International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:9, 2013

Upper Bound of the Generalize p-Value for the Behrens-Fisher Problem with a Known Ratio of Variances

Rada Somkhuean, Suparat Niwitpong, and Sa-aat Niwitpong

Abstract—This paper presents the generalized *p*-values for testing the Behrens-Fisher problem when a ratio of variance is known. We also derive a closed form expression of the upper bound of the proposed generalized p-value.

Keywords—Generalized p-value, hypothesis testing, ratio of variances, upper bound.

I. Introduction

SCHECHTMAN and Sherman [1] described a situation with a known ratio of variances arises in practice when two instruments reports (averaged) response of the same object based on a difference number of replicates. If the two instruments have the same precision for a single measurement, then the ratio of the variance of the responses is known and it is simply the ratio of the number of replicates going into each response. They proposed a t-test statistic, which has an exact t-distribution with $n+m-2$ degrees of freedom compared to the Satterthwaite's t-test statistic [2]. They found that their proposed test has more power than the existing Satterthwaite's test. However, they did not investigate the coverage probability and the expected length of the confidence interval for the difference of two normal population means when the ratio of variances is known. Niwitpong and Niwitpong [3] derived analytic expressions to find the coverage probabilities and expected lengths of two confidence intervals, the Schechtman-Sherman confidence interval and the Welch-Satterthwaite (WS) confidence interval [4], in comparison with each other. In this paper, following Weerahandi [5], we propose the generalized *p*-value to test the hypothesis $H_0 : \theta < \theta_0$ vs $H_1 : \theta > \theta_0$ where θ is the parameter of interest, and, $\theta = \mu_1 - \mu_2$ and θ_0 is fixed and with know a ratio of variances.

II. Generalized p-values for the Behrens-Fisher problem

Let $X_1, ...X_n$ and $Y_1, ..., Y_m$ be random samples from two independent normal distributions with means μ_x, μ_y and standard deviations σ_x and σ_y , respectively.

Rada Somkhuean is with Department of Applied Statistic, Faculty of Applied Science, King Mongkuts University of Technology North Bangkok, Bangkok 10800, Thailand (e-mail: rada_m_1@hotmail.com).

Sa-aat Niwitpong is with Department of Applied Statistic, Faculty of Applied Science, King Mongkuts University of Technology North Bangkok, Bangkok 10800, Thailand (e-mail: snw@kmutnb.ac.th).

Supatar Niwitpong is with Department of Applied Statistic, Faculty of Applied Science, King Mongkuts University of Technology North Bangkok, Bangkok 10800, Thailand (e-mail: suparat8@gmail.com).

Let $\theta = \mu_x - \mu_y$ be the parameter of interest. The problem is to test the hypothesis $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_a : \theta > \theta_0$ for some fixed θ_0. The sufficient statistic of this problem is $(\bar{X}, \bar{Y}, S^2_{xs}, S^2_{ys})$ (Tsui and Weerahandi [6]) where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i, \bar{Y} = m^{-1}\sum_{j=1}^{m} Y_j$,

$S^2_{xs} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ and $S^2_{ys}\dfrac{\sum_{j=1}^{m}(Y_j - \bar{Y})^2}{m}$.

The probability distributions of the statistics, $\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$, $\bar{Y} \sim N(\mu_y, \frac{\sigma_y^2}{m}), V = \frac{nS^2_{xs}}{\sigma_x^2} \sim \chi^2_{n-1}$ and $U = \frac{mS^2_{ys}}{\sigma_y^2} \sim \chi^2_{m-1}$ are independent of one another. Tsui and Weerahandi[6] proposed the generalized p-value for the above hypothesis as follow:

Suppose a random quantity $T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ can be expressed as

$$T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2) = T(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2) - \theta$$

where

$$T(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2) = \frac{\bar{X} - \bar{Y} - \theta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sqrt{\frac{\sigma_x^2 s^2_{xs}}{nS^2_{xs}} + \frac{\sigma_y^2 s^2_{ys}}{mS^2_{ys}}}$$

and $T(x, y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2) = \bar{x} - \bar{y} - \theta_0$. It is straightforward to see that $T(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is free from nuisance parameters σ_x^2 and σ_y^2 and has the same distribution $Z\sqrt{\frac{s^2_{xs}}{V} + \frac{s^2_{ys}}{U}}$ where $Z \sim N(0, 1)$.

$T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is defined to be a generalized test variable and $T(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is defined to be a generalized pivot statistic and $T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is required to satisfy the following conditions:

C1. For a fixed x and y, the probability distribution of $T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is free of the unknown parameters.

C2. The observed value of $T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$, namely $T^*(x, y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is simply θ.

C3. For fixed x, y and $\delta = (\sigma_x^2, \sigma_y^2)$,
$T^*(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is stochastically monotone in θ.

The generalized pivot statistic $T(X, Y, x, y, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ is also required to satisfy the following conditions:

C4. For a fixed x and y, the probability distribution of $T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)$ is free of the unknown parameters θ and $\delta=(\sigma_x^2,\sigma_y^2)$.

C5. The observed valued of $T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)$, namely
$T(x,y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)$ is simply equal to θ.
A $100(1-\alpha/2)\%$ generalized lower confidence limit for θ is then given by $T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)_{1-\alpha}$, the $100(1-\alpha)th$ percentiles of $T(x,y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)$.

Further, given the observed value x, let t_1 and t_2 be such values that
$P(t_1 < T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2) < t_2|\theta) = 1-\alpha$ for chosen significant level $\alpha\in(0,1)$ than the confidence interval for parameter θ defined by
$\{\theta : t_1 < T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2) < t_2\}$ is a $100(1-\alpha)\%$ generalized confidence interval for θ.

For the one-sided hypothesis given above they defined a data-based extreme region $C_{x,y}$ of the form

$$C_{x,y}(\theta,\sigma_x^2,\sigma_y^2) = \{(X,Y) : T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2) - T(x,y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2) \geq 0.$$

For the one-sided Behrens-Fisher problem, the generalized *p*-value is

$$p^* = Pr(T(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2) - T(x,y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)|\theta=\theta_0).$$

III. Main results for Behrens-Fisher problem with one variance unknown

Following Schechtman and Sherman [1], we suppose a ratio of variances is known i.e. $\frac{\sigma_y^2}{\sigma_x^2}=c$, where c is a constant. According to Tsui and Weerahandi [6], one of the potential pivotal quantity can be defined as

$$\begin{aligned}
&Q(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}+\theta\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}\frac{\sigma_x^2}{\sigma_x^2}}+\theta\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\sigma_x^2\left(\frac{1}{n}+\frac{c}{m}\right)}+\theta\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\sigma_x^2\frac{(ns_x^2)}{nS_x^2}\left(\frac{m+nc}{nm}\right)}+\theta\\
&= Z\sqrt{\frac{s_x^2}{V}\left(\frac{m+nc}{m}\right)}+\theta \qquad (1)
\end{aligned}$$

For the one-side Behrens-Fisher problem as stated,
$H_0:\theta<\theta_0$ against $H_a:\theta>\theta_0$, we can assume $\theta_0=0$ without loss of generality, and the generalized *p*-value for the one-sided Behrens-Fisher problem is $p(q)$ which is

$$\begin{aligned}
&Pr(Q(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)\geq q_{obs}=0)\\
&= Pr\left(Z\sqrt{\frac{s_x^2}{V}\left(\frac{m+nc}{m}\right)}\geq \bar{x}-\bar{y}\right)\\
&= Pr\left(Z\geq(\bar{x}-\bar{y})\left(\frac{s_x^2}{V}\left(\frac{m+nc}{m}\right)\right)^{-\frac{1}{2}}\right)\\
&= Pr\left(Z\leq(\bar{y}-\bar{x})\left(\frac{s_x^2}{V}\left(\frac{m+nc}{m}\right)\right)^{-\frac{1}{2}}\right)\\
&= E_V\left(\Phi\left((\bar{y}-\bar{x})\left(\frac{s_x^2}{V}\left(\frac{m+nc}{m}\right)\right)^{-\frac{1}{2}}\right)\right) \qquad (2)
\end{aligned}$$

where $\Phi(.)$ is a *cdf* of the standard normal distribution and $E_V(.)$ is an expectation operator with respect to V.

To find the upper bound of $p(q)$, we need Theorems 1-2 based on Tang and Tsui [7] as follows:

Theorem 1. *Define*

$$h(v)=\Phi\left(z\sqrt{\frac{vm}{t}}\right) \qquad for \quad v\in(0,1).$$

Then for fixed $z<0$, $h(v)$ is a convex function of v.

Proof: Letting

$$h(v)=z\sqrt{\frac{vm}{t}} \quad ,$$

we have $f(v)=\Phi(h(v))$. Let ϕ be the probability density function of standard normal distribution.
Then

$$\begin{aligned}
f''(v)=(f'(v)')&=(\phi(h(v))h'(v))'\\
&=\phi'(h(v))(h'(v))^2+\phi(h(v))h''(v)
\end{aligned}$$

For $Z<0$, $h(v)<0$. Hence $\phi'(h(v))\geq 0$. Obviously, $\phi(h(v))\geq 0$. Moreover,

$$\begin{aligned}
h''(v)&=\left[z\left(\frac{1}{2}\right)\left(\frac{vm}{t}\right)^{-\frac{1}{2}}\left(\frac{m}{t}\right)\right]'\\
&=-\frac{z}{4}\left(\frac{vm}{t}\right)^{-\frac{3}{2}}\left(\frac{m}{t}\right)^2\\
&=-\frac{z}{4}\frac{\left(\frac{m}{t}\right)^2}{\left(\frac{vm}{t}\right)^{\frac{3}{2}}}>0
\end{aligned}$$

Hence $h(v)\geq 0$, and $h(v)$ is convex in v. ■

Theorem 2. *Let*

$$g(a)=P\left[\Phi\left(z\sqrt{\frac{(n-1)m}{C_{n-1}a(m+nc)}}\leq r\right)\right],$$

where z, C_{n-1} independent random variables such that $z\sim N(0,1)$, $C_{n-1}\sim\chi^2_{n-1}$. Then $g(a)$ is a convex function in a.

Proof:

$$\begin{aligned} g(a) &= P\left[\Phi\left(z\sqrt{\frac{(n-1)m}{C_{n-1}a(m+nc)}} \le r\right)\right] \\ &= P\left[z\sqrt{\frac{(n-1)m}{C_{n-1}a(m+nc)}} \le \Phi^{-1}(r)\right] \\ &= P\left[z\sqrt{\frac{(n-1)}{C_{n-1}}} \le \sqrt{\frac{a(m+nc)}{m}}\left(\Phi^{-1}(r)\right)\right] \\ &= E_T\left[\Psi_{n-1}\left(\sqrt{\frac{a(m+nc)}{m}}\left(\Phi^{-1}(r)\right)\right)\right] \end{aligned}$$

$E_T(.)$ is an expectation operator with respect to T and $\Psi(.)_{n-1}$ is a cdf of t distribution with $(n-1)$ degree of freedom, denote

$$h_1(a) = \sqrt{\tfrac{a(m+nc)}{m}}\Phi^{-1}(r) \text{ and } g_1(a) = \Psi_{n-1}(h_1(a))$$

Let ψ_{n-1} be the probability density function of *t* distribution with $n-1$ degees of freedom. we have

$$\begin{aligned} g_1''(a) &= (g_1'(a))' \\ &= (\psi_{n-1}(h_1(a))h_1'(a))' \\ &= \psi_{n-1}'(h_1(a))(h_1'(a))^2 + \psi_{n-1}(h_1(a))h_1''(a). \end{aligned}$$

For $r \le 0.5$, $h_1(a) \le 0$, and consequently, $\psi_{n-1}'(h_1(a)) \ge 0$. Morever,

$$\begin{aligned} h_1''(a) &= \left[\frac{1}{2}\left(\frac{a(m+nc)}{m1}\right)^{-\frac{1}{2}}\Phi^{-1}(r)\left(\frac{(m+nc)}{m}\right)\right]' \\ &= -\frac{1}{4}\Phi^{-1}(r)\left(\frac{a(m+nc)}{m}\right)^{-\frac{3}{2}}\left(\frac{(m+nc)}{m}\right)^2 \ge 0. \end{aligned}$$

Hence $g_1''(a) \ge 0$. That is $g_1(a)$ is convex in a. As a result, $g(a) = E_T(g_1(a))$ is convex in a. ■

Theorem 3. *For the one-sided Behrens Fisher problem , when a ratio of variances is known with $H_0 : \mu_1 - \mu_2 \le \theta_0$ and any $0 < r < 0.5$. The generalized p-value , $p(q)$ in (2), has the following property under H_0:*

$$P_q(p(q) \le r) < \Psi_{n-1}(k\Phi^{-1}(r)) \quad k = \sqrt{\frac{m+nc}{m}}$$

Where $\Psi_{n-1}(.)$ is a cdf of t distribution with $n-1$ degrees of freedom,$\Phi(.)$ is cdf of the standard normal distribution, and Φ^{-1}inverse function of $\Phi(.)$.

Proof: Denote

$$A = \frac{\frac{\sigma_x^2}{n}}{\frac{\sigma_n^2}{n} + \frac{\sigma_m^2}{m}} \quad z = \frac{\bar{y} - \bar{x}}{\sqrt{\frac{\sigma_n^2}{n} + \frac{\sigma_m^2}{m}}} \quad C_{n-1} = \frac{ns_x^2}{\sigma_x^2}$$

From (2)

$$\begin{aligned} p(q) &= E_V\left[\Phi\left((\bar{y} - \bar{x})\left(\frac{s_x^2}{V}\left(\frac{m+nc}{m}\right)\right)^{-\frac{1}{2}}\right)\right] \\ &= E_V\left[\Phi\left(\frac{(\bar{y} - \bar{x})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}\left(\sqrt{\frac{1}{\frac{s_x^2}{n} + \frac{\sigma_y^2}{m}\left(\frac{m+nc}{m}\right)\frac{1}{V}}}\right)\right)\right] \\ &= E_V\left[\Phi\left(Z\left(\sqrt{\frac{Vm}{C_{n-1}A(m+nc)}}\right)\right)\right] \end{aligned}$$

For any $r < 0.5$ and $p(q) < r$, we must have. Hence by theorem 1

$f(v) = E_V\left[\Phi\left(Z\left(\sqrt{\frac{Vm}{C_{n-1}A(m+nc)}}\right)\right)\right]$ is convex in V.

By Jensens Inequality,

$$\begin{aligned} p(q) &= E_V(f(V) \ge f(E(V))) = f(n-1) \\ p(q) &= \phi\left(Z\left(\sqrt{\frac{(n-1)m}{C_{n-1}A(m+nc)}}\right)\right) \equiv p_1(q) \end{aligned}$$

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

$$P_q(\{q : p(q) \le r\} \le P_q\{p_1(q) \le r\} = g(A)$$

. where $g(a)$ is a defined in theorem 2. Next by theorem 2 for $0 < r < 0.5$, $g(A)$, is convex in A.

$$\begin{aligned} g(A) &\le max\{g(0), g(1)\} \\ &= \Phi\left(Z\left(\sqrt{\frac{(n-1)m}{C_{n-1}A(m+nc)}}\right) \le r\right) \\ &= \left(z\sqrt{\frac{n-1}{C_{n-1}}} \le \Phi^{-1}(r)\sqrt{\frac{m+nc}{m}}\right) \\ &= \Psi_{n-1} \le (k\Phi^{-1}(r)) \end{aligned}$$

where $k = \sqrt{\frac{m+nc}{m}}$ ■

IV. Conclusion

In this paper, we derive an expression of the upper bound of the generalized *p*-value for the Behrens-Fisher problem with a know ratio of variances used the method described by Tang and Tsui [7]. This upper bound can be easily computed by R program with command: pnorm(k*qnorm(r)), when r is a fixed real value between 0 to 0.5.

References

[1] E. Schechtman, and M. Sherman, "The two-sample t-test with a known ratio of Variances", *Statistical Methodology*, Vol.4, pp. 508-514, 2007.
[2] F.E. Satterthwaite, "An approximate distribution of estimates of variance components", *Biometric Bulletin*, Vol.6, pp. 110-114, 1946.
[3] S. Niwitpong, and Sa. Niwitpong, "Confidence interval for the difference of two normal population means with a known ratio of variances", *Applied Mathematical Sciences*, Vol.4, pp. 347359, 2010.

[4] B.L. Welch, "The significance of the difference between two means when the population variances are unequal", *Biometrika*, Vol.29, pp. 350-362, 1983.
[5] S. Weerahandi, "Exact Statistical Methods for Data Analysis", *Springer*, NewYork, 1995.
[6] K-W. Tsui, and S. Weerahandi, "Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters", *J. Amer Statist Assoc*, Vol.84, pp. 60207, 1989.
[7] S. Tang,and K-W. Tsui, "Distributional properties for the generalized p-value for the BehrensFisher problem", *Statistics Probability Letters*, Vol.77, pp. 18, 2007.