Chapter 4: False Discovery Rate Control

Benjamini & Hochberg FDR Control

Under the null

$$H_{0i}: P_i \sim \text{Unif}(0,1)$$

which we can order

$$P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(N)}$$

Let R_D be the number of cases rejected then the false discovery proportion with respect to D is

$$Fdp_D = \frac{a_D}{R_D}$$

where

Decision

	Null	Non Null	.
Null	[illegible]	[illegible]	[illegible]

Actual		$N_0 - a$	a	N_0
	Nonnull	$N_1 - b$	b	N_1
		$N - R$	R	N

So we have a Type I errors

and $N - R$ Type II errors.

Here we have $R = a + b$ hypothesis

<u>Task:</u> How many of R are true discoveries.

<u>Def:</u> Family wise error rates (FWER)

$$\mathbb{P}(a > 0)$$

Check: When $N = 1$

$$\mathbb{P}(a = 1 \mid N_0 = 1) = \alpha$$

$$\mathbb{P}(b = 1 \mid N_1 = 1) = \beta$$

Under independence with rejection region R

$$a \mid R \sim \text{Binom}(R, \phi(\mathcal{Z}))$$

counts number of type II

$$\phi(\mathcal{Z}) = \mathbb{P}(\text{reject} \mid \text{null is true})$$

Suppose $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ is the marginal dist. of the test stat.

$$\phi(\mathcal{Z}) = \frac{\mathbb{P}_0(z \in \mathcal{Z})\, \pi_0}{\mathbb{P}(z \in \mathcal{Z})}$$

$$= \frac{\pi_0 \int_{\mathcal{Z}} f_0(z)\, dz}{\int_{\mathcal{Z}} f(z)\, dz}$$

$$= FDR(z)$$

What is the dist. of a p-value?

$$\mathbb{P}(P_i \le u) = \mathbb{P}(F(Z_i) \le u)$$

$$= \mathbb{P}(Z_i \le F^{-1}(u))$$

$$= F_Z(F_Z(u)) = u$$

So the false discovery prop.

$$Fdp_D = \frac{a_D}{R_D} \leftarrow$$ # false discovery for region/decision rule D.

Benjamini Hochberg

For a fixed value $q \in (0,1)$ let

$$i_{max} = \operatorname{argmax} \quad p_{(i)} \leq \frac{i}{N} q$$

Decision: reject $H_{0(i)}$ if $i \leq i_{max}$

and the adjusted p-value is given by

$$\frac{N}{i} p_{(i)} < q$$

Thrm: Under independence the BH algorithm controls the expected false discovery prop. at q

$$\mathbb{E}\left(Fdp_{BH(q)}\right) = \pi_0 q \leq q \qquad \pi_0 = \frac{N_0}{N}$$

Pf: $t \in [0,1]$ $\quad R(t) = \# p_i \leq t$

$a(t) = \#$ of false discoveries

$$Fdp(t) = \frac{a(t)}{\max(R(t), 1)}$$

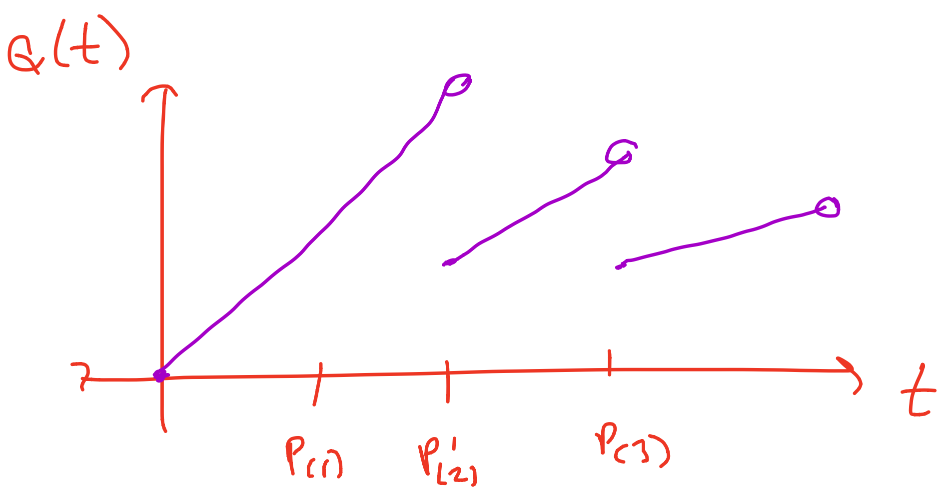
$$Q(t) = \frac{\#\text{ rejects}}{\max(R(t),1)}$$

$$\overset{H_0}{=} \frac{Nt}{\max(R(t),1)}$$

In the BH procedure

$$P_i \le \frac{i}{N} q \iff \frac{N}{i} P_i \le q$$

$$\iff Q(P_{(i)}) \le q$$



Using a generalized inference,

$$t = \sup\{Q(t) \le q\}$$

$$t_q = \sup_t (a(t) - t)$$

So the decision becomes

$$P_{(i)} \le t_q .$$

Now define $A(t) = \dfrac{a(t)}{t}$

We claim this is a martingale.

$$\frac{1}{s} \mathbb{E}\left(a(s) \mid a(t) = y_2 \right)$$

$$\mathbb{P}(a(s) = y_1 \mid a(t) = y_2) \qquad y_1 \le y_2$$

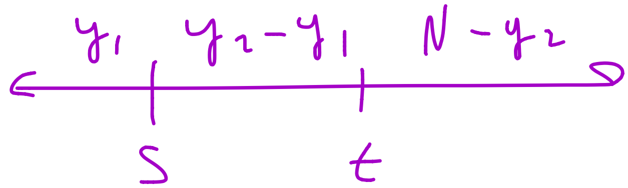
$$= \frac{\mathbb{P}\left(a(s) = y_1 ,\ a(t) = y_2 \right)}{\mathbb{P}(a(t) = y_2)}$$

$$= \frac{\mathbb{P}\left(a(s) = y_1 ,\ a(t) - a(s) = y_2 - y_1 \right)}{\mathbb{P}(a(t) = y_2)}$$

Now $a(t) \sim \mathrm{Binom}(N_0 , t)$

by looking at the rows

$$= \frac{\binom{N_0}{y_1 \; y_2-y_1} s^{y_1} (t-s)^{y_2-y_1} (1-t)^{N_0-y_2}}{\binom{N_0}{y_2} t^{y_2} (1-t)^{N_0-y_2}}$$



$$= \frac{\binom{N_0}{y_1, y_2-y_1}}{\binom{N_0}{y_2}} \left(\frac{s}{t}\right)^{y_1} \left(\frac{(t-s)}{t}\right)^{y_2}$$

$$= \frac{\dfrac{\cancel{N_0!}}{y_1! \; y_2-y_1! \; \cancel{N_0-y_2}!}}{\dfrac{\cancel{N_0!}}{N_0! \; \cancel{N_0-y_2}!}} \left(s/t\right)^{y_1} \left(1-s/t\right)^{y_2}$$

$$= \binom{y_2}{y_1}\left(\frac{s}{t}\right)^{y_1}\left(1-s/t\right)^{y_2-y_1}$$

$$\frac{1}{s}\mathbb{E}\left(a(s) \mid a(t)=y_2\right) = \frac{1}{s} y_2 \frac{s}{t} = \frac{y_2}{t}$$

$$\mathbb{E}\left(A(s) \mid A(t)=t_2\right)$$

$$= \mathbb{E}\left(A(s) \mid a(t)=t t^2\right)$$

$$= \frac{t t_2}{t} = t_2$$

So $A(s)$ is a decreasing martingale with stopping time t_z. So by optimal sampling therm

$$\mathbb{E}\left(A(t_z)\right) = \mathbb{E}\left(A(1)\right) = \mathbb{E}\left(a(1)\right)$$

$$= N_0$$

$$\max\{R(t_q), 1\} = \frac{N t_q}{Q(t_q)} = \frac{N t_q}{q}$$

So the false discovery prop.

$$\mathbb{E}(FDP_D) = \frac{a(t_q)}{\max(1, R(t_q))} = \frac{q\, a(t_q)}{N t_q}$$

$$\mathbb{E}(FDP(t_q)) = \frac{q}{N} \mathbb{E}(A(t_q))$$

$$= \frac{q}{N} N_0$$

$$= \pi_0 q \le q$$

- On expect. we control the rate.
  What about variability
- How should q be chosen?
- Is the theoretical null

$p_i \sim \text{Unif}(0,1)$

correct?

Empirical Bayes Interp.

$$p_{(i)} = F_0(z_{(i)})$$

$$z_{(1)} \le \cdots \le z_{(N)}$$

$$\overline{Fdr}(z) = \frac{\pi_0 F_0(z)}{\overline{F}(z)}$$

BH is equivalent to

$$\frac{\pi_0 F_0(z_{(i)})}{\overline{F}(z_i)} \le \pi_0 \xi$$