技术与方法 DOI:10.3969/j.issn.0253-3626.2011.12.010

错误发现率及其扩展和应用

王 婷,曾 平,黄水平,赵华硕
(徐州医学院公共卫生学院流行病与卫生统计学教研室,徐州 221002)

【摘 要】目的:介绍多重假设检验中的错误发现率控制和在此基础上发展起来的 q 值和局部错误发现率。方法:将 Benjamini 和 Hochberg(BH)控制程序、q 值和局部错误发现率应用于一个前列腺癌微阵列数据的基因差别表达分析,控制和估计错误发现率。结果:当错误发现率为 0.10 时通过 BH 程序得到 60 个差别表达基因,q 值小于 0.10 的有 74 个基因,局部错误发现率小于 0.10 的有 31 个基因。结论:高维数据多重假设检验中能同时控制或估计错误发现率,BH 控制程序、q 值和局部错误发现率的联合应用能提供更多的信息。

【关键词】微阵列数据;多重假设检验;错误发现率;q值;局部错误发现率;前列腺癌
【中国图书分类法分类号】R195.1 【文献标志码】A 【收稿日期】2011-10-08

False discovery rate and its extension and application

WANG Ting, ZENG Ping, HUANG Shui-ping, ZHAO Hua-shuo
(*Epidemiology and Health Statistics Staff Room, School of Public Health, Xuzhou Medical College*)

【Abstract】*Objective*: To introduce the multiple hypothesis testing based on the false discovery rate and its extension to q value and local false discovery rate. *Methods*: False discovery rate, q value and local false discovery rate were applied to the analysis of differentially expressed genes concerning a prostate cancer microarray data to control and estimate the false discovery rate. *Results*: It identified 60 differentially expressed genes with the procedure of Benjamini and Hochberg when controlling the false discovery rate below 0.10, 74 genes with q value less than 0.10 and 31 genes with local false discovery rate less than 0.10. *Conclusion*: We can control and estimate false discovery rate simultaneously in multiple comparisons of high-dimensional data. It can provide more information for analysis to combine false discovery rate, q value and local false discovery rate.

【Key words】microarray data; multiple hypothesis testing; false discovery rate; q value; local false discovery rate; prostate cancer

以微阵列技术(Microarray)和功能磁共振成像(Functional magnetic resonance imaging,fMRI)为代表的现代生物技术给医学科学研究带来了革命性的影响,给人们提供了一种从没有过的医学实践方式[1,2]。由此产生的大规模数据也给统计分析带来了前所未有的机遇和挑战,研究者常常需要同时进行上千次的多重假设检验,错误发现率(False discovery rate,FDR)为大规模数据多重检验中的错误控制提供了十分合适的测量指标,受到越来越多的重视[3]。自 Benjamini 和 Hochberg 于 1995 年的开创性论文以来,关于错误发现率的理论研究和应用得到迅速发展[4],本文主要以微阵列数据为例介绍多重假设检验中错误发现率的控制和估计,以及在Benjamini和Hochberg 基础上的扩展及其应用。

1 方法和原理

1.1 微阵列数据和多重假设检验

微阵列数据一般表示为 $m \times n$ 的矩阵形式,m 表示基因数,n 表示生物样本,通常只有几个或者几十个。表 1 给出了一个关于前列腺癌的微阵列实验数据(Singh *et al.*,2002 年),包括 50 个正常对照和 52 个前列腺癌患者 6 033 个基因的表达水平,以 6 033×102 的矩阵排列,即 m=6 033,n=102。研究者一个主要目的是,哪些基因在正常者和前列腺癌患者中的表达水平不同?需要识别出这些基因以进一步分析。

表 1 前列腺癌微阵列数据形式

基因	正常对照			前列腺癌患者			t值	p值
编号	1	…	50	1	…	52		
gene1	-0.93	…	-1.09	-1.10	…	3.43	1.484	0.141 0
gene2	-0.84	…	-0.83	-0.83	…	-1.17	3.670	0.000 4
…	…	…	…	…	…	…	…	…
gene6 032	-0.90	…	1.68	-0.06	…	-0.71	0.104	0.917 5
gene6 033	-0.25	…	0.44	-0.81	…	-0.16	-0.913	0.363 2

设 m 次多重检验的零假设为 $H_{0i}, i=1,\cdots,m$,对应的统计量和 p 值分别为 t_i, p_i,表 2 给出了 m 个假设检验可能出现的结果,H_0=0 表示零假设成立,即正常对照和前列腺癌患者的基因表达水平无差异,H_0=1 表示零假设不成立,即 2 组人群的基因表达水平有差异,m_0和 m_1分别表示在 m 次多重检验中真实 H_0和非真实 H_0的个数。V 表示在所有 R 次拒绝 H_0的决

作者介绍:王 婷(1984-),女,实习研究员,硕士,
研究方向:微阵列数据分析和贝叶斯统计研究。

基金项目:徐州医学院公共卫生学院课题资助项目。

定中实际上拒绝了原本真实的 H_0的次数，除 m、R 和 W 外，表 2 中的其它量均为没有观察到的随机变量。根据表 2，Ⅰ和Ⅱ型错误可分别表示为 $E(V)/m_0$和 $E(T)/m_1$，功效为 $1-E(T)/m_1$。总体错误率(Family wise error rate，FWER)是经典的多重检验错误控制指标，定义为 Prob(V>0)，即至少犯 1 次Ⅰ型错误的概率[6]。常用的 FWER 控制程序包括 Bonferroni、Holm、Hochberg 和 Hommel 等方法，但在大规模数据中总体错误率控制过于严格，随着检验次数 m 的增加其效能明显降低。

表 2　m 次多重检验的可能结果

	不拒绝 H_0	拒绝 H_0	合计
H_0=0	U	V	m_0
H_0=1	T	S	m_1
合计	W	R	m

1.2　错误发现率

Benjamini 和 Hochberg(1995 年)根据在 R 次拒绝中错误拒绝次数 V 的比例定义错误发现率为 $E(V/R)$。其主要的考虑在于，相对于所有 R 次拒绝，如果错误拒绝次数 V 的比例足够小，那么这种错误就是可以接受的。

设预先的错误发现率控制水平为 $\alpha\in(0,1)$，Benjamini 和 Hochberg(BH)给出了一个基于 p 值的逐步向下控制程序[3]：

①将 p 值排序：$p_{(1)}\leqslant p_{(2)}\leqslant\cdots\leqslant p_{(m)}$，$H_{0(1)},H_{0(2)},\cdots,H_{0(m)}$ 为对应的零假设；

②计算 k=argmax{$1\leqslant l\leqslant m: p_{(l)}\leqslant l\alpha/m$}；

③如果 k 存在，拒绝 $H_{0(1)},\cdots,H_{0(k-1)},H_{0(k)}$，否则不拒绝任何一个 H_0。

Benjamini 和 Hochberg 证明了在统计量独立和满足一般弱相关的条件下 BH 程序控制 FDR=$\pi_0\alpha\leqslant\alpha$，$\pi_0=m_0/m$ 为微阵列数据中无差别表达基因的比例。

总体错误率从表 2 的行“$m_0\rightarrow V$”考虑错误控制，而错误发现率从表 2 的列“$R\rightarrow V$”方向控制错误，这种看似简单的转换具有突破性的意义，给大规模数据的多重比较提供了十分恰当的错误测量指标。以 FDR=0.05 为例，其含义在于按照 BH 程序如一共拒绝 H_0 100 次(R=100)，那么平均而言大约有 5=100×0.05 次属于错误拒绝。错误发现率控制提高了多重检验的以效能，以 Bonferroni 为例，其检验水准为 α/m，BH 程序的标准比 Bonferroni 宽松 i(i=1,⋯,m)倍。

1.3　pFDR 和 q 值

对错误发现率的定义是所有 R 次拒绝中错误发现的期望比例 $E(V/R)$，由此导致的后果是，当 R=0 时 $E(V/R)$=0/0 不能定义错误发现率，因此 Benjamini 和 Hochberg 技术性地将错误发现率表达为 FDR=$E(V/R|R>0)P(R>0)$，并额外定义 R=0 时FDR=0。在微阵列分析中，对数以千计的基因同时假设检验往往会或多或少地拒绝 H_0，也即是几乎肯定有 $P(R>0)\approx1$，此时可将错误发现率表达为：

$$\mathrm{FDR}=E(V/R|R>0)P(R>0)\approx E(V/R|R>0)$$

Storey 将 $E(V/R|R>0)$ 称为阳性错误拒绝率(Positive false discovery rate，pFDR)[7]，称为阳性是基于至少拒绝 1 个 H_0的事实，也即是只关心在发生拒绝 H_0的事件之后的错误拒绝。设 Γ_α为事先选定的拒绝域，pFDR 可表示为[7]：

$$\mathrm{pFDR}=E\left(V(\Gamma_\alpha)/R(\Gamma_\alpha)|R>0\right)$$

$V(\Gamma_\alpha)=\#\{t_i\in\Gamma_\alpha|H_0=0\}$表示错误发现次数，$R(\Gamma_\alpha)=\#\{t_i\in\Gamma_\alpha\}$表示所有拒绝 H_0的次数。按照和 p 值类似的定义，Storey 定义 q 值为[7]：

$$q(t)=\mathrm{pFDR}(\Gamma_\alpha)=\inf_{\{\Gamma_\alpha:t\in\Gamma_\alpha\}}\mathrm{Prob}(H_0=0|t\in\Gamma_\alpha)$$

q 值量化了以 $\Gamma_\alpha=\{|T|\geqslant|t|\}$作为拒绝域时所犯的最小 pFDR。根据 p 值做出决定时，例如 $p_i\leqslant\gamma(0\leqslant\gamma\leqslant1)$，pFDR 为：

$$\mathrm{pFDR}(\gamma)=\frac{\pi_0\mathrm{Prob}(p\leqslant\gamma|H_0=0)}{\mathrm{Prob}(p\leqslant\gamma)}=\frac{\pi_0\gamma}{\mathrm{Prob}(p\leqslant\gamma)}$$

$\mathrm{Prob}(p\leqslant\gamma|H_0=0)=\gamma$ 是因为在 H_0=0 的条件下，p 服从 0-1的均匀分布。阳性错误拒绝率估计的算法如下：

① 设 λ 为一个合适选择的值，称为调整参数(Tuning parameter)，合适是指能够认为几乎所有在[λ,1]区域的 p_i都来自 H_0=0，如 λ=0.50，此时 $E(W)=m\pi_0(1-\lambda)$，用 $\#\{p_i>\lambda\}$估计 $E(W)$，则$\hat{\pi}_0(\lambda)=\#\{p_i>\lambda\}/(m-m\lambda)$。显然$\hat{\pi}_0$随 λ 而变化，Storey 给出了选择最优 λ 值的 bootstrap 方法[8]；当 π_0=1 或 λ=0 时 pFDR 估计和前文的 BH 控制等价；

② pFDR 和 FDR 的估计值为：

$$\widehat{\mathrm{pFDR}}(\gamma)=\frac{\hat{\pi}_0(\lambda)\gamma}{\#\{p_i\leqslant\gamma\}\left(1-(1-\gamma)^m\right)},\ \widehat{\mathrm{FDR}}(\gamma)=\frac{\hat{\pi}_0(\lambda)\gamma}{\#\{p_i\leqslant\gamma\}}$$

m 很大时$\left(1-(1-\gamma)^m\right)\rightarrow1$，则 pFDR≈FDR。

p_i值对应的 q 值为 $q(p_i)=\inf_{\gamma\geqslant p_i}\{\mathrm{pFDR}(\gamma)\}$，其计算过程为：

① 对 p_i值排序：$p_{(1)}\leqslant p_{(2)}\cdots\leqslant p_{(m)}$；

② 使得$\hat{q}(p_{(m)})=\mathrm{pFDR}(p_{(m)})$；

③ 计算$\hat{q}(p_{(i)})=\min\{\mathrm{pFDR}(p_{(i)}),\hat{q}(p_{(i+1)})\}$，$i=m-1,m-2,\cdots,1$。

第三步中 min 函数的作用在于使得 q 值具有嵌套的性质，即如果 $p_i\leqslant p_j$则有 $q_i\leqslant q_j$。

1.4　局部错误发现率

阳性错误拒绝率是定义在 1 个拒绝域如 Γ_α或[γ,1]上的错误拒绝率估计过程，但是在很多时候研究者可能关心根据单个统计值做出拒绝时的错误发现率。例如，如果根据 t_i=3 拒绝 H_0，那么此时的错误拒绝率为多少呢？大规模数据使得根据单个统计量估计对应的错误发现率变得可能。设 π_0为 H_0=0 的先验概率，具有和前文 $\pi_0=m_0/m$ 类似的含义，H_0=1 的先验概率为 $1-\pi_0$，密度函数分别为 f_0和 f_1，则 t 的边际密度为：

$$f(t)=\pi_0f_0(t)+(1-\pi_0)f_1(t)$$

根据贝叶斯定理定义：

$$\text{locfdr}(T=t)=p_0(H_0=0|T=t)$$
$$=\pi_0 f_0(t)/f(t)$$

locfdr 称为局部错误发现率（Local false discovery rate，locfdr）[9,10]。对局部错误发现率的完全贝叶斯分析需要进一步指定 π_0、f_0和 f_1的先验，但大规模数据为 π_0、f_0和 f_1的经验估计提供了便利，在微阵列数据中 π_0多在 0.9 以上，因此 1 可作为 π_0的保守估计值，f_0为 H_0=0 条件下的零分布（Null distribution），前列腺癌数据中 f_0为自由度 100 的 t 分布，因此实际上只需要估计 $f(t)$即可。不同的方法可用来估计边际密度 $f(t)$，如Grenander密度估计和 Poisson 回归方法[10,11]。为了方便，常常将 t 值转换为 z 值，其转换关系为 $z=\Phi^{-1}(T(t|\nu))$，$T(\cdot|\nu)$ 表示自由度为 ν 的 t 分布累积函数，Φ^{-1}表示正态分布累积函数，此时 f_0=N(0,1)。

2 实例应用

本文接下来通过前列腺癌微阵列数据展示上述的方法，所有统计分析在 R2.13.0 中完成[12]。图 1 展示了前列腺癌数据 z 值的直方图，其中实线为 z 值的边际密度曲线。

2.1 BH 程序控制

图 2 展示了 BH 控制过程，选择 α=0.10 得到 60 个基因，图中的直线过原点，斜率为 0.10/6 033，参考线以左的 p 值被认为有差别，此时对应的 p 值为 1.3×10^{-4}。α=0.10 的含义是，平均而言大约 3=60×0.05 个基因属于错误识别。如果选择 α=0.05 或 0.15，分别得到 21 个或 86 个表达差别基因，同理，α=0.05 或 0.15 的含义为，在 21(86)个差别基因中约有 1(13)个属于错误发现。

2.2 q 值

根据 bootstrap 抽样选择的最优 λ 值为 0.92，也即大于 0.92的 p_i被认为来自无差别表达的基因，$\#\{p_i>0.92\}$=391，对应 π_0的估计值为 0.810=391/(6 033-6 033×0.92)。图 3 展示了 q 值的直方图，q 值小于 0.10 的基因数为 74，比 0.10 水平的 BH 控制多识别出 14 个基因，当然如果取 π_0=1 则有 60 个 q 值小于 0.10，图4 给出了以不同 q 值为临界点的差别表达基因数，图 5 给出了差别基因数和错误识别的关系。

2.3 局部错误发现率

假设 $f_0(z)$=N(0,1)，采用 Efron 提出的 Poisson 回归方法估计 $f(z)$[10]，见图 1。例如，在 t=1.484 时 z=1.472，f_0(1.472)=0.135，$\hat{f}$(1.472)=0.134，则局部错误发现率的估计值为1.004 π_0，因此需要进一步估计 π_0。这里为了简单我们给出 π_0的一个保守估计方法，局部错误发现率中隐含着限制条件：locfdr=$p_0(z)\leqslant1$，即 $\pi_0\leqslant\min(f(z)/f_0(z))$，该式给出了关于 π_0估计值的一个上限，本文数据中 $\pi_0\leqslant$0.943。局部错误发现率的估计为 $0.943f_0(z)/f(z)$。图 6 给出了 locfdr 和 z 值的关系，小于 0.10 的有 31 个。表 3 给出了根据 BH 程序(α=0.05)识别出的 21 个基因的 z 值、q 值和 locfdr。

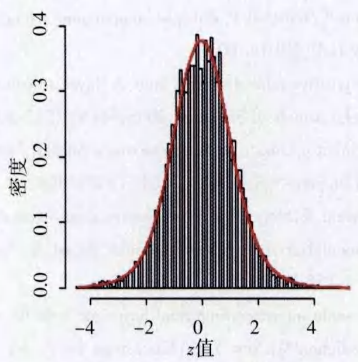


图 1 前列腺癌数据 z 值的直方图

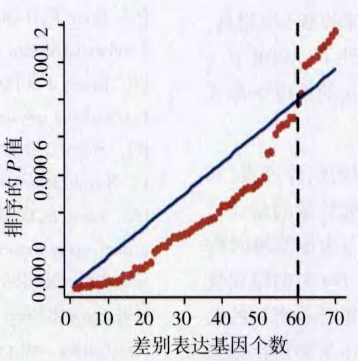


图 2 BH 程序控制过程

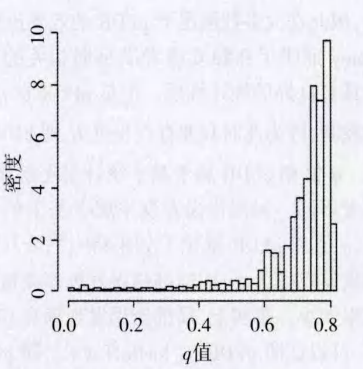


图 3 q 值的直方图

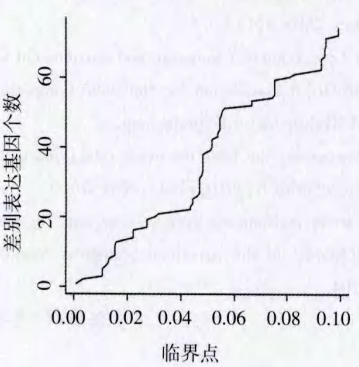


图 4 不同 q 值下的差别基因数

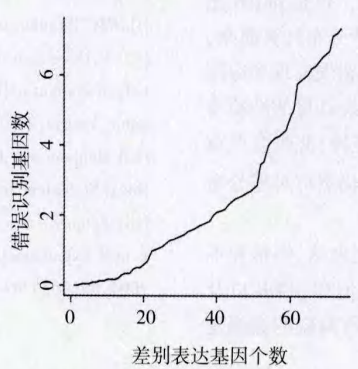


图 5 基因数和错误识别

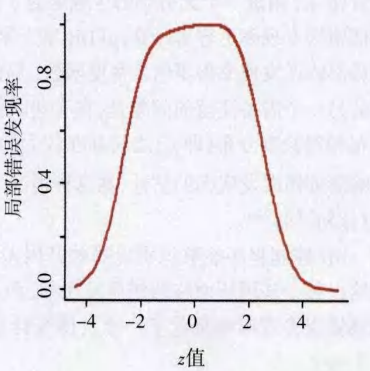


图 6 z 值和对应的 locfdr

表 3 最大的 21 个 |z| 值所对应的 q 值和 locfdr

z值	q值	locfdr	z值	q值	locfdr	z值	q值	locfdr
5.292 0	0.000 6	0.000 6	4.247 0	0.013 2	0.020 3	3.978 7	0.022 0	0.041 5
4.834 4	0.003 3	0.003 3	4.187 5	0.015 2	0.024 0	3.969 5	0.022 0	0.042 4
4.473 5	0.010 7	0.010 5	4.135 4	0.015 2	0.027 6	3.875 4	0.027 4	0.053 5
−4.423 9	0.010 7	0.012 3	−4.144 7	0.015 2	0.027 7	−3.884 6	0.027 4	0.055 0
4.398 4	0.010 7	0.013 1	4.123 6	0.015 2	0.028 5	−3.877 8	0.027 4	0.055 9
−4.334 4	0.011 9	0.016 1	4.099 8	0.015 5	0.030 3	−3.839 8	0.030 1	0.061 4
−4.287 5	0.012 6	0.018 5	3.979 1	0.022 0	0.041 4	−3.815 6	0.031 6	0.065 2

3 讨 论

传统多重检验的总体错误率控制过于严格，在 m（基因数）很大时效率很低，因此不适合微阵列数据的多重比较。作为一种探索性的过程，微阵列实验分析中研究者更关心的是尽可能多地筛选出候选基因，只要将错误识别的比例控制在允许的范围内。由 Benjamini 和 Hochberg 提出的错误发现率为这一分析目的提供了恰当的平衡，他们 1995 年的论文十分流行的另一个重要的原因在于，BH 控制程序简单实用。在独立的假设下，BH 控制程序保证了 $FDR=\pi_0\alpha\leqslant\alpha$，事实上即使是非独立的情形时，BH 程序仍然有效。BH 程序中选择 $\pi_0=1$，但是显然已知 π_0 的信息可进一步提高检验功效[13]。

在 Benjamini 和 Hochberg 开创性的工作基础上，后来的研究者提出了不同的错误发现率控制程序和估计过程，q 值和 pFDR 即是在此背景下提出的，由于 pFDR 考虑了 π_0 的信息，因此在大多数情况下 pFDR 的效率比单纯的 BH 控制高。Storey 证明了在独立或者满足弱相关的条件下，pFDR 和 q 值具有良好的统计性质。但是 pFDR 在 $m=m_0$ 的情况下却无法控制，因为此时只要存在拒绝 H_0 则 FDR=1。

q 值和 pFDR 属于基于统计量尾部面积的估计过程，和前者不同，局部错误发现率属于基于单个统计量的估计过程，q 值和 pFDR 量化了如果将 $\{|T|\geqslant|t|\}$ 作为拒绝域时的错误发现率大小，而局部错误发现率测量了 $T=t$ 时的错误发现率大小。事实上，局部错误发现率和 pFDR 有着密切的关系，可以证明 $pFDR=E[locfdr|T\geqslant|t|]$，即 pFDR 是局部错误发现率在 $\{|T|\geqslant|t|\}$ 条件下的期望值，或者可将局部错误发现率看作在 t 附近一个无穷小的区域内的 q 值。但是 pFDR 比局部错误发现率更容易计算，pFDR 基于累计分布计算而来，而局部错误发现率需要估计密度函数。局部错误发现率应用中的另一个需要注意的问题是，在有些时候统计量的经验零分布和理论零分布（即 f_0）之间存在较大的差异，此时会严重影响局部错误发现率的估计，在这种情况下仍然可对零分布进行经验估计[14]。

BH 控制程序能够识别出哪些基因差别表达，但是并不能给出每个基因所面临的错误发现率，而 q 值和 pFDR 以及局部错误发现率都量化了一个具体统计量所对应的错误发现率，因此在实际中推荐三者的联合应用，采用 BH 控制程序识别出差别表达基因后同时给出每个基因对应的 q 值和局部错误发现率，以观察该基因对 H_0 的支持程度。

参 考 文 献

[1] Schena M.Microarray analysis[M].New York:John Wiley & Sons, 2003.1-10.

[2] Lazar N.The statistical analysis of functional MRI data[M].New York:Springer,2008.165-210.

[3] Benjamini Y,Hochberg Y.Controlling the false discovery rate:A practical and powerful approach to multiple testing[J].Journal of the Royal Statistical Society:Series B,1995,57(1):289-300.

[4] Dudoit S,Shaffer J P,Boldrick J C.Multiple hypothesis testing in microarray experiments[J].Statistical Science,2003,18(1):71-103.

[5] Singh D,Febbo P G,Ross K,et al.Gene expression correlates of clinical prostate cancer behavior[J].Cancer Cell,2002,1(2):203-209.

[6] Bretz F,Hothorn T,Westfall P.Multiple comparisons using R[M].London:Chapman & Hall,2010.1-60.

[7] Storey J D.The positive false discovery rate:A Bayesian interpretation and the q-value[J].Annals of Statistics,2003,31(6):2013-2035.

[8] Storey J D.A direct approach to false discovery rates[J].Journal of the Royal Statistical Society:Series B,2002,64(3):479-498.

[9] Efron B,Tibshirani R,Storey J D,et al.Empirical bayes analysis of a microarray experiment[J].Journal of the American Statistical Association,2001,96(456):1151-1160.

[10] Efron B.Large-scale inference:Empirical bayes methods for estimation,testing,and prediction[M].New York:Cambridge University Press, 2010.46-69.

[11] Strimmer K.A unified approach to false discovery rate estimation [J].BMC Bioinformatics,2008,3(1):1-14.

[12] R Development Core Team.R:A language and environment for statistical computing[EB/OL].R Foundation for Statistical Computing,Vienna,Austria,2007.URL http://www.R-project.org.

[13] Benjamini Y.Discovering the false discovery rate[J].Journal of the Royal Statistical Society:Series B,2010,72(4):405-416.

[14] Efron B.Large-scale simultaneous hypothesis testing:The choice of a null hypothesis[J].Journal of the American Statistical Association, 2004,99(465):96-104.

（责任编辑：冉明会）