(z) &= \pi_0 F_0^{(c)}(z)/\bar{F}^{(c)}(z) \end{aligned}$$

where $z_i|H_{0i} \sim F_0$, $z_i \sim F$ and the complementary CDF $F^{(c)}(z) = 1 - F(z)$.
Here, for left-sided tests, when z is positively large $(z > 3)$,

$$\bar{F}(z) \to F_0(z); \overline{Fdr}(z) \to \pi_0$$

Similarly, for right-sided tests, when z is negatively large $(z < -3)$,

$$\bar{F}^{(c)}(z) \to F_0^{(c)}(z); \overline{Fdr}^{(c)}(z) \to \pi_0$$

When setting $\pi_0 = 1$, we should see that

$$\begin{aligned} \overline{Fdr}(z) &\simeq 1, \ z > 3 \\ \overline{Fdc}^{(c)}(z) &\simeq 1, \ z < -3 \end{aligned}$$

as shown up in Figure 4.3.

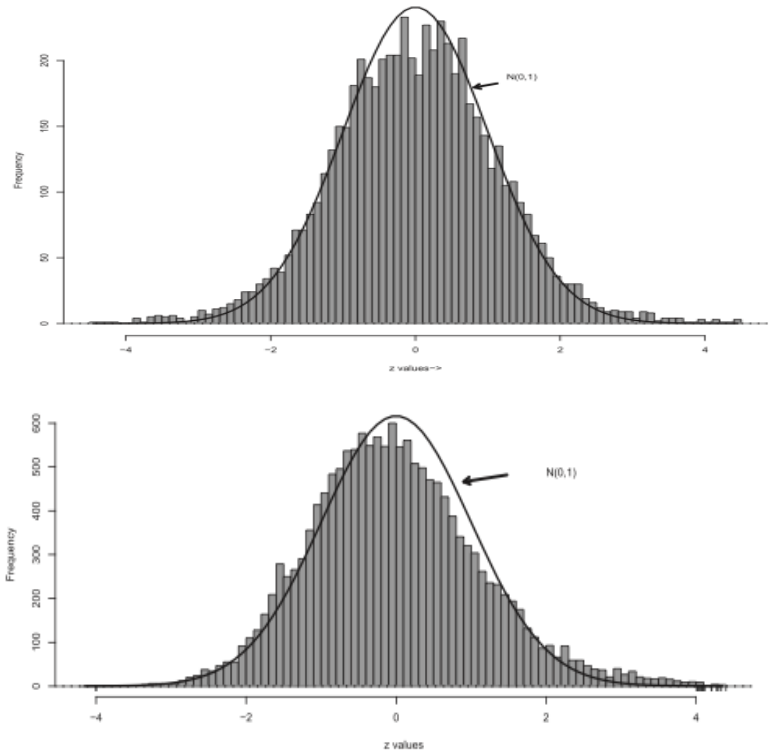

Figure 1: Efon Figure2.1 (top) and Figure3.1 (bottom)

4.6

From the z-value histogram of Efron Figure 2.1 and Figure 3.1, we can see that the prostate data is more *symmetrical* than the DTI data, with respect to null standard normal distribution. The DTI data actually has two half-brain data, which is more heavily tailed to the right than to the left, making the two-sided testing results rejecting fewer cases. So, the results summerized in the table show that the two-sided testing of the prostate data are less wasteful.

$R_{BH(0.1)}(Z)$	left-tailed	right-tailed	two-tailed
prostate	$32(z_i \leq -3.26)$	$28(z_i \geq 3.36)$	$60(\|z_i\| \geq 3.29)$
DTI	0	$192(z_i \geq 3.02)$	110

Prostate microarray data

We have the prostate microarray data from the Efron book, a 6033×102 matrix $X = \{x_{ij}\}$.

$$x_{ij} = \text{expression of gene } i \text{ on patient } j,$$

where $i = 1, .., N$, $N = 6033$; normal patients $j = 1, ..., 50$ vs. cancer patients $j = 51, ..., 102$.

Efron Figure 2.1

The large-scale hypothesis testing that we perform on this dataset use the two-sample t-statistic. For testing gene i,

$$t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i}$$

$$\bar{x}_i(1) = \frac{1}{50}\sum_{j=1}^{50} x_{ij}; \quad \bar{x}_i(2) = \frac{1}{52}\sum_{j=51}^{102} x_{ij}$$

$$s_i^2 = \frac{\sum_{j=1}^{50}(x_{ij} - \bar{x}_i(1))^2 + \sum_{j=51}^{102}(x_{ij} - \bar{x}_i(2))^2}{100} \cdot (\frac{1}{50} + \frac{1}{52})$$

Then, we transform to the z-values, where Φ is the standard normal CDF and F_{100} is the Student-t CDF with 100 degrees of freedom,

$$z_i = \Phi^{-1}(F_{100}(t_i))$$

Finally, we reproduce Efron Figure 2.1, histogram of z-values testing $N = 6033$ genes for possible involvement with prostate cancer.

Efron Figure 4.2

We implement Benjamini and Hochberg's FDR control algorithm here. Our right-sided testing for the prostate data produces p-value p_i for each case,

$$p_i = F_{100}(-t_i)$$

After ordering and choosing $q = 0.1$, we reproduce Efron Figure 4.2.

Here, stars indicate p-values for the 50 largest z_i, and the solid line (slope=q/N) intersection gives us the 27 rejections of the null cases under BH(q) control. So among the 27 non-null genes, the expected number of false discoveries should be 2.7, which is quite good.

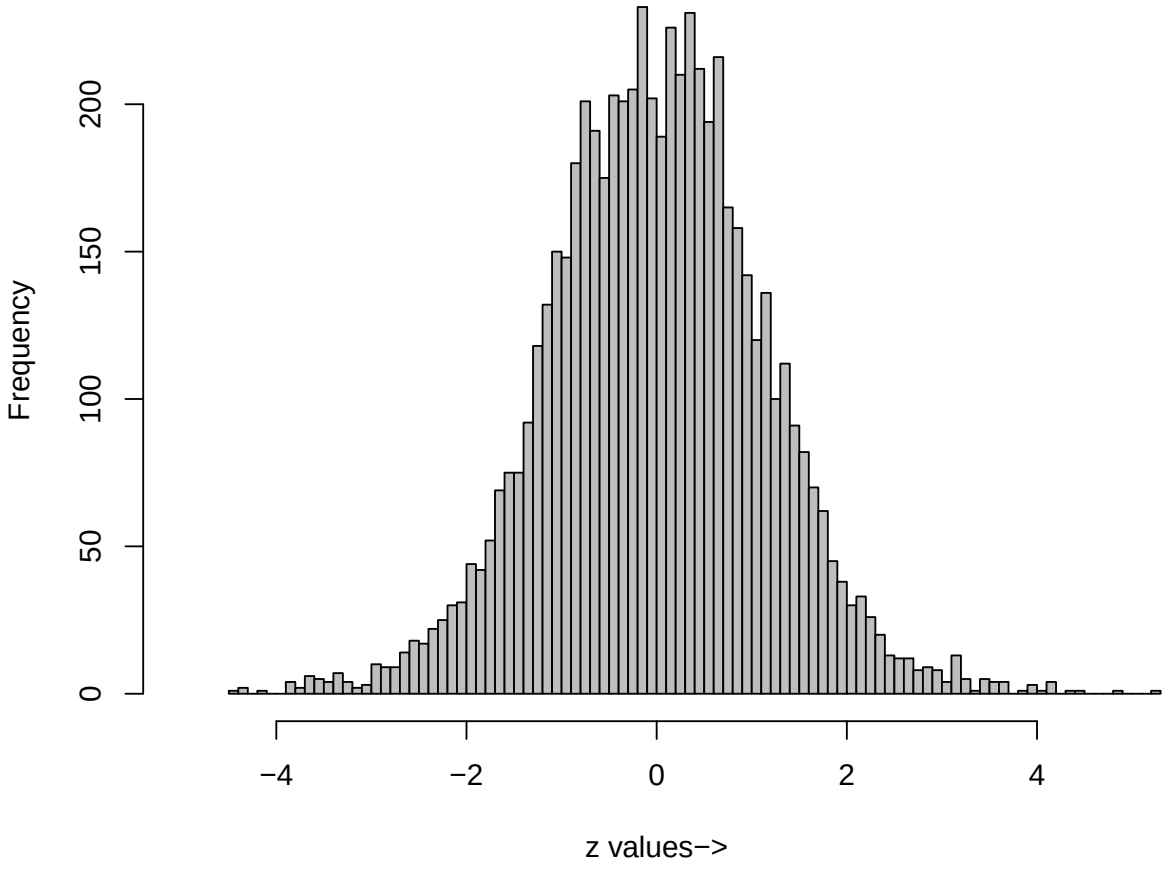

Figure 2: Reproduced Efron Figure 2.1

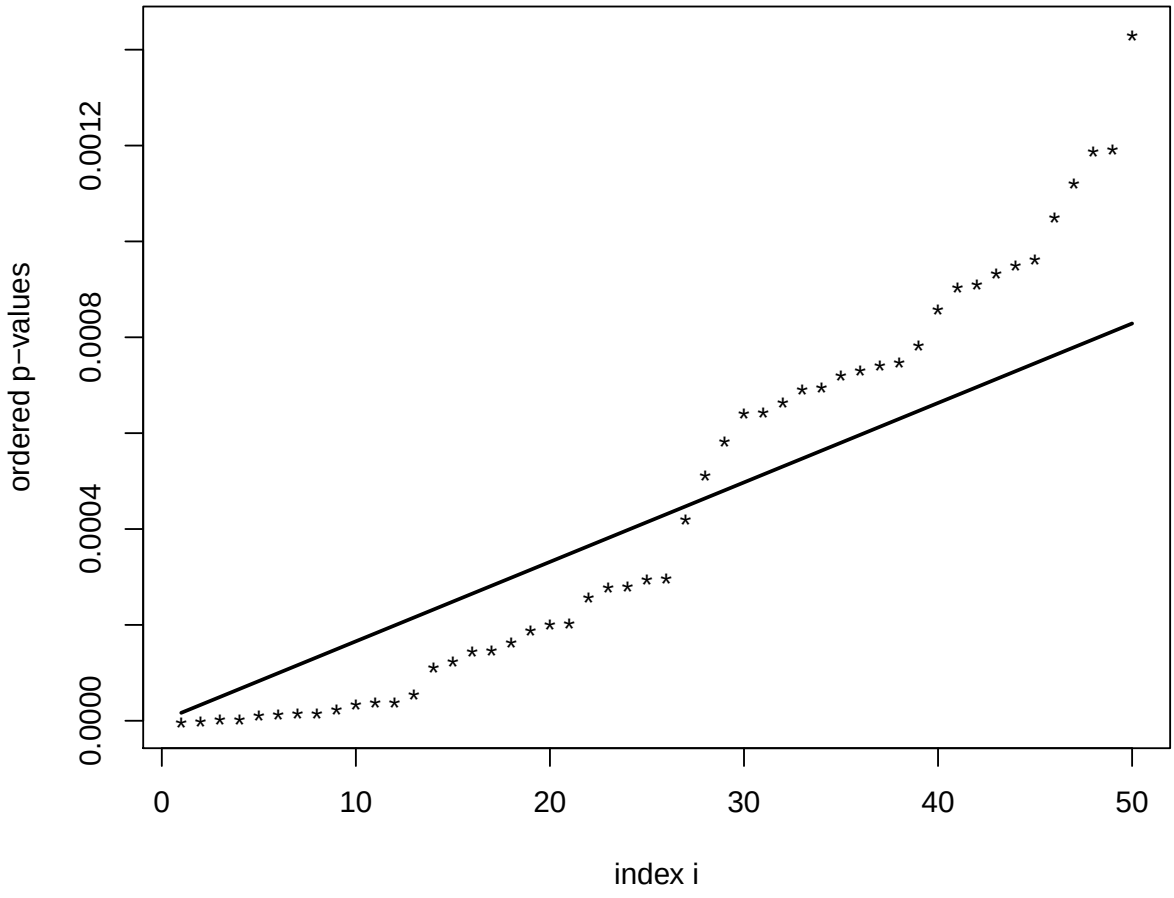

Figure 3: Reproduced Efron Figure 4.2