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

Behrens-Fisher Problem with One Variance Unknown

Sa-aat Niwitpong, Rada Somkhuean, and Suparat Niwitpong

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

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

I. Introduction

MAITY and Sherman [1] mentioned that the situation of the hypothesis testing for the difference of two normal population means with one variance unknown, arises in practice. For example, when one is interested in comparing a standard treatment with a new treatment. A known variance comes from the standard treatment while an unknown variance comes from the new treatment. Maity and Sherman found that their proposed t-test has more power than the existing Satterthwaite's test [2], [3]. 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 one variance is unknown. Niwitpong [4] also derived analytic expressions to find coverage probabilities and expected lengths of the confidence interval using the pivotal statistic t-statistic proposed by Maity and Sherman compared to Welch-Satterthwaite (WS) [5] confidence interval. In this paper, following Weerahandi [6], we propose the gerneralized p-value to test the hypothesis $H_0 : \theta \leq \theta_0$ 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 when one of variance is unknown.

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.
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 [7]) where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i, \bar{Y} = m^{-1}\sum_{j=1}^{m} Y_j$,

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).
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).
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).

$S^2_{xs} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ and $S^2_{ys} \dfrac{\sum_{j=1}^{m}(Y_j - \bar{Y})^2}{m}$.
The probability distributions of the statistics, $\bar{X} \sim N(\mu_x, \frac{\sigma^2_x}{n})$, $\bar{Y} \sim N(\mu_y, \frac{\sigma^2_y}{m}), V = \frac{S^2_{xs}}{\sigma^2_x} \sim \chi^2_{n-1}$ and $U = \frac{mS^2_{ys}}{\sigma^2_y} \sim \chi^2_{m-1}$ are independent of one another. Tsui and Weerahandi [7] proposed the generalized p-value for the above hypothesis as follow:

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

$$T^*(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y) = T(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y) - \theta$$

where

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

and $T(x, y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y) = \bar{x} - \bar{y} - \theta_0$. It is straightforward to see that $T(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ is free from nuisance parameters σ^2_x and σ^2_y and has the same distribution $Z\sqrt{\frac{s^2_{xs}}{V} + \frac{s^2_{ys}}{U}}$ where $Z \sim N(0, 1)$.
$T^*(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ is defined to be a generalized test variable and $T(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ is defined to be a generalized pivot statistic and $T^*(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ is required to satisfy the following conditions:

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

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

C3. For fixed x, y and $\delta = (\sigma^2_x, \sigma^2_y)$,
$T^*(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ is stochastically monotone in θ.
The generalized pivot statistic $T(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ is also required to satisfy the following conditions:

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

C5. The observed valued of $T(X, Y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$, namely
$T(x, y, x, y, \mu_x, \mu_y, \sigma^2_x, \sigma^2_y)$ 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. 3. Main results for Behrens-Fisher problem with one variance unknown

Following Maity and Sherman [1], we suppose one of variances is unknown i.e., σ_y^2. According to Tsui and Weerahandi [7], one of the potential pivotal quantity can be defined as

$$\begin{aligned}
&W(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2)\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}+\theta\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}\frac{s_x^2}{S_x^2}}+\theta\\
&= \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{U}}+\theta\\
&= Z\sqrt{\frac{\sigma_x^2}{n}+\frac{s_y^2}{U}}+\theta \qquad (1)
\end{aligned}$$

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 $p(w)$ which is

$$\begin{aligned}
&Pr(W(X,Y,x,y,\mu_x,\mu_y,\sigma_x^2,\sigma_y^2) \geq w_{obs}|\theta>0)\\
&= Pr\left(Z\sqrt{\frac{\sigma_x^2}{n}+\frac{s_y^2}{U}} \geq \bar{x}-\bar{y}\right)\\
&= Pr\left(Z \geq (\bar{x}-\bar{y})\frac{1}{\sqrt{\frac{\sigma_x^2}{n}+\frac{s_y^2}{U}}}\right)\\
&= Pr\left(Z \leq (\bar{y}-\bar{x})\frac{1}{\sqrt{\frac{\sigma_x^2}{n}+\frac{s_y^2}{U}}}\right)\\
&= E_U\left(\Phi\left((\bar{y}-\bar{x})\frac{1}{\sqrt{\frac{\sigma_x^2}{n}+\frac{s_y^2}{U}}}\right)\right) \qquad (2)
\end{aligned}$$

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

Now to find the upper bound of $p(w)$ using the method described by Tang and Tsui [8], we need Theorems 1 and 2 as following,

Theorem 1. *Define*

$$f(u) = \Phi\left(z\sqrt{\frac{1}{t_1+\frac{t_2}{u}}}\right) \qquad for \quad u \in (0,1).$$

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

Proof: Letting

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

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

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

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

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

Hence $f(u) \geq 0$, and $f(u)$ is convex in u. ■

Theorem 2. *Let*

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

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

Proof:

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

$E_{C_{m-1}}(.)$ is an expectation operator with respect to C_{m-1} with $(n-1)$ degree of freedom and $\Phi(.)$ is a cdf of the standard normal distribution, denote

$h_1(a) = \sqrt{\frac{a(m-1)+(1-a)C_{m-1}}{m-1}}(\Phi^{-1}(r))$
and $g_1(a) = \Phi(h_1(a))$ 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 $r \leq 0.5$, $h_1(a) \leq 0$, and consequently, $\phi'(h_1(a)) \geq 0$. Morever,

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

Hence $g_1''(a) \geq 0$. That is $g_1(a)$ is convex in a. As a result, $g(a) = E_{C_{m-1}}(g_1(a))$ is convex in a. ■

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

$$P_w(p(w) \leq r) < \Phi(\Phi^{-1}(r))$$

Where $\Phi(.)$ is a cdf of the standard normal distribution and $\Phi^{-1}(.)$ is the inverse function of $\Phi(.)$.

Proof: Denote

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

From (2)

$$\begin{aligned} p(w) &= E_U\left[\Phi\left((\bar{y}-\bar{x})\frac{1}{\sqrt{\frac{\sigma_n^2}{n}+\frac{\sigma_y^2}{U}}}\right)\right] \\ &= E_U\left[\Phi\left(\frac{(\bar{y}-\bar{x})}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\frac{1}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{U}}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}\right)\right] \\ &= E_U\left[\Phi\left(Z\frac{1}{\sqrt{\frac{\sigma_x^2/n+s_y^2/U}{\sigma_x^2/n+\sigma_y^2/m}}}\right)\right] \\ &= E_U\left[\Phi\left(Z\frac{1}{\sqrt{\frac{\sigma_x^2/n}{\sigma_x^2/n+\sigma_y^2/m}+\frac{\frac{1}{U}(C_{m-1}\sigma_y^2/m)}{\sigma_x^2/n+\sigma_y^2/m}}}\right)\right] \\ &= E_U\left[\Phi\left(Z\frac{1}{\sqrt{A+\frac{(1-A)C_{m-1}}{U}}}\right)\right] \end{aligned}$$

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

$f(U) = \Phi\left(Z\frac{1}{\sqrt{A+\frac{(1-A)C_{m-1}}{m-1}}}\right)$ is convex in U.

By Jensens Inequality,

$$\begin{aligned} p(w) &= E_U(f(U) \geq f(E(U))) = f(m-1) \\ p(w) &= \phi\left(Z\frac{1}{\sqrt{A+\frac{(1-A)C_{m-1}}{m-1}}}\right) \equiv p_1(w) \end{aligned}$$

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

$$P_w(\{w : p(w) \leq r\} \leq P_w\{p_1(w) \leq r\} = g(A)$$

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

$$\begin{aligned} g(A) &\leq max\{g(0), g(1)\} \\ &= max\left\{P(\Phi(Z \leq r)), P\left(\Phi\left(Z\sqrt{\frac{1}{\frac{C_{m-1}}{m-1}}} \leq r\right)\right)\right\} \\ &= max\left\{P\left(Z \leq \Phi^{-1}(r)\right), P\left(Z\sqrt{\frac{1}{\frac{C_{m-1}}{m-1}}} \leq \Phi^{-1}(r)\right)\right\} \\ &= max(\Phi(\Phi^{-1}(r)), \Psi_{m-1}(\Psi^{-1}(r))) \\ &= \Phi(\Phi^{-1}(r)) \end{aligned}$$

where $\Phi(.)$ is *cdf* of standard normal distribution. ■

IV. Conclusion

In this paper, we derive an expression of the upper bound of the generalized p-value for the Behrens-Fisher problem with one unknown variance used the method described by Tang and Tsui [8]. This upper bound can be easily computed by R program with command: pnorm(qnorm(r)), when r is a fixed real value between 0 to 0.5.

References

[1] A. Maity, and M. Sherman, "The Two Sample T-test with One Variance Unknown", *The American Statistician*, Vol. 60, No.2, pp. 163-166, 2006.
[2] F.E. Satterthwaite, "Synthesis of variance", *Psychometrik*, Vol. 6, pp. 309-316, 1941.
[3] F.E. Satterthwaite, "An approximate distribution of estimates of variance components", *Biometric Bulletin*, Vol. 6, pp. 110-114, 1946.
[4] S. Niwitpong, "Confidence intervals for the difference of two normal population means with one variance unknown", *Thailand Statistician*, Vol. 7, pp. 161-177, 2009.
[5] B.L. Welch, "The significance of the difference between two means when the population variances are unequal", *Biometrika*, Vol. 29, pp. 350-362, 1983.
[6] S. Weerahandi, "Exact Statistical Methods for Data Analysis", Springer, NewYork, 1995.
[7] 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.
[8] 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.