Bioinformation

by Biomedical Informatics Publishing Group

Consider testing m hypothesis pairs (H_{0i}, H_{Ai}), $i = 1, \ldots, m$. In most applications of microarray gene expression analyses, m is typically on the order of 10^5–10^6. Suppose m P values, $P_1, \ldots, P_m$, one for each hypothesis pair, are calculated, and a decision on whether to reject H_{0i} is to be made. Let m_0 be the number of true null hypotheses, and let $m_1 := m - m_0$ be the number of true alternative hypotheses. The outcome of testing these m hypotheses can be tabulated as in Table 1. **[6]**

Here V is the number of null hypotheses erroneously rejected, S is the number of alternative hypotheses correctly captured, and R is the total number of rejections. Conceptually these quantities are random variables. Clearly only m is known and only R is observable. An important parameter is m_0, or equivalently, the *null proportion* $\pi_0 := m_0/m$. This parameter will appear frequently in the subsequent sections, and its estimation will be discussed in **Section 4.**

Multiple hypotheses tests and related error measurements can be well understood as an estimation problem, which is described below in the frequentist framework. First for two probability distributions P_1 and P_2 on R with respective cumulative distribution function (cdf) $F_1(\cdot)$ and $F_2(\cdot)$, P_1 is said *stochastically less than* P_2, written as $\mathsf{P}_1 <_{st} \mathsf{P}_2$, if $F_1(t) \geq F_2(t)$ for all $t \in \mathsf{R}$. Next define the parameter $\Theta = [\theta_1, \ldots, \theta_m]$ as $\theta_i = 1$ if H_{Ai} is true, and $\theta_i = 0$ if H_{0i} is true ($i = 1, \ldots, m$). The data consist of the P values $\{P_1, \ldots, P_m\}$, and under the assumption that each test is exact and unbiased, the population is described by the following probability model:

$$P_i \sim \mathsf{P}_{i,\theta i}$$
$$\mathsf{P}_{i,0} \text{ is } U(0,1), \text{ and } \mathsf{P}_{i,1} <_{st} U(0,1); \qquad (1)$$

each distribution $\mathsf{P}_{i,1}$ has a continuously differentiable cdf $F_i(\cdot)$, $i = 1, \ldots, m$. The P values are dependent in general and have a joint distribution on $[0, 1]^m$. The marginal cdf of P_i can be written as $G_i(t) = (1 - \theta_i)t + \theta_i F_i(t)$. Note $F_i(t) \geq t$ and $G_i(t) \geq t$ for $t \in [0, 1]$.

A test procedure is an estimator of $\Theta : \hat{\Theta} = \hat{\Theta}(P_1, \ldots, P_m) = [\hat{\theta}_1, \ldots, \hat{\theta}_m] \in \{0, 1\}^m$, where $\hat{\theta}_i = 1$ indicates rejecting H_{0i} in favor of H_{Ai}, $i = 1, \ldots, m$. With this notation, the random variables in Table 1 can be expressed as

$$V = V_\Theta\left(\hat{\Theta}\right) = \sum_{i=1}^{m} (1 - \theta_i)\hat{\theta}_i$$
$$S = S_\Theta\left(\hat{\Theta}\right) = \sum_{i=1}^{m} \theta_i \hat{\theta}_i \qquad (2)$$
$$R = R\left(\hat{\Theta}\right) = \sum_{i=1}^{m} \hat{\theta}_i .$$

A natural and perhaps the simplest procedure is the "hard-thresholding" (HT) estimator $\hat{\Theta} = \hat{\Theta}(\alpha)$ defined as

$$HT(\alpha): \hat{\theta}_i = 1 \text{ iff } P_i \leq \alpha , \qquad (3)$$

Where $\alpha \in (0, 1)$ is a significance threshold common to all tests. Clearly for this procedure the distributions of the random variables V, S, and R all depend on α.

2.1 False discovery rate

At least one *family-wise type-I error* is committed if $V > 0$, and procedures for multiple hypothesis testing have traditionally been produced for solely controlling the family-wise type-I error probability $\Pr(V > 0)$. It is well-known that such procedures are often lack of statistical power. In an effort to develop more powerful procedures, **[6]** approached the multiple testing problem from a different perspective and introduced the concept of *false discovery rate* (FDR), which is, loosely speaking, the expected value of the ratio *V/R*. Rigorously, the FDR is defined as $FDR = E\left[V/R \mid R > 0\right]\Pr(R > 0)$. Note that if no alternative hypothesis is true, i.e., $m_0 = m$, then $V = R$ and $E[V/R|R>0] = 1$ with probability one; therefore $FDR = \Pr(V > 0)$, the family-wise type-I error probability.

$$FDR_{\Theta}(\hat{\Theta}) = E\left[\frac{\sum_{i=1}^{m}\hat{\theta}_i(1-\theta_i)}{\sum_{i=1}^{m}\hat{\theta}_i + \Pi_{i=1}^{m}(1-\hat{\theta}_i)}\right] \qquad (4)$$

Benjamini and Hochberg (1995) aim at determining an α based on the P values so that the FDR of the *HT*(α) procedure (3) is controlled below a *pre-specified* level. **[7]**

2.2 Positive FDR and *q*-value

For more discovery-oriented applications the FDR level is often not specified *a priori*, but rather determined after one sees the data (P values), and it is often determined in a way allowing for some "discovery" (rejecting one or more null hypotheses). Hence the *positive false discovery rate* (pFDR; **[7, 8]**, defined as $pFDR := E[V/R|R>0]$, is a more appropriate error measurement. Storey (2002) develops estimators of FDR and pFDR and introduces the concept of *q*-value in a Bayesian framework. **[7]** Assuming that each θ_i is a Bernoulli random variable with $\Pr(\theta_i = 1) = \Pr(H_{0i}) = 1 - \pi_0$ (prior probability), all test statistics have the same null distribution, all test statistics have the same alternative distribution, and all tests are performed with identical rejection regions **[7]**, the pFDR of the *HT*(α) procedure is $pFDR(\alpha) = \pi_0\alpha/\Pr(P \leq \alpha)$, where *P* is the random P value resulted from any test. Storey (2002) uses the phrase "identical tests" to describe the set of assumptions. **[7]**

To understand the *q*-value, first consider the P value. Suppose there are *m* two-sample Student-t tests with a common degrees of freedom *d* and observed statistics $t_1, \ldots, t_m$. For a single test, say the *i*th test, the P value is $P_i = \Pr_{H_{oi}}\left(\left|T_d\right| \geq \left|t_i\right|\right)$, where T_d is a random variable following the *t* distribution with *d* degrees freedom. If a threshold $t^* > 0$ is applied to make the decision whether to reject the null hypothesis, i.e., reject the *i*th null if and only if $|t_i| \geq t^*$ or equivalently, $|t_i|$ is in the rejection region $[t^*,\infty)$, then the P value at $|t_i|$ is $P_i = \inf_{t^* \geq |t_i|}\{\Pr_{H_{0i}}(|T_d| \geq t^*)\}$, that is, the minimum probability over all the rejection regions less stringent than $|t_i|$ under the *i*th null hypothesis. Note the P value is defined for a single test. The *q*-value is defined for all *m* tests as a whole, using pFDR in lieu of the probability distribution under the null hypothesis. Storey (2002) gives a general definition of the *q*-value **[7]**; for the *HT*(α) procedure (3) the *q*-value at α is defined as $q(\alpha) := \inf_{\gamma \geq \alpha}\{pFDR(\gamma)\}$, and $q(\alpha) = \inf_{\gamma \geq \alpha}\{\pi_0\gamma/\Pr(P \leq \gamma)\}$ under the Bayesian model. So the *q*-value at α is the minimum pFDR over all the rejection regions less stringent than α. Thus the *q*-value is an error measurement related to the positive FDR, but it is neither the pFDR nor the FDR. The *q*-value can only be meaningfully interpreted in the Bayesian framework. **[7]** Storey (2003) shows that in the Bayesian framework the *q*-value $q(\alpha)$ can be interpreted as the posterior probability of the null hypothesis given $P \leq \alpha$. **[9]** Estimation of the pFDR and *q*-value will be reviewed in **Section 5.1.**

2.3 Erroneous rejection ratio

As discussed by Benjamini and Hochberg (1995, 2000), the FDR criterion has many desirable properties not possessed by other intuitive alternative criteria for multiple tests. **[6, 10]** However, methodological and theoretical developments and extensions of the FDR approach require to assume certain weak dependence conditions **[9, 11, 12]** or positive dependence structure **[13]** among the test statistics. These conditions may be too strong for genome-wide tests of gene expression–phenotype associations, in which a substantial proportion of the tests can be strongly dependent. **[14]** In such applications it may not be even reasonable to assume that the tests of the true null hypotheses are independent, an assumption often used in FDR research. Without these assumptions however, the FDR becomes difficult to handle analytically. Cheng (2006) defines an analytically simple error measurement in the same spirit of FDR **[15]**, called the *erroneous rejection ratio* (ERR): With notation given in Equation (2),

$$ERR_{\Theta}(\hat{\Theta}) := \frac{E[V_{\Theta}(\hat{\Theta})]}{E[R(\hat{\Theta})]}\Pr(R(\Theta) > 0). \qquad (5)$$

Just like FDR, when all null hypotheses are true $ERR = \Pr(R(\Theta) > 0)$, which is the family-wise type-I error probability because now $V_\Theta(\hat{\Theta}) = R(\hat{\Theta})$ with probability one. An advantage of ERR is that it can be handled under arbitrary dependent relationships among the tests; this will be elaborated later. Denote by $V(\alpha)$ and $R(\alpha)$ respectively the V and R random variables in Table 1 and by $ERR(\alpha)$ the ERR of the $HT(\alpha)$ procedure. Then

$$ERR(\alpha) = \frac{E[V(\alpha)]}{E[R(\alpha)]} \Pr(R(\alpha) > 0).$$

Let $FDR(\alpha) := E[V(\alpha) / R(\alpha) | R(\alpha) > 0] \Pr(R(\alpha) > 0)$. $ERR(\alpha)$ is essentially $FDR(\alpha)$. Under the hierarchical (or random effect) model employed in several papers **[7, 8, 9, 12, 16],** $FDR(\alpha) = ERR(\alpha)$ for all $\alpha \in (0, 1]$, following from Lemma 2.1 of Genovese and Wasserman (2004). **[12]** More generally $ERR/FDR = \{E[V] / E[R]\} / E[V/R | R > 0]$ provided $\Pr(R > 0) > 0$. Asymptotically as $m \to \infty$, if $\Pr(R > 0) \to 1$ then $E[V/R | R > 0] \cong E[V/R]$; if furthermore $E[V/R] \cong E[V]/E[R]$, then $ERR / FDR \to 1$. The last condition is approximately satisfied for the $HT(\alpha)$ procedure if α is close to zero **[8]**, which is often true in microarray applications.

Similar to pFDR is the *positive ERR*, $pERR := E[V] / E[R]$. It is well-defined provided $\Pr(R > 0) > 0$. The relationship between pERR and pFDR is the same as that between ERR and FDR described above.

It is instructive to examine each component of $ERR(\alpha)$. Let $P_{1:m}$ be the smallest P value. First, under model (1)

$$E[V(\alpha)] = \sum_{i=1}^{m} (1 - \theta_i) \Pr(\hat{\theta}_i = 1) = m_0 \alpha$$

$$E[R(\alpha)] = \sum_{i=1}^{m} \Pr(\hat{\theta}_i = 1) = m_0 \alpha + \sum_{j:\theta_j = 1} F_j(\alpha)$$

$$\Pr(R(\alpha) > 0) = \Pr(P_{1:m} \le \alpha).$$

Define

$$F_m(t) := m^{-1} \sum_{i=1}^{m} G_i(t) = \pi_0 t + (1 - \pi_0) H_m(t),$$

$$H_m(t) := m_1^{-1} \sum_{j:\theta_j = 1} F_j(t),$$

$t \in \mathsf{R}$. Then

$$ERR(\alpha) = \frac{\pi_0 \alpha}{F_m(\alpha)} \Pr(P_{1:m} \le \alpha). \qquad (6)$$

Note the functions $F_m(\cdot)$ and $H_m(\cdot)$ both are cdf's with $F_m(0) = H_m(0) = 0$ and $F_m(1) = H_m(1) = 1$. $F_m(\cdot)$ is the average of all P value individual (marginal) cdf's. It describes the ensemble behavior of all P values, hence will be called the *ensemble P value cdf*. $H_m(\cdot)$ is the average of the P value marginal cdf's corresponding to the true alternative hypotheses, and describes the ensemble behavior of the P values corresponding to the true alternative hypotheses; hence will be called the *ensemble P value alternative cdf*. Next, these functions are linked to the actual data (i.e., observed P values) by the Empirical Distribution Function (EDF) of the P values defined as $\tilde{F}_m(t) := m^{-1} \sum_{i=1}^{m} I(P_i \le t)$, $t \in \mathsf{R}$. Simple calculations show that under model (1)

$$E[\tilde{F}_m(t)] = F_m(t) = \pi_0 t + (1 - \pi_0) H_m(t), \; t \in [0, 1]. \qquad (7)$$

This link provides opportunities to develop estimators of the FDR and data-driven significance criteria which will be reviewed in **Sections 4, 5, and 6.**

The false positive error behavior of a given multiple test procedure can be investigated in terms of either FDR (pFDR) or ERR (pERR). The ratio $pERR(\alpha) = E[V(\alpha)] / E[R(\alpha)]$ can be handled easily under arbitrary dependence among the tests

because $E[V]$ and $E[R]$ are simply means of sums of indicator random variables. Cheng (2006) **[15]** develops a data-driven significance threshold criterion to determine an α for the hard-thresholding $HT(\alpha)$ procedure (3) so that its ERR and pERR are guaranteed to diminish asymptotically as the number of tests m goes to infinity, for arbitrarily dependent tests; see **Section 6.**

2.4 Other error measurements

The expected number of type-II errors (false negatives) is $E[m_1 - S]$. For the $HT(\alpha)$ procedure, under model (1) $E[m_1 - S] = m_1 - \sum_{i=1}^{m} I(\theta_i = 1)G_i(\alpha) = m_1 - m_1 Hm(\alpha)$. The *false negative proportion* is $m^{-1}E[m_1 - S] = (1 - \pi_0)(1 - H_m(\alpha))$. This quantity will be further considered in Section 6.2.

Symmetric to FDR, the false non-discovery rate (FNR) can be defined as $FNR = E[(m_1 - S) / (m - R) \mid R < m]$. **[11]**

Lehmann and Ramano (2005) introduced the *generalized family-wise error rate* (gFWER) which is $\Pr(V > k)$ for a specified k. **[17]** The traditional FWER corresponds to $k = 0$. In a series of papers van der Laan and colleagues develop resampling and augmentation procedures of controlling gFWER and the probability $\Pr(V/R > k)$ for a specified k.