Large scale hypothesis testing and false discovery rate (FDR)

$K \geq 1$ hypotheses to be tested simult.

$H_0: H_{01} \text{ & } H_{02} \text{ & } H_{03} \dots \text{ & } H_{0K}$ vs $H_1: H_{11}$ or H_{12} or $\dots$ or H_{1K}

$H_0: H_{0i} \quad i = 1, \dots, K$ vs $H_1: H_{1i} \quad i = 1, \dots K$

For $i = 1, \dots, K$ let p_i (p-value) of H_{0i} vs H_{1i}

Bonferroni strategy: if $\alpha \in (0,1)$ reject H_{0i} if $p_i < \frac{\alpha}{K}$

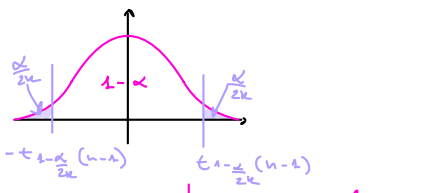
Example

$X_1, \dots, X_n$ iid $\sim N_p(\mu, \Sigma)$

$a_1, \dots, a_K \in \mathbb{R}^p$

$H_{0i}: a_i' \mu = \delta_i$ vs $H_{1i}: a_i' \mu \neq \delta_i$

Bonf: $\alpha \in (0,1)$ reject H_{0i} if $\frac{|a_i' \bar{X} - \delta_i|}{\sqrt{a_i' \hat{\Sigma} a_i}} > t_{1-\frac{\alpha}{2K}}(n-1)$



Bonf strategy: $\mathbb{P}\left[\bigcup_{i=1}^{K} \{\text{rejecting } H_{0i}\} \,\middle|\, \bigcap_{i=1}^{K} H_{0i}\right] \leq \sum_{i=1}^{K} \mathbb{P}[\text{rejecting } H_{0i} \mid H_{0i}] \leq$

$\leq \sum_{i=1}^{K} \frac{\alpha}{K} = \alpha$

FDR: Benjamini & Hochberg

$H_0: H_{0i} \quad i = 1, \dots, K$ vs $H_1: H_{1i} \quad i = 1, \dots K$

Let D be any strategy for running H_0 vs H_1

Truth \ Decision following D	Do not RS H_0	Reject H_0	
H_0	U	false discovery V	K_0
H_1	missed discovery T	true discovery S	$K - K_0$
	$K - R$	R observables	K

Assume H_{0i} is true iff $i \in I_0$ $\quad |I_0| = K_0$

D = Bonf.

$\mathbb{P}[V \geq 1] = \mathbb{P}[\text{At least one false discovery}] = \mathbb{P}\left[\bigcup_{j \in I_0} \{\text{rejecting } H_{0j}\} \,\middle|\, \bigcap_{j \in I_0} H_{0j}\right] \leq$

$\leq \sum_{j \in I_0} \mathbb{P}[\text{rejecting } H_{0j} \mid H_{0j}] = \sum_{j \in I_0} \frac{\alpha}{K} = K_0 \frac{\alpha}{K} \leq \alpha$

Family wise Error rate: Bonf FWER $\leq \alpha$

Consider $\frac{V}{R}$ with the convention $\frac{V}{R} = 0$ if $R = 0$, $\quad Q = \begin{cases} 0 & \text{if } R = 0 \\ \frac{V}{R} & \text{if } R > 0 \end{cases}$

Definition (FDR)

False discovery rate: $FDR = \mathbb{E}[Q]$

Obs: 1. Assume $K_0 = K$ (no discovery to be made) $\Longrightarrow S = 0 \Longrightarrow V = R$

$Q = \begin{cases} 0 & \text{if } R = 0 \iff V = 0 \\ 1 & \text{if } R > 0 \iff V > 0 \end{cases}$

$FDR = \mathbb{E}[Q] = \mathbb{P}[V > 0] = \mathbb{P}[V \geq 1] = FWER$

2. Assume $K_0 < K$:

if $V = 0 \Longrightarrow Q = 0 = V$

if $V > 0 \Longrightarrow Q = \frac{V}{R} < 1$

$\mathbb{1}[V > 0] = \begin{cases} 0 & V = 0 \\ 1 & V > 0 \end{cases} \Longrightarrow Q \leq \mathbb{1}[V > 0]$

$FDR = \mathbb{E}[Q] \leq \mathbb{E}[\mathbb{1}[V > 0]] = \mathbb{P}[V \geq 1] = FWER$

$\Rightarrow$ FDR $\leq$ FWER

Hence: controlling FDR is less restrictive than controlling FWER

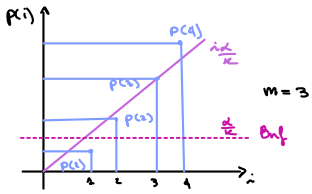
A strategy for controlling FDR

Let p_i p-value of H_{0i} vs H_{1i}

Consider the K p-values: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(K)}$

$p_{(1)}$: $H_{0(1)}$ vs $H_{1(1)}$; $p_{(K)}$: $H_{0(K)}$ vs $H_{1(K)}$

Let $\alpha \in (0,1)$. $m = \max\{ i = 1, \dots, K : p_{(i)} \leq i \frac{\alpha}{K} \}$



$m = 3$ Reject $H_{0(1)}, H_{0(2)}, H_{0(3)}$

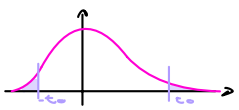
Theorem

If $p_1, \dots, p_K$ are independent. The strategy D_α that rejects $H_{0(i)}$ if $i \leq m$ controls FDR at level α

Obs: p-value: $p = p(x_1, \dots, x_n)$

Example

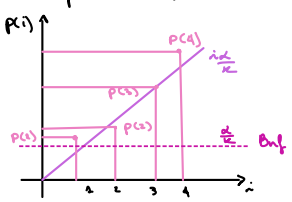
$p(x_1, \dots, x_n) = P\left[\frac{|a'\bar{x} - \delta|}{\sqrt{a' S a}} \sqrt{n} \leq |t| \right]$ $\overset{t_0}{=}$



What if $p_1, \dots, p_K$ are NOT independent? (Benjamini & Yekutieli, 2001)

1. $p_1, \dots, p_K$ are positively correlated $\Longrightarrow$ BH (1995) controls FDR
2. $p_1, \dots, p_K$ are negatively correlated BY consider this strategy:
   reject $H_{0(i)}$ if $i \leq m^* = \max\{ i \in \{1, \dots, K\} : p_{(i)} \leq \frac{i}{K C(K)} \alpha \}$ with $C(K) = \sum_{j=1}^{K} \frac{1}{j}$
   $\Longrightarrow$ Controls FDR at level α

obs: It might happen



BH reject level α $H_{0(1)}, H_{0(2)}, H_{0(3)}$

Stray version of BH: reject $H_{0(i)}$ if $p_{(i)} \leq \frac{i\alpha}{K}$

Comparing means of multivariate Gaussian dist.

Paired data: n statistical units obs. twice:

$\mathbb{R}^p \ni x_{1i}$, $x_{2i} \in \mathbb{R}^p$ $\quad i = 1, \dots, n$

$$\mathbb{R}^p \ni \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}, \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}, \dots, \begin{pmatrix} x_{1n} \\ x_{2n} \end{pmatrix} \quad \text{iid} \sim N_{2p}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma_{2p \times 2p} \right)$$

Attention: x_{1i} and x_{2i} might be dependent!

Typical inf. questions: $H_0: \mu_1 - \mu_2 = \delta$ vs $H_1: \mu_1 - \mu_2 \neq \delta$ (e.g. $\delta = 0$)

$CR_{1-\alpha}(\mu_1 - \mu_2)$

Consider $D_i = x_{1i} - x_{2i}$ iid $N_p(\delta, \Sigma_D)$ $\Longrightarrow$ $\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$

$$S_D = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})(D_i - \bar{D})'$$

Note: $n(\bar{D} - \delta)' S_D^{-1} (\bar{D} - \delta) \sim \frac{(n-1)p}{(n-p)} F(p, n-p)$ pivotal

$\alpha \in (0,1)$ $CR_{1-\alpha}(\mu_1 - \mu_2) = CR_{1-\alpha}(\delta) = \left\{ \delta \in \mathbb{R}^p : n(\bar{D}-\delta)' S_D^{-2} (\bar{D}-\delta) \leq \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \right\}$

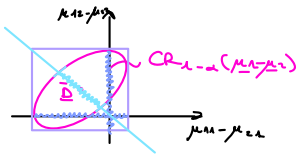
$H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_1: \mu_1 - \mu_2 \neq \delta_0$

$\alpha \in (0,1)$ reject H_0 if $n(\bar{D}-\delta_0)' S_D^{-2} (\bar{D}-\delta_0) > \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p)$

$\delta = \mu_1 - \mu_2$

$Sim\, CI_{1-\alpha}(a'\delta) = \left[a'\bar{D} \pm \sqrt{\frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p)} \sqrt{\frac{a' S_D a}{n}} \right]$

Example



$\mu_1 - \mu_2 = \delta \in \mathbb{R}^2$

$H_0: \delta = 0$ vs $H_1: \delta \neq 0$

Reject H_0 at level α : $\mu_1 \neq \mu_2$

$a'\mu_1 = a'\mu_2$ $\forall a$

Repeated Measurements

Univariate case

$X_1, \dots, X_n$ iid $\sim N_q(\mu, \Sigma)$

$X_i = (x_{i1}, \dots, x_{iq})' \in \mathbb{R}^q$ q measurments repeated on the same unit $i = 1, \dots, n$

$\mu = (\mu_1, \dots, \mu_q) \in \mathbb{R}^q$

$H_0: \mu_1 = \mu_2 = \dots = \mu_q$ vs $H_1: \exists i,j \in \{1, \dots, q\}$ st. $\mu_j \neq \mu_i$

$H_0: \mu \in \mathcal{L}(1)$ vs $H_1: \mu \notin \mathcal{L}(1)$

Contrast matrix

C $(q-1) \times q$ is contrast matrix if $C = \begin{bmatrix} c_1' \\ \vdots \\ c_{q-1}' \end{bmatrix}$ $c_i \in \mathbb{R}^q$

st. 1. $c_1, \dots, c_{q-1}$ are linearly indep $\iff$ $\mathcal{L}^\perp(1) = Span(c_1, \dots, c_{q-1})$

2. $c_i' \cdot 1 = 0$ for $i = 1, \dots, q-1$

Then we can rewrite the hypotesis: $H_0: C\mu = 0$ vs $H_1: C\mu \neq 0$

- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is unbiased for μ
  $C\bar{X}$ is unbiased for $C\mu$
- $\bar{X} \sim N_q(\mu, \frac{1}{n}\Sigma) \implies C\bar{X} \sim N_{q-1}(C\mu, \frac{1}{n} C\Sigma C')$
- S estimate for Σ : $(n-1)S' \sim Wish(\Sigma, n-1)$
  $(n-1)CS'C' \sim Wish(C\Sigma C', n-1) \perp\!\!\!\perp C\bar{X}$

Hotelling's thm $\implies$ $\underbrace{n(C\bar{X}-C\mu)'(CSC')^{-1}(C\bar{X}-C\mu)}_{\text{pivotale}} \sim \frac{(n-1)(q-1)}{n-q+1} F(q-1, n-q+1)$

$\alpha \in (0,1)$ $T_0^2 = n(C\bar{X})'(CSC')^{-1}(C\bar{X})$

reject at level α if $T_0^2 > \frac{(n-1)(q-1)}{n-q+1} F_{1-\alpha}(q-1, n-q+1)$

Obs: C and $\tilde{C}$ two different contrast matrices:

$\implies$ B $(q-1)\times(q-1)$, $Det(B) \neq 0$ and $C = B\tilde{C}$

$T_0^2 = n(C\bar{X})'(CSC')^{-1}(C\bar{X}) = n(B\tilde{C}\bar{X})'(B\tilde{C}S\tilde{C}'B')^{-1}(B\tilde{C}\bar{X}) =$

$= n(\tilde{C}\bar{X})'B'(B')^{-1}(\tilde{C}S\tilde{C}')^{-1}B^{-1}B(\tilde{C}\bar{X}) = \tilde{T}_0^2$

Ex

$$C = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 & -1 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & \dots & 0 & -1 \end{bmatrix}$$

Obs: $x_1, \dots x_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$

$H_0: \underline{\mu} \in \mathcal{L} \subseteq \mathbb{R}^p$ $\quad H_1: \underline{\mu} \notin \mathcal{L}$ $\quad \mathcal{L}$ = linear subspace of $\mathbb{R}^p$, $\dim(\mathcal{L}) = k \geq 1$

Assume: $\mathcal{L}^\perp = \text{span}(c_1, \dots, c_{p-k})$

Let $C = \begin{bmatrix} c_1' \\ \vdots \\ c_{p-k}' \end{bmatrix}$ $\quad (p-k) \times p$

$H_0: C\underline{\mu} = \underline{0}$ $\quad$ vs $\quad H_1: C\underline{\mu} \neq \underline{0}$

$C\bar{x}$ unbiased for $C\underline{\mu}$

Hotelling's thm: if H_0 is true $\quad T_0^2 = n(C\bar{x})'(CSC')^{-1}(C\bar{x}) \sim \frac{(n-1)(p-k)}{n-p+k} F(p-k, n-p+k)$

Multivariate cases

unit i: $\begin{pmatrix} x_{i1}(h) \\ x_{i1}(w) \end{pmatrix}, \begin{pmatrix} x_{i2}(h) \\ x_{i2}(w) \end{pmatrix}, \dots, \begin{pmatrix} x_{ip}(h) \\ x_{ip}(w) \end{pmatrix}$ $\quad i = 1, \dots, n$ $\quad$ units are independent

$\underline{\mu}_1 \quad \underline{\mu}_2 \quad \underline{\mu}_p \in \mathbb{R}_2$

$\begin{cases} H_0: \mu_1(h) = \mu_2(h) = \dots = \mu_p(h) \text{ and } \mu_1(w) = \mu_2(w) = \dots = \mu_p(w) \\ H_1: H_0^c \end{cases}$

Change of rep: $x_i = (x_{i1}(h), x_{i2}(h), \dots, x_{ip}(h), x_{i1}(w), \dots, x_{ip}(w))' \in \mathbb{R}^{2p}$