STAT540

Wednesday, February 25, 2015

Dr. Gabriela Cohen Freue

Multiple Testing Correction: Review

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	true negatives	false positives Type I errors	# nulls
H_A holds "interesting"	false negatives Type II errors	true positives	# alts
		discoveries	# genes

false positive rate = P(stat sig test stat or pvalue) for a truly null gene, i.e. one of the m_0

if you threshhold at a p-value $\leq \alpha$, for example $\alpha = 0.05$, then you control the false positive rate

then the expected number of false positives is αm_0 .... which can be quite large!

example: 0.05 * 5000 = 250

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	m_0 - F	F	m_0
H_A holds "interesting"	m_1 - T	T	m_1
		S	m

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	true negatives	false positives Type I errors	# nulls
H_A holds "interesting"	false negatives Type II errors	true positives	# alts
		discoveries	# genes

family-wise error rate (FWER) = $P(F > 1)$

probability at least one null gene is called a hit

if you use Bonferroni, then you control FWER

very very conservative approach

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	m_0 - F	F	m_0
H_A holds "interesting"	m_1 - T	T	m_1
		S	m

Bonferroni correction

Used to control FWER, i.e. to ensure that

FWER = $P(F > 1) = P(\text{at least one false positive}) \leq \alpha$

Viewpoint #1: adjust the p-values

$\tilde{p}_i = mp_i$ (or, more technically correct, $\min(mp_i, 1)$)

Then threshhold the $\tilde{p}_i$ at α.

Viewpoint #2: adjust the threshhold

$\tilde{\alpha} = \alpha / m$

Then threshhold the p_i at $\tilde{\alpha}$.

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	true negatives	false positives Type I errors	# nulls
H_A holds "interesting"	false negatives Type II errors	true positives	# alts
		discoveries	# genes

false discovery rate (FDR) = $E\left(\frac{F}{S}\right)$

expected proportion of false positives among the hits

if you use q-values, you control FDR

"call" based on obs. data true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	m_0 - F	F	m_0
H_A holds "interesting"	m_1 - T	T	m_1
		S	m

3. FALSE DISCOVERY RATE CONTROLLING PROCEDURE

3.1. *The Procedure*

Consider testing $H_1, H_2, \ldots, H_m$ based on the corresponding p-values $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leqslant P_{(2)} \leqslant \ldots \leqslant P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following Bonferroni-type multiple-testing procedure:

$$\text{let } k \text{ be the largest } i \text{ for which } P_{(i)} \leqslant \frac{i}{m} q^*;$$

$$\text{then reject all } H_{(i)}\ i = 1, 2, \ldots, k. \tag{1}$$

Theorem 1. For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.

Proof. The theorem follows from the following lemma, whose proof is given in Appendix A.

Lemma. For any $0 \leqslant m_0 \leqslant m$ independent p-values corresponding to true null hypotheses, and for any values that the $m_1 = m - m_0$ p-values corresponding to the false null hypotheses can take, the multiple-testing procedure defined by procedure (1) above satisfies the inequality

$$E(\mathbf{Q} \mid P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leqslant \frac{m_0}{m} q^*. \tag{2}$$

Now, suppose that $m_1 = m - m_0$ of the hypotheses are false. Whatever the joint distribution of $P_1'', \ldots, P_{m_1}''$ which corresponds to these false hypotheses is, integrating inequality (2) above we obtain

$$E(\mathbf{Q}) \leqslant \frac{m_0}{m} q^* \leqslant q^*,$$

and the FDR is controlled.

Remark. Note that the independence of the test statistics corresponding to the false null hypotheses is not needed for the proof of the theorem.

From Benjamini and Hochberg 1995. Q is the false discovery proportion. E(Q)=FDR.

let k be the largest i for which $P_{(i)} \leqslant \frac{i}{m} q^*$;

then reject all $H_{(i)}$ $i = 1, 2, \ldots, k$.

Call a "hit" if p-value $\leq \dfrac{\text{rank of p-value}}{m} q^*$

Let's try to get in a more practical form:

Call a "hit" if q-value $\leq q^*$

That implies this definition of a q-value:

$$\text{q-value} \equiv \text{p-value} \frac{m}{\text{rank of p-value}}$$

Theorem 1. For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.

$$E(\mathbf{Q}) \leqslant \frac{m_0}{m} q^* \leqslant q^*,$$

Statistical significance for genomewide studies. Storey JD, Tibshirani R. Proc Natl Acad Sci USA 2003 Aug 5100(16):9440-5

Storey coined the term q-value.

q-value (feature) = expected proportion of false positives if this feature is called significant

= expected proportion of false positives among all features as or more extreme than this feature

$$FDR(t) = E\left(\frac{F(t)}{S(t)}\right) \cong \frac{E(F(t))}{E(S(t))}$$

$$FDR(t) = E\left(\frac{F(t)}{S(t)}\right) \cong \frac{E(F(t))}{E(S(t))}$$

Substitute our estimates of numerator and denominator to get:

$$\widehat{FDR}(t) = \frac{\hat{\pi}_0 mt}{\#\{p_i \leq t\}}$$

Consider all threshholds t.

Basically, obtain the q-value of each feature by plugging its p-value into the formula above in the place of t.

More correctly,

$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{FDR}(t)$$

q-value is basically this:

$$\hat{q}(p_i) = \frac{\hat{\pi}_0 m p_i}{\#\{p_j \le p_i\}} = \frac{\hat{\pi}_0 m}{\text{rank of } p_i} p_i$$

"call" based on obs. data / true state of nature	"not hit"	reject H_0 "hit"	
H_0 holds	m_0 - F	F	m_0
H_A holds	m_1 - T	T	m_1
		S	m

With the q-values computed, one can get a "hit list" with estimated FDR of q^* by calling all genes significant with q-values $\le q^*$.

Let's revisit the proto-q-value we computed from Benjamini-Hochberg:

$$q_{BH}(p_i) = \frac{m}{\text{rank of } p_i} p_i$$

So the only difference is a multiplicative factor of $\hat{\pi}_0$, which should be near one in the relevant applications. Pretty close!

The approach in Storey 2003 / q-value pkg is a bit less conservative than plain vanilla BH and reduces to BH if one takes $\hat{\pi} = 1$.

Take home message:

BH adjustment (q-values) controls the FDR

q-value (feature) = expected proportion of false positives if this feature is called significant

= expected proportion of false positives among all features as or more extreme than this feature