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最大概似估計量

計畫類別：個別型計畫
計畫編號：NSC93-2118-M-004-003-
執行期間：93 年 08 月 01 日至 94 年 07 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 10 月 31 日

Maximum Likelihood Estimation of the Number of True Null Hypotheses in a Multiple Testing Problem

Huey-Miin Hsueh
Department of Statistics, National Cheng-Chi University, Taipei, Taiwan
hsueh@nccu.edu.tw

ABSTRACT

When there are many hypotheses to be tested, the risk of a false positive finding in the simultaneous inference increases severely. A multiple comparison procedure(MCP), which uses a conservative adjustment in significance level of individual test, is suggested for controlling a familywise error rate. However, the conservativeness becomes substantial when the number of hypotheses is large. If the number of true null hypotheses, m_0, is known, it can be used to improve the power for a MCP. This paper proposes two maximum likelihood estimators of m_0. One is based on the observed p-value under a mixture model. The estimator is found by the E-M algorithm. The second one is based on the number of rejection at some given level α. The two estimators are compared with the mean-difference method(MD) in a simulation study. The performance of the maximum likelihood estimators depends on the model assumption. When no violation of assumption presents, the maximum likelihood estimators are better than the MD method. The analysis of a real data set from a microarray experiment is given as an illustration.

1. INTRODUCTION

We consider the problems of simultaneously testing m null hypotheses. If the number of true null hypotheses,m_0, is known or its estimate is obtained, the knowledge can be used to improve a conventional MCP. For example, the Bonferroni-type MCP, in which all p-values are compared with the adjusted level α/m, can be shown to control its FWER at level α. If an overestimated estimate of m_0 exists, that is, $\hat{m}_0 \geq m_0$, the revised MCP which uses the cut-off $\alpha/\hat{m}_0$ has its FWER maintained at α by the Bonferroni inequality. Because a less strict cut-off value is employed, the revised MCP has a greater power. Such technique can be applied in an FDR-controlled MCP as well. For example, in the linear step-up procedure, which was developed by Benjamini and Hochberg (1995), the $i-$th ordered p-value is compared with $i\alpha/m$. Let

$$k = \max\{i : p_{(i)} \leq i\alpha/m\}.$$

The null hypotheses which are corresponding to $p_{(1)}, p_{(2)}, \cdots, p_{(k)}$ are rejected in the procedure. Benjamini and Hochberg(1995) show that the FDR of this procedure is bounded by $m_0\alpha/m \leq \alpha$. If a $\hat{m}_0 \leq m_0$ exists, the adaptive step-up procedure with the modified cut-off $i\alpha/\hat{m}_0$ becomes more powerful. Several adaptive procedures have been developed in recent literatures, see Benjamini and Hochberg(2000), Benjamini, Krieger and Yekutieli(2002).

In this paper, two maximum likelihood estimators(MLEs) based on different realization of the data are studied. The first one uses the full information of observed p-values and assumes a mixture model. The Expectation-Maximization algorithm is considered. The second one uses a summarized information of the data set, the number of rejections at a cut-off. The two methods will be described in Section 2. Section 3 will present the result of an empirical study. A real example will be analyzed in Section 4. Section 5 will give some concluding remarks.

2. MLE OF m_0

We consider testing m simultaneous one-sided hypotheses,

$$H_{0i} : \mu_i = 0, \text{ vs. } H_{1i} : \mu_i > 0, \ i = 1, \cdots, m.$$

Each null hypothesis H_{0i} is rejected if the corresponding test statistic, Z_i, is too large. Assume that the test statistics $Z_1, \cdots, Z_m$ are independently normally distributed,

$$Z_1, \cdot, Z_m \sim i.i.d.N(\Delta_i, 1), \tag{1}$$

where Δ_i is the true effective size of Z_i. $\Delta_i = 0$, if H_{0i} is true, $\Delta_i > 0$, otherwise. For simplicity, we assume $\Delta_1 = \cdots = \Delta_m$ in this paper. Further, let $I_i, J_i, i = 1, \cdots, m$ be the indicators of true null hypotheses and significance, that is, for $i = 1, \cdots, m$,

$$I_i = \begin{cases} 1, & \text{if } H_{0i} \text{ is true;} \\ 0, & \text{if } H_{0i} \text{ is false,} \end{cases}$$

and

$$J_i = \begin{cases} 1, & \text{if } H_{0i} \text{ is rejected;} \\ 0, & \text{if } H_{0i} \text{ is not rejected.} \end{cases}$$

Then $m_0 = \sum_{i=1}^m I_i$ is our target parameter.

2.1 P-VALUES-BASED MLE

Given observed $z_1, \cdots, z_m$, the p-values can be calculated by

$$p_i = P(Z \geq z_i) = 1 - \Phi(z_i), \;\; i = 1, \cdots, m,$$

where $\Phi(z)$ is the distribution of N(0,1). Under (1), p_i follows the distribution,

$$U(0,1) \cdot I_i + F(p) \cdot (1 - I_i),$$

where

$$F(p) = 1 - \Phi(\Phi^{-1}(1-p) - \Delta).$$

Under the assumption of independence, the log-likelihood based on the p-values is

$$L_p = \log\{\prod_{i=1}^m 1^{I_i} \cdot f(p_i)^{(1-I_i)}\}, \qquad (2)$$

where $f(p)$ is the density function corresponding to $F(p)$. In the likelihood, there are $m+1$ unknown parameters which includes $I_1, \cdots, I_m$ and Δ. With only m observations, the estimation is impossible. Thus a mixture model is considered to overcome the problem. In the mixture model, I_i's are no longer unknown parameters but are regarded as missing Bernoulli-distributed random variables, i.e.

$$I_1, \cdots, I_m \sim i.i.d.\ Bernoulli(\pi).$$

The success probability, π, is the proportion of true nulls. While an estimate of π, $\hat{\pi}$, is obtained, $m \times \hat{\pi}$ can be used to estimate m_0. Under the mixture model,$i = 1, \cdots, m$,

$$p_i | I_i \sim \; I_i U(0,1) + (1 - I_i) F(p), \;\; I_i \sim Bernoulli(\pi),$$

the p_i's follows

$$p_i \sim i.i.d.\ \pi \cdot U(0,1) + (1-\pi) \cdot F(p).$$

Using the Expectation-Maximization algorithm, the MLE of π, $\hat{\pi}$, can be solved and the p-value-based MLE of m_0 is then given by $m \times \hat{\pi}$.

2.2 R-BASED MLE

From literature review, we found that many proposed methods are based on the relationship between rejection number and the significance level, see Hsueh et al. (2003). The main advantage in using only the information of rejection number rather than the full information of p-values is to increase the robustness of the estimation. In this subsection, we consider the MLE based on the number of rejection, $R = \sum_{i=1}^m J_i$. Given a cut-off level α for individual test, the true state of nature and the result of the m hypotheses testing can be summarized by the following 2×2 table,

Table I. The result and probability of occurrence in testing m hypotheses.

	Conclusion		Total
	Significant	Non-Significant	
H_0 is True	V	S	m_0
	(α)	$(1-\alpha)$	
H_1 is True	U	T	m_1
	$(1-\beta)$	(β)	
Total	R	$m-R$	m

The quantities $V = \sum_{i=1}^m I_i J_i, S = \sum_{i=1}^m I_i(1 - J_i), U = \sum_{i=1}^m (1-I_i)J_i$ and $T = \sum_{i=1}^m (1-I_i)(1-J_i)$ involve the unknown indicators I_i's and are unobservable. Only the marginal column total, $R = \sum_{i=1}^m J_i, m - R = \sum_{i=1}^m (1 - J_i)$, are observable. Here R is the number of rejection. Given m_0 and m_1, V and U have binomial distributions with "success" probabilities α and $(1-\beta)$, respectively. The probability $(1-\beta)$ is exactly the power of the test and depends on the level α and the true effective size Δ. Explicitly,

$$1 - \beta = 1 - \Phi(z_\alpha - \Delta),$$

where z_α is the $100(1-\alpha)-$th percentile of N(0,1). The probability function of R can be consequently derived by the convolution formula. Given the observed rejection number $R = r$, the likelihood function is

$$L_R \;\; = \;\; \sum_{v=v_L}^{v_U} \binom{m_0}{v} \alpha^v (1-\alpha)^{(m_0 - v)} \cdot \binom{m_1}{r-v} (1-\beta)^{(r-v)} \beta^{(m_1 - r + v)},$$

where $v_L = max(0, r - m_1), v_U = min(m_0, r)$. To

easy the computation, the power is estimated by

$$(1-\tilde{\beta}) \quad = \quad \frac{\sum_{i=1}^{m}\{1-\Phi(z_\alpha-\tilde{\Delta}_i)J_i\}}{\sum_{i=1}^{m}J_i} \\ \frac{\sum_{i=1}^{m}\{1-\Phi(z_\alpha-z_i)J_i\}}{\sum_{i=1}^{m}J_i}$$

in advance. The estimate is then plugged in the likelihood function, L_R. For a conservative estimate, the value, $\hat{m}_0$, which satisfies

$$L_R(m_0) \leq \cdots \leq L_R(\hat{m}_0+1) \leq L_R(\hat{m}_0) \geq L_R(\hat{m}_0-1),$$

is considered as the R-based MLE.

3. A SIMULATION STUDY

In the article of Hsueh et. al. (2003), several estimators for m_0 were compared. The authors concluded that the mean-difference method(MD) is the preferred method. The MD method is derived by the fact that, among the i.i.d. true null distributions, the differences between each adjacent ordered p-values are i.i.d. beta-distributed random variates with mean $1/(m_0+1)$. The estimate of m_0 is inversely proportional to the sample mean of differences. Here the two MLEs are compared with the MD estimate through an empirical study.

We consider testing $m = 500, 1000$ right-tailed hypotheses on the means of normal populations,

$$H_0 : \mu = 0 \ \text{ versus } \ H_1 : \mu > 0,$$

The number of true null hypotheses is $m_0 = m\pi$. Three choices for true null proportion, $\pi = 0.5, 0.7, 0.9$, are considered. The Z-test is considered and generated from the following distribution,

$$Z_i \sim \begin{cases} N(0,1) & \text{if } H_{0i} \text{ is true;} \\ N(3.168,1) & \text{if } H_{0i} \text{ is not true.} \end{cases}$$

The effective size $\Delta = 3.168$ is determined by achieving power 80% at level 1% for individual test. In additional to the independent case, an equi-correlated case is also simulated. According to the true state, all tests are classified into two groups, null group and non-null group. Any two test statistics are correlated as long as they are in the same group. A moderate correlation coefficient, $\rho = \sqrt{0.2} = 0.4472$, is used. We use a mixture model in the p-values-based MLE. To study the influence of this model assumption, the mixture model is also considered in the simulation. In the fixed nulls model, the total number of true nulls, $m_0 = \sum_{i=1}^{m} I_i$, are kept fixed and equal to $m \times \pi$. In the mixture model,

$$I_i \sim \ i.i.d. \ Bernoulli(\pi)$$

and thus m_0 varies in every replication. The empirical mean and standard deviation of each estimator are calculated from $1,000$ replications. As stated in previous section, the estimate can be used to improve an MCP. To be conservative, overestimation is preferred. Thus the proportion of overestimating within $1,000$ replications are also calculated and denoted by $Overest$.

Table 1 and Table 2 are the results of independent case and Table 3 and Table 4 are of dependent case. In the independent case, one can find that the P-values-based MLE has the best performance in the sense that it has least bias and variation. However, there is about a half chance of underestimation due to the consistency of this estimator. Three levels are used for the R-based MLE. Obviously, the performance of this method depends on the level. With an adequate level, this method can be comparable with the P-values-based MLE. On the other hand, a loose cut-off, for example $\alpha = 5\%$ here, may lead to unsatisfactory results.

To study the robustness of these methods, dependent data are generated and the results are presented in Table 3 and Table 4. From these tables, one observes that all methods are affected. Especially, the moderate dependency increases their variations. The P-values-based MLE is least robust than other methods. Because using full information of the data, this method strongly relies on the distributional assumptions. In general, the difference in the results between a fixed nulls model and a mixture model is minor. The MD method always overestimates the true value.

4. AN EXAMPLE

The example data is from a triple flip-dye microarray experiment, see Kerr *et al.*, Martinez *et al.*. The compound 2,3,7,8-tetrachlordibenzo-*p*-dioxin(TCDD) was studied. The expression of 1907 genes from TCDD-treated(T) and control(C) cell lines were compared. 6 arrays were conducted in the experiment. Kerr *et al.*(2002) in their paper considered the analysis of variance (ANOVA) method to identify the variations from different sources. Several models were used and compared. Using their model (2), the target parameters were the differences of the interactions effects between treatment and control group for each gene, $(VG)_{Tg} - (VG)_{Cg}$, $k = T, C, g = 1, \cdots, 1907$. Using partial mean estimates, the other effects in the model were estimated and subtracted. The residuals then approximately

follow the model,

$$e_{kgl} \approx (VG)_{kg} + \epsilon_{kgl}, \quad k = T, C; \; g = 1, \cdots, 1907; \; l = 1, \cdots, 6.$$

The differential expressive genes, which may be either up-regular or down-regular expressed, were to be found. The two-sided alternative hypotheses were of interest and tested.

Under the assumptions of normality and constant variance, the two-sample t-test statistic is considered in a parametric analysis. The p-values and the common critical values are determined by a t-distribution with ten degrees of freedom. The t-test statistic has a non-central t-distribution under the non-null distribution. On the other hand, its power function can be derived through the joint distribution of $(\bar{X}, S)$ from a normal population distribution. For some critical value $t > 0$, at μ_0,

$$(1-\beta) \quad = \quad E\left\{1 - \Phi(t\frac{S}{\sigma} - \Delta) + \Phi(-t\frac{S}{\sigma} - \Delta)\right\},$$

where $Z \sim N(0,1), \Delta = \mu_0/(\sigma/\sqrt{n})$. Given a significance level α, and using S to estimate σ, the power is estimated by

$$(1-\tilde{\beta}) = \frac{\sum_{i=1}^{1907}\{1 - \Phi(t_{\alpha/2} - t_i) + \Phi(-t_{\alpha/2} - t_i)\}J_i}{\sum_{i=1}^{1907} J_i}.$$

Here J_i is the indicator of the significance of the $i-$th test and $t_{(p)}$ is the $100(1-p)^{th}$ percentile of t-distribution with 10 degrees of freedom. In additional to the parametric tests, nonparametric p-values were also obtained by 10,000 permutations. However, the difference between the parametric and non-parametric p-values are limited. In Figure 1, the X-axis is the parametric p-value and the Y-axis is the nonparametric p-value of every gene. The points mostly cluster around the 45^0 line. We conclude that the marginal normality assumption is quite reasonable.

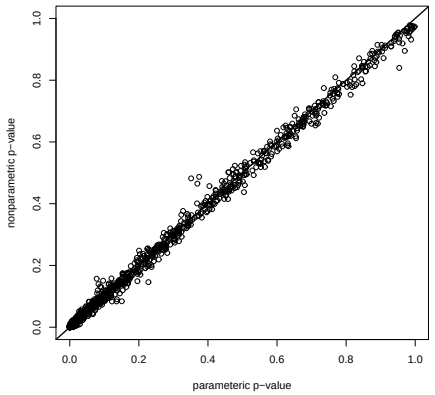

Fig 1. The parametric versus nonparametric p-values.

All pairwise correlations between the 1,907 genes are calculated. The following table is the summary statistics of these correlations and Figure 2 is the kernel fit of these correlations.

Table II. Summary statistics of the pairwise correlations

Min	Q1	Q2	Mean	Q3	Max
-0.95	-0.06	0.30	0.26	0.61	1.00

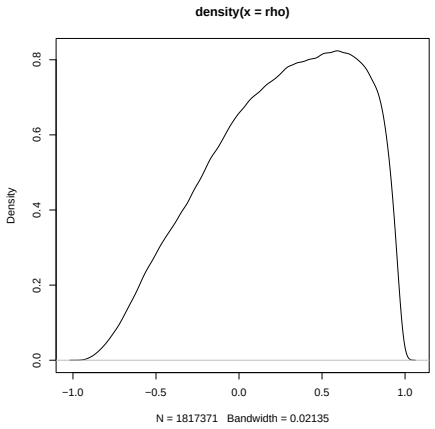

Fig 2. The kernel fit of all pairwise correlations.

From Table II, a quarter of the correlation are greater than 0.61. We conclude that severe correlation exists between these tests. From Section 3, the P-values-based MLE is found to be sensitive to the dependency. Thus this method is not applied to the current data set.

Table III gives the resulting parametric and nonparametric estimates of m_0 at different α levels. No difference between the parametric method and the

nonparametric method. The R-based MLE of m_0 at levels, $1\%, 5\%$ and 10% are solved. Using a strict level, $\alpha = 1\%$, a hypothesis tends to be classified as true null and an overestimating result, $\hat{m}_0 = 865$, is obtained. On the contrary, if one uses a loose level, $\alpha = 10\%$, underestimation likely occurs.

Table III. Estimate of m_0 for NIEHS data set

Method	Parametric	Nonparametric
MD	740	740
R-based MLE		
$\alpha = 1\%$	865	865
$\alpha = 5\%$	642	642
$\alpha = 10\%$	560	560

5. CONCLUDING REMARKS

In this paper, we propose two maximum likelihood estimators, P-values-based and R-based, for the number of true null hypotheses, m_0. The independence and constant effective size are assumed. Through empirical studies, the methods performs well when there is no violation of assumption. Because using more information in the data set, the P-values-based MLE is more sensitive to the model assumption. The R-based MLE depends on the level chosen for individual test. With an adequate level, this method, which uses only summarized information, can be comparable with the P-values-based MLE. In general, a strict level results in a conservative estimate. In an exploratory analysis, this estimate can provide useful knowledge about the current data set. A liberal estimate may be reasonable to prevent any false negative finding in this initial screening stage. On the other hand, this estimate can be considered to improve the statistical power of a conventional MCP. To be conservative, an overestimate is more adequate.

REFERENCE

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289-300.

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, **25**, 60-83.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2002) Adaptive linear step-up false discovery rate controlling procedures. Draft.

Hsueh, H., Chen, J.J. and Kodell, R.L. (2003) Comparison of methods for estimating number of true null hypothesis in multiplicity testing. *Journal of Biopharmaceutical Statistics*, **13**, 675-689.

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J., and Churchill, G.A. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, **12**(1), 203-217.

Martinez, J. M., Afshari, C. A., Bushel, P. R., Masuda, A., Takahashi, T. and Walker, N. J.(2002) Differential toxicogenomic responses to 2,3,7,8-Tetrachloroethane-*p*-dioxin in malignant and nonmalignant human airway epithelial cells. *Toxicological sciences*, **69**, 409-423.

Table 1. $m = 500, \gamma_0 = 3.1680$, independent case

			Fixed nulls			Mixture model		
m_0			overest	mean	s.d.	overest	mean	s.d.
250	MD		996	270.9	22.5	994	271.2	22.4
	MLE :P-based		490	249.2	5.9	485	249.7	12.6
	R-based	$\alpha = 0.005$	997	274.8	26.5	997	275.5	26.9
		0.010	821	257.5	11.3	837	258.4	11.8
		0.050	17	233.9	17.6	11	234.1	17.7
350	MD		999	368.4	19.7	995	368.9	19.9
	MLE : P-based		498	349.3	5.8	473	349.4	11.5
	R-based	$\alpha = 0.005$	981	365.1	16.8	965	365.0	16.6
		0.010	762	354.5	8.2	746	354.7	8.3
		0.050	50	337.4	14.5	47	337.6	14.7
450	MD		982	459.3	10.4	984	459.4	10.3
	MLE :P-based		485	448.9	5.0	456	448.9	8.1
	R-based	$\alpha = 0.005$	858	454.6	6.6	864	454.9	6.8
		0.010	592	450.6	4.5	582	450.7	4.7
		0.050	132	441.9	10.6	125	441.9	10.7

Table 2. $m = 1000, \gamma_0 = 3.1680$, independent case

			Fixed nulls			Mixture model		
m_0			overest	mean	s.d.	overest	mean	s.d.
500	MD		1000	550.4	52.6	1000	549.4	51.1
	MLE : P-based		483	499.2	8.3	483	499.4	8.6
	R-based	$\alpha = 0.005$	1000	550.0	51.7	1000	550.0	51.4
		0.010	910	515.7	19.9	895	515.6	19.7
		0.050	2	467.5	34.0	1	467.5	34.4
700	MD		1000	742.6	44.3	1000	742.8	44.2
	MLE : P-based		487	699.0	8.2	478	699.5	8.4
	R-based	$\alpha = 0.005$	997	729.1	30.9	996	729.7	31.5
		0.010	809	707.9	12.6	815	708.5	12.9
		0.050	4	675.1	26.9	12	675.3	26.9
900	MD		999	921.9	23.0	997	921.7	23.3
	MLE : P-based		490	898.7	7.4	489	898.4	7.7
	R-based	$\alpha = 0.005$	929	909.0	11.1	915	908.5	11.2
		0.010	599	901.1	6.5	584	900.3	6.9
		0.050	66	884.2	18.8	78	883.8	19.1

Table 3. $m = 500, \gamma_0 = 3.1680$, dependent case.

			Fixed nulls			Mixture model		
m_0			overest	mean	s.d.	overest	mean	s.d.
250	MD		857	266.1	35.5	872	269.8	33.9
	MLE: P-based		469	233.7	44.8	498	238.1	42.3
	R-based	$\alpha = 0.005$	701	276.4	46.4	713	277.6	47.7
		0.010	508	258.4	31.2	512	260.2	32.9
		0.050	152	232.3	29.4	170	235.3	27.0
350	MD		846	363.5	32.2	868	364.0	31.8
	MLE: P-based		512	328.2	66.9	500	330.5	60.7
	R-based	$\alpha = 0.005$	676	364.4	28.2	720	364.7	27.6
		0.010	491	354.2	21.3	500	353.7	20.2
		0.050	383	336.5	31.9	368	336.6	31.0
450	MD		85	455.4	29.0	842	453.9	30.7
	MLE: P-based		58	406.7	134.6	582	406.8	135.4
	R-based	$\alpha = 0.005$	714	455.0	12.3	732	454.1	11.9
		0.010	615	450.7	13.3	623	449.9	13.8
		0.050	553	442.0	36.1	544	440.0	38.9

Table 4. $m = 1000, \gamma_0 = 3.1680$, dependent case.

			Fixed nulls			Mixture model		
m_0			overest	mean	s.d.	overest	mean	s.d.
500	MD		880	540.3	68.5	879	541.8	65.4
	MLE: P-based		457	470.9	84.0	461	471.1	83.2
	R-based	$\alpha = 0.005$	678	547.6	87.6	689	546.4	86.5
		0.010	475	513.4	58.4	453	512.6	57.2
		0.050	127	464.9	55.7	110	464.8	54.8
700	MD		885	735.3	62.5	900	733.6	64.5
	MLE: P-based		512	662.9	118.5	525	659.3	128.9
	R-based	$\alpha = 0.005$	693	728.4	54.5	694	728.6	53.2
		0.010	487	707.4	40.0	490	706.8	38.3
		0.050	386	673.4	60.9	385	672.4	64.3
900	MD		872	912.2	53.7	890	916.7	38.7
	MLE: P-based		568	806.7	280.5	593	815.5	264.9
	R-based	$\alpha = 0.005$	686	907.6	22.0	706	908.0	20.2
		0.010	597	899.0	26.2	614	900.5	22.6
		0.050	537	878.7	78.6	557	883.9	69.0