Advances in Applied Mathematics 应用数学进展, 2021, 10(10), 3532-3538
Published Online October 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1010372

Hans汉斯

多重检验技术在大数据分析中的应用

杜　欢，刘瑞银*，周志慧

沈阳师范大学数学与系统科学学院，辽宁　沈阳

收稿日期：2021年9月25日；录用日期：2021年10月18日；发布日期：2021年10月27日

摘　要

在对大数据进行假设检验时，为了控制假阳性，需要采用多重检验技术。多重检验技术有多种，本文通过对大数据进行实际分析，比较各种算法的优缺点，给出不同方法的适用场合，从而对数据分析人员给以理论上的指导。文章首先阐述多重检验的必要性以及多重检验的相关概念；其次分别介绍对总体错误率和错误发现率进行控制的两类方法；最后将这几种多重检验方法应用到基因大数据中对基因的表达与否进行判断。实验结果表明，控制错误发现率的方法优于控制总体错误率的方法，在控制错误发现率的方法中，*q*值法的结果最好。原因在于*q*值法考虑了原假设的先验信息，能很好地控制错误发现率的大小，因此具有较高的精确性和检验功效。

关键词

大数据，多重假设检验，总体错误率，错误发现率，*q*值

Application of Multiple Test Techniques in Big Data Analysis

Huan Du, Ruiyin Liu*, Zhihui Zhou

School of Mathematics and Systems Science, Shenyang Normal University, Shenyang Liaoning

Received: Sep. 25th, 2021; accepted: Oct. 18th, 2021; published: Oct. 27th, 2021

Abstract

In the hypothesis test of big data, in order to control false positives, multiple test technology needs to be used. There are many kinds of multiple test techniques. This paper makes a practical analysis of big data, compares the advantages and disadvantages of various algorithms, and gives the

*通讯作者。

文章引用: 杜欢, 刘瑞银, 周志慧. 多重检验技术在大数据分析中的应用[J]. 应用数学进展, 2021, 10(10): 3532-3538.
DOI: 10.12677/aam.2021.1010372

application occasions of different methods, so as to give theoretical guidance to data analysts. Firstly, this paper expounds the necessity and the related concepts of multiple testing; Secondly, two kinds of methods to control the family-wise error rate and false discovery rate are introduced respectively; Finally, these multiple test methods are applied to gene big data to judge whether the genes are expressed or not. The experimental results show that the method of controlling the false discovery rate is better than the method of controlling the family-wise error rate. Among the methods of controlling the false discovery rate, the *q*-value method has the best result. The reason is that the *q*-value method considers the prior information of the original hypothesis and can well control the false discovery rate, so it has high accuracy and power.

Keywords

Big Data, Multiple Hypothesis Testing, Family-Wise Error Rate, False Discovery Rate, *q*-Value

1. 引言

在当今大数据时代，对于数据的搜集和获得不再是难题，难题出现在对数据的分析上。大数据的出现不仅仅是数据量的增多，它也使得数据分析方法发生了根本的变化，传统的统计分析方法不再适用，统计从业人员需要谨慎的分析数据，并利用合适的方法去分析，以免出现决策性的失误。多重假设检验就是相应于经典的单个假设检验提出的一种适应于大数据的假设检验方法。例如，对基因的选择问题，需要根据基因表达数据对大量的基因进行检测从而筛选出需要的基因。这就需要对大量基因进行检验，即多重检验问题。

在 Lobenhofer *et al.* (2003) [1]中，作者利用 ORIOGEN 算法[2]对表达的基因进行了挑选。本文利用多重检验技术，对 Lobenhofer *et al.* (2003)中的基因重新进行了挑选，结果显示，在控制错误率的前提下，我们提高了检验的功效。

2. 多重假设检验概述

在 N-P 假设检验中，单个假设检验问题要求犯第一类错误的概率在可接受的范围内时，使犯第二类错误的概率控制在最小。在同时对多个假设进行检验时，对每个单独的假设检验而言第一类错误都在控制范围内，但对于整体而言犯第一类错误的概率将随着检验次数的增多而增大[3]，超出可控范围。例如，取显著性水平 $\alpha=0.05$，当进行两次独立检验时，对总体而言犯第一类错误的概率为 $1-(1-\alpha)^2$；进行 m 次检验时总体犯第一类错误的概率为 $1-(1-\alpha)^m$，随着检验次数的增加总体错误率增大，如图 1，当假设检验的次数超过 100 时，总体错误率接近于 1。因此，在多重假设检验问题中，不能像单个假设检验一样控制第一类错误。

多重假设检验方法有很多种，它们都需要控制总体错误率(FWER)或错误发现率(FDR) [4]，才能使得第一类错误整体控制在 α 水平内。考虑 m 个假设检验 H_i, $i=1,2,\cdots,m$，当原假设 H_0 为真时，记为 $H_i=0$，否则 $H_i=1$。设 m_0 和 m_1 分别表示 m 个假设检验中 H_0 和 H_1 为真的个数。对 m 个假设检验结果的分类如表 1 所示。

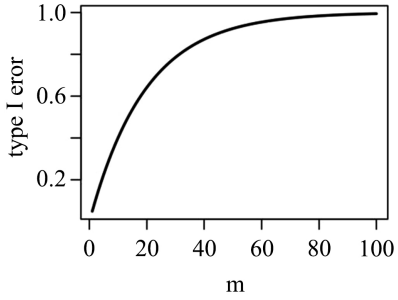

Figure 1. Relationship between multiple tests on times m and type I error
图 1. 多重检验次数 m 与 I 型错误的关系

Table 1. Possible results of m multiple tests
表 1. m 次多重检验可能出现的结果

假设检验	接受 H_0	拒绝 H_0	合计
H_0 为真	U	V	m_0
H_0 非真	T	S	m_1
	m-R	R	m

总体错误率表示在 m 次检验中，至少出现一次错误拒绝 H_0 的概率，即 $\mathrm{FWER}=\mathrm{Prob}(V\geq 1)$ 且 $\mathrm{FWER}\leq\alpha$。错误发现率表示在 R 次拒绝 H_0 的检验中，错误拒绝所占比例的数学期望，即 $\mathrm{FDR}=E(V/R)$，当 $R=0$ 时，$\mathrm{FDR}=0$ 且 $\mathrm{FDR}\leq\alpha$。针对不同问题，选取的错误率控制指标也不同。在经典的多重检验中，通常 m 的取值较小，一般采取控制总体错误率的方法；但是在现在的大数据分析中，m 的取值较大，此时采用控制总体错误率的话就过于严格，一般选择控制错误发现率的方法。

3. 多重假设检验方法

多重假设检验方法的具体算法可以分为两大类，一类为控制总体错误率的方法，另一类为控制错误发现率的方法，下面分别介绍这两类方法。

3.1. 控制总体错误率的算法

[Bonferroin 算法] [5]

给定显著性水平 α，对 m 个假设进行检验。采用单步法进行算法流程，每个假设各自的显著性水平选为 α/m，如果第 j 个假设检验的 p 值 $p_j\leq\alpha/m$，则拒绝 H_{j0}，$j=1,2,\cdots,m$。因此调整过的 p 值为 $\tilde{p}_j=\min(mp_j,1)$。

[Homl 算法] [6]

Homl 算法与 Bonferroin 算法相比保守降低并且提高了功效。首先将 m 个假设检验的 p 值从小到大排序 $p_{(1)}\leq p_{(2)}\leq\cdots\leq p_{(m-1)}\leq p_{(m)}$，对应的检验为 $\{H_{0(1)},\cdots,H_{0(m)}\}$。如果存在

$$\hat{j}=\min\left\{j:p_{(j)}>\frac{\alpha}{m-j+1}\right\}$$

则拒绝 $H_{0(j)}$, $j=1,\cdots,\hat{j}$。如果不存在 $\hat{j}$，则拒绝所有原假设。Homl 过程调整过的 p 值为：

$$\tilde{p}_{(i)}=\max_{k=1,\cdots,i}\left\{\min\left((m-k+1)p_{(k)},1\right)\right\}。$$

Homl 算法为 Bonferroin 算法的改进，其他改进方法还包括 Hommel、Hochberg 算法[7] [8]。针对大数据问题，我们关注的是在能够允许 R 次拒绝中发生少量的错误识别时，尽可能多地检验出显著的个体，即在控制错误发现率的同时，尽可能地提高检验的检验功效。

3.2. 控制错误发现率的算法

[BH 算法] [9]

在显著性水平 α 下，控制 FDR 的过程如下：将原来的 m 个 p 值从小到大进行排序，即 $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m-1)} \le p_{(m)}$。如果存在

$$\hat{k} = \max\left\{ j : p_{(j)} \le \frac{j}{m}\alpha \right\}$$

则拒绝 $H_{0(j)}, j = 1, \cdots, \hat{k}$；如果不存在 $\hat{k}$，则不拒绝任何原假设。

BH 算法控制 FDR 时，FDR 满足关系： $\text{FDR} \le \frac{m_0}{m}\alpha \le \alpha$。

[BY 算法] [9]

根据 BH 算法，修改 FDR 的上界为 $m_0\alpha \Big/ \left(m\sum_{i=1}^{m}\frac{1}{i} \right)$，将原来的 m 个 p 值从小到大进行排序，即 $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m-1)} \le p_{(m)}$。如果存在

$$\hat{k} = \max\left\{ j : m_0\alpha \le m\sum_{i=1}^{m}\frac{1}{i} \right\}$$

则拒绝 $H_{0(j)}, j = 1, \cdots, \hat{k}$。如果不存在 $\hat{k}$，则不拒绝任何原假设。

相对于 BH 算法而言，BY 算法得出的上界值过于保守，所以 j 值减小，拒绝原假设的个数减少。在表 1 中表示为假设 R 值一定时，V 值减小，S 值增大。BH 算法和 BY 算法的基本思想是给定显著性水平 α，找到拒绝域，从而将错误水平控制在 α 以下。Storey (2002)提出一种控制 FDR 的直接方法，其基本思想是先凭借经验给出拒绝域，再估计错误率。

[q 值法]

Storey (2002) [10] [11]将 $E\left(V/R \middle| R>0\right)$ 称为阳性错误拒绝率(pFDR)。设 Γ_α 为事先给定的拒绝域，则 pFDR 可表示为：

$$\text{pFDR} = E\left[V\left(\Gamma_\alpha\right) / R\left(\Gamma_\alpha\right) \middle| R>0 \right]$$

其中 $V\left(\Gamma_\alpha\right) = \#\left\{ H_i = 0 : T_i = \Gamma_\alpha \right\}$ 表示错误发现次数， $R\left(\Gamma_\alpha\right) = \#\left\{ T_i \in \Gamma_\alpha \right\}$ 表示所有拒绝 H_0 的次数。

定理 1 [12]对 m 个完全相同的假设进行检验，检验统计量为 $T_1, T_2, \cdots, T_m$，显著区域为 Γ。假设 $\left(T_i, H_i\right)$ 是独立同分布的随机变量， $T_i \middle| H_i \sim \left(1-H_i\right)\cdot F_0 + H_i \cdot F_1$，其中 F_0 为统计量在原假设的分布， F_1 为统计量在备择假设的分布， $H_i \sim \text{Bernoulli}\left(\pi_1\right)$， $i = 1, 2, \cdots, m$。则：

$$\text{pFDR} = \text{Prob}\left(H_0 = 0 \middle| T \in \Gamma_\alpha \right) = \frac{\pi_0 \text{Prob}\left(T \in \Gamma_\alpha \middle| H_0 = 0 \right)}{\text{Prob}\left(T \in \Gamma_\alpha \right)}$$

$$\text{Prob}\left(T \in \Gamma_\alpha \right) = \pi_0 \text{Prob}\left(T \in \Gamma_\alpha \middle| H_0 = 0 \right) + \left(1-\pi_0\right)\text{Prob}\left(T \in \Gamma_\alpha \middle| H_0 = 1 \right)$$

其中 $\pi_0 = 1 - \pi_1 = \text{Prob}\left(H_0 = 0 \right)$ 为 H_0 的先验概率。 $\text{pFDR} = \text{Prob}\left(H_0 = 0 \middle| T \in \Gamma_\alpha \right)$ 反映了在已经拒绝 H_0 的条件下，该假设为真的概率，可认为 pFDR 是贝叶斯后验 p 值。

Storey 对 q 值的定义如下：

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

即：统计量落入拒绝域时原假设为真的最小概率。q 值不过是 p 值定义的一个逆过程，q 值是在 $T\in\Gamma_\alpha$ 的条件下 $H_0=0$ 的概率。q 值法与 BH 算法恰好相反，即通过选定拒绝域 Γ_α 去估计对应的 q 值，当 $q\le\alpha$ 时可保证 FDR $\le\alpha$。在文献[13]中 Storey 给出了 π_0 和 q 值的具体估计算法。

4. 在基因表达数据分析中的应用

在当今的大数据时代，处处需要用到多重检验技术。下面我们以基因大数据为例，介绍几种多重检验方法的具体应用并进行比较，详细列举了每种方法的优缺点，以求能对实际数据工作者以相应的指导。

4.1. 实例

Lobenhofer *et al*. (2003)的微阵列实验中，评价了 17-β 对于 MCF-7 胸癌细胞的影响。该实验分别在 6 个时间点同时对 1900 个基因进行观测，每个时间点上有 8 个观测值。我们需要根据这些观测数据，判断这 1900 个基因在观测时间内是否表达。

检验问题描述如下：

$$H_0:\mu\in C_0,\ H_1:\mu\in\bigcup_{i=1}^{h}C_i$$

其中 $C_0=\left\{\mu\in R^6:\mu_1=\mu_2=\cdots=\mu_6\right\}$，$C_1=\left\{\mu\in R^6:\mu_1\ge\cdots\ge\mu_i\ge\cdots\ge\mu_6\right\}$，
$C_{10}=\left\{\mu\in R^6:\mu_1\le\cdots\le\mu_i\le\cdots\le\mu_6\right\}$，$C_i=\left\{\mu\in R^6:\mu_1\le\cdots\le\mu_i\ge\cdots\ge\mu_6\right\}$，$i=2,3,4,5$，
$C_i=\left\{\mu\in R^6:\mu_1\ge\cdots\ge\mu_i\le\cdots\le\mu_6\right\}$，$i=6,7,8,9$

Peddada *et al*. (2003)提出了 ORIOGEN 算法对该实验中的表达基因进行了选择，在该算法中为了降低假阳性，把假设检验中的 p 值调整为 $\tilde{p}_j=h\cdot p_j$，$j=1,2,\cdots,m$，h 为曲线模式的个数。

本文在 ORIOGEN 算法的基础上，利用多重假设检验技术对基因表达数据进行了分析。具体算法如下：

下面假定对某个选定的基因 g 进行分析，

第 1 步：选取所关注的表达曲线模式，将这些曲线形式记为 $C_1,C_2,\cdots,C_h$。

第 2 步：利用 PAVA 算法[14]求出该基因在每个模式 C_i，$i=1,2,\cdots,h$ 下的均值的估计值。

第 3 步：在每个曲线模式 C_i，$i=1,2,\cdots,h$ 下，分别计算 l_∞ 模。找到 r 满足 $l_\infty^g=l_\infty^{g(r)}=\max_i l_\infty^{g(i)}$。(某个模式的 l_∞ 模为其中任意两个参数估计值之差的最大值，参见 Peddada *et al*. (2003)。)

第 4 步：对该基因根据其观测表达数据抽取 N 个 bootstrap 样本。对每个抽取的样本进行步骤 2 和 3，从而获得该基因的统计量 l_∞^g 在原假设下 H_0 的分布，从而根据样本统计量 $l_\infty^{g(r)}$ 的值求出该基因的 p_g 值。

对所有的基因重复以上步骤，得到 1900 个 p 值。

第 5 步：进行多重假设检验：

$$H_0:\mu\in C_0,\ H_1:\mu\in\bigcup_{i=1}^{h}C_0$$

给定显著性水平 α，对所有基因进行多重检验，挑选出表达显著的基因。

4.2. 实验结果

本文利用 R 统计软件，对 1900 个基因表达数据进行了分析，结果见下表。

Table 2. Analysis results of gene expression data controlled by FWER and FDR
表 2. FWER 控制和 FDR 控制的基因表达数据分析结果

FWER 控制	表达基因个数	FDR 控制	表达基因个数
Bonferroni	17	BH	151
Hochberg	17	BY	31
Holm	17	q 值	199
Hommel	17		

根据控制错误率的不同，我们在两类不同的多重假设检验方法下分别进行了分析。表 2 报告了在 $N=1000000$，显著性水平 $\alpha=0.05$ 的情况下，利用上述多种算法控制 FWER 和 FDR，得到的 1900 个基因中表达的基因个数。在原始的 1900 个 p_i 值中，有 423 个 p_i 值小于 0.05。利用 ORIOGEN 算法识别出 170 个表达的基因[15]。对于控制总体错误率，利用多重假设检验算法进行分析，Bonferroin 算法识别出了 17 个表达的基因，Holm、Hochberg、Hommel 算法也都识别出了 17 个表达基因；对于控制错误发现率，BH 算法识别出 151 个表达的基因，由于 BY 算法的上界值过于保守，仅识别出 31 个表达的基因，q 值法识别出 199 个表达的基因。显然，在假阳性水平相同的情况下，控制 FDR 的多重检验算法挑选出的表达基因个数远高于控制 FWER 的算法，其中 q 值法挑选出的表达基因最多，比 ORIOGEN 算法挑出的基因还要多，而且 q 值法的检验功效比较大。因此，在实际数据分析中，推荐使用 q 值法进行多重假设检验分析。

5. 总结

在大数据的假设检验分析中，需要使用多重检验技术来控制错误率。从实例数据分析中可以看出，在使用多重检验算法时，控制 FWER 的意义并不大。研究者更关心的问题是当错误识别个数控制在可以接受的范围内时，尽可能多地识别出显著的基因。因此推荐控制 FDR 的多重检验算法。在控制 FDR 的算法中，从本文实例分析的结果可以看出 q 值法比 BH 算法的检验功效更大，因为其考虑了先验信息。在当前的大数据时代，数据量的变化也对传统的统计理论提出了挑战。在检验问题中，利用 q 值法解决大数据的多重检验问题，具有很强的实际意义。

基金项目

国家自然科学基金项目 11401393。
辽宁省教育厅自然科学基金项目 LJC201914。

参考文献

[1] Peddada, S.D., Lobenhofer, E.K., Li, L., *et al*. (2003) Gene Selection and Clustering for Time-Course and Dose-Response Microarray Experiments Using Order-Restricted Inference. *Bioinformatics*, **19**, 834-841. https://doi.org/10.1093/bioinformatics/btg093

[2] Simmons, S.J. and Peddada, S.D. (2007) Order-Restricted Inference for Ordered Gene Expression (ORIOGEN) Data under Heteroscedastic Variances. *Bioinformatics*, **1**, 414-419. https://doi.org/10.6026/97320630001414

[3] © Silicon Genetics. Multiple Testing Corrections.

[4] 杨柳. 多重假设检验中错误率控制过程的分析[D]: [硕士学位论文]. 哈尔滨: 黑龙江大学, 2009.

[5] 刘遵雄, 陈昊. 多重相关检验中错误发现率的控制算法[J]. 井冈山大学学报(自然科学版), 2016, 37(3): 35-40.

[6] Holm, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, **6**, 65-

70.

[7] Hommel, G. (1988) A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. *Biometrika*, **75**, 383-386. https://doi.org/10.1093/biomet/75.2.383

[8] Hochberg, Y. (1988) A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*, **75**, 800-802. https://doi.org/10.1093/biomet/75.4.800

[9] 裴艳波. 多重假设检验问题中关于三种错误测度-FWER, FDR 和 pFDR 的讨论[D]: [硕士学位论文]. 长春: 东北师范大学, 2005.

[10] Storey, J.D. (2002) A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **64**, 479-498. https://doi.org/10.1111/1467-9868.00346

[11] Storey, J.D. (2003) The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics*, **31**, 2013-2035. https://doi.org/10.1214/aos/1074290335

[12] 王婷, 曾平, 黄水平, 等. 错误发现率和q值及其微阵列数据分析的应用[J]. 现代预防医学, 2013, 40(5): 811-814.

[13] Storey, J.D., Tibshirani, R., Storey, J.D. and Tibshirani, R. (2003) Statistical Significance for Genomewide Studies. *Proceedings of the National Academy of Sciences*, **100**, 9440-9445. https://doi.org/10.1073/pnas.1530509100

[14] Robertson, T., Wright, F.T. and Dykstra, R.L. (1990) Order Restricted Statistical Inference. *Journal of the American Statal Association*, **85**, 398-409. https://doi.org/10.2307/2289813

[15] 刘瑞银. 基于趋势性的剂量反应研究[D]: [博士学位论文]. 长春: 东北师范大学, 2011.