

Distributional Property of the Generalized p -value for
the Behrens-Fisher Problem with Applications to
Multiple Testing

Kam-Wah Tsui and Shijie Tang

Department of Statistics

University of Wisconsin, Madison, WI 53706

email: kwtsui@stat.wisc.edu

October 31, 2005

Technical Report NO. 1111

Author's Footnote:

Kam-Wah Tsui is Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (email: kwtsui@stat.wisc.edu); and Shijie Tang is Ph.D Student, Department of Statistics, University of Wisconsin, Madison, WI 53706 (email: tangs@stat.wisc.edu). The authors thank Professor Kjell Doksum for useful suggestions, and thank Wei Zheng for providing the microarray data and helpful discussions.

Abstract

The generalized p -value method introduced by Tsui and Weerahandi (1989) has been used to successfully provide finite sample solutions for many hypothesis testing problems when no solutions are available using the usual approach. Although simulation studies show that generalized p -values have similar distributional properties as ordinary p -values, no one has yet developed theoretical properties of generalized p -values. This paper provides some theoretical distributional properties of the generalized p -value for the Behrens-Fisher problem of testing the difference of two independent normal distribution means with possibly unequal distributional variances. We show why the probability that the generalized p -value for the Behrens-Fisher problem is less than a given value r in $(0, 0.5)$, is approximately r . We use this theoretical result to obtain a multiple testing procedure based on generalized p -values, which controls the false discovery rate (FDR). In particular, we introduce a theoretically valid testing procedure that controls FDR in simultaneously testing several Behrens-Fisher problems. Our procedure to control FDR is an extension of the important work of Benjamini and Hochberg (1995), which is based on ordinary p -values. Our proof is constructive and concise and it shows clearly the essential step that makes the FDR result holds. It should be useful for development of further results. Data from a microarray experiment are used to illustrate our method.

KEYWORDS: Generalized p -value; Behrens-Fisher problem; False Discovery Rate

1. INTRODUCTION

Tsui and Weerahandi (1989) outline some basic steps to construct generalized p -values for hypothesis testing problems when nuisance parameters are present. Their generalized p -value method has been used to successfully provide finite sample solutions for many hypothesis testing problems when no solutions are available using the usual approach. Peterson and Weerahandi (2003) give a comprehensive review of the generalized p -value method. A detailed development and extensive studies and references appear in two books: Weerahandi (1995*b*) and Weerahandi (2004).

We first briefly review the basic rationale of the generalized p -value approach and then outline the presentation of our paper. Suppose we are interested in testing the null hypothesis, H_0 , about the parameter θ . Denote $\xi = (\theta, \eta)$, where η is a nuisance parameter. Let D represent a random sample for this testing problem and let d be a particular observed sample. The key idea in the generalized p -value method is to construct an appropriate generalized test variable, $W(D, d, \xi)$, a function of three components, to define an extreme region C consisting of all the samples D that are as extreme as the observed d . Usually, C is of the form $C = \{D : W(D, d, \xi) \geq 0\}$.

Given the observed sample d , the generalized p -value, $p(d)$, is defined to be the largest probability that a sample D is in C under the null hypothesis H_0 . That is,

$$p(d) = \sup_{\theta \in H_0} P_D(D \in C|\theta) = \sup_{\theta \in H_0} P_D(\{D : W(D, d, \xi) \geq 0\}|\theta), \quad (1.1)$$

where P_D denotes the probability with respect to the probability distribution of D .

By requiring that the generalized test variable W satisfy some general conditions, the generalized p -value, $p(d)$, in (1.1) is free of any parameters and can be used to test the hypothesis, H_0 . For example, one can reject H_0 if $p(d)$ is less than 0.05.

The expression for $p(d)$ is usually very complicated. Hence, analytic properties of $p(d)$ are not easy to obtain. However, computer simulations have been used to study the repeated sampling performance of $p(d)$ for many hypothesis testing problems. For example, in

repeated sampling of d under the null hypothesis, various computer simulation studies show the following result:

$$P_d(p(d) \leq r) \leq r, \text{ approximately,} \quad (1.2)$$

where P_d denotes the probability with respect to the probability distribution of d . However, no one has yet developed theoretical properties of generalized p -values. In this paper, we provide some theoretical distributional properties of the generalized p -value for the Behrens-Fisher problem of testing the difference of two independent normal distribution means with possibly unequal distributional variances, as described in Tsui and Weerahandi (1989). We explain why (1.2) is expected, analytically, in the Behrens-Fisher problem. We believe our result for the Behrens-Fisher problem provides the first and important step in understanding the theoretical properties of generalized p -values. We present this result in Section 2.

The problem of simultaneously testing a large number of hypotheses has generated a great amount of interest. Benjamini and Hochberg (1995) introduced the concept of False Discovery Rate (FDR), given in (3.1) and (3.2). Roughly speaking, FDR is defined to be the expected value of the ratio of the number of incorrectly rejected hypotheses and the total of number of rejected hypotheses. Assume a usual p -value is available for each hypothesis. Based on the p -values of the hypotheses, Benjamini and Hochberg (1995) provided a multiple testing procedure that guarantees the FDR to be less than or equal to a prefixed value q . In Section 3, we consider the problem of simultaneously testing many independent Behrens-Fisher problems using the corresponding generalized p -values of the hypotheses. Using the analytic result in Section 2, we obtain a multiple testing procedure based on generalized p -values, which controls FDR . As a special case, we have a theoretically valid testing procedure that controls FDR in simultaneously testing several Behrens-Fisher problems. Our procedure to control FDR is an extension of the important work of Benjamini and Hochberg (1995), which is based on ordinary p -values. Our proof is constructive and concise. It shows clearly the essential step that makes the FDR result holds. It should be useful for development of future results. In Section 4, we use a dataset from a microarray experiment

to illustrate our multiple testing procedure based on generalized p -values.

2. DISTRIBUTIONAL PROPERTY OF GENERALIZED P -VALUE FOR BEHRENS-FISHER PROBLEM

Denote $N(\mu, \sigma^2)$ to be a normal distribution with mean μ and variance σ^2 , χ_v^2 to be a Chi-square distribution with v degrees of freedom. Let $\{X_1, \dots, X_m\}$ and $\{Y_1, \dots, Y_n\}$ be two independent samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Here σ_1^2 and σ_2^2 are not necessarily equal. The problem of interest is to test the null hypothesis, $H_0 : \theta \equiv \mu_1 - \mu_2 \leq \delta_0$ against the alternative hypothesis, $H_1 : \theta = \mu_1 - \mu_2 > \delta_0$, for some fixed δ_0 . This testing problem is called the Behrens-Fisher problem. In this problem, $\eta = (\sigma_1^2, \sigma_2^2)$ is the nuisance parameter. Denote $\xi = (\theta, \eta)$. The sufficient statistics of this problem is $(\bar{X}, \bar{Y}, S_1^2, S_2^2)$ where

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_1^2 = \frac{1}{m} \sum_{i=1}^m (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2. \quad (2.1)$$

The probability distributions of the statistics are independent of one another and are given as follows:

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right), \quad \frac{mS_1^2}{\sigma_1^2} \sim \chi_{m-1}^2, \quad \frac{nS_2^2}{\sigma_2^2} \sim \chi_{n-1}^2. \quad (2.2)$$

In order to simplify our notation, we denote

$$D = (X_1, \dots, X_m, Y_1, \dots, Y_n), \quad d = (x_1, \dots, x_m, y_1, \dots, y_n). \quad (2.3)$$

Here d is the vector of the observed sample vector. Let $(\bar{x}, \bar{y}, s_1^2, s_2^2)$ be the observed value of the sufficient statistics $(\bar{X}, \bar{Y}, S_1^2, S_2^2)$. Note that in repeated sampling, $(\bar{x}, \bar{y}, s_1^2, s_2^2)$ follows the same probability distributions as in (2.2). Define

$$T(D, d, \xi) = (\bar{X} - \bar{Y} - \delta_0) \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \right)^{-\frac{1}{2}} \left[\frac{\sigma_1^2 s_1^2}{m S_1^2} + \frac{\sigma_2^2 s_2^2}{n S_2^2} \right]^{\frac{1}{2}}. \quad (2.4)$$

Then $T(d, d, \xi) = \bar{x} - \bar{y} - \delta_0$.

The generalized test variable for the above Behrens-Fisher problem, given in Tsui and Weerahandi (1989), can be written as $W(D, d, \xi) = T(D, d, \xi) - T(d, d, \xi)$.

The resulting extreme region $C = \{D : W(D, d, \xi) \geq 0\}$ describes the sample D that is as extreme or more extreme than the observed sample d . In this paper, we denote the cumulative distribution function of a t -distribution with v degrees of freedom to be Ψ_v . For a given observed sample d , the generalized p -value for the one-sided Behrens-Fisher problem is as described in (1.1), and it can be expressed as:

$$p(d) = E_B \left\{ \Psi_{m+n-2} \left[(\bar{y} - \bar{x} + \delta_0) \sqrt{\frac{m+n-2}{\frac{s_1^2}{B} + \frac{s_2^2}{1-B}}} \right] \right\}, \quad (2.5)$$

where $B \sim \text{beta}(\frac{m-1}{2}, \frac{n-1}{2})$.

The probability distribution of $p(d)$ and its corresponding cumulative probability distribution G are difficult to describe explicitly. However, extensive computer simulation indicate that $G(r) \approx r$ for $0 \leq r \leq 1$. That is, $p(d)$ appears to be approximately uniformly distributed on $[0, 1]$. Theorem 1 below gives two tight upper bounds, (2.6) and (2.7), for $G(\alpha)$.

Theorem 1. *For the one-sided Behrens-Fisher problem with $H_0 : \mu_1 - \mu_2 \leq \delta_0$, let $A = (\sigma_1^2/m)/(\sigma_1^2/m + \sigma_2^2/n)$. Then, the generalized p -value, $p(d)$, given in (2.5), has the following property under H_0 : for any $0 < r \leq 0.5$, the probability of the event $\{d : p(d) \leq r\}$ is bound above by:*

$$g(A) \equiv P\left(\Psi_{m+n-2} \left[z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \right] \leq r \right), \quad (2.6)$$

where z, C_{m-1} and C_{n-1} are independent random variables such that $z \sim N(0, 1)$, $C_{m-1} \sim \chi_{m-1}^2$, and $C_{n-1} \sim \chi_{n-1}^2$.

Furthermore, (2.6) is bounded above by

$$\Psi_{\min\{m-1, n-1\}}(\Psi_{m+n-2}^{-1}(r)), \quad (2.7)$$

where Ψ_{m+n-2}^{-1} is the inverse function of Ψ_{m+n-2} .

The following two lemmas, which are proved in the appendix, are needed for the proof of Theorem 1.

Lemma 1. *Let Ψ_v be the cumulative distribution function of a t -distribution with v degrees of freedom. Define*

$$f(b) = \Psi_v \left[z \sqrt{\frac{1}{\frac{t_1}{b} + \frac{t_2}{1-b}}} \right], \text{ for } b \in (0, 1)$$

Then for fixed $z \leq 0$, $f(b)$ is a convex function of b .

Lemma 2. *The function $g(A)$ defined in (2.6) is a convex function of A .*

Proof of Theorem 1: Without loss of generality, we can assume $\delta_0 = 0$. Denote

$$z = \frac{\bar{y} - \bar{x}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}, \quad C_{m-1} = \frac{ms_1^2}{\sigma_1^2}, \quad C_{n-1} = \frac{ns_2^2}{\sigma_2^2}.$$

Then (2.5) becomes

$$\begin{aligned} p(d) &= E_B \left\{ \Psi_{m+n-2} \left[\frac{\bar{y} - \bar{x}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sqrt{\frac{m+n-2}{\frac{C_{m-1}\sigma_1^2/m}{B(\sigma_1^2/m + \sigma_2^2/n)} + \frac{C_{n-1}\sigma_2^2/n}{(1-B)(\sigma_1^2/m + \sigma_2^2/n)}}} \right] \right\} \\ &= E_B \left\{ \Psi_{m+n-2} \left[z \sqrt{\frac{m+n-2}{\frac{AC_{m-1}}{B} + \frac{(1-A)C_{n-1}}{1-B}}} \right] \right\}, \end{aligned}$$

with $B \sim \text{beta}(\frac{m-1}{2}, \frac{n-1}{2})$.

For any $r < 0.5$ and $p(d) < r$, we must have $z < 0$. Hence by Lemma 1,

$$f(B) = \Psi_{m+n-2} \left[z \sqrt{\frac{m+n-2}{\frac{AC_{m-1}}{B} + \frac{(1-A)C_{n-1}}{1-B}}} \right]$$

is convex in B . By Jensen's inequality:

$$\begin{aligned}
p(d) &= E_B(f(B)) > f(E(B)) = f\left(\frac{m-1}{m+n-2}\right) \\
&= \Psi_{m+n-2} \left[z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \right] \equiv p_1(d)
\end{aligned}$$

Now observe that under $\mu_1 - \mu_2 = 0$, $z \sim N(0, 1)$, $C_{m-1} \sim \chi_{m-1}^2$, $C_{n-1} \sim \chi_{n-1}^2$, and that z , C_{m-1} , C_{n-1} are independent of one another. For $0 < r \leq 0.5$,

$$\begin{aligned}
P_d(\{d : p(d) \leq r\}) &\leq P_d(\{d : p_1(d) \leq r\}) \\
&= P(\Psi_{m+n-2} \left[z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \right] \leq r)
\end{aligned}$$

as given in (2.6).

Next, by Lemma 2, for $0 < r \leq 0.5$, $g(A)$ is convex in A . We have,

$$\begin{aligned}
g(A) &< \max\{g(0), g(1)\} \\
&= \max \left\{ P(\Psi_{m+n-2} \left[Z \sqrt{\frac{1}{\frac{C_{m-1}}{m-1}}} \right] \leq r), P(\Psi_{m+n-2} \left[Z \sqrt{\frac{1}{\frac{C_{n-1}}{n-1}}} \right] \leq r) \right\} \\
&= \Psi_{\min\{m-1, n-1\}}(\Psi_{m+n-2}^{-1}(r))
\end{aligned}$$

which is the expression given in (2.7).

Remark If $\sigma_1^2/\sigma_2^2 = (m^2 - m)/(n^2 - n)$, or equivalently, $A = (m-1)/(m+n-2)$, then $g(A)$ is exactly equal to r . The other upper bound, (2.7), which is free of parameter, is obtained based on the convexity of $g(A)$ and the resulting inequality $g(A) < \max\{g(0), g(1)\}$. Hence (2.7) is a conservative upper bound because it is achieved only when $\sigma_1 = 0$ or $\sigma_2 = 0$. Nevertheless it is very close to r for moderately large sample sizes m and n . These results are of significantly interest, because they show why the probability of the event $\{d : p(d) \leq r\}$ is approximately r , analytically.

For the two-sided Behrens-Fisher problem $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$, the

extreme region can be defined as:

$$C = \{D : |T(D, d, \xi)| - |T(d, d, \xi)| \geq 0\}$$

The generalized p -value can be shown to be:

$$p_2(d) = P(D \in C | \mu_1 - \mu_2 = 0) = 2E_B \left\{ \Psi_{m+n-2} \left[-|\bar{y} - \bar{x}| \sqrt{\frac{m+n-2}{\frac{s_1^2}{B} + \frac{s_2^2}{1-B}}} \right] \right\}, \quad (2.8)$$

where $B \sim \text{beta}(\frac{m-1}{2}, \frac{n-1}{2})$.

A theoretical property of (2.8) is given in Theorem 2 below.

Theorem 2. *Under the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$, the generalized p -value for the two-sided Behrens-Fisher problem given by (2.8) satisfies the following inequalities:*

$$P_d(\{d : p_2(d) \leq r\}) \leq 2P(\Psi_{m+n-2} \left[z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \right] \leq \frac{r}{2}) \quad (2.9)$$

$$< 2\Psi_{\min\{m-1, n-1\}}(\Psi_{m+n-2}^{-1}(\frac{r}{2})) \quad (2.10)$$

for arbitrary $r \leq 1$, where z , C_{m-1} , C_{n-1} are as given in Theorem 1.

Proof: Let $p(d)$ be as given in (2.5). Observe that, for $c \leq 0.5$, $P(p(d) \leq c) = P(p(d) \geq 1 - c)$. Then (2.8) becomes:

$$p_2(d) = \begin{cases} 2p(d) & \text{if } p(d) \leq 0.5; \\ 2 - 2p(d) & \text{if } p(d) > 0.5. \end{cases}$$

Under H_0

$$\begin{aligned}
P(\{d : p_2(d) \leq r\}) &= P(p(d) \leq \frac{r}{2} \text{ or } p(d) \geq 1 - \frac{r}{2}) \\
&= P(p(d) \leq \frac{r}{2}) + P(p(d) \geq 1 - \frac{r}{2}) \\
&= 2P(p(d) \leq \frac{r}{2}) \\
&\leq 2P(\Psi_{m+n-2} \left[Z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \right] \leq \frac{r}{2}) \\
&< 2\Psi_{\min\{m-1, n-1\}}(\Psi_{m+n-2}^{-1}(\frac{r}{2}))
\end{aligned}$$

The last two inequalities follows from the results of Theorem 1.

3. MULTIPLE TESTING PROBLEMS USING GENERALIZED P -VALUES

Benjamini and Hochberg (1995) introduce the concept of False Discovery Rate in testing simultaneously several independent null hypotheses. They propose a procedure that is based on the availability of a p -value in testing each individual hypothesis. Their procedure guarantees that the FDR cannot exceed a prefixed rate q . In this section, we extend the multiple testing problem to the case when ordinary p -values are not available, but generalized p -value can be constructed. One such example is simultaneously testing several independent Behrens-fisher problems, which appear to be useful in many applications.

Let H_1, \dots, H_m be m independent hypotheses to be tested. Let R_T be the number of true hypotheses that are incorrectly rejected, and let R_N be the number of 'not true' hypotheses that are rejected. The total number of hypotheses rejected is $R = R_T + R_N$. Define Q to be the false discovery proportion:

$$Q = \begin{cases} \frac{R_T}{R_T + R_N} = \frac{R_T}{R} & \text{if } R \neq 0; \\ 0 & \text{if } R = 0. \end{cases} \quad (3.1)$$

Benjamini and Hochberg (1995) define the FDR to be expected value of Q .

$$FDR = E(Q) \tag{3.2}$$

The following theorem describes a procedure that controls the FDR in testing several independent null hypotheses based on generalized p -values.

Theorem 3. *Assume that H_1, \dots, H_m are independent null hypotheses to be tested. Let p_1, \dots, p_m be the corresponding generalized p -values. For a given $q > 0$, suppose that there exists a cumulative distribution function F_i such that:*

$$P(p_i \leq r | H_i) \leq F_i(r), \quad \text{for } r \leq r_0, \tag{3.3}$$

for some r_0 satisfying $q \leq F_i(r_0)$, assuming that H_i is a true hypothesis. Denote $p_i^* = F_i(p_i)$, for $i = 1, \dots, m$. Let $p_{(1)}^* \leq p_{(2)}^* \leq \dots \leq p_{(m)}^*$ be the ordered values of p_i^* 's, and let $H_{(1)}, \dots, H_{(m)}$ be the corresponding hypotheses. Define $q_i = \frac{iq}{m}$ and

$$k^* = \max\{i : p_{(i)}^* \leq q_i\} \tag{3.4}$$

Then, the procedure that rejects $H_{(i)}$ for $i \leq k^*$ guarantees that $FDR \leq q$.

Proof: Let m_0 be the number of H_i 's that are the true null hypotheses. Without loss of generality, we can assume the hypotheses H_1, \dots, H_{m_0} to be the true hypotheses and H_{m_0+1}, \dots, H_m to be the hypotheses that are not true.

Denote $D_i = 1$, if H_i is rejected; $D_i = 0$, otherwise, for each $i = 1, \dots, m$. Let $T_i = \max\{l : \text{at least } l-1 \text{ of the indexes } j, j \neq i, \text{ such that } p_j^* \leq q_l\}$. Notice that the events $\{p_i^* \leq q_k, T_i = k\}$ and $\{D_i = 1, R = k\}$ are equivalent. Then

$$\begin{aligned}
FDR &= E(Q) = E\left(\frac{\sum_{i=1}^{m_0} D_i}{R}\right) \\
&= \sum_{k=1}^m E\left(\frac{\sum_{i=1}^{m_0} D_i}{R} \mid R = k\right) P(R = k) \\
&= \sum_{k=1}^m \sum_{i=1}^{m_0} E\left(\frac{D_i}{R} \mid R = k\right) P(R = k) \\
&= \sum_{k=1}^m \sum_{i=1}^{m_0} E\left(\frac{D_i}{R} \mid D_i = 1, R = k\right) P(D_i = 1, R = k) \\
&= \sum_{k=1}^m \sum_{i=1}^{m_0} \frac{1}{k} P(D_i = 1, R = k) \\
&= \sum_{i=1}^{m_0} \sum_{k=1}^m \frac{1}{k} P(p_i^* \leq q_k, T_i = k) \\
&= \sum_{i=1}^{m_0} \sum_{k=1}^m \frac{1}{k} P(p_i^* \leq q_k) P(T_i = k \mid p_i^* \leq q_k). \tag{3.5}
\end{aligned}$$

Since $q_k \leq q \leq F_i(r_0)$, we have $F_i^{-1}(q_k) \leq r_0$. So for $i \leq m_0$, we have,

$$P(p_i^* \leq q_k) = P(F_i(p_i) \leq q_k) \leq F(F^{-1}(q_k)) = q_k. \tag{3.6}$$

From (3.5),

$$FDR \leq \frac{q}{m} \sum_{i=1}^{m_0} \sum_{k=1}^m P(T_i = k \mid p_i^* \leq q_k). \tag{3.7}$$

If the generalized p -values are independent, then $P(T_i = k \mid p_i^* \leq q_k) = P(T_i = k)$ From (3.7),

$$FDR \leq \frac{q}{m} \sum_{i=1}^{m_0} \sum_{k=1}^m P(T_i = k) = \frac{q}{m} \sum_{i=1}^{m_0} 1 = \frac{m_0 q}{m} \leq q.$$

Remark From Theorem 1 in Section 2, the generalized p -value for the Behrens-Fisher problem given in (2.5) satisfies condition (3.3) in Theorem 3 if we use $F(r) = \Psi_{\min\{m-1, n-1\}}^{-1}(\Psi_{m+n-2}^{-1}(r))$ and $r_0 = 0.5$. If p_i has a uniform distribution on $[0, 1]$ under the null hypotheses H_i , then we have $F(r) \equiv r$, and Theorem 3 reduces to the seminal result given in Benjamini and Hochberg (1995). From the remark of Theorem 1, (3.3) is expected to hold with $F_i(r) = r$ for a large range of values of σ_1^2/σ_2^2 .

4. MULTIPLE TESTING OF A MICROARRAY EXPERIMENT

Dr. Michael Culbertson's lab at the University of Wisconsin Madison conducted a microarray experiment to identify the individual gene of mutant yeast strains that has higher expression level than that of the wild-type yeast strains (Lelivelt and Culbertson 1999). Each array provides measurements representing individual gene expression levels in the yeast genome. There are 15 arrays in the experiment: $m = 3$ of the arrays are replicates for the wild type yeast strains and $n = 12$ arrays are replicates of the mutant yeast strains. From another independent study (manuscript in preparation), Dr. Culbertson's lab identified 435 genes that are likely to be differentially expressed, we call this group of genes the suspicious group. In addition, from studies of Dr. Culbertson's lab and other researchers in this area (He, Li, Spatrnick, Casillo, Dong and Jacobson 2003), 600 genes are strongly believed to be equally expressed for the mutant yeast stains and the wild-type yeast strains. We call this group of equally expressed genes as the control group. Using this preliminary biological information about the suspicious group and the control group, we examine the performance of our proposed multiple Behrens-Fisher testing procedure based on the generalized p -values given in (2.5). We denote (1) GP to be the procedure using (2.5) with $F_i(r)$ given by (2.7), (2) GPU to be the procedure using (2.5) with $F_i(r) = r$. We also compare the performance with the multiple testing procedure that uses the familiar Welch's t-test (Welch 1938). Using the notations in Section 2, (2.1), (2.2), (2.3), the Welch's t-test uses the test statistic:

$$T(D) = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/m + S_2^2/n}}, \quad (4.1)$$

with the observed values:

$$T(d) = \frac{\bar{x} - \bar{y}}{\sqrt{s_1^2/m + s_2^2/n}}. \quad (4.2)$$

The p -value for the one-sided null hypothesis $H_0 : \mu_1 - \mu_2 \leq 0$, is

$$p_w(d) = \sup_{\theta \in H_0} P(T(D) \geq T(d) | \theta) \quad (4.3)$$

where $T(D)$ has a t-distribution with degrees of freedom:

$$d.f. = \frac{(s_1^2/m + s_2^2/n)^2}{(s_1^2/m)^2/(m-1) + (s_2^2/n)^2/(n-1)}$$

We denote WT to be the multiple testing procedure that uses p -values defined in (4.1), (4.2), (4.3) with $F_i(r) = r$. We calculate the one-sided generalized p -value using the equation (2.5) with $\delta_0 = 0$, and the one-sided Welch's test p -value using (4.3), (4.2) and (4.1), for the suspicious group and the control group. For each gene, we test whether we should reject the hypothesis that expression level of the mutant yeast strains is no more than that of the wild type yeast strains. We denote the number of rejection to be n_g , n_w and n_{gu} for the procedures GP, WT and GPU , respectively.

For various values of FDR , q , Table 1 below reports the values of n_g , n_w and n_{gu} for the suspicious group, the control group, and the combined group. The results in the table show that GP and GPU are more conservative than WT .

Table 1: n_w : number of rejection by WT procedure, n_g : number of rejection by GP procedure, n_{gu} : number of rejection by GPU procedure.

q	0.01	0.05	0.2	0.3	0.5
n_w :suspicious group	144	217	326	366	403
n_g :suspicious group	0	0	203	343	402
n_{gu} :suspicious group	16	169	314	358	402
n_w :control group	3	5	13	15	23
n_g :control group	0	0	0	0	0
n_{gu} :control group	0	0	0	5	9
n_w :combined group	116	191	324	389	525
n_g :combined group	0	0	0	0	388
n_{gu} :combined group	4	97	251	346	495

For the control group, both the GP and the GPU procedure do not reject any hypothesis for FDR controlled to be under $q = 0.2$. In fact, the GP procedure does not reject any hypothesis even for $q \leq 0.5$. This result agrees with the result of biological research cited.

On the other hand, for the suspicious group, both *GPU* and *WT* indicate that, for most of genes, the mean expression level of the mutant yeast strains is higher than that of the wild-type yeast strains.

Figure 1 displays the first 200 sorted generalized p -values for the *GPU* procedure, and the first 200 sorted p -values for the *WT* procedure for the genes in the combined group, with $q = 0.05$. The straight lines are plotted according to $f(i) = iq/mk$. Based on the *GPU* and the *WT* procedure, the hypotheses corresponding to the generalized p -values or p -values that are to the left of the intersection of the straight line and the dotted line are to be rejected. Figure 1 explains why we should be cautious in using the *WT* procedure to control *FDR*. For $q = 0.05$, the *WT* procedure rejects 191 hypotheses in the combined group, 16 of which are in control group. Presumably, at least 16 hypothesis was wrongly rejected. That is, the actual false discovery proportion $Q \geq 16/191 \approx 0.08$. Moreover, if $q = 0.01$, the *WT* procedure rejects 116 hypotheses, 7 of which are in control group. Thus, $Q \geq 7/116 \approx 0.06$. For the *GPU* and the *GP* procedures, we always have $Q \leq q$, for any values of q considered.

One reason that the *WT* procedure, using the p -value in (4.3), fails to control the *FDR* correctly, is likely due to the fact that the Welch's p -value given by 4.3 is not uniformly distributed. Indeed, the p -value $p_w(d)$ is known to depend on the ratio of two normal distribution variances σ_1^2/σ_2^2 . For $m = 3$, $n = 12$ and various values of σ_1^2/σ_2^2 , our simulation results show that $p_w(d)$ is not uniformly distributed in the interval $[0, 1]$. Note that if we control *FDR* under $q = 0.5$, we reject 402 hypotheses by using *GP* and 403 hypotheses by using *WT*.

[Figure 1 about here.]

5. DISCUSSION

In this paper, we provide some theoretical distributional properties of the generalized p -value of the Behrens-Fisher problem. We then use these theoretical properties to derive a generalized multiple testing procedure that can control the *FDR* using generalized p -value,

as described in Theorem 3. Separately, we also obtain similar results for the problem of testing the difference of corresponding parameters of two independent linear regressions, with possibly different error variances. However, we do not report the result here. Future work is needed to study distributional property of the generalized p -value for many other testing problems, such as Weerahandi (1995a) on the *ANOVA* problem, and Weerahandi (1991) on testing of variance components in mixed model.

For the one-sided test problem $H_0 : \mu_1 - \mu_2 \leq \delta_0$ versus $H_1 : \mu_1 - \mu_2 > \delta_0$, Jeffreys (1961) showed that under the non-informative prior $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2}\sigma_2^{-2}$, the posterior probability that $H_0 : \mu_1 - \mu_2 \leq \delta_0$ is true has the same expression as (2.5). Jeffreys' result follows from the fact that posterior probability can be expressed as

$$P(\mu_1 - \mu_2 \leq \delta_0 | d) = \int P(\mu_1 - \mu_2 \leq \delta_0 | d, \sigma_1^2, \sigma_2^2) p(\sigma_1^2, \sigma_2^2 | d) d\sigma_1^2 d\sigma_2^2, \quad (5.1)$$

where $p(\sigma_1^2, \sigma_2^2 | d) = p(\sigma_1^2 | d) p(\sigma_2^2 | d)$ is the posterior probability density function of (σ_1^2, σ_2^2) , and,

$$\sigma_1^2 | d \sim \frac{ms_1^2}{\chi_{m-1}^2}, \quad \sigma_2^2 | d \sim \frac{ns_2^2}{\chi_{n-1}^2}. \quad (5.2)$$

Thus results in Theorem 1 and Theorem 3 remain true if we replace the generalized p -value by posterior probability that the one-side null hypothesis is true. Weerahandi and Tsui (1996) also observe the relationship between the generalized p -values and the posterior probabilities.

Recall Q given in (3.1) is the false discovery proportion. Instead of requiring that the expected value of Q to be less than or equal to a given value q , Lehmann and Romano (2005) recently propose a new criterion to control Q . For given γ and α in $(0, 1)$, their criterion is

$$P(Q > \gamma) \leq \alpha. \quad (5.3)$$

They propose a procedure based on usual p -values that can guarantee (5.3) to hold. We

believe that their result can be extended to the case where generalized p -values are used instead.

The research in the area of multiple testing has generated a great deal of interest. See for example, Storey (2003), Storey, Taylor and Siegmund (2004), Pacifico, Genovese, Verdinelli and Wasserman (2004), and the references cited therein. Obtaining a multiple testing procedure when the hypotheses are not independent is an important problem. Benjamini and Yekutieli (2001), for example, propose very interesting and important results in this area. We will examine the situation where the generalized p -value are not independent. We believe appropriate conditions can be found under which the result in Theorem 3 still holds. Our proof of Theorem 3 allows us focus on the crucial step that would make the dependence case go through.

Appendix

Proof of Lemma 1: Denote

$$h(b) = z \sqrt{\frac{1}{\frac{t_1}{b} + \frac{t_2}{1-b}}},$$

we have $f(b) = \Psi_v(h(b))$. Let ψ_v be the probability density function of a t-distribution with v degrees of freedom. Then

$$f''(b) = (f'(b))' = (\psi_v(h(b))h'(b))' = \psi'_v(h(b))(h'(b))^2 + \psi_v(h(b))h''(b).$$

For $z \leq 0$, $h(b) < 0$. Hence $\psi'_v(h(b)) \geq 0$. Obviously, $\psi(h(b)) \geq 0$. Moreover

$$\begin{aligned}
h''(b) &= z \left[\left(\frac{t_1}{b} + \frac{t_2}{1-b} \right)^{-3/2} \cdot (-1/2) \left(-\frac{t_1}{b^2} + \frac{t_2}{(1-b)^2} \right) \right]' \\
&= -\frac{z}{2} \left[\frac{-\frac{t_1}{b^2} + \frac{t_2}{(1-b)^2}}{\left(\frac{t_1}{b} + \frac{t_2}{1-b} \right)^{3/2}} \right]' \\
&= -\frac{z}{2} \frac{\left(\frac{2t_1}{b^3} + \frac{2t_2}{(1-b)^3} \right) \left(\frac{t_1}{b} + \frac{t_2}{(1-b)} \right)^{3/2} - \frac{3}{2} \left(\frac{t_1}{b} + \frac{t_2}{(1-b)} \right)^{1/2} \left(-\frac{t_1}{b^2} + \frac{t_2}{(1-b)^2} \right)^2}{\left(\frac{t_1}{b} + \frac{t_2}{1-b} \right)^3} \\
&= -\frac{z}{2} \frac{\left(\frac{2t_1}{b^3} + \frac{2t_2}{(1-b)^3} \right) \left(\frac{t_1}{b} + \frac{t_2}{(1-b)} \right) - \frac{3}{2} \left(-\frac{t_1}{b^2} + \frac{t_2}{(1-b)^2} \right)^2}{\left(\frac{t_1}{b} + \frac{t_2}{1-b} \right)^{5/2}} \\
&= -\frac{z}{2} \frac{\frac{t_1^2}{2b^4} + \frac{t_2^2}{2(1-b)^4} + \frac{2t_1t_2}{b(1-b)^3} + \frac{2t_1t_2}{b^3(1-b)} + \frac{3t_1^2t_2}{b^2(1-b)^2}}{\left(\frac{t_1}{b} + \frac{t_2}{1-b} \right)^{5/2}} \\
&> 0.
\end{aligned}$$

Hence $f''(b) > 0$, and $f(b)$ is convex in b . $f(b)$ is strictly convex if $z < 0$.

Proof of Lemma 2:

$$\begin{aligned}
g(A) &= P \left[\Psi_{m+n-2} \left[Z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \right] \leq r \right] \\
&= P \left[Z \sqrt{\frac{1}{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}}} \leq \Psi_{m+n-2}^{-1}(r) \right] \\
&= P \left[Z \leq \sqrt{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}} \left(\Psi_{m+n-2}^{-1}(r) \right) \right] \\
&= E_{C_{m-1}, C_{n-1}} \left[\Phi \left(\sqrt{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}} \left(\Psi_{m+n-2}^{-1}(r) \right) \right) \right]
\end{aligned}$$

where Φ is the *c.d.f* of a standard normal distribution. For fixed C_{m-1}, C_{n-1} , denote

$$h_1(A) = \sqrt{\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1}} \Psi_{m+n-2}^{-1}(r), \text{ and } g_1(A) = \Phi(h_1(A))$$

Let ϕ be the probability density function of a standard normal distribution. We have

$$\begin{aligned} g_1''(A) &= (g_1'(A))' = (\phi(h_1(A))h_1'(A))' \\ &= \phi'(h_1(A))(h_1'(A))^2 + \phi(h_1(A))h_1''(A). \end{aligned}$$

For $r \leq 0.5$, $h_1(A) \leq 0$, consequently, $\phi'(h_1(A)) \geq 0$.

Moreover

$$h_1''(A) = -\frac{1}{2} \Psi_{m+n-2}^{-1}(r) \left(\frac{AC_{m-1}}{m-1} + \frac{(1-A)C_{n-1}}{n-1} \right)^{-3/2} \left(\frac{C_{m-1}}{m-1} - \frac{C_{n-1}}{n-1} \right)^2 \geq 0.$$

Hence $g_1''(A) \geq 0$. That is, $g_1(A)$ is convex as a function of A . As a result, $g(A) = E_{C_{m-1}, C_{n-1}}(g_1(A))$ given in (2.6) is convex as a function of A .

REFERENCES

- 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. Series B (Methodological)*, 57, 289–300.
- Benjamini, Y., and Yekutieli, D. (2001), ‘‘The Control of the False Discovery Rate in Multiple Testing under Dependency,’’ *Annals of Statistics*, 29, 1165–1188.
- He, F., Li, X. R., Spatrick, P., Casillo, R., Dong, S. Y., and Jacobson, A. (2003), ‘‘Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast,’’ *Molecular Cell*, 12, 1439–1452.
- Jeffreys, H. (1961), *Theory of Probability(3rd ed.)*, Oxford, U.K.: Oxford University Press.

- Lehmann, E. L., and Romano, J. P. (2005), “Generalizations of the familywise error rate,” *Annals of Statistics*, 33, 1138–1154.
- Lelivelt, M. J., and Culbertson, M. R. (1999), “Yeast Upf proteins required for RNA surveillance affect global expression of the yeast transcriptome,” *Molecular and Cellular Biology*, 19, 6710–6719.
- Pacifico, M. P., Genovese, C., Verdinelli, I., and Wasserman, L. (2004), “False Discovery Rates for Random Fields,” *Journal of the American Statistical Association*, 99, 1002 – 1014.
- Peterson, J., and Weerahandi, S. (2003), “Generalized p-values and confidence intervals: Their role in statistical methods for pharmaceutical research and development,” *manuscript*, .
- Storey, J. D. (2003), “The positive false discovery rate: A Bayesian interpretation and the q-value,” *Annals of Statistics*, 31, 2013–2035.
- Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), “Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach,” *Journal of the Royal Statistical Society, Series B*, 66, 187–205.
- Tsui, K.-W., and Weerahandi, S. (1989), “Generalized p -Values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters,” *Journal of the American Statistical Association*, 84(406), 602–607.
- Weerahandi, S. (1991), “Testing Variance Components in Mixed Models With Generalized p Values,” *Journal of the American Statistical Association*, 86, 151–153.
- Weerahandi, S. (1995a), “ANOVA under Unequal Error Variances,” *Biometrics*, 51, 589–599.
- Weerahandi, S. (1995b), *Exact Statistical Methods for Data Analysis*, New York: Springer.

Weerahandi, S. (2004), *Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models*, New Jersey: John Wiley & Sons.

Weerahandi, S., and Tsui, K.-W. (1996), “comment on ‘Posterior predictive assessment of model fitness via realized discrepancies’ by Gelman, Meng, and Stern,” *Statistica Sinica*, 6(4), 792–796.

Welch, B. L. (1938), “The Significance of the Difference Between Two Means when the Population Variances are Unequal,” *Biometrika*, 29, 350–362.

List of Figures

- 1 Control *FDR* to be under $q = 0.05$ by the *GPU* and *WT* procedures. Small dots are the sorted generalized p -values for the *GPU* procedure, and the sorted p -values for the *WT* procedure. Large dots are the corresponding generalized p -values and p -values for genes in the control group, which presumably are not be rejected. 24

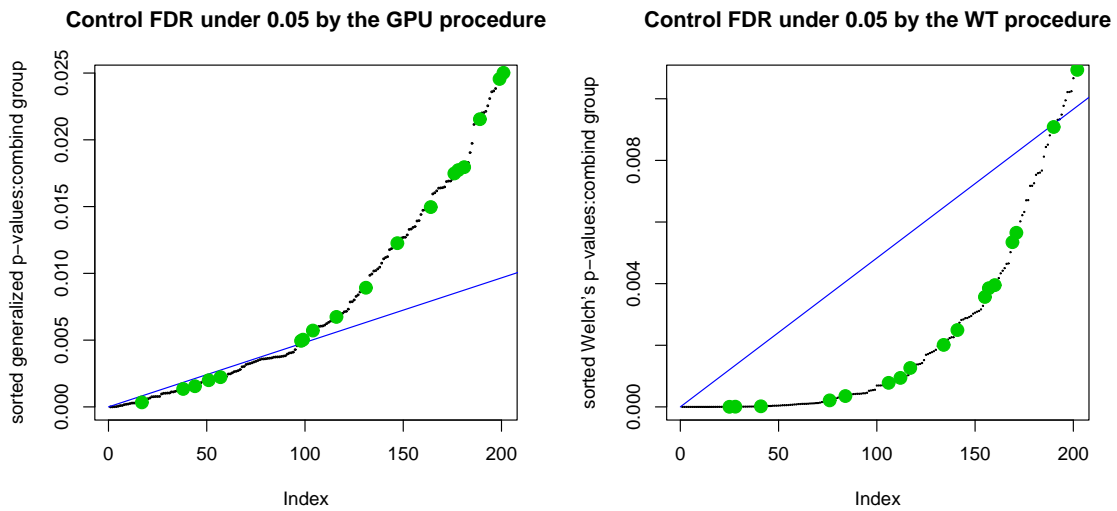


Figure 1: Control FDR to be under $q = 0.05$ by the GPU and WT procedures. Small dots are the sorted generalized p -values for the GPU procedure, and the sorted p -values for the WT procedure. Large dots are the corresponding generalized p -values and p -values for genes in the control group, which presumably are not be rejected.