# Multiple testing

March 7, 2024

# Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of $p$ genes for $n_1$ healthy individuals

$Y_{n_2 \times p}$ - expressions of $p$ genes for $n_2$ cancer patients

Assumption: $X_{ij}$ for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

$Y_{ij}$ for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene $j$ is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

We test $H_{0j} : \mu_{1j} = \mu_{2j}$ with a t-test $t_j = \frac{\bar{X}_{.j} - \bar{Y}_{.j}}{S(\bar{X}_{.j} - \bar{Y}_{.j})}$, where $S(\bar{X}_{.j} - \bar{Y}_{.j})$ is the estimate of the standard deviation of $\bar{X}_{.j} - \bar{Y}_{.j}$

If $n_1$ and $n_2$ are large enough then $t_j \sim N(\mu_j, 1)$ with $\mu_j = \frac{\mu_{1j} - \mu_{2j}}{\sigma_{1j}/\sqrt{n_1} + \sigma_{2j}/\sqrt{n_2}}$ and $H_{0j} : \mu_j = 0$

# Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \dots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

Reject $H_{0i}$ when $|X_i| > c$

Multiple comparison problem: if all $\mu_i$s are equal to zero than $max(|X_1|, \dots, |X_p|) = \sqrt{2 \log p}(1 + o_p)$

Thus to separate signal from noise we need $c = c(p) \to \infty$ as $p \to \infty$.

# Testing for global null, Bonferroni procedure

$X_i \sim N(\mu_i, 1), \quad i = 1, \dots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

$$H_0 : \bigcap_{i=1}^{p} H_{0i}$$

Bonferroni procedure: Reject $H_0$ when

$\max(|X_1|, \dots, |X_p|) \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = c_{Bon}$

Probability of type I error:

$$P_{H_0}\left(\bigcup_{j=1}^{p} \{|X_j| > c_{Bon}\}\right) \leq \sum_{j=1}^{p} P(\{|X_j| > c_{Bon}\}) = \alpha$$

# Exact type I error of Bonferroni

Due to independence

$$
\begin{aligned}
P(\mathit{Type\ I\ Error}) &= 1 - P_{H_0}\left(\bigcap_{j=1}^{p}\{|X_j| < c_{Bon}\}\right) \\
&= 1 - \left(1 - \frac{\alpha}{p}\right)^{p} \to 1 - e^{-\alpha} = \alpha + o(\alpha)
\end{aligned}
$$

$\alpha = 0.05$ , $n = 30000$, $P(\mathit{Type\ I\ Error}) \approx 0.0488$

# Multiple testing

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

| | $H_0$ accepted | $H_0$ rejected | |
|---|---|---|---|
| $H_0$ true | U | V | $p_0$ |
| $H_0$ false | T | S | $p_1$ |
| | W | R | p |

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

$$\alpha = 0.05, p_0 = 5000 \rightarrow E(V) = 250$$

# Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject $H_{0i}$ if $|X_i| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2\log p}(1 + o(1))$

Benjamini-Hochberg (1995) procedure:

(1) $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$

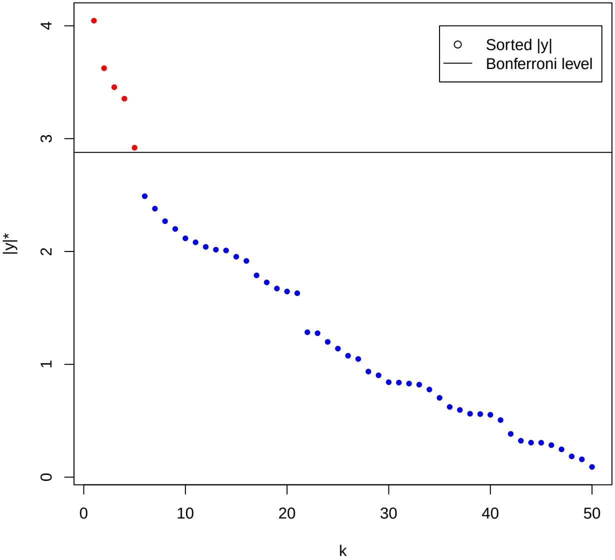
(2) Find the largest index $i$ such that

$$|X|_{(i)} \geq \Phi^{-1}(1 - \alpha_i), \quad \alpha_i = \alpha \frac{i}{2p}, \tag{1}$$

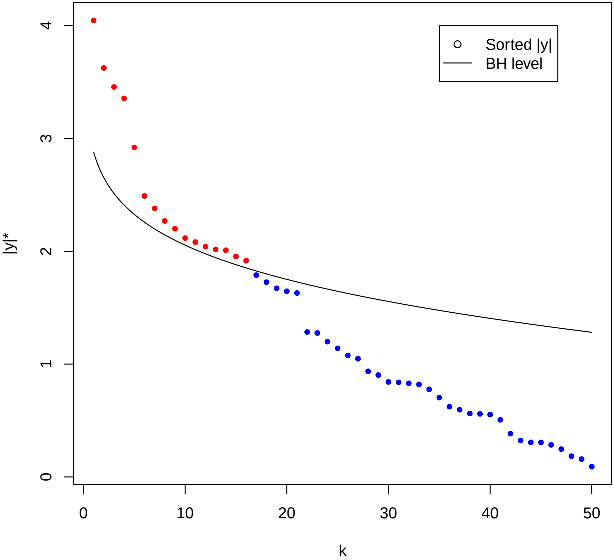
Call this index $i_{\text{SU}}$.

(3) Reject all $H_{(i)}$'s for which $i \leq i_{\text{SU}}$

# Bonferroni correction



# Benjamini and Hochberg correction



# FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR at:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}, \tag{2}$$

where $p_0$ is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

(Benjamini, Yekutieli, 2001) When test statistics are "positively correlated" then BH controls FDR at or below the level $\alpha \frac{p_0}{p}$. Independently of the correlation structure FDR is controlled at or below the level $\alpha \frac{p_0}{p}$ if $|X|_{(j)}$ is compared to $\Phi^{-1}\left(1 - \frac{j\alpha}{2p\sum_{i=1}^{p}\frac{1}{i}}\right)$.