# AN ALTERNATIVE TO THE SATTERSWAITE-WELCH DEGREES OF FREEDOM AND ITS COMPUTER INTENSIVE CONFIDENCE INTERVALS

Yamashita, Takuto
The Graduate School of Medicine, Kurume University

Yanagawa, Takashi
The Biostatistics Center, Kurume University

https://doi.org/10.5109/1563531

出版情報：Bulletin of informatics and cybernetics. 45, pp.45-57, 2013-12. Research Association of Statistical Sciences
バージョン：
権利関係：

# AN ALTERNATIVE TO THE SATTERSWAITE-WELCH DEGREES OF FREEDOM AND ITS COMPUTER INTENSIVE CONFIDENCE INTERVALS

by

**Takuto Yamashita** and **Takashi Yanagawa**

---

*Reprinted from the Bulletin of Informatics and Cybernetics*
*Research Association of Statistical Sciences, Vol.45*

FUKUOKA, JAPAN
2013

*Bulletin of Informatics and Cybernetics, Vol. 45, 2013*

# AN ALTERNATIVE TO THE SATTERSWAITE-WELCH DEGREES OF FREEDOM AND ITS COMPUTER INTENSIVE CONFIDENCE INTERVALS

**By**

**Takuto Yamashita**[*] and **Takashi Yanagawa**[†]

**Abstract**

An alternative to the Satterswaite-Welch degrees of freedom (*df*) is developed in this paper based on unbiased estimating equations. The method enables us to evaluate the impact of sample variation due to using estimated *df* instead of the true *df* by computer intensive method. It is shown by simulation that the alternative *df* is close to the Welch *df* and the method developed could be also used for evaluating the impact of sample variation due to using estimated *df* in the Welch test.

*Key Words and Phrases:* Behrens-Fisher problem, heterogeneity of population variances, Satterswaite test, Student $t$-test, Welch test

## 1. Introduction

We express the intention of this paper in a framework of two-sample problem. Let $X_{i1}, X_{i2}, \ldots, X_{im_i}$ be a random sample from a population with normal distribution $\mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2$. We consider statistical test for testing $H_0$: $\mu_1 = \mu_2$ against $H_1$: $\mu_1 \neq \mu_2$. If population variances are homogeneous, routine practice is to apply the Student $t$-test. The test is based on statistic

$$T_p = (\bar{X}_1 - \bar{X}_2) \,/\, S_p \sqrt{\frac{1}{m_1} + \frac{1}{m_2}},$$

where $S_p$ is a pooled estimator of $\sigma = \sigma_1 = \sigma_2$ defined by

$$S_p^2 = \frac{1}{m_1 + m_2 - 2} \left( \sum_{j=1}^{m_1} (X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{m_2} (X_{2j} - \bar{X}_2)^2 \right).$$

$T_p$ follows a $t$-distribution with $(m_1 + m_2 - 2)$ degree of freedom (*df*) under $H_0$. If $\sigma_1 \neq \sigma_2$, $T_p$ does not follow a $t$-distribution. A reasonable alternative is to employ statistic

$$T = (\bar{X}_1 - \bar{X}_2) \,/\, \sqrt{\frac{S_1^2}{m_1} + \frac{S_2^2}{m_2}},$$

[*] The Graduate School of Medicine, Kurume University, 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan. tel +81–942–31–7835 a207gm023y@std.kurume-u.ac.jp

[†] The Biostatistics Center, Kurume University 67 Asahi-machi, Kurume-city, Fukuoka 830-0011, Japan. tel +81–942–31–7835 yanagawa _ _ takashi@kurume-u.ac.jp

where $S_i^2$ is the group sample variance given by

$$S_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (X_{ij} - \bar{X}_i)^2.$$

The distribution of $T$ was first discovered by Behrens (1929) and later developed by Fisher (1939). The problem of constructing the optimum test for comparing two normal means when the unknown population variances are unequal have been called the Behrens-Fisher problem (Lehman and Romano, 2005).

A common alternative to the Behrens-Fisher sampling distribution that is complicated to apply in practice is to approximate the distribution of $T$ by $t$-distribution with adjusted *df*. Smith (1936) and later Welch (1938) found that such adjusted *df* was given by

$$\nu_w = (U_1 + U_2)^2 \ / \ \left( \frac{U_1^2}{m_1 - 1} + \frac{U_2^2}{m_2 - 1} \right),$$

where $U_i = S_i^2/m_i$. The $\nu_w$ is often called the Welch *df* and the test that uses $T$ by considering it follows $t$-distribution with the adjusted *df* is called the Welch test. The test is installed in standard statistical packages and suggested to use when $\sigma_1 \neq \sigma_2$. The Welch test is later extended to $K$-sample problems by Satterswaite (1946).

In subsequent years the robustness of the Student $t$-test was studied extensively. It was shown by numerous authors (Boneau, 1960; Box, 1953; Cochran, 1947; Posten, Yeh, and Owen, 1982; Srivastava, 1958 among others) that $t$-test was robust against considerable departure from its theoretical assumption. In particular, it was shown that if the sample sizes were approximately equal ($m_1 \approx m_2$), the $t$-test based on $T_p$ was fairly robust against violations of the variance homogeneity up to ratio of variances about 4.0 (Millar, 1986 Section 7.3). Thus the usefulness of the Welch test is limitted in the situation where sample sizes are substantially imbalanced and/or ratio of variances is more than 4.

We look at the fact in this study that the Welch *df* is subject to sampling variation via $S_i^2$. We show by simulation in the next section that impact of the variation on $p$-values of the Welch test could be substantial, even if the sample sizes are not small.

We develop an unbiased estimating equation of the *df* that approximates the distribution of $T$ by $t$-distribution. We call the *df* the *tdf* and the estimator of *tdf* from the unbiased estimating equation is called the u-estimator of the *tdf*. We replace the Welch *df* with the u-estimator of *tdf* in this paper and evaluate the impact of its sampling variation, by first constructing computer intensive confidence intervals of *tdf* and then calculating p-values using the upper and lower limits of the confidence intervals.

In Section 3 we formulate the problem in a framework of $K$-sample problems and define *tdf*, the degree of freedom that approximates the distribution of $T$ by $t$-distribution. In Section 4 we develop an unbiased estimating equation for the *tdf* based on samples generated in computer by assuming those observed sample means and variances from populations are as if given constants. Also c-intensive approximate 90% confidence intervals of the *tdf* and $p$-value are constructed by using the unbiased estimating equation in the same section. The $U$-test, an alternative to the Welch-Satterswaite test, is defined in Section 5. Behavior of the c-intensive confidence intervals are studied by simulation in Section 6. Section 7 states the conclusion of the study in this paper and finaly discussion is given in Section 8.

## 2. Impact of sampling variation of Welch *df*

The Welch *df* does depend on samples. We conducted simulation to evaluate its impact on the levels of the Welch test in the following way. (i) Construct two samples by generating $m_1$ random digits that follow normal distribution with mean zero and variance $\sigma_1^2$, and $m_2$ random digits that follow normal distribution with mean zero and variance $\sigma_2^2$. (ii) Compute the values of Welch *df* and $T$ based on the two samples. (iii) Repeat (i)$\sim$(ii) 10,000 times. (iv) Denoting by $(f_i, T_i)$ the computed values of Welch *df* and $T$ in the $i$-th run, stratify $(f_i, T_i)$, $i = 1, 2, \ldots, 10,000$, into four strata by means of the first, second and third quadrants of $f$'s, and compute empirical tail probabilities of the Welch test by

$$R_i = \frac{1}{2500} \sum_{j=1}^{2500} I\big(|t_{ij}| > t_{f_{ij}}(0.025)\big),$$

where $I(A) = 1(0)$, if $A$ is true (false), $t_{ij}$ and $f_{ij}$ are the $j$-th values of $T$ and *df* in the $i$-th stratum, and $t_f(0.025)$ is the upper 2.5% point of the $t$-distribution with $f$ degrees of freedom. $R_i$ represents empirical tail probabilities of the Welch test under $H_0$ when the nominal level is 0.05.

Table 1 Tail probabilities of the Welch test when the nominal value is 0.05.

| id | $\sigma_1^2$ | $\sigma_2^2$ | $m_1$ | $m_2$ | $R_1$ | $R_2$ | $R_3$ | $R_4$ | RSS |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 9 | 5 | 8 | 0.024 | 0.050 | 0.061 | 0.055 | 0.029 |
| 2 | 6 | 9 | 15 | 20 | 0.046 | 0.048 | 0.045 | 0.056 | 0.008 |
| 3 | 6 | 9 | 50 | 60 | 0.047 | 0.049 | 0.047 | 0.050 | 0.004 |
| 4 | 6 | 9 | 8 | 5 | 0.012 | 0.034 | 0.060 | 0.101 | 0.066 |
| 5 | 6 | 9 | 20 | 15 | 0.032 | 0.042 | 0.052 | 0.068 | 0.027 |
| 6 | 6 | 9 | 60 | 50 | 0.044 | 0.043 | 0.054 | 0.053 | 0.011 |
| 7 | 6 | 24 | 5 | 8 | 0.047 | 0.040 | 0.047 | 0.058 | 0.014 |
| 8 | 6 | 24 | 15 | 20 | 0.039 | 0.050 | 0.048 | 0.066 | 0.020 |
| 9 | 6 | 24 | 50 | 60 | 0.042 | 0.046 | 0.050 | 0.061 | 0.015 |
| 10 | 6 | 24 | 5 | 21 | 0.013 | 0.032 | 0.062 | 0.083 | 0.054 |
| 11 | 6 | 24 | 15 | 53 | 0.036 | 0.051 | 0.057 | 0.069 | 0.025 |
| 12 | 6 | 24 | 50 | 160 | 0.043 | 0.051 | 0.052 | 0.066 | 0.018 |
| 13 | 6 | 24 | 8 | 5 | 0.008 | 0.020 | 0.048 | 0.146 | 0.109 |
| 14 | 6 | 24 | 20 | 15 | 0.026 | 0.038 | 0.053 | 0.080 | 0.040 |
| 15 | 6 | 24 | 60 | 50 | 0.036 | 0.042 | 0.054 | 0.070 | 0.027 |

Table 1 summarizes values of $R_i$, $i = 1, 2, 3, 4$, when $(\sigma_1^2, \sigma_2^2)$=(6,9), (6,24) and $(m_1, m_2)$=(5,8), (8,5), (15,20), (20,15), (50,60), (60,50), (5,21), (15,53), (50,160). Those values should be in the vicinity of 0.05 if the Welch *df* is valid. The last column of the table gives the value of

$$RSS = \sqrt{\sum_{i=1}^{4} (R_i - 0.05)^2}$$

to evaluate it. The table shows that when $\sigma_1^2 < \sigma_2^2$ and $m_1 < m_2$, RSS decreases as the increase of sample sizes, so long as $(\sigma_2^2/\sigma_1^2)/(m_2/m_1)$ is constant; but that it is not the

case when $\sigma_1^2 < \sigma_2^2$ and $m_1 > m_2$, or when $\sigma_1^2 < \sigma_2^2$, $m_1 < m_2$ and $(\sigma_2^2/\sigma_1^2)/(m_2/m_1)$ is not constant; that tail probabilities of the Welch test on the 2nd and 3rd strata are close to the nominal value, 0.05, when $(\sigma_2^2/\sigma_1^2)$ is small, but it is not the case when $(\sigma_2^2/\sigma_1^2)$ is large; that those probabilities on the first and the fourth strata are far from the nominal level when $\sigma_1^2 < \sigma_2^2$ and $m_1 > m_2$, for example, $R_1 = 0.036$ when $m_1 = 60$, $m_2 = 50$, $\sigma_1^2 = 6$ and $\sigma_2^2 = 24$. Furthermore, when $\sigma_1^2 < \sigma_2^2$ and $m_1 > m_2$ and $m_1$ and $m_2$ are small, values of $R_1$ are far away from 0.05; for example, $R_1 = 0.012$ when $\sigma_1^2 = 6, \sigma_2^2 = 9, m_1 = 8, m_2 = 5$.

## 3. Mathematical development

### 3.1. $K$-sample problem

We shall formulate the problem in a framework of $K$-sample problem. Suppose that only summary statistics such as sample size, sample mean and variance, denoted by $m_i$, $\bar{X}_i$ and $S_i^2$, $i = 1, 2, \ldots, K$, are available. We assume in this paper that these statistics are obtained from the $i$-th sample from the population with normal distribution $\mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, K$.

Consider statistical test for testing hypothesis $H_0$:$c_1\mu_1 + c_2\mu_2 + \cdots + c_K\mu_K = 0$ against the alternative hypothesis $H_1$ that negates $H_0$, where $c's$ are given contrasts such that $\sum_{i=1}^K c_i = 0$. We employ statistic

$$T = \sum_{i=1}^K c_i \bar{X}_i \ / \ \sqrt{\sum_{i=1}^K c_i^2 S_i^2/m_i}$$

for testing the hypothesis. If $\sigma_1 = \sigma_2 = \cdots = \sigma_k$, $T$ follows a $t$-distribution with $\sum_{i=1}^K (m_i - 1)$ *df* under $H_0$. However, if it is not the case $T$ does not follow a $t$-distribution under $H_0$. Satterswaite (1946), extending the idea of Smith (1936) and Welch (1938), proposed to approximate the disribution of $T$ by $t$-distribution with adjusted *df* given by

$$\hat{\nu}_w = \left(\sum_{i=1}^K \frac{c_i^2 S_i^2}{m_i}\right)^2 \ / \ \left(\sum_{i=1}^K \frac{1}{m_i - 1}(\frac{c_i^2 S_i^2}{m_i})^2\right). \tag{1}$$

The $\hat{\nu}_w$ is called the Satterswaite-Welch *df*. It reduces to the Welch *df* when $K = 2$.

### 3.2. Definition of the *tdf*

We first define *tdf*, the degree of freedom that approximates the distribution of $T$ by $t$-distribution. In order for $T$ to follow $t$-distribution, the squared denominator of $T$ multiplied by constant must follows a chi-squared distribution. We define *tdf* as the *df* of chi-squared distribution whose first two moments agree with those moments of the distribution of the statistic. More precisely, putting, $A = \sum_{i=1}^K c_i^2 S_i^2/m_i$, and letting $Q_\nu$ be the statistic that follows a chi-squared distribution with $\nu$ *df*, *tdf* is defined as $\nu$ that satisfies

$$E(A) = E(\xi Q_\nu), \qquad Var(A) = Var(\xi Q_\nu), \tag{2}$$

where $\xi$ is an unknown parameter. We need following lemmas to get explicit form of *tdf*. The proofs of lemmas are straightforward and are omitted.

**Lemma 1.** *Let $S_i^2$ be the sample variance of data from a population with normal distribution $N(\mu_i, \sigma_i^2)$. Then*

$$\textit{(i)}\ E(S_i^2) = \sigma_i^2, \quad \textit{(ii)}\ V(S_i^2) = \frac{2\sigma_i^4}{(m_i - 1)}, \quad \textit{(iii)}\ E(S_i^4) = \frac{(m_i + 1)\sigma_i^4}{(m_i - 1)}.$$

**Lemma 2.** *Assume the assumption of Lemma 1, then (2) is represented by*

$$\begin{aligned} \sum_{i=1}^{K} \frac{c_i^2 \sigma_i^2}{m_i} &= \xi\nu, \\ \sum_{i=1}^{K} \frac{1}{m_i - 1} (\frac{c_i^2 \sigma_i^2}{m_i})^2 &= \xi^2\nu. \end{aligned}$$

From Lemma 2 and definition, we have

$$tdf = \left( \sum_{i=1}^{K} \frac{c_i^2 \sigma_i^2}{m_i} \right)^2 \left( \sum_{i=1}^{K} \frac{1}{m_i - 1} (\frac{c_i^2 \sigma_i^2}{m_i})^2 \right)^{-1}, \tag{3}$$

and

$$\xi = \left( \sum_{i=1}^{K} \frac{1}{m_i - 1} (\frac{c_i^2 \sigma_i^2}{m_i})^2 \right) / \left( \sum_{i=1}^{K} \frac{c_i^2 \sigma_i^2}{m_i} \right). \tag{4}$$

**Theorem 1.** *Assume the assumption of Lemma 1, then the distribution of $T$ under $H_0$ is approximated by a t-distribution with tdf degree of freedom.*

**Proof**. The proof of the theorem is given in Appendix.

Note that the Satterswaite-Welch *df* may be obtained by replacing unknown parameter $\sigma^2$ in formula (3) with $S_i^2$, namely from

$$\left( \sum_{i=1}^{K} \frac{c_i^2 S_i^2}{m_i} \right)^2 - \nu \sum_{i=1}^{K} \frac{1}{m_i - 1} \left( \frac{c_i^2 S_i^2}{m_i} \right)^2 = 0. \tag{5}$$

This estimating equation is not unbiased for *tdf*. An unbiased estimating equation for *tdf* is given in the following theorem.

**Theorem 2.** *Assume the assumption of Lemma 1, then an unbiased estimating equation for tdf is given by*

$$\left( \sum_{i=1}^{K} \frac{c_i^2 S_i^2}{m_i} \right)^2 - 2 \sum_{i=1}^{K} \frac{1}{m_i + 1} \left( \frac{c_i^2 S_i^2}{m_i} \right)^2 - \nu \sum_{i=1}^{K} \frac{1}{m_i + 1} \left( \frac{c_i^2 S_i^2}{m_i} \right)^2 = 0. \tag{6}$$

**Proof**. The proof of the theorem is given in Appendix.

Note that when $m = m_i$ for all $i = 1, 2, \ldots, K$ it follows that

$$\hat{\nu}_u = \hat{\nu}_w + \frac{2}{m-1}\Big(\hat{\nu}_w - (m-1)\Big),$$

where $\hat{\nu}_u$ and $\hat{\nu}_w$ are solutions of equations (6) and (5), respectively, thus $\hat{\nu}_u \geq \hat{\nu}_w$ since $\hat{\nu}_w \geq m-1$. We target $\hat{\nu}_u$, the estimator of *tdf* obtained from (6), in this paper. $\hat{\nu}_u$ is called the $u$-estimator in the introduction. The advantage of unbiasedness of estimating equations is discussed in Yanagaimoto and Yamamoto (Chapter 6 in Godambe, 1999).

## 4. Sampling from population with estimated mean and variance

### 4.1. Unbiased estimating equation

Generate $n_i$ random digits, denoted by $X^*_{i1}, X^*_{i2}, \ldots, X^*_{in_i}$, from $\mathrm{N}(\bar{x}_i, s_i^2)$ by assuming that $\bar{X}_i = \bar{x}_i$ and $S_i^2 = s_i^2$ are given constants, where $n_i$ may be different from $m_i$, $i = 1, 2, \ldots, K$. Put

$$\bar{X}^*_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X^*_{ij} \quad \textit{and} \quad S^{*2}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X^*_{ij} - \bar{X}^*_i)^2.$$

Note that $\bar{X}^*_i$ and $S^{*2}_i$ are conditional sample mean and sample variance conditioned on $\bar{X}_i = \bar{x}_i$ and $S_i^2 = s_i^2$. We have the following lemma.

**Lemma 3.** *Assume the assumption of Lemma1. Then unconditional expectations of* $S^{*2}_i$ *and* $S^{*4}_i = (S^{*2}_i)^2$ *are given by*

$$\text{(i)}\ \ E(S^{*2}_i) = \sigma_i^2, \qquad \text{(ii)}\ \ E(S^{*4}_i) = \frac{(m_i+1)(n_i+1)}{(m_i-1)(n_i-1)}\sigma_i^4.$$

(iii) *Furthermore, the unconditional variance of* $S^{*2}_i$ *is given by*

$$V(S^{*2}_i) = \frac{2(m_i+n_i)}{(m_i-1)(n_i-1)}\sigma_i^4.$$

**Proof**. The proof of the lemma is given in Appendix.

**Theorem 3.** *Assume the assumption of Lemma 1. Then an unbiased estimating equation for tdf based on* $\{X^*_{ij}\}_{j=1}^{n_i}$, $i = 1, 2, \ldots, K$, *is given by*

$$\left(\sum_{i=1}^{K}\frac{c_i^2 S^{*2}_i}{m_i}\right)^2 - 2\sum_{i=1}^{K}\left(\frac{c_i^2 S^{*2}_i}{m_i}\right)^2 \frac{(m_i+n_i)}{(m_i+1)(n_i+1)} = \nu\sum_{i=1}^{K}\left(\frac{c_i^2 S^{*2}_i}{m_i}\right)^2 \frac{(n_i-1)}{(m_i+1)(n_i+1)}. \quad (7)$$

**Proof.** The expectation of the left hand side of the equation is

$$\begin{aligned} & E\left(\sum_{i=1}^{K}\frac{c_i^2 S^{*2}_i}{m_i}\right)^2 - 2\sum_{i=1}^{K}\left(\frac{c_i^2}{m_i}\right)^2 \frac{(m_i+n_i)}{(m_i+1)(n_i+1)} E(S^{*4}_i) \\ = \ & E\left(\sum_{i=1}^{K}\frac{c_i^2}{m_i} S^{*2}_i\right)^2 - 2\sum_{i=1}^{K}\left(\frac{c_i^2}{m_i}\right)^2 \frac{(m_i+n_i)}{(m_i-1)(n_i-1)}\sigma_i^4 \end{aligned}$$

$$
\begin{aligned}
&= E\left(\sum_{i=1}^{K}\frac{c_i^2}{m_i}S_i^{*2}\right)^2 - \sum_{i=1}^{K}\left(\frac{c_i^2}{m_i}\right)^2 V(S_i^{*2}) \\
&= \left(\sum_{i=1}^{K}\frac{c_i^2\sigma_i^2}{m_i}\right)^2.
\end{aligned}
$$

On the other hand, the expectation of the right hand side of the equation in the theorem is

$$
\begin{aligned}
&\nu\sum_{i=1}^{K}\left(\frac{c_i^2}{m_i}\right)^2\frac{(n_i-1)}{(m_i+1)(n_i+1)}E(S_i^{*4}) \\
= \;&\nu\sum_{i=1}^{K}\left(\frac{c_i^2}{m_i}\right)^2\frac{1}{m_i-1}\sigma_i^4.
\end{aligned}
$$

Therefore the expectations of left and right hand sides agree with if $\nu$ is equal to the *tdf* given in formula (3). This completes the proof of the theorem.

It follows from Theorem 3 that the $u$-estimator for *tdf* based on $\{X_{ij}^*\}_{j=1}^{n_i}$ is given by

$$
\hat{\nu}^* = \left(\left(\sum_{i=1}^{K}U_i^*\right)^2 - 2\sum_{i=1}^{K}(U_i^*)^2\frac{(m_i+n_i)}{(m_i+1)(n_i+1)}\right)\left(\sum_{i=1}^{K}(U_i^*)^2\frac{(n_i-1)}{(m_i+1)(n_i+1)}\right)^{-1}, \quad (8)
$$

where $U_i^* = c_i^2 S_i^{*2}/m_i$.

### 4.2. Computer intensive confidence intervals for *tdf* and $p$-value

$\hat{\nu}^*$ is an estimator of *tdf*, but its distribution is not easy to develop. We construct an approximate 90% empirical confidence interval (CI) of *tdf* and $p$-value from the empirical distribution of $\hat{\nu}^*$ in this section.

First, we set up the *true* empirical CI of *tdf* as follows. (1-1) Give the values of $\mu_i$ and $\sigma_i^2$, generate random degits of size $m_i$ from $\mathrm{N}(\mu_i, \sigma_i^2)$, and compute $\bar{x}_i$ and $s_i^2$ for $i = 1, 2, \ldots, K$. (1-2) Compute $\hat{\nu}_u$ from the unbiased estimating equation given in Theorem 2. (1-3) Repeat this process, get the empirical distribution of $\hat{\nu}$, and obtain 5% and 95% points of the empirical distribution. Denote these points by $\nu_{0.05}$ and $\nu_{0.95}$. We call interval $(\nu_{0.05}, \nu_{0.95})$ the true 90% empirical CI of *tdf*.

Next, we construct the true CI of $p$-value as follows. Let $p_{low}$ and $p_{upp}$ be the values of $P(T > |t|)$ computed by assuming the distribution of $T$ being $t$-distribution with $\nu_{0.05}$ and $\nu_{0.95}$ degrees of freedom, respectively. Then it follows that

$$
p_{low} \leq p\text{-value} \leq p_{upp},
$$

since when $T$ follows a $t$-distribution with $\nu$ df the tail probability $P(|T| > x)$ is a decreasing function of $\nu$ for fixed $x > 0$. Furthermore, it follows that $(p_{low}, p_{upp})$ is a 90% empirical CI of $p$-value given $T = t$. We call it the true empirical conditional 90% CI of the $p$-value given $T = t$. Note that $(\nu_{0.05}, \nu_{0.95})$ and $(p_{low}, p_{upp})$ are available when $\mu_i$ and $\sigma_i^2$ are known.

Finally, when $\mu_i$ and $\sigma_i^2$ are unknown, we construct approximate 90% CIs of the *tdf* and $p$-value based on the empirical distribution of $\hat{\nu}^*$ as follows. (2-1) Putting $n_i = hm_i$ and denoting a given class of $h$ by $H$, where $m_i$ is the size of the original sample from the $i$-th population, generate $n_i$ random digits, $X_{i1}^*, X_{i2}^*, \ldots, X_{in_i}^*$, from $\mathrm{N}(\bar{x}_i, s_i^2)$, using $\bar{x}_i$ and $s_i^2$ that were given in step (1-1), $i = 1, 2, \ldots, K$. (2-2) Compute the estimator given in formula (8). (2-3) Repeat steps (2-1) $\sim$ (2-2) $L$ times and find the 5% and 95% points of the empirical distribution of $\hat{\nu}^*$; denote them by $\hat{\nu}_{low,h}^*$ and $\hat{\nu}_{upp,h}^*$. (2-4) Obtain $\hat{p}_{low,h}^*$ and $\hat{p}_{upp,h}^*$ from $\hat{\nu}_{low,h}^*$ and $\hat{\nu}_{upp,h}^*$ similarly as above. (2-5) Compare $(\hat{p}_{low,h}^*,\ \hat{p}_{upp,h}^*)$ with the true 90% empirical conditional CI of $p$-value, and select the $h \in H$ that makes $(\hat{p}_{low,h}^*,\ \hat{p}_{upp,h}^*)$ the closest to the true 90% empirical conditional CI of $p$-values. Since the bounds of the interval changes monotonically with the increase of $h$, it will not take time to find the best $h$. The intervals with the best $h$, denoted by $(\hat{\nu}_{low}^*,\ \hat{\nu}_{upp}^*)$ and $(\hat{p}_{low}^*,\ \hat{p}_{upp}^*)$, are called the c-intensive approximate 90% CIs of *tdf* and $p$-value, respectively.

## 5. $U$-test

For testing $H_0$:$c_1\mu_1 + c_2\mu_2 + \cdots + c_K\mu_K = 0$ with $T$, we approximate the distribution of $T$ under $H_0$ by $t$-distribution with $\hat{\nu}$ degree of freedom, where $\hat{\nu}$ is the $u$-estimator of *tdf* obtained from the unbiased estimating function given in Theorem 2. We call this test the $U$-test. The $p$-value of the $U$-test is given by $P(|T| > |t||H_0)$, where $t$ is the observed value of $T$. As the Welch test, the $U$-test employs the estimated *tdf*, i.e., $\hat{\nu}$, in computing $p$-value and it depends on the variability of both $\hat{\nu}$ and $t$. We suggest to express the test results by $p$-value with its c-intensive approximate 90% CI. It could be useful for a cautious evaluation of the $p$-value. For example, we may reject $H_0$ if $\hat{p}_{upp}^* < 0.05$, but not reject, or retain hypothesis $H_0$ if $\hat{p}_{upp}^* \geq 0.05$, even if $p$-value $< 0.05$.

## 6. Simulation

Simulation was conducted to investigate the behavior of the proposed method when $K = 2$. First we established the true empirical CI of *tdf* as follows. (1) Generating random digits of size $m_i$ from $\mathrm{N}(\mu_i, \sigma_i^2)$ for given values of $\mu_i$ and $\sigma_i^2$, and computed $\bar{x}_i$ and $s_i^2$ for $i = 1, 2$ when $\mu_1 = 2.0, 2.25, 2.3, 2.5, 2.75, 2.8, 2.9, 3.3, 3.4, 3.8, 3.9, 4.4$, $\mu_2 = 0$; $\sigma_1^2 = 3$, $\sigma_2^2 = 6, 12, 18, 24$; $m_1 = 5$ and $m_2 = 7, 10, 20$. (2) Obtained $\hat{\nu}_u$ from the unbiased estimating equation given in Theorem 2. (3) Repeated this process 5,000 times, and obtained the true empirical CI of *tdf*. (4) Then for $H = \{1, 2, 3\}$ generated random digits of size $n_i = hm_i$ from $\mathrm{N}(\bar{x}_{ij},\ s_{ij}^2)$, $i = 1, 2$ and $h \in H$, where $(\bar{x}_{ij},\ s_{ij}^2)$ was the value of $(\bar{x}_i,\ s_i^2)$ obtained in step (1) in the $j$-th run, $j = 1, 2, \ldots, 5,000$. (5) Repeated the step (4) 1,000 times (i.e. $L = 1,000$) and obtained 90% interval $(\hat{\nu}_{j,low,h}^*, \hat{\nu}_{j,upp,h}^*)$ for each $j$ ($j = 1, 2, \ldots, 5000$) and $h \in H$. (6) Finally, $\hat{\nu}_{low,h}^*$ and $\hat{\nu}_{upp,h}^*$ were obtained by averaging those 5,000 lower and upper bounds for each $h \in H$.

Table 2 summarizes the results of the simulation. The first six columns in the table are values of $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$, $m_1$ and $m_2$. The 7th and 8th columns list the values of the true 90% empirical CI. The remaining columns list c-intensive intervals of *tdf* for $h \in H = \{1, 2, 3\}$. Table 3 lists the values of conditional $(\hat{p}_{low},\ \hat{p}_{upp})$ and $(\hat{p}_{low,h}^*,\ \hat{p}_{upp,h}^*)$ given $t = 2.2$ for each $h \in H$. These intervals are computed from corresponding intervals of *tdf* listed in Table 2. Table 3 shows that those conditional lower and upper bounds of $p$-value given $t = 2.2$ are fairly stable throughout the experimental conditions and that

the values of the lower band of the c-intensive CIs for $p$-value are not decreasing with the increase of $h$, and those of the upper band do not increase as the increase of $h$; that many values of the lower bound listed under $h = 1$, $h = 2$ and $h = 3$ are fairly close to the true bounds; and that those values of the upper bounds listed under $h = 1$ are the closest to those true values among $h = 1, 2, 3$. Thus we suggest to use $(\hat{p}^*_{low}, \hat{p}^*_{upp})$ obtained when $h = 1$ as the c-intensive approximate 90% CI of $p$-values.

Table 2 The true CI and c-intensive CI of *tdf* when $h = 1, 2, 3$

| | | | | | | true CI | | c-intensive CI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean, SD and sample size | | | | | | | | $h=1$ | | $h=2$ | | $h=3$ | |
| $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $m_1$ | $m_2$ | $\hat{\nu}_{low}$ | $\hat{\nu}_{upp}$ | $\hat{\nu}^*_{low}$ | $\hat{\nu}^*_{upp}$ | $\hat{\nu}^*_{low}$ | $\hat{\nu}^*_{upp}$ | $\hat{\nu}^*_{low}$ | $\hat{\nu}^*_{upp}$ |
| 2.9 | 0 | 3 | 6 | 5 | 7 | 7.00 | 11.99 | 6.88 | 14.58 | 7.85 | 12.84 | 8.29 | 12.25 |
| 3.4 | 0 | 3 | 12 | 5 | 7 | 6.84 | 11.98 | 6.93 | 14.20 | 7.77 | 12.38 | 8.15 | 11.77 |
| 3.9 | 0 | 3 | 18 | 5 | 7 | 6.59 | 11.94 | 6.81 | 13.77 | 7.49 | 11.86 | 7.80 | 11.23 |
| 4.4 | 0 | 3 | 24 | 5 | 7 | 6.45 | 11.87 | 6.69 | 13.35 | 7.25 | 11.39 | 7.51 | 10.74 |
| 2.3 | 0 | 3 | 6 | 5 | 10 | 7.35 | 14.98 | 8.61 | 17.57 | 9.56 | 15.77 | 10.01 | 15.07 |
| 2.8 | 0 | 3 | 12 | 5 | 10 | 9.81 | 14.99 | 9.59 | 17.42 | 10.63 | 15.78 | 11.11 | 15.23 |
| 3.3 | 0 | 3 | 18 | 5 | 10 | 9.98 | 14.98 | 9.82 | 17.15 | 10.78 | 15.50 | 11.22 | 14.95 |
| 3.8 | 0 | 3 | 24 | 5 | 10 | 9.85 | 14.97 | 9.84 | 16.85 | 10.69 | 15.17 | 11.08 | 14.61 |
| 2.0 | 0 | 3 | 6 | 5 | 20 | 6.19 | 24.27 | 9.23 | 26.45 | 9.81 | 21.62 | 10.18 | 19.49 |
| 2.25 | 0 | 3 | 12 | 5 | 20 | 8.52 | 24.91 | 12.62 | 27.45 | 13.49 | 24.76 | 14.02 | 23.36 |
| 2.5 | 0 | 3 | 18 | 5 | 20 | 10.87 | 24.96 | 14.87 | 27.57 | 15.90 | 25.60 | 16.49 | 24.64 |
| 2.75 | 0 | 3 | 24 | 5 | 20 | 13.13 | 24.98 | 16.41 | 27.55 | 17.52 | 25.83 | 18.12 | 25.10 |

Table 3 $(\hat{p}_{low}, \hat{p}_{upp})$ and $(\hat{p}^*_{low}, \hat{p}^*_{upp})$ when $h = 1, 2, 3$

| | | | | | | true CI | | c-intensive CI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mean, SD and sample size | | | | | | | | $h=1$ | | $h=2$ | | $h=3$ | |
| $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $m_1$ | $m_2$ | $\hat{p}_{low}$ | $\hat{p}_{upp}$ | $\hat{p}^*_{low}$ | $\hat{p}^*_{upp}$ | $\hat{p}^*_{low}$ | $\hat{p}^*_{upp}$ | $\hat{p}^*_{low}$ | $\hat{p}^*_{upp}$ |
| 2.9 | 0 | 3 | 6 | 5 | 7 | 0.048 | 0.064 | 0.044 | 0.064 | 0.047 | 0.060 | 0.048 | 0.058 |
| 3.4 | 0 | 3 | 12 | 5 | 7 | 0.048 | 0.065 | 0.045 | 0.064 | 0.047 | 0.060 | 0.049 | 0.058 |
| 3.9 | 0 | 3 | 18 | 5 | 7 | 0.048 | 0.066 | 0.045 | 0.065 | 0.048 | 0.061 | 0.050 | 0.060 |
| 4.4 | 0 | 3 | 24 | 5 | 7 | 0.048 | 0.067 | 0.046 | 0.065 | 0.049 | 0.062 | 0.051 | 0.061 |
| 2.3 | 0 | 3 | 6 | 5 | 10 | 0.044 | 0.062 | 0.041 | 0.057 | 0.043 | 0.054 | 0.044 | 0.052 |
| 2.8 | 0 | 3 | 12 | 5 | 10 | 0.044 | 0.053 | 0.042 | 0.054 | 0.043 | 0.051 | 0.044 | 0.050 |
| 3.3 | 0 | 3 | 18 | 5 | 10 | 0.044 | 0.052 | 0.042 | 0.053 | 0.043 | 0.051 | 0.044 | 0.050 |
| 3.8 | 0 | 3 | 24 | 5 | 10 | 0.044 | 0.053 | 0.042 | 0.053 | 0.044 | 0.051 | 0.044 | 0.050 |
| 2.0 | 0 | 3 | 6 | 5 | 20 | 0.038 | 0.069 | 0.037 | 0.055 | 0.039 | 0.053 | 0.040 | 0.052 |
| 2.25 | 0 | 3 | 12 | 5 | 20 | 0.037 | 0.057 | 0.036 | 0.047 | 0.037 | 0.046 | 0.038 | 0.045 |
| 2.5 | 0 | 3 | 18 | 5 | 20 | 0.037 | 0.050 | 0.036 | 0.044 | 0.037 | 0.043 | 0.037 | 0.042 |
| 2.75 | 0 | 3 | 24 | 5 | 20 | 0.037 | 0.046 | 0.036 | 0.042 | 0.037 | 0.041 | 0.037 | 0.041 |

## 7. Conclusion

Results of the simulation suggest that the $p$-value of the $U$-test for testing $H_0$:$c_1\mu_1 + c_2\mu_2 + \cdots + c_K\mu_K = 0$ is better to be evaluated by using the c-intensive approximate 90% CI of $p$-value constructed by the the size of sample $n_i = m_i$, $i = 1, 2. \ldots, K$.

## 8. Discussion

The u-estimator of *tdf* is not substantially different from the Welch *df*. Table 4 lists the values of the Welch *df* and u-estimator of *tdf* obtained from the simulation that

was carried out by 5,000 run by generating random samples of size $m_1$ from $\mathrm{N}(\mu_1, \sigma_1^2)$ and of size $m_2$ from $\mathrm{N}(\mu_2, \sigma_2^2)$, where values of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, $m_1$ and $m_2$ employed are listed in Table 4. The table shows relatively large differences between the Welch *df* and u-estimator of the *tdf*; for example, 8.71 vs. 10.21 in id 1, or 18.78 vs. 21.46 in id 12. However those differences are not substantial when evaluated by means of the right tail probabilities. Table 5 lists the right tail probabilities of $t$-distribution in the range of $t \geq 1.0$ when the values of *df* are 8.71 (id 1 Welch), 10.21 (id 1 u-estimator), 18.78 (id 12 Welch) and 21.46 (id 12 u-estimator). The table indicates that differences of $p$-values are small between two *df*'s.

Looking at Table 5 the referee of the present paper pointed out that the tail probabilities of t-distribution with 8.71 *df* and 21.46 *df* were not substantially different, indicating the stability of p-values and thus no need of constructing the CI of $p$-values regardless of the variation of estimated *df*. Our answer to the comment is as follows. The finding from Table 1 is true; namely, the impact of the variation of estimated *df* on the level of the test, and also on the p-value, could be substantial. Readers might have the feeling of contradiction between the comment and this statement. To resolve it they are asked to understand that the right tail probability of t-distribution could undergoes significant fluctuation due to small shift of *df* in the range of small values, for example, when $t = 2.2$ the right tail probabilities of $df = 4$ and $df = 8$ are 0.046 and 0.029, respectively. To express the point more clearly we obtained the empirical distribution of Welch's *df* by generating 5,000 sets of normal random digits when $\sigma_1^2 = 3$, $\sigma_2^2 = 6$, $m_1 = 5$ and $m_2 = 20$. The summary statistics of the distribution were given as follows.

min = 4.35, max = 23.00, mean=10.82, SD=4.95, 90% CI=(5.47, 21.73).

When $t = 2.00$, the upper tail probabilities of t-test with the *df* corresponding to the above summary statistics are obtained as

min = 0.029, max = 0.055, mean=0.036 and 90% CI=(0.029, 0.049).

The referee also pointed out that over-all behavior of the Welch test was not so bad. Look at the id.13 in Table 1. The empirical levels of the test on $R_1$, $R_2$, $R_3$ and $R_4$ are 0.008, 0.020, 0.048, 0.146, respectively; the levels on $R_1$ and $R_4$ are far from the nominal level. However, the average of those levels is 0.055, fairly close to 5%, indicating that if we apply the Welch test many times repeatedly, then we get over-all satisfaction. However, we stress that the Welch test is often applied to data that repetition is never conceived. For example, the repetition is not conceived in many clinical studies and yet decision is done based on $p$-value from each study. Such is the case that a cautious evaluation of the $p$-values would be absolutely-required.

Table 4 Values of the Welch *df* and u-estimator.
†: u-estimator for *tdf*.

| | mean, SD and sample size | | | | | | Welch *df* | *udf*† |
|---|---|---|---|---|---|---|---|---|
| id | $\mu_1$ | $\mu_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $m_1$ | $m_2$ | $\hat{\nu}_w$ | $\hat{\nu}_u$ |
| 1 | 2.9 | 0 | 3 | 6 | 5 | 7 | 8.71 | 10.21 |
| 2 | 3.4 | 0 | 3 | 12 | 5 | 7 | 8.61 | 9.83 |
| 3 | 3.9 | 0 | 3 | 18 | 5 | 7 | 8.28 | 9.27 |
| 4 | 4.4 | 0 | 3 | 24 | 5 | 7 | 7.99 | 8.81 |
| 5 | 2.3 | 0 | 3 | 6 | 5 | 10 | 10.49 | 12.31 |
| 6 | 2.8 | 0 | 3 | 12 | 5 | 10 | 11.58 | 13.21 |
| 7 | 3.3 | 0 | 3 | 18 | 5 | 10 | 11.74 | 13.09 |
| 8 | 3.8 | 0 | 3 | 24 | 5 | 10 | 11.64 | 12.78 |
| 9 | 2 | 0 | 3 | 6 | 5 | 20 | 10.82 | 13.05 |
| 10 | 2.25 | 0 | 3 | 12 | 5 | 20 | 14.79 | 17.58 |
| 11 | 2.5 | 0 | 3 | 18 | 5 | 20 | 17.22 | 22.05 |
| 12 | 2.75 | 0 | 3 | 24 | 5 | 20 | 18.78 | 21.46 |

Table 5 Right tail probabilities of *t*-distribution when the values of *df* are selected from Table 4.
Selected are those *df* of id.1 and of id.12 in Table 4.

†: u-estimator

| | id 1 in Table 4 | | id 12 in Table 4 | |
|---|---|---|---|---|
| | Welch *df* | *udf*† | Welch *df* | *udf*† |
| t | 8.71 | 10.21 | 18.78 | 21.46 |
| 1.0 | 0.173 | 0.170 | 0.166 | 0.164 |
| 1.2 | 0.132 | 0.129 | 0.123 | 0.121 |
| 1.4 | 0.100 | 0.096 | 0.090 | 0.088 |
| 1.6 | 0.074 | 0.070 | 0.064 | 0.062 |
| 1.8 | 0.055 | 0.051 | 0.045 | 0.043 |
| 2.0 | 0.040 | 0.037 | 0.031 | 0.029 |
| 2.2 | 0.029 | 0.026 | 0.021 | 0.019 |
| 2.4 | 0.022 | 0.019 | 0.014 | 0.013 |
| 2.6 | 0.016 | 0.013 | 0.009 | 0.008 |
| 2.8 | 0.012 | 0.009 | 0.006 | 0.005 |
| 3.0 | 0.009 | 0.007 | 0.004 | 0.003 |
| 3.2 | 0.006 | 0.005 | 0.003 | 0.002 |
| 3.4 | 0.005 | 0.003 | 0.002 | 0.001 |
| 3.6 | 0.003 | 0.002 | 0.001 | 0.001 |

**Acknowledgement**

The authors are grateful to the anonymous referee for his careful reading and helpful comments.

## References

Boneau, C.A. (1960). The effects of violations of assumptions underlying the $t$-test, *Psychological Bulletin* , **57**, 49–64.

Box, G.E.P. (1953). Non-normality and tests of variances, *Biometrika* , **40**, 318–335.

Cochran,W.G. (1947). Some consequences when the assumptions for analysis of variance are not satisfied, *Biometrics* , **3**, 22–38.

Godambe, V.P. Edited (1991). Estimating Functions, *Oxford Science Publications, Clarrendon Press, Oxford.*

Lehmann, E.L. and Romano, J.P. (2005). Testing Statistical Hypotheses-Third Edition, *Springer, New York. Oxford Science Publications, Clarrendon Press, Oxford.*

Miller, R.G. (1986). Beyond Anova, *John Wiley, New York.*

Posten, H.O., H.C. Yeh, and D.B. Owen (1982). Robustness of the two-sample $t$-test under violations of the homogeneity of variance assumption, *Communication in Statistics-Theory and Method* , **11**, 2109–126.

R. A. FISHER (1939). The comparison of samples with possibly unequal variances, *Annals of Eugenics* , **9**, 174–180.

Satterswaite, F.E. (1946). An approximate distribution of estimates of variance component, *Biometrics Bulletin* , **2**, 110–114.

Smith, R.A. (1936). The problem of comparing the results of two experiments with unequal errors, *J. Council Sci. Industrial Res* , **9**, 211–212.

Srivastava, A.B. (1958). Effects of non-normality on the power function of the $t$-test, *Biometrika* , **45**, 421–430.

Welch, B.L. (1938). The significance of the difference between two means when the population variance are unequal, *Biometrika* , **29**, 350–362.

W. V. Behrens. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen, *Landwirtschaftliche Jahrbücher*, **68**, 807–837.

## Appendix: Proofs of Theorems

**Proof of Theorem 1**

First we note that $\sum c_i \bar{X}_i$ and $\sum (c_i^2 \sigma_i^2 / m_i)$ are mutually independent, since $\bar{X}_i$ and $S_i^2$ are independent from the assumption of normality of underlying population distribution. Next we note that the distribution of $T$ is approximated by the distribution of

$$T^* = \sum c_i \bar{X}_i / \sqrt{\xi Q_{tdf}},$$

where *tdf* and $\xi$ are given in formulae (3) and (4). Now $Z = \sum c_i \bar{X}_i / \sqrt{\sum c_i^2 \sigma_i^2 / m_i}$ follows standard normal distribution under $H_0$, and by using formulae (3) and (4) we may represent $T^*$ by $T^* = Z/\sqrt{Q_{tdf}/tdf}$. Thus $T^*$ follows $t$-distribution with *tdf* degrees of freedom. This completes the proof of Theorem 1.

**Proof of Theorem 2**

It follows from Lemma 1 that

$$\begin{aligned} E\left(\sum \frac{c_i^2 S_i^2}{m_i}\right)^2 &= Var\left(\sum \frac{c_i^2 S_i^2}{m_i}\right) + \left(E\left(\sum \frac{c_i^2 S_i^2}{m_i}\right)\right)^2 \\ &= 2\sum \frac{1}{m_i - 1}\left(\frac{c_i^2 \sigma_i^2}{m_i}\right)^2 + \left(\sum \frac{c_i^2 \sigma_i^2}{m_i}\right)^2 \end{aligned}$$

and

$$\begin{aligned} E\left(\sum \frac{1}{m_i+1}\left(\frac{c_i^2 S_i^2}{m_i}\right)^2\right) &= \sum \frac{1}{m_i+1}\left(Var\left(\frac{c_i^2 S_i^2}{m_i}\right) + \left(E\left(\frac{c_i^2 S_i^2}{m_i}\right)\right)^2\right) \\ &= 2\sum \frac{1}{(m_i+1)(m_i-1)}\left(\frac{c_i^2 \sigma_i^2}{m_i}\right)^2 + \sum \frac{1}{m_i+1}\left(\frac{c_i^2 \sigma_i^2}{m_i}\right)^2. \end{aligned}$$

Therefore the expectation of the left hand side of the equation is

$$\begin{aligned} &E\left(\left(\sum \tfrac{c_i^2 S_i^2}{m_i}\right)^2 - 2\sum \tfrac{1}{m_i+1}\left(\tfrac{c_i^2 S_i^2}{m_i}\right)^2 - \nu \sum \tfrac{1}{m_i+1}\left(\tfrac{c_i^2 S_i^2}{m_i}\right)^2\right) \\ &\quad = \left(\sum \tfrac{c_i^2 \sigma_i^2}{m_i}\right)^2 - \nu \sum \tfrac{1}{m_i-1}\left(\tfrac{c_i^2 \sigma_i^2}{m_i}\right)^2. \end{aligned}$$

Thus the expectation is zero if $\nu = tdf$, where *tdf* is given in formula (3). This complete the proof of Theorem 2.

**Proof of Lemma 3**

(i) From Lemma 1

$$E(S_i^{*2}) = E\left(E(S_i^{*2})|s_i^2\right) = E(S_i^2) = \sigma^2.$$

(ii) Also from Lemma 1

$$\begin{aligned} E(S_i^{*4}) &= E\left(E(S_i^{*4})|s_i^2\right) = E\left(\frac{n_i+1}{n_i-1} S_i^4\right) \\ &= \frac{(n_i+1)(m_i+1)}{(n_i-1)(m_i-1)} \sigma_i^4. \end{aligned}$$

*Received October 19, 2012*
*Revised August 20, 2013*