Multiple testing issues

Motivation: differential gene expression

Suppose a biologist is interested in identifying genes which are *differentially expressed* under different biological treatments. The biologist observes 10 subjects under treatment A, and 10 subjects under treatment B. Gene expression measurements $X_{i,j}$ (treatment A) and $Y_{i,j}$ (treatment B) are recorded for 1000 different genes ($i = 1, \ldots, 1000$, $j = 1, \ldots, 10$).

For each gene i, the biologist tests $H_0 : \mu_{i,A} = \mu_{i,B}$, rejecting when the p-value is below a threshold α.

> If H_0 is actually true for all 1000 genes, how many false positives do we expect?

Motivation: multiple testing

> In what other settings might we test many hypotheses?

Outcomes for multiple hypothesis tests

Suppose we test m null hypotheses $H_{0,1}, \dots, H_{0,m}$, of which m_0 are true. Ideally we would fail to reject the m_0 true nulls, and reject the $m - m_0$ others.

Possible outcomes:

	H_0 true	H_0 false	
Reject	V		
Fail to reject			
	m_0	$m - m_0$	m

We want to control V, the # of false positives (type I errors). One option is the family-wise error rate (FWER):

$$\text{FWER} = P(V > 0)$$

Family-wise error rate

Definition: Suppose we test m null hypotheses $H_{0,1}, \ldots, H_{0,m}$. The *family-wise error rate* is the probability of making *at least one* type I error:

$$FWER = P\left(\bigcup_{i:H_{0,i} \text{ is true}} \{\text{reject } H_{0,i}\}\right)$$

> Suppose all m tests are independent, and $H_{0,i}$ is true for all tests. For each test, we reject if the corresponding p-value $p_i < \alpha$. What is the FWER?

$FWER = P\left(\bigcup_{i:H_{0,i} \text{ true}} \text{reject } H_{0,i}\right) = P\left(\bigcup_{i=1}^{m} \text{reject } H_{0,i}\right)$ (all $H_{0,i}$ true)

$= 1 - P(\text{fail to reject all } H_{0,i})$

$= 1 - (1-\alpha)^m$ (independence)

The Sidak correction

Suppose we test m null hypotheses $H_{0,1}, \dots, H_{0,m}$. The *family-wise error rate* is the probability of making *at least one* type I error:

$$FWER = P\left(\bigcup_{i:H_{0,i} \text{ is true}} \{\text{reject } H_{0,i}\}\right)$$

> If all m hypotheses are independent, at what threshold α^* should we reject each test, such FWER $\leq \alpha$?

$FWER \leq 1-(1-\alpha^*)^m$ (upper bound achieved when all $H_{0,i}$ true)

$\Rightarrow 1-(1-\alpha^*)^m = \alpha \Rightarrow \alpha^* = 1-(1-\alpha)^{\frac{1}{m}}$

Sidak correction: Reject $H_{0,i}$ if $p_i < 1-(1-\alpha)^{\frac{1}{m}}$

Union bound: $P(\cup_i A_i) \leq \sum_i P(A_i)$

The Bonferroni correction

Suppose we test m null hypotheses $H_{0,1}, \ldots, H_{0,m}$. The *family-wise error rate* is the probability of making *at least one* type I error:

$$FWER = P\left(\bigcup_{i:H_{0,i} \text{ is true}} \{\text{reject } H_{0,i}\}\right)$$

Suppose we reject each $H_{0,i}$ if $p_i < \alpha^*$

$$P\left(\bigcup_{i:H_{0,i} \text{ is true}} \text{reject } H_{0,i}\right) \leq \sum_{i:H_{0,i} \text{ true}} P(\text{reject } H_{0,i}) \leq \sum_{i:H_{0,i} \text{ true}} \alpha^* \leq m\alpha^*$$

Bonferroni correction: Reject when $p_i < \frac{\alpha}{m}$

$$\Rightarrow FWER \leq m\left(\frac{\alpha}{m}\right) = \alpha$$

$\frac{\alpha}{m} < 1-(1-\alpha)^{\frac{1}{m}}$ so Bonferroni is more conservative (rejects less often) than Sidak

Holm's procedure

Suppose we test 5 hypotheses, and observe p-values 0.4, 0.01, 0, 0, 0. Does it still seem reasonable to use the Bonferroni cutoff $\alpha/5$ for each test?

No — want to use other p-values to make a decision for each hypothesis

Idea: order p-values $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$

First test: if $p_{(1)} < \frac{\alpha}{m}$ (Bonferroni threshold), reject $H_{0(1)}$

If reject first test, consider $p_{(2)}$. There are now $m-1$ tests left, so reject $H_{0(2)}$ if $p_{(2)} < \frac{\alpha}{m-1}$

⋮

continue procedure: if $p_{(i)} < \frac{\alpha}{m-i+1}$, reject $H_{0(i)}$

As soon as $p_{(i)} > \frac{\alpha}{m-i+1}$ for some i, stop the procedure (and fail to reject $H_{0(i)}, \ldots, H_{0(m)}$)

Holm's procedure

Suppose we test m null hypotheses $H_{0,1}, \ldots, H_{0,m}$. Let p_i be the corresponding p-value for test i.

- Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
- Let $i^* = \min\left\{i : p_{(i)} > \frac{\alpha}{m-i+1}\right\}$
- Reject $H_{0,(i)}$ for all $i < i^*$

Claim: Holm's procedure controls FWER at level α

Proof: Let $I_0 = \{i : H_{0(i)} \text{ is true}\}$. Let $m_0 = \#\{i : H_{0(i)} \text{ is true}\}$ (number of true nulls)

Let $j = \min(I_0)$. Holm's procedure compares $p_{(j)}$ to $\frac{\alpha}{m-j+1}$

if $p_{(j)} > \frac{\alpha}{m-j+1}$, fail to reject all true nulls

Since $j = \min(I_0)$ and there are m_0 elements in I_0,

$m-j+1 \geq m_0$ $\Rightarrow$ fail to reject all true nulls if $p_{(j)} > \frac{\alpha}{m_0}$

$\Rightarrow$ FWER $\leq P\left(\min_{i \in I_0} p_{(i)} < \frac{\alpha}{m_0}\right) \leq m_0\left(\frac{\alpha}{m_0}\right) = \alpha$ (union bound)

Class activity

https://sta711-s23.github.io/class_activities/ca_lecture_28.html

- Sidak is slightly more powerful than Bonferroni, but requires independence
- Holm's procedure is more powerful than Bonferroni, and requires no additional assumptions

  why? Bonferroni rejects if $p_i < \frac{\alpha}{m}$

  For Holm, each p-value $p_{(i)}$ is compared to $\frac{\alpha}{m-i+1} \geq \frac{\alpha}{m}$

  So the threshold is less stringent for Holm than for Bonferroni