行政院國家科學委員會專題研究計畫　成果報告

多重檢定問題中真實虛無假設個數的估計方法之研究

計畫類別：個別型計畫
計畫編號：NSC92-2118-M-004-002-
執行期間：92 年 08 月 01 日至 93 年 10 月 31 日
執行單位：國立政治大學統計學系

計畫主持人：薛慧敏

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94 年 2 月 17 日

Comparison of Methods for Estimating the Number of True Null Hypotheses in Multiplicity Testing

Huey-miin Hsueh,[1] James J. Chen,[2] and Ralph L. Kodell,[2]

[1] Department of Statistics
National Chengchi University
Taipei, Taiwan

[2] Division of Biometry and Risk Assessment
National Center for Toxicological Research
Food and Drug Administration
Jefferson, Arkansas 72079

Send correspondence to:

Prof. Huey-miin Hsueh
Department of Statistics,
National Chengchi University,
Taipei, Taiwan 116
Tel : 886-2-29393091 ext. 81138
Fax: 886-2-29398024
E-mail : hsueh@nccu.edu.tw

Abbreviated title: Estimating the number of true null hypotheses

Abstract

When a large number of statistical tests is performed, the chance of false positive findings could increase considerably. The traditional approach is to control the probability of rejecting at least one true null hypothesis, the faimilywise error rate (FWE). To improve the power of detecting treatment differences, an alternative approach is to control the expected proportion of errors among the rejected hypotheses, the false discovery rate (FDR). When some of the hypotheses are not true, the error rate from either the FWE- or the FDR-controlling procedure is usually lower than the designed level. This paper compares five methods to estimate the number of true null hypotheses over a large number of hypotheses. The estimated number of true null hypotheses is then used to improve the power of FWE- or FDR-controlling methods. Monte Carlo simulations are conducted to evaluate the performance of these methods. The lowest slope method, by Benjamini and Hochberg,[1] and the mean of differences method appear to perform the best. These two methods control the FWE properly when the number of non-true null hypotheses is small. A data set from a toxicogenomic microarray experiment is used for illustration.

KEY WORDS: Comparison-wise error rate (CWE); False discovery rate (FDR); Familywise error rate (FWE); Multiple endpoints; Number of true number hypotheses.

1. Introduction

It has been well recognized that performing many tests without appropriately accounting for the multiple testing effect can inflate the overall Type I error rate or familywise error rate (FWE), where FWE is the probability of rejecting at least one true null hypothesis in the given family of the hypotheses. A number of multiple comparison procedures (MCP) have been proposed for controlling the FWE (Hochberg and Tamahane;[2] Westfall and Young[3]). One difficulty with the use of a MCP approach to controlling the FWE is that it often substantially reduces the power to detect a difference when the number of comparisons is large. Some FWE-controlling methods have unsatisfactory performances if the test statistics are highly correlated. Benjamini and Hochberg[4] proposed controlling the false discovery rate (FDR) as an alternative to controlling the FWE. The FDR is defined as the expected proportion of errors among the rejected hypotheses. Benjamini and Hochberg,[4] hereafter BH, and Benjamini and Liu[5] proposed sequential procedures to control the FDR. Weller, Song, Heyen, Lewin and Ron[6] adopted the BH procedure to the multiplicity problem in the genetic dissection of complex traits. They argued that control of the FDR is a more appropriate approach than control of the FWE for multiple marker quantitative trait detection. However, Zaykin,Young and Westfall[7] demonstrated that the BH procedure only controls the FDR in an unconditional manner; the method cannot control the FDR, conditional on having rejected one or more hypotheses.

Kwong, Holland and Cheung[8] proposed an alternative FDR controlling procedure by incorporating the known distribution and dependence structure of the test statistics. Their procedure is more powerful than the BH procedure. Rather than controlling FDR, Tusher, Tibshirani and Chu[9] introduced a procedure to estimate an FDR, called the SAM, through permutations. Another way to improve the conservativeness of MCP methods is adaptively using some estimate of m_0, e.g., Holland and Cheung[10], Benjamini and Hochberg,[1] Storey and Tibshirani,[11] described below.

Schweder and Spjøtvoll[12] proposed a graphical method to estimate the number of true hypotheses to help in deciding on the number of null hypotheses to be rejected. The estimated number of true null hypotheses can be applied to the Bonferroni-type FWE-controlling methods to improve the power and reduce the false negative rate (Hochberg and Benjamini[13]). Benjamini and Hochberg[1] provided a procedure based on the slope of the p-values to estimate the number of true null hypotheses, m_0. Recently, Storey[14] proposed a method to estimate m_0 and then used the estimate of m_0 to calculate a (*positive*) FDR for a fixed rejection region. This paper presents two other methods to estimate m_0, a mean difference method (discussed by Benjamini and Hochberg[1]) and a full least squares method.

The purpose of this paper is to compare methods for estimating number of true null hypotheses. Five estimations methods are evaluated: Schweder and Spjøtvoll,[12] Benjamini and Hochberg,[1] Storey,[14] the mean of differences, and the least squares methods. Section 2 describes five estimation methods. Section 3 contains Monte Carlo simulations to compare the means and standard deviations of the estimated m_0 from the five methods. In addition, the empirical FWEs using the Bonferroni adjustment method based on the estimated m_0 are also evaluated. Section 4 contains an illustration of the use of the estimated m_0 to improve power of identifying differentially expressed genes in a microarray data set from existing FWE- and FDR-controlling procedures.

2. Estimating number of true null hypotheses

2.1 Estimating methods

We consider the problem of simultaneously testing m null hypotheses. Assume there are only two possibilities in the parameter space. According to the true state of nature, either the null or non-null is true; the results and the probability (in parentheses) from m independent tests can be summarized as a 2×2 table (here we modify the notation of Benjamini and Hochberg[4]):

	Declared Significant	Declared Non-Significant	Total
Null True	V(α)	S($1-\alpha$)	m_0
Alternative True	U($1-\beta$)	T(β)	$m-m_0$
Total	R	$m-R$	m

The random variables V and U are unobservable; but, the random variable $R = U + V$, the number of significances declared, is observable. The FWE is the probability of rejecting at least one true null hypothesis in the given family of the hypothesis tests, $\Pr(V \geq 1)$. The number of true null hypotheses, m_0, is fixed but unknown.

For the m null hypotheses, assume that m_0 are from the true null population with zero mean and $(m - m_0)$ are from the alternative non-true null population each with effect size $\gamma_j > 0$, $j = (m_0 + 1), \cdots, m$. Define $\gamma_j = 0$, for $j = 1, \cdots, m_0$. Let α be the comparison-wise error rate for each individual test among true null hypotheses and $(1 - \beta_j)$ be the power for the j-th non-true alternative hypothesis. Given m_0 and $m - m_0$, V and U are binomially distributed with "success" probabilities

$$\alpha = Pr(\text{Declared significant} \mid \text{True null}, \gamma_j = 0)$$

and

$$1 - \beta_j = Pr(\text{Declared significant} \mid \text{True alternative}, \gamma_j > 0),$$

respectively. Let the observed ordered p-values for the m hypotheses be denoted as $p_{(1)}, \cdots, p_{(m)}$.

Schweder and Spjøtvoll's Method

For a relatively small α, the expected number of non-significant hypotheses (Equation 1 of Schweder and Spjøtvoll[12]) can be approximated as

$$m - r(\alpha) \approx E((m - R)|\alpha) \approx m_0(1 - \alpha), \tag{1}$$

where $r(\alpha)$ is the observed number of rejections at level α. Several methods of estimating m_0 have been proposed based on that the p-values are uniformly distributed under the

null hypothesis. The number of rejections at the level $p_{(i)}$ is exactly i, that is, $r(p_{(i)}) = i$. Schweder and Spjøtvoll[12] first considered the cumulative plot of observed $1 - p_{(i)}$ against $m-i$, $i = 1 \cdots, m$. The left-hand part of the data, $\{(1-p_{(1)}, 1), (1-p_{(2)}, 2), \cdots, (1-p_{(i)}, i)\}$, was used for finding an estimate of m_0. The procedure starts from $j = m$ and decreases one in each successive calculation. In each step, the least squares estimate of the slope, $\hat{m}_0^{SS(j)}$ is computed. The slope of each successive line will decrease. The calculation stops when the slope estimate increases, that is, the first time $\hat{m}_0^{SS(j^0)} \leq \hat{m}_0.^{SS(j^0-1)}$ The estimate is then $\hat{m}_0^{SS} = \hat{m}_0.^{SS(j^0)}$

Storey's Method

Instead of the graphical approach with the least squares method, Storey[14] proposed to estimate the slope directly (Equation 1)

$$\hat{m}_0^{ST} = \{m - r(\lambda)\}/(1 - \lambda),$$

where λ ideally is the change point of the p-values between true null and true alternative hypotheses. A bootstrapping procedure was suggested for the optimal λ in his paper. Later, $\lambda = 0.5$ will be used in empirical studies as in the evaluation of the Storey (ST) method.

Benjamini and Hochberg's Lowest Slope Method

Benjamini and Hochberg[1] proposed the Lowest Slope (LSL) estimator with a slightly different approach. They considered simply the slope of the line passing through the points $(m+1, 1)$ and $(i, p_{(i)})$. Starting from $i = 1$, each slope $S_i = (1-p_{(i)})/(m+1-i)$ is calculated. Proceeding towards larger i, the procedure stops the first time $S_{j^0} < S_{j^0-1}$. The estimate is then $\hat{m}_0^{LSL} = \min[(1/S_{j^0} + 1), m]$.

Mean of Differences Method

As mentioned in Benjamini and Hochberg,[1] the LSL estimator can be derived in view of the gap $d_{(i)} = p_{(i)} - p_{(i-1)}, i = m - m_0 + 2, \cdots, m + 1, p_{(0)} = 0, p_{(m+1)} = 1$. Under the independence assumption, the largest $(m + 1 - m_0)$ p-values, $p_{(m_0+1)}, \cdots, p_{(m+1)}$, are very likely from the true null hypotheses. The gaps $d_{(i)}s$ are independently and identically

Beta(1,m_0) distributed and have the common mean $E(D) = 1/(m_0 + 1)$. Thus, m_0 can be estimated as

$$\hat{m}_0^{MD} = 1/\bar{d}_{m_0} - 1 \approx 1/E(D) - 1,$$

where $\bar{d}_{m_0} = \sum_{i=m+2-m_0}^{m+1} d_i/m_0 = \{1 - p_{(m-m_0+1)}\}/m_0$. To have a conservative estimate, the search starts from $j = m$ with $\bar{d}_m$ and j proceeds downward. As the contamination of non-true nulls becomes less and less, the estimate of m_0 decreases. The search stops when the first time $\hat{m}_0^{MD-1} \geq \hat{m}_0^{MD}$. The difference between the mean of differences (MD) method and the LSL method is minor. Mainly, the MD method is derived mathematically based on the distribution of the differences of the ordered p-values.

Least Squares Method

The above four methods use partial data presumed all coming from the null population. To bring in information from the non-true null distributions, an alternative least squares estimation is described. At a given significance level α, the probability of rejection of the i-*th* null hypothesis is {1- $\beta(\gamma_i, \alpha)$}. The expected number of significances declared is

$$\mathrm{E}\{\mathrm{R}(\alpha)\} = \sum_{i=1}^{m_0}\{1 - \beta(\gamma_i, \alpha)\} + \sum_{i=m_0+1}^{m}\{1 - \beta(\gamma_i, \alpha)\} = m_0\alpha + (m - m_0)\bar{\beta}(\alpha),$$

where $\bar{\beta}(\alpha) = \sum\{1 - \beta(\gamma_i, \alpha)\}/(m - m_0)$ is the average power among the $(m - m_0)$ non-true null hypotheses tested at the α level. Each $p_{(i)}$ can be treated as a realization of a significance level in which the observed number of significances declared is i. Given $p_{(1)}, \cdots, p_{(m)}$, and $\bar{\beta}(p_{(i)})$, the expected number of rejected hypotheses at significance level $p_{(i)}$ is $E(i) = m_0 p_{(i)} + (m - m_0)\bar{\beta}(p_{(i)})$. Thus, an estimate of m_0 can be obtained by minimizing the sum of squares

$$\sum_{i=1}^{m}\left(i - m_0 p_{(i)} - (m - m_0)\bar{\beta}(p_{(i)})\right)^2. \tag{2}$$

given as

$$\hat{m}_0^{LS} = \sum_{i=1}^{m} x_i y_i / \sum_{i=1}^{m} x_i^2,$$

where $y_i = i - m\bar{\beta}(p_{(i)})$ and $x_i = p_{(i)} - \bar{\beta}(p_{(i)})$. It can be shown that conditional on $\bar{\beta}(p_{(i)})$, the estimate $\hat{m}_0^{LS}$ is unbiased.

2.2 Applying $\hat{m}_0$ to multiple testing procedures

The estimates of m_0 can be used to improve the power of Bonferroni-type multiple testing procedures. If m_0 is the number of true null hypotheses and all hypotheses are independent, then testing each individual hypothesis at level α will have the FWE of

$$\text{FWE} = Pr(V \geq 1) = 1 - Pr(V = 0) = 1 - (1 - \alpha)^{m_0} = \alpha_0.$$

Conversely, for a FWE-controlling test set at the level α_0, each individual test will be required to be rejected at the level $\alpha = 1 - (1 - \alpha_0)^{1/m_0}$. Without the independence assumption, testing each individual hypothesis at the level α_0/m_0 will have the FWE $\leq \alpha_0$.

Benjamini and Hochberg[4] proposed a FDR-controlling method by comparing each p_i to $q^* i/m$, where q^* is the designed significance FDR level. The procedure begins with $i = m$, the procedure proceeds to the smaller p_i as long as $p_i > q^* i/m$. The test stops and all $p_1, .., p_k$ are rejected, when $p_k \leq q^* k/m$. The estimated $\hat{m}_0$ can be used to improve the power by comparing each p_i with $q^* i/\hat{m}_0$ (Benjamini and Hochberg[1]). The procedure controls the FDR at q^*, i.e. $FDR = E(V/R) \leq q^*$, the expected number of false rejections can be approximately estimated as $E(V) \approx r \cdot q^*$.

3. Simulation Study

We formulated the problem in terms of testing m univariate means of a m-variate normal random vector, $H_{oi} : \mu_i = 0$ versus $H_{1i} : \mu_i \neq 0$, $i = 1, .., m$. The number of hypotheses considered was $m = 20$, 100, 500 and 1000 among which the number of true null hypotheses was $m_0 = 0.7m, 0.9m$ and m, the complete null case. Two multivariate normal models were considered for the underlying distribution of the parameters: 1) independent model, the univariate normal random variables are independent; 2) equi-correlated model, the pairwise correlations of the normal random variables for the true null models are $\sqrt{0.2}$, as are the pairwise correlations for the non-true hypotheses. The correlations between the null variables and the non-true variables are 0. Each univariate normal random variable has variance 1.

The m_o true null variables were generated from a m_o-variate normal random vector with mean vector 0. For the non-true variables, a non-zero effect size γ was added to each random variate. Two alternative models were considered for the effect sizes: A) a simple alternative model, in which the effect size is constant $\gamma = \gamma_0$; B) a multiplicity alternative model, in which the effective size γ is generated from a truncated normal distribution $C \cdot N(\gamma_0, 1) I\{\gamma > 0\}$ with some normalizing constant $C > 0$.

The simulations consist of two parts. The first part includes a comparison of the five methods to estimate m_0. The parameter γ_0 was designed to have 80% power, $1 - \beta(\gamma_0, \alpha) = 1 - \Phi(\Phi^{-1}(1-\alpha) - \gamma_0)$, for the individual CWE $= \alpha$. The CWE α was set to ensure FWE $= 0.25$ under independence. That is, $\alpha = 1 - (1 - .25)^{1/m_0}$. For each simulated data set, the number of true null hypotheses was estimated using five estimating methods introduced in Section 2. Table 1 contains the sample means and standard deviations of the 10,000 estimates from the five methods. Table 2 contains the results of a similar simulation with γ_0 fixed at 2 for $m = 20$ and 100. The results for $m = 500$ and 1,000 are similar (not shown).

For complete null cases, given a constraint that $\hat{m}_0 \leq m$ in solving each estimate, all methods are underestimated. The multiplicity alternatives do not appear to have significant effects on the mean and standard deviation estimates. However, the dependencies among the hypotheses dramatically increase the variations in all five estimators. Generally speaking, the LSL and the MD estimators have the most desired performance, upward bias and least variation. The LS estimation is closer to the true m_0 on the average. However, in subsequent application to FWE- or FDR-controlling procedures, a more conservative method is preferred. The LS method has better performance than either the SS or ST method. The SS has the worst performance; it not only has a large variation but also is severely under-estimated.

The second simulation is to evaluate a modification of the Bonferroni method based on the estimated numbers of true hypotheses. The parameter γ_0 was set such that the power 1-$\beta(\gamma_0, \alpha) = 1 - \Phi(\Phi^{-1}(1-\alpha) - \gamma_0) = .8$ at CWE $= \alpha$, where α was set such that FWE

= 0.05, 0.10, and 0.25 under the independent model. We evaluated the empirical FWE's at the CWE nominal level $FWE/\hat{m}_0$ with various estimates of m_0 for $m = 100$ and 1,000. The empirical FWEs for the five multiple testing methods are shown in Table 3 and Table 4. Each value is based on ten thousand replications.

Due to under-estimation of the true m_0 (Table 1), the empirical FWEs from the SS and ST methods often exceed the nominal level, in particular, for FWE = 0.05. On the other hand, the LSL and MD methods appear to well control the FWE, except in a few cases for $m = 100$. The LS method also has the empirical FWE generally less than the nominal level. In general, the LSL, MD, and LS methods perform well except for $m = m_0 = 100$ and FWE = 0.05. All three methods have smaller empirical FWE under the correlated model than under the independent model with very few exceptions for the LS method. Also, the differences between simple and multiplicity alternative alternatives appears small.

4. Example

The example is a cDNA two-channel experiment from a toxicogenomic study of gene expression levels of kidney samples from rats dosed with a drug. The experiment includes 6 replicate arrays (arrays A1-A6) from a 700 gene rat Phase-1 chip, Molecular Toxicology (Santa Fe, NM). On the arrays A1-A3, the control samples were assigned to the red dye and treated samples were assigned to the green dye. The dye assignments to the control and treated samples were reversed on the arrays A4-A6. In addition, sequences of five genes from other species different from the one of 700 genes were also spotted on the array to monitor non-specific background binding of labelled RNA. Chen, Kodell, Sistare, Thompson, Morris and Chen[15] described several normalization methods for this data set. Briefly, let y_{ijk} denote the base-2 logarithm of the intensity for the i-th gene on the array j in the t-th treatment and k-th dye, $i = 1, .., g, \ , j = 1, .., r, \ k = 1, 2$, and $t = 1, 2$. For a given array, let s denote the number of disjoint subsets (partitions) in the array, which is based on the spotting pattern matrix generated by a single pin with the size 14×14. Denote $L_{l(j)}$ as

the l-th subset (location) on the array j, $l = 1, .., s$. Chen *et al.*(4) proposed the subset normalization model

$$y_{ijk,l} = m + G_i + L_{l(j)} + I_j + D_k + (AD)_{jk} + e_{ijk,l},$$

where m is the overall average signal, G_i represents the effect of the i-th gene, $L_{l(j)}$ represents the effect of the location l on the j-th array, I_j represents the effect of the intensity on the array j, D_k represents the effect of the k-th dye and $(AD)_{jk}$ accounts for the effect of array j and dye k. This model is a generalization of the Kerr's global ANOVA model (12,13); the array effects A_j are decomposed into location and intensity components, $L_{l(j)} + I_j$. In this paper, the $L_{l(j)}$ is estimated by the median, I_j is estimated using the *lowess* fit and the other effects are estimated using the least-squares estimates. The residuals (normalized intensities) removing the overall effects from the fitted model correspond to the Treatment$\times$Gene interactions as the effect of interest.

Identifying differentially expressed genes between the control and treatment groups can be done by computing the two-sample t-test under the equal variance model one gene at a time. Figure 1 gives the p-value plot with ordered (1-p-value) as x-axis and the number of insignificance as y-axis. Applying the procedures, the estimated numbers of true hypotheses are all listed in Table 5. In calculating the LS method, we assumed a constant effect model. That is, $\beta_i = \bar{\beta}(\alpha)$ for the non-true null hypotheses; the power $1 - \bar{\beta}(\alpha)$ was estimated by $\hat{\bar{\beta}}(\alpha)$. The estimated values range from 410 to 524, in which, Schweder and Spjøtvoll's method has the smallest m_0 estimate (the most liberal estimate). Storey's method, with $\lambda = 0.5$, also gives a liberal estimate. On the other hand, the LSL and MD methods are very similar; both methods are conservative. The LS method gives the estimate closer to the LSL and MD methods.

These estimates are applied to the FWE-controlling Bonferroni adjustment. Table 5 (upper panel) shows the observed number of rejections $r(\alpha)$, at the FWE level $\alpha_0 = 0.05, 0.10, 0.25$. All methods show an increase in the number of rejections ranging from 2 to 9. For com-

parison purpose the number of rejections from resampling-based analysis (by performing all possible permutations) is also given. It can be seen that the resampling-based analysis, which accounts for correlations among the 705 genes, appears to be more powerful than using the estimated m_0 with the Bonferroni correction for this data set.

These estimates are applied to the BH FDR-controlling procedure. Table 5 (lower panel) shows the observed number of rejections r and estimated number of false rejections for the FDR levels $q^* = 0.001, 0.01, 0.05$. Among these procedures, similar trends appear as seen in the FWE-controlled analysis. The BH procedure set at the FDR = 0.001 is quite equivalent to the Bonferroni procedure set at FWE = 0.05. These procedures are sensitive to the FDR level (in the range considered: 0.001 to 0.05); great increment occurs in the observed number of rejections when the FDR level increases. At a specific FDR level, with an observed number of rejections, r, the expected number of false rejections is $EV \approx r \cdot q^*$ (shown in the parenthesis). With the FDR level $q^* \leq 0.01$, every procedure has less than one expected false rejection.

5. Summary

The FWE and FDR approaches have been proposed to control an "error rate" in multiple hypothesis testing. In the analysis of a large number of endpoints such as identifying differentially expressed genes in a microarray experiment, the FWE-controlling procedures may fail to identify some significant genes because of lack of power. The closed testing procedures using the permutation distribution of the min P statistic have been proposed to improve the power (e.g., Westfall, Zaykin and Young[16] ; Dudoit,Yang, Callow and Speed[17]). Alternatively, a FDR-controlling procedure can be applied (the second column $m = 705$ of Table 5). The power of either approach can be further improved, when the number of true hypotheses is known.

This paper evaluates five methods for estimating the number of true null hypotheses. To properly control the FWE or FDR, an conservatively over-estimated method is desired. The LSL and MD give the most satisfactory empirical results among the five methods considered.

On the other hand, as a point estimation, the LS method performs well with less bias and variability in some cases. Finally, the estimated number of true hypotheses can be used for screening (data filtering) purposes. For example, the hypotheses with the largest $\hat{m}_0$ p-values can be eliminated from further analyses. In this case, a method that gives a lower limit estimate is preferable.

Acknowledgements

The authors wish to thank Drs. Frank Sistare and Karol Thompson for providing the example data set. The authors also thank Dr. Y. Benjamini for pointing out some manuscripts on this topic, Dr. P. H. Westfall and the referees for their helpful comments on the draft.

References

1. 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, 60-83.

2. Hochberg, Y. ; Tamhane, A.C. Multiple Comparison Procedures. John Wiley & Sons : New York, 1987.

3. Westfall, P.H. ; Young, S.S. Resampling-Based Multiple Testing. John Wiley & Sons : New York, 1993.

4. Benjamini, Y. ; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B **1995**, 57, 289-300.

5. Benjamini, Y. ; Liu, W. A step down multiple hypotheses testing procedure that controls the false discovery rate under independence. J. Statist. Plann. Inference **1999**, 82, 163-170.

6. Weller, J.I.; Song, J.Z.; Heyen, D.W.; Lewin, H.A.; Ron, M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics **1998**, 150, 1699-1706.

7. Zaykin, D.V.; Young, S.S.; Westfall, P.H. Letter to Editor. Using the false discovery rate in the genetic dissection of complex traits. Genetics **2000**, 154, 1917-1918.

8. Kwong, K. S.; Holland, B.; Cheung, S. H. A modified Benjamini-Hochberg multiple comparisons procedure for controlling the false discovery rate. J. Statist. Plann. Inference **2002**, 104, 351-362.

9. Tusher, V. G.; Tibshirani, R. ; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. **2001**, 98, 5116-5121.

10. Holland, B.; Cheung, S. H. Familywise robustness criteria for multiple-comparison procedures. J. R. Statist. Soc. **2002**, 64, 63-77.

11. Storey, J. D. ; Tibshirani, R. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. To appear in *The Analysis of Gene Expression Data : Methods and Software*; Parmigiani, G., Garrett, E. S., Irizarry, R. A., Zegger, S. L., Eds. ; Springer : New York, 2003.

12. Schweder, T. ; Spjøtvoll, E. Plots of p-values to evaluate many tests simultaneously. Biometrika **1982**, 69, 493-502.

13. Hochberg, Y. ; Benjamini, Y. More powerful procedures for multiple significance testing. Statistics in Medicine **1990**, 9, 811-818.

14. Storey, J.D. A direct approach to false discovery rates. J. R. Statist. Soc. B **2002**, 64, 479-498.

15. Chen, Y-J.; Kodell, R.L.; Sistare, F.; Thompson, K.; Morris, S.; Chen, J.J. Normalization methods for cDNA microarray data Analysis. Journal of Biopharmaceutical Statistics **2003**, to appear.

16. Westfall, P.H.; Zaykin, D.V.; Young, S.S. Multiple tests for genetic effects in association studies. Methods in Molecular Biology. In *Biostatistical Methods*; Looney, S. Ed.; Humana Press : Toloway, New Jersey, 2001; 143-168.

17. Dudoit, S.; Yang, Y.H.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differential expressed genes in replicated cDNA microarray experiments. Statistica Sinica **2002**, 12, 111-139.

Table 1: Comparisons of various estimations for m_0, the number of true null hypotheses. The parameter r_0 is set to have 80% power at the FWE = 25%. The means and standard deviations (s.d.) are based on 10,000 replicates.

			Simple Alternative				Multiplicity Alternative			
			Independent		Correlated		Independent		Correlated	
m	m_0	method	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
20	20	SS[a]	17.64	3.55	15.34	6.01	17.56	3.65	15.25	6.05
		ST[b]	18.24	2.65	16.59	4.81	18.21	2.64	16.55	4.80
		LSL[c]	19.83	0.70	19.03	2.87	19.84	0.67	19.06	2.81
		MD[d]	19.44	1.11	18.88	2.63	19.46	1.08	18.89	2.62
		LS[e]	18.02	2.22	16.85	3.85	18.03	2.20	16.84	3.83
	18	SS	16.37	3.72	14.71	5.99	16.35	3.76	14.75	6.04
		ST	17.20	3.12	15.78	5.07	17.16	3.14	15.90	5.04
		LSL	19.49	1.06	18.64	3.17	19.50	1.09	18.72	3.06
		MD	18.18	1.54	17.82	2.94	18.25	1.59	17.93	2.91
		LS	16.43	2.06	16.13	3.44	16.51	2.11	16.27	3.42
	14	SS	12.68	3.28	11.92	5.12	12.87	3.42	12.09	5.26
		ST	13.96	3.59	13.55	5.32	14.11	3.61	13.67	5.31
		LSL	16.47	1.90	16.12	3.48	16.84	1.98	16.42	3.55
		MD	14.53	1.96	14.73	3.13	14.91	2.04	15.06	3.12
		LS	13.09	1.37	13.05	2.27	13.49	1.43	13.48	2.36
100	100	SS	95.18	9.21	73.21	34.10	95.09	9.31	72.86	33.94
		ST	96.11	5.83	84.56	21.46	96.01	5.88	84.72	21.00
		LSL	99.83	0.00	97.82	8.19	99.84	0.00	98.01	7.60
		MD	99.41	1.00	97.48	8.31	99.44	0.96	97.67	7.71
		LS	96.65	4.28	89.03	15.28	96.63	4.34	89.06	15.24
	90	SS	86.80	10.01	70.26	33.71	86.70	10.52	70.53	33.70
		ST	89.35	8.27	80.64	22.58	89.37	8.37	80.79	22.69
		LSL	92.50	2.03	91.52	9.02	93.21	2.11	92.06	9.07
		MD	90.51	2.02	89.65	8.79	91.21	2.11	90.17	8.86
		LS	87.09	4.15	85.83	13.54	87.54	4.23	86.04	13.57
	70	SS	67.11	8.96	57.14	27.76	67.53	9.54	57.81	28.30
		ST	70.04	8.40	68.47	23.71	70.45	8.37	69.01	23.75
		LSL	73.22	2.38	72.17	8.66	75.18	2.92	73.92	9.20
		MD	71.22	2.37	70.25	8.49	73.18	2.91	72.04	8.92
		LS	70.74	2.61	70.50	9.17	72.25	2.66	72.27	9.02

- continued -

Table 1 - continued -

			Simple Alternative				Multiplicity Alternative			
			Independent		Correlated		Independent		Correlated	
m	m_0	method	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
500	500	SS	489.46	22.29	338.22	187.89	489.59	21.92	331.96	190.46
		ST	491.00	13.92	425.43	104.15	491.13	14.00	422.94	105.46
		LSL	499.79	6.56	494.88	26.00	499.79	6.62	494.92	24.66
		MD	499.38	5.45	494.55	26.24	499.39	5.53	494.57	24.98
		LS	493.35	10.47	448.36	76.89	493.35	10.47	448.36	76.89
	450	SS	443.35	23.02	325.80	187.53	443.74	23.76	324.63	186.83
		ST	449.97	21.13	405.29	111.73	450.31	21.18	405.58	110.19
		LSL	454.33	2.68	450.82	24.50	457.88	3.66	454.05	25.33
		MD	452.33	2.86	448.82	24.51	455.88	3.47	452.05	25.32
		LS	440.36	8.87	434.35	63.34	441.79	8.90	435.58	62.78
	350	SS	343.86	20.29	264.41	152.98	344.26	23.69	264.99	153.85
		ST	349.56	18.63	341.91	115.78	350.64	18.91	343.77	115.05
		LSL	357.17	3.99	353.38	22.73	366.04	5.55	362.45	24.76
		MD	355.17	4.02	351.44	22.52	364.04	5.57	360.46	24.76
		LS	358.63	5.26	357.18	42.08	363.66	5.52	363.19	41.59
1000	1000	SS	985.27	33.56	647.84	392.97	984.54	35.30	648.30	391.42
		ST	987.64	21.17	849.00	209.21	987.24	21.25	851.46	206.01
		LSL	999.75	13.19	992.63	42.17	999.76	13.12	993.08	39.61
		MD	999.35	10.89	992.30	42.29	999.35	10.89	992.78	39.66
		LS	990.65	16.39	898.04	153.38	990.65	16.39	898.04	153.38
	900	SS	891.13	34.53	626.62	386.52	890.42	35.46	623.09	388.29
		ST	900.23	30.17	810.47	219.36	899.63	30.16	810.81	220.87
		LSL	906.43	3.94	901.17	36.41	913.12	4.68	906.52	44.38
		MD	904.43	3.43	899.16	36.47	911.12	4.79	904.52	44.43
		LS	882.03	12.34	870.20	124.73	884.37	12.40	872.22	123.74
	700	SS	691.80	29.45	509.43	317.77	691.91	33.54	507.64	318.85
		ST	699.91	26.27	684.63	231.89	700.60	26.22	687.95	229.64
		LSL	711.41	4.87	705.12	37.10	728.04	7.87	723.28	36.59
		MD	709.40	5.26	703.13	37.07	726.05	7.65	721.28	36.62
		LS	717.64	7.49	715.38	83.30	726.36	7.61	725.93	82.21

a. SS : Schweder and Spjøtvoll[12] ; b. ST : Storey[14]; c. LSL : Benjamini and Hochberg[1]; d. MD : mean difference method; e. LS : lease square method.

Table 2: Comparisons of various estimations for m_0, the number of true null hypotheses. The parameter $r_0 = 2$. The means and standard deviations (s.d.) are based on 10,000 replicates.

			Simple Alternative				Multiplicity Alternative			
			Independent		Correlated		Independent		Correlated	
m	m_0	method	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
20	20	SS	17.62	3.58	15.31	6.00	17.61	3.59	15.41	5.99
		ST	18.24	2.61	16.57	4.79	18.25	2.61	16.64	4.75
		LSL	19.85	0.63	19.04	2.81	19.84	0.64	19.07	2.78
		MD	19.47	1.08	18.85	2.69	19.45	1.09	18.90	2.64
		LS	18.04	2.19	16.82	3.84	18.02	2.20	16.91	3.82
20	18	SS	16.37	3.87	14.74	6.05	16.49	3.82	14.77	6.06
		ST	17.22	3.16	15.88	5.05	17.28	3.08	15.97	4.97
		LSL	19.53	1.10	18.69	3.13	19.57	1.07	18.77	3.05
		MD	18.48	1.63	18.06	2.98	18.54	1.59	18.18	2.87
		LS	16.54	2.28	16.20	3.64	16.70	2.25	16.34	3.57
20	14	SS	12.94	3.65	12.21	5.45	13.38	3.79	12.56	5.57
		ST	14.18	3.60	13.74	5.29	14.64	3.65	14.03	5.24
		LSL	17.32	2.14	16.78	3.67	17.71	2.07	17.16	3.58
		MD	15.45	2.22	15.48	3.34	15.89	2.23	15.86	3.24
		LS	13.43	1.70	13.45	2.81	13.99	1.79	13.98	2.82
100	100	SS	95.11	9.38	72.38	34.37	94.96	9.74	73.55	33.89
		ST	96.09	5.81	84.21	21.56	96.08	5.76	85.02	21.08
		LSL	99.83	0.00	97.78	8.28	99.84	0.00	97.91	8.36
		MD	99.39	1.02	97.43	8.48	99.43	0.94	97.61	8.28
		LS	96.70	4.29	88.66	15.36	96.68	4.25	89.16	15.17
100	90	SS	87.08	11.34	69.97	34.11	87.89	11.23	71.17	33.63
		ST	89.71	8.28	80.91	22.69	90.36	8.24	81.87	21.91
		LSL	96.43	2.68	94.61	9.46	96.57	2.51	95.11	8.93
		MD	94.52	2.81	93.00	9.40	94.63	2.63	93.42	8.87
		LS	88.15	5.09	85.76	14.91	89.19	5.01	86.96	14.12
100	70	SS	67.85	11.61	58.06	30.21	70.56	11.73	60.78	30.41
		ST	71.22	8.53	69.42	23.70	74.08	8.81	71.93	23.23
		LSL	82.61	4.77	81.03	12.51	84.27	4.32	82.98	10.77
		MD	80.61	4.77	79.18	12.30	82.27	4.32	81.05	10.58
		LS	72.89	3.55	72.83	11.93	75.73	3.63	75.98	11.41

Table 3: Empirical familywise error rates at 5%, 10% and 25% nominal levels from four multiple testing methods under a simple alternative model.

			Independent			Correlated		
m	m_0	method	fwe=0.050	0.100	0.250	fwe=0.050	0.100	0.250
100	100	SS	0.058	0.107	0.243	0.134	0.193	0.289
		ST	0.057	0.105	0.238	0.078	0.126	0.222
		LSL	0.056	0.101	0.229	0.050	0.086	0.176
		MD	0.056	0.102	0.230	0.050	0.087	0.178
		LS	0.055	0.105	0.233	0.059	0.101	0.199
	90	SS	0.053	0.103	0.241	0.121	0.180	0.287
		ST	0.052	0.098	0.233	0.074	0.119	0.214
		LSL	0.048	0.095	0.225	0.047	0.083	0.167
		MD	0.049	0.096	0.230	0.047	0.084	0.169
		LS	0.053	0.101	0.235	0.053	0.093	0.189
	70	SS	0.051	0.101	0.228	0.122	0.183	0.287
		ST	0.048	0.096	0.218	0.075	0.124	0.223
		LSL	0.046	0.091	0.208	0.049	0.087	0.175
		MD	0.047	0.093	0.213	0.050	0.089	0.179
		LS	0.049	0.093	0.220	0.050	0.088	0.180
1000	1000	SS	0.050	0.100	0.227	0.167	0.225	0.317
		ST	0.050	0.100	0.226	0.067	0.113	0.198
		LSL	0.049	0.098	0.223	0.039	0.071	0.151
		MD	0.049	0.098	0.223	0.039	0.071	0.151
		LS	0.048	0.094	0.222	0.051	0.087	0.164
	900	SS	0.050	0.096	0.228	0.153	0.216	0.306
		ST	0.049	0.095	0.226	0.066	0.106	0.185
		LSL	0.048	0.094	0.223	0.040	0.070	0.142
		MD	0.049	0.095	0.224	0.040	0.070	0.142
		LS	0.047	0.098	0.230	0.052	0.084	0.163
	700	SS	0.048	0.092	0.219	0.157	0.218	0.315
		ST	0.048	0.091	0.216	0.068	0.106	0.191
		LSL	0.047	0.090	0.212	0.039	0.070	0.142
		MD	0.047	0.090	0.213	0.039	0.070	0.142
		LS	0.045	0.091	0.213	0.044	0.079	0.156

Table 4: Empirical familywise error rates at 5%, 10% and 25% nominal levels from four multiple testing methods under a multiplicity alternative model.

			Independent			Correlated		
m	m_0	method	fwe=0.050	0.100	0.250	fwe=0.050	0.100	0.250
100	100	SS	0.049	0.101	0.232	0.129	0.194	0.297
		ST	0.048	0.098	0.229	0.073	0.121	0.221
		LSL	0.046	0.096	0.221	0.047	0.084	0.173
		MD	0.046	0.096	0.222	0.048	0.085	0.174
		LS	0.049	0.099	0.232	0.058	0.098	0.198
	90	SS	0.053	0.101	0.235	0.128	0.185	0.284
		ST	0.051	0.097	0.227	0.072	0.121	0.219
		LSL	0.048	0.093	0.216	0.045	0.081	0.170
		MD	0.049	0.095	0.220	0.045	0.083	0.173
		LS	0.053	0.105	0.240	0.057	0.097	0.193
	70	SS	0.053	0.098	0.235	0.122	0.183	0.286
		ST	0.049	0.092	0.224	0.074	0.122	0.221
		LSL	0.046	0.086	0.210	0.051	0.085	0.171
		MD	0.047	0.087	0.215	0.052	0.086	0.175
		LS	0.052	0.097	0.215	0.052	0.098	0.191
1000	1000	SS	0.056	0.105	0.233	0.166	0.223	0.312
		ST	0.055	0.104	0.232	0.068	0.109	0.191
		LSL	0.054	0.103	0.229	0.038	0.072	0.143
		MD	0.054	0.103	0.229	0.038	0.072	0.143
		LS	0.048	0.094	0.222	0.051	0.087	0.164
	900	SS	0.049	0.096	0.225	0.163	0.220	0.314
		ST	0.049	0.095	0.222	0.068	0.108	0.189
		LSL	0.047	0.093	0.220	0.042	0.072	0.144
		MD	0.047	0.093	0.220	0.042	0.072	0.144
		LS	0.048	0.093	0.219	0.048	0.080	0.159
	700	SS	0.047	0.094	0.231	0.156	0.216	0.314
		ST	0.046	0.092	0.229	0.070	0.109	0.195
		LSL	0.044	0.089	0.220	0.042	0.073	0.146
		MD	0.045	0.090	0.220	0.042	0.073	0.146
		LS	0.043	0.086	0.211	0.044	0.075	0.151

Table 5: Estimated number of true null hypotheses and number of rejections in m=705 genes of FWE-controlled procedure and FDR-controlled procedure at various levels.

Estimation of m_0		Method				
	m	SS	ST	LSL	MD	LS
	705	410	432	524	522	509

FWE-controlled procedure			No. of rejections : r				
	Bonf.	Resamp.	SS	ST	LSL	MD	LS
FWE : α_0=0.05	43	50	51	51	47	47	48
0.10	51	56	56	55	53	53	53
0.25	57	76	66	65	62	62	62

BH FDR-controlled procedure		No. of rejections : r (Number of false rejections : $r \cdot q^*$)				
	BH	SS	ST	LSL	MD	LS
FDR : $q^* = 0.001$	42	51	51	46	46	47
	(0.042)	(0.051)	(0.051)	(0.046)	(0.046)	(0.047)
0.010	77	93	91	85	85	85
	(0.77)	(0.93)	(0.91)	(0.85)	(0.85)	(0.85)
0.050	147	191	183	172	172	172
	(7.35)	(9.55)	(9.15)	(8.6)	(8.6)	(8.6)

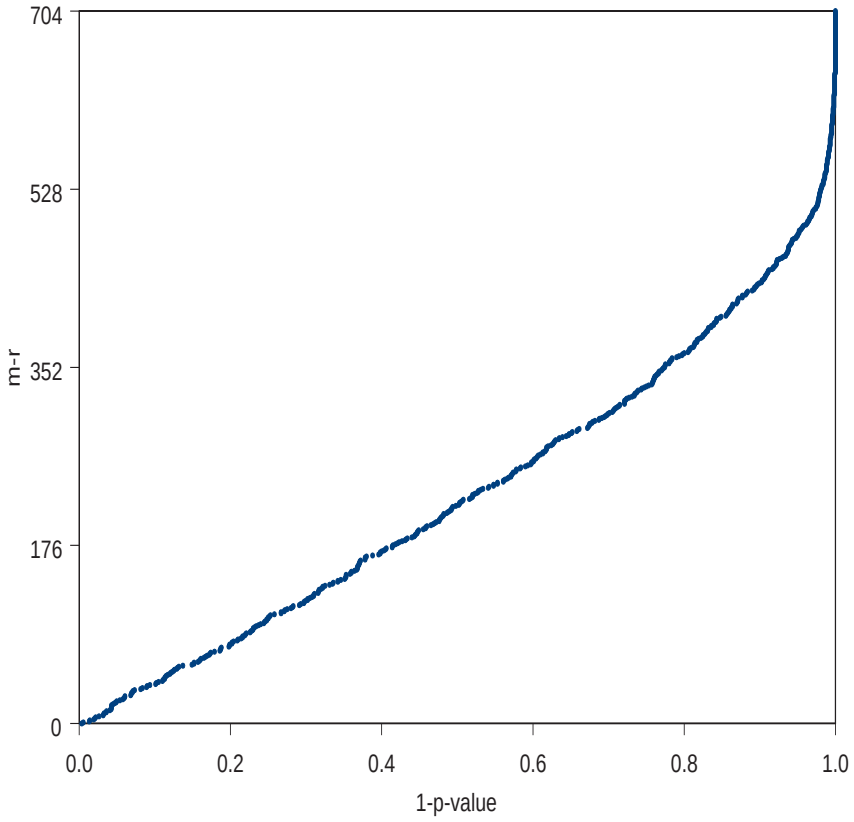

Figure 1: P-value plot of the example data.