·综述·

# 多重假设检验中 FDR 的控制与估计方法*

哈尔滨医科大学卫生统计学教研室(150081) 刘 晋 张 涛 李 康△

近年来,基因组学、蛋白组学和代谢组学等高通量检测技术得到迅速发展[1-4],由此产生变量数目巨大的数据(如 $m>2\ 000$),而样品数目较小(如 $10\leqslant n\leqslant 100$),用传统的统计检验方法对生物标志物进行鉴别会产生大量的假阳性结果(如检验水准取 $\alpha=0.05$ 或 $\alpha=0.01$ 等),存在阳性发现错误率(false discovery rate,FDR)问题。对于多重检验,若规定检验水准为 $\alpha$,则对于 $m$ 次检验,至少犯一次假阳性错误的概率为 $\alpha_m=1-(1-\alpha)^m$,当 $m$ 增加时,$\alpha_m$ 趋于 1。多重检验阳性发现错误率(FDR)的控制及估计方法对分析高维数据具有重要的意义,譬如研究中选择了 50 个潜在的生物标志物,如果能够估计出其中有多少具有研究价值的标志物其意义不言而喻。

理论和实际应用表明:传统的多重检验方法,如 Bonferroni 法、sidak 法等不能有效地解决高维微阵列海量数据的多重检验问题,从而使这一问题成为近年国外统计学方法研究中的热点之一[5]。研究主要集中在两个方面:一是对阳性发现错误率的控制,二是对阳性发现错误率的估计。本文主要在这两个方面介绍其统计学方法的研究进展。

## FWER 控制面临的问题

在多重检验中,需要对整体的错误率进行控制,目前广泛使用的错误测度指标是族错误率(family wise error rate,FWER),其他的一些错误测度还包括 $K$ 族错误率($K$ family wise error rate,KFWER)[6]、平均比较错误率(per-comparison error rate,PCER)[6]、平均族错误率(per-family error rate,PFER)[6]等。为说明这些指标的意义,给出表 1。

表 1 多重假设检验四种结果的频数

| 真实情况 | 不拒绝 $H_0$ | 拒绝 $H_0$ | 合计 |
|---|---|---|---|
| $H_0$ 为真 | $U$ | $V$ | $m_0$ |
| $H_1$ 为真 | $T$ | $S$ | $m_1$ |
| 合计 | $W$ | $R$ | $m$ |

FWER 定义为拒绝真实无效假设的个数大于等于 1 的概率(记作 $P(V\geqslant 1)$),对此通常使用 Bonferroni 法对其进行检验,即对每一个假设都在显著性水平 $\alpha/m$ 下进行检验,保证 $\mathrm{FWER}=P(V\geqslant 1)$ 小于或等于事先给定的 $\alpha$ 检验水准。这种传统方法的主要问题是:$V>0$ 这一条件过于严格地控制了假阳性结果,使得多重检验效能降低,同时 FWER 的实际意义也不够直观和容易理解。KFWER 定义为 $P(V>K)$ 的概率,即为拒绝真实无效假设的个数大于等于 $K$ 的概率,在一定程度上克服了传统检验方法的缺点。实际中,更多需要的是估计多重检验为阳性结果时,其中可能包含有多少假阳性结果。

## FDR 的提出与定义

1995 年 Benjamini 和 Hochberg 首次提出了 FDR 的概念,并给出了在多重检验中对它的控制方法(简称 BH 方法)[7]。然而,当时组学海量数据尚未大量出现,开始并未受到重视,甚至因为考虑了 64 个假设检验而受到质疑[7]。数年之后,伴随着微阵列检测技术的发展、海量数据的大量出现使得 FDR 有了应用。目前为止,Benjamini 和 Hochberg 的文章引用次数已经达到上万次,FDR 的理论和应用研究也在不断走向成熟。FDR(false discovery rate)的定义如下:

$$FDR=\begin{cases}E(V/R), & R\neq 0\\ 0, & R=0\end{cases} \quad (1)$$

其中 $E(\cdot)$ 为数学期望。同理,我们可以得到假阴性发现率(false negative discovery rate,FNDR)的定义:

$$FNDR=\begin{cases}E(T/W), & W\neq 0\\ 0, & W=0\end{cases} \quad (2)$$

FDR 的含义是阳性检验结果中判断错误的比例。FDR 具有以下优点:①可以灵活调整其取值,作为假设检验错误率的控制指标,其控制值可以根据需要灵活选取,而传统的假设检验(FWER)的取值则较为固定,通常定为 0.05;②FDR 的意义明确,可以作为筛选出的差异变量的评价指标,而 FWER 则主要是用来控制Ⅰ类错误的。FDR 与 FWER 两者的关系:当所有无效假设为真时,控制 FDR 和控制 FWER 等价;当 $m_0<m$ 时($m_0$ 为真实无差异变量的数目),控制 FDR 相当于弱控制 FWER。

## 多重检验的 FDR 的控制方法

* 国家自然科学基金资助(30872185)

△通讯作者:李康,E-mail:liking@ems.hrbmu.edu.cn

控制是指决定一个显著性水平的界值，从而使FDR被限制在某一固定水平，类似于FWER的控制，对此可以采用线性向上的控制方法，分两步进行：首先将所有检验的 $p$ 值进行排序，即 $p_{(1)} \leqslant p_{(2)} \leqslant p_{(3)} \leqslant \cdots \leqslant p_{(m)}$；然后逐步后退比较 $p_{(i)} \leqslant \frac{i}{m} q(i = m, m-1, m-2, \cdots, 1)$，取第一个满足条件的 $p_{(k)}(k \geqslant 1)$，理论上可以证明在此情况下可以将FDR控制在 $q(0 \leqslant q \leqslant 1)$ 水平下。

上述方法需要满足各变量假设检验间是独立的条件。在此基础上，1999年Yekutieli和Benjamini[8]给出了一种改进的方法，但其估计的FDR值略为保守，其思想是利用重复抽样的方法来计算 $p$ 值，可以在变量相关条件下控制FDR值。同年Benjamini和Liu则提出了一种逐步向下(step-down)的控制方法，过程与BH基础方法相近，只是对 $p_{(k)}$ 的控制方法不同[9]。2000年Benjamini和Hochberg提出了两阶段的FDR控制，以改进原有方法的保守性[10]；2001年Benjamini和Yekutieli对算法进行了进一步的改进[11]，可用于不同变量检验间独立和相关不同条件下的FDR的控制，不足之处是其检验效能较低。Benjamini和Hochberg在2005年提出了一种自适应线性向上的控制方法(adaptive linear step-up，ALSU)，这种方法的特点是在不同的显著水准下两次使用上述介绍的基础过程，特别是在变量相关条件下得到的FDR估计较为稳健[12]。

## 多重检验的FDR估计方法

FDR估计是指在某一“有统计学意义”阈值情况下所有选入的有差异变量的假发现率估计值。对此，可以从贝叶斯角度给出解释，近期对于FDR的研究多数在此框架下进行[13]。贝叶斯定义使用两成分模型来构建 $p$ 值的分布函数，其中 $p_0$ 为真实无效假设所占总检验次数的比例，$F_0(p)$ 为无效假设下 $p$ 的分布函数；$p_1$ 为实际有差异变量在所有变量中所占的比例，$F_1(p)$ 为备则假设成立下 $p$ 值的分布函数，即

$$F(p) = p_0 F_0(p) + p_1 F_1(p) \tag{3}$$

根据贝叶斯公式，可得

$$FDR = \Pr\{\text{null} \mid p\} = \frac{p_0 F_0(p)}{F(p)} \tag{4}$$

若用概率密度函数代替分布函数，则可以得到局部FDR(记作 $fdr$)

$$f(p) = p_0 f_0(p) + p_1 f_1(p) \tag{5}$$

$$far = \Pr\{\text{null} \mid p\} = \frac{p_0 f_0(p)}{f(p)} \tag{6}$$

其定义为在某 $p$ 值时，若认为差异有统计学意义，结果为假阳性的概率。相对于FDR使用尾侧概率的定义，局部FDR更适合从贝叶斯角度进行解释。局部FDR的提出，使得我们能够估计出任意一次检验为假阳性的概率，通常情况下有 $FDR \leqslant fdr$。

目前，FDR估计方法已经有数十种，以下介绍几种有代表性的FDR估计方法。

1. $q$-value法

该方法由Storey(2003)提出[14]，其基本思想是在最初的FDR控制过程基础上，估计了无差异变量占总变量个数的比例 $p_0$，以提高原有方法的检验效能。具体方法如下：

(1) 假设 $p$ 值服从0-1间的连续分布，有

$$\#\{p > \lambda\} = m_0(1-\lambda) - m_0 p_0(1-\lambda)$$

$$p_0 = \frac{\#\{p > \lambda\}}{m(1-\lambda)} \tag{7}$$

其中“#”表示满足括号中条件的变量个数。选择 $\lambda(0 < \lambda < 1)$ 需要在对 $p_0$ 的无偏估计和变异度之间取得平衡，对此可以使用boostrap抽样方法计算一个合理的取值。

(2) 假设在无差异情况下 $p$ 值服从均匀分布，则有 $F_0(p) = p$；采用经验估计 $F(p)$，则 $F(p) = order(p_i)/m = i/m$。

(3) 实际中 $p_0 \leqslant 1$，若保守估计将 $p_0 = 1$ 代入Bayes公式，则得 $FDR = P_i m/i$，结果与前述BH基础方法等价，其算法实际上是在贝叶斯定义下的非参数保守估计。

(4) 给出FDR的估计，称其为 $q$ 值，$q(p_i) = \min\limits_{t \geqslant p_i}(FDR(t))$

2. SAM法

目前广泛使用的FDR估计过程是Tusher等(2001)提出的SAM(significance analysis of microarray)方法[15]。这种方法不使用 $p$ 统计量，而是使用类似于 $z$ 的 $d$ 统计量，在无效假定下 $d$ 统计量值分布在0附近。$p_0 = \#\{d_i \in (P_{25}, P_{75})\}/(0.5m)$，其中 $(P_{25}, P_{75})$ 为四分位间距，在一定条件下SAM和 $q$-value法对 $p_0$ 的估计等价($\lambda = 0.5$)。我们可以根据 $d$ 统计量及其分布计算两端的界值 $[\Delta_{low}, \Delta_{up}]$，$d$ 统计量在无效假设下的分布通过permutation方法进行估计，FDR由下式计算：

$$\hat{FDR}(\Delta_{up}, \Delta_{low}) = \hat{\pi} \frac{\hat{F}_0(\Delta_{low}) + 1 - \hat{F}_0(\Delta_{up})}{\hat{F}_0(\Delta_{low}) + 1 - \hat{F}_0(\Delta_{up})} \tag{8}$$

3. 经验贝叶斯方法

Efron注意到在FDR的估计方法中，许多方法假定无效假设的 $p$ 值服从均匀分布，这在“理想情况下”是成立的。然而，由于实际数据的复杂性，通常不能满足所要求的条件(如 $t$ 检验要求的方差齐性，正态性)，从而导致在无效假设下 $p$ 值不再服从均匀分布，使FDR的估计出现偏差，而即使利用permutation方法也并非总能得到无效假设下的真实分布[16]，由此得到的

FDR 也不一定准确。许多估计方法，如 $q$-value、BUM、kerfdr 等方法皆依赖于均匀分布的假设，SAM 则是基于 permutation 抽样的 FDR 估计方法。Efron 提出采用经验贝叶斯法直接对无效假设下检验统计量的分布进行估计，如对 $z$ 统计量的均数和方差进行经验估计，而不用标准正态分布进行检验，重新计算 $p$ 值。有研究者采用这种方法对乳腺癌和艾滋病数据进行了分析[16-18]，对于这两个数据，无效假设的理论 $z$ 值分布明显不合理，而经验分布则能够很好地拟合实际数据。下面是三种经验贝叶斯方法。

(1) locfdr 算法[19]　与许多方法直接对 $p$ 值进行建模不同，这种方法直接对 $z$ 统计量进行建模。如可以采用 Poisson 广义线性模型对 $f(z)$ 进行估计，这里假设在无效假设下 $z$ 统计量服从均值为 $\mu$、标准差为 $\sigma$ 的正态分布，并用几何图形法和数值分析法对 $f_0(z)$ 进行估计。几何分析法假定在 $z$ 接近于 0 时，无效假设下 $z$ 的分布近似于实际数据分布，据此估计出 $z$ 的均值和方差。数值分析法则规定在某一区域内的 $z$ 值为无效假设的拒绝域，采用极大似然法估计，在估计无效假设分布的同时，也估计出 $p_0$。数值分析法通常得到更为稳健的结果，但比几何法有更大的偏差。

(2) fdrtool 算法　Korbinian Strimmer(2007) 结合经验贝叶斯估计和非参估计方法(简称 GRENDER 法)，提出了 fdrtool 方法[20]。这种方法可以利用各种不同统计量的值对 FDR 进行估计，该方法具有两个优点：①可以同时估计分布函数和概率密度函数；②可以确保估计的 FDR 值和 $p$ 值具有一致的单调性，同时不需要广义线性模型(GLM)过程需要设定自由度等参数。模拟实验结果显示，在各变量假设检验独立的条件下，这种方法能够准确地估计无差异变量所占的比例。实际数据分析表明，它与 locfdr 方法具有相似的特性，而在稳定性上优于 $q$-value 和另外一种 nFDR 方法。

(3) PRML 算法　Ryan Martin(2010) 提出了一种半参数的经验贝叶斯估计方法[21]，这种方法在无效假设下计算出的统计量分布采用经验贝叶斯法进行估计，并对备择假设下统计量的分布函数也进行了定义，并采用 PRML(predictive recursion marginal likelihood) 估计方法对其分布函数进行估计。模拟实验和实际数据分析结果显示：一般情况，该方法与其他方法得到的结果相近，但对一些复杂结构数据，在其他方法估计参数出现明显不合理的情况时(如出现 $p_0 > 1$)，该方法仍能够获得稳健的参数估计结果。

## FDR 几种方法的比较

对于 FDR 的控制与估计，使用者主要关心其估计的无偏性、稳定性、检验效能和不同数据结构(如相关性)对 FDR 控制或估计的影响。模拟实验证实，在大部分理想情况下 FDR 的各类控制方法能够将错误控制在指定水平下(BH 控制方法偏于保守)，并且是无偏估计。在变量间存在“弱相关”条件下，SAM、$q$-value 等大部分方法对 FDR 的估计依然是稳健的，在变量存在简单正相关条件下，BH 基础算法依然保持变量独立条件下的性质。但是在“任意相关”条件下，模拟实验证实 SAM 和 $q$-value 等算法对 $p_0$ 和 FDR 的估计将产生较大的变异和偏性，而此时自适应 BH 法则显现出较好的稳定性[22]。经验贝叶斯法在理论上对于无效假设能够进行更好的拟合，解决任意相关对 FDR 估计的影响，但对实际数据分析发现这种方法有时并不能得到理想的结果。如在对一项乳腺癌与正常对照数据的分析中，将 FDR 控制在一定水平，使用 locfdr 算法未发现有差异基因，而使用简单的 BH 算法却发现了 107 个“差异基因”[16]。需要注意的是，许多方法如 $q$-value、locfdr、BUM、kerfdr 等用于实际复杂数据分析时，可能由于无法满足适用条件使 $p_0$ 和 FDR 的估计明显超出合理范围[21]。

## FDR 计算软件

目前，有多种计算 FDR 的软件可供研究者使用，其中大部分是基于 R 语言编写的软件包。一些常用的算法如 $q$-value、SAM、BH 等都可以方便地供研究者调用，使用时通常只需要给出假设检验的 $p$ 值或统计量值(如 $z$ 统计量)，就可以直接计算出 FDR。表 2 列出了常用的 FDR 控制与估计计算的软件包。

表 2　常用 FDR 控制估计软件包(R 语言)

| 软件包名称 | FDR 类型 | 统计量 | 使用方法 |
|---|---|---|---|
| fdrtool | fdr Fdr | $p, z, r, t$ | 调整的 GRENANDER 函数估计方法，对无效分布采用截断的最大似然估计[20] |
| mixfdr | fdr Fdr | $z$ | 正态混合模型的密度函数估计[23] |
| BUM | fdr Fdr | $p$ | 对于参数模型的最大似然估计[24] |
| SAGx | fdr Fdr | $p$ | Grenander 密度估计[25] |
| qvalue | Fdr | $p$ | 最优截断点的最大似然估计[14] |
| nFDR | Fdr | $p$ | Bernstein 多项式概率函数估计[26] |
| multtest | Fdr | $p$ | BH 算法[7] |
| LBE | Fdr | $p$ | 基于位置的估计[27] |
| locfdr | local fdr | $z$ | 泊松回归的概率函数估计，对经验贝叶斯模型采用截断的最大似然估计[19] |
| nomi | local fdr | $z$ | 正态混合模型[28] |
| LocalFDR | local fdr | $p$ | 局部加权回归(LOESS)平滑[29] |
| kerfdr | local fdr | $p$ | 核密度估计[30] |
| twilight | local fdr | $p$ | KS 法选择截断点[31] |
| localFDR | local fdr | $p$ | stochastic order 模型[32] |

## 目前存在的主要问题

1. 无效假设的选择

在 FDR 估计中，对无效假设下统计量的分布 $f_0$ 的估计十分重要，由于许多 FDR 控制和估计方法是基于 $p$ 值计算的，因此 $p$ 值的准确性最终影响 FDR 估计的准确性。由于 $p$ 值的计算依赖于一定的假设，而实际

数据通常无法满足这些假设，从而无法得到准确的 $p$ 值。为此，有些方法直接使用样本统计量的经验分布对 $p$ 值进行估计，如 SAM 法直接使用 $d$ 统计量，通过重复抽样得到无效假设的统计量分布。另外，经验贝叶斯法也在一定程度上解决了估计 $p$ 值无效假设分布的可靠性问题。然而，并非在所有数据情况下这些方法都能得到理想的结果，如何选择合理的无效假设，构建更合适的统计量或得到准确的 $p$ 值，需要进一步的研究。

2. 变量相关问题

组学问题如基因组研究中，许多基因位点和微阵列变量通常具有一定相关性，对于目前大多数方法，在“弱相关”，或“简单正相关”情况下，能够进行合理的 FDR 估计[22]，然而还没有在任意相关条件下的合理估计方法。在实际数据的处理中，如果变量高度相关，将会导致 FDR 的估计的变异性增大，同时对 $p_0$ 的估计将非常不稳定。模拟实验表明，SAM、$q$-value 等方法在强相关下计算出的 FDR 具有较大变异和一定的偏性；另外，使用 permutation 方法得到的统计量分布也不一定准确，从而无法正确地估计 $p$ 值和 FDR。这一问题也有待于进一步研究。

## 参 考 文 献


1. Dudoit S, van der Laan MJ. Multiple testing procedures with applications to genomics. New York: Springer, 2008.
2. Juliane Schäfer, Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics, 2005, 21(6): 754-764.
3. Dettmer K, Almstetter MF, Appel IJ, et al. Comparison of serum versus plasma collection in gas chromatography-mass-spectrometry-based metabolomics. Electrophoresis, 2010, 31(14): 2365-2373.
4. Debashis Ghosh. Assessing signifcance of peptide spectrum matches in pro-teomics: a multiple testing approach. Statistics in Biosciences, 2009, 1(2): 199-213.
5. Yoav Benjamini. Discovering the false discovery rate. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010, 72(4): 405-416.
6. 刘乐平，张龙，蔡正高假设检验及其在计量经济学中的应用. 统计研究，2007，24(4)：26-30.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B, 1995, 57(1): 289-300.
8. Yekutieli D, Benjamini Y. Resampling based False Discovery Rate controlling procedure for dependent test statistics. J. Statist. Planng Inf, 1999, 82(1-2): 171-196.
9. Benjamini Y, Liu W. A step-down multiple testing procedure that controls the false discovery rate under independence. J. Statist. Plann. inference, 1999, 82: 163-170.
10. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Statist, 2000, 25(1): 60-83.
11. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Statist, 2001, 29(4): 1165-1188.
12. Benjamini, Yoav, Abba M. Krieger et al. Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 2006, 93(3): 491-507.
13. Alessio Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat. Methods. Med. Res, 2008, 17(4): 347-388.
14. Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat, 2003, 31(6): 2013-2035.
15. I R, . Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 2001, 98(9): 5116-5121.
16. Efron B. Microarrays, Empirical Bayes, and the Two-Groups Model. statist Sci, 2008, 23(1): 1-22.
17. Hedenfalk I, Duggen D, Chen Y, et al. Gene expression profles in hereditary breast cancer. New Engl Jour. Medicine, 2001, 344(8): 539-548.
18. Van't Wout A, Lehrma G, et al. Cellular gene expression upon human immunodefciency virus type 1 infection of CD $ + T-Cell lines, 2003, 77(2): 1392-1402.
19. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Amer Statist Assoc, 2004, 99(465): 96-104.
20. Korbinian Strimmer. A unified approach to false discovery rate estimation. BMC bioinformati cs, 2008, 9: 303.
21. Ryan Martin. A nonparametric empirical Bayes framework for large-scale signifcance testing . http://www.stat.duke.edu/~st118/Publication/MT-test.pdf.
22. Kim, Kyung In. False Discovery Rate Procedures for High-Dimensional Data, 2008.
23. Muralidharanan O. An empirical Bayes mixture method for effect size and false discovery rate estimation, 2010, 4(1): 422-438.
24. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p values. Bioinformatics, 2003, 19(10): 1236-1242.
25. Langaas M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J R Statist Soc B, 2005, 67(4): 565-572.
26. Guan Z, Wu B, Zhao H. Nonparametric estimator of false discovery rate based on Bernstein polynomials. Statistica Sinica, 2008, 18: 905-923.
27. Dalmasso C, Bröet P, Moreau T. A simple procedure for estimating the false discovery rate. Bioinformatics, 2005, 21(5): 660-668.
28. McLachlan GJ, Bean RW, Jones LBT. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics, 2006(13): 1608-1615.
29. Determination of the differentially expressed genes in microarray experiments using local FDR. BMC Bioinformatics, 2004, 5: 125.
30. Robin S, Bar-Hen A, Daudin JJ, Pierre L. A semi-parametric approach for mixture models: application to local false discovery rate estimation. Comput Statist Data Analysis, 2007, 51(12): 5483-5493.
31. Scheid S, Spang R. A stochastic downhill search algorithm for estimating the local false disovery rate. IEEE T Comp Biol Bioinf, 2004, 1(3): 98-108.
32. Liao JG, Lin Y, Selvanayagam ZR, Shih WJ. A mixture model for estimating the local false discovery rate in DNA microarray analysis. Bioinformatics, 2004, 20(16): 2694-2701.