The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 775 - 785

Robustness for Pairwise Multiple Comparison Procedures with Trimmed Means under Violated Assumptions : Bonferroni, Shaffer, and Welsch Procedure

Hyunchul Kim[1)]

Abstract

Robustness rates for repeated measures pairwise multiple comparison procedures were investigated in a split plot design with one between- and one within-subjects factor using untrimmed and trimmed data. Five factors were manipulated in the study: distribution, sphericity, variance-covariance heteroscedasticity, total sample size, and sample size ratio. The Welsch test (W) and the Welsch test on trimmed data (W_{RT}) performed better than the other procedures, but had a liberal tendency. The trimmed difference score Bonferroni procedure (B_{DT}) was a good choice in some conditions.

1. Introduction

There has been a limited amount of research on the robustness of pairwise multiple comparison procedures when the normality and/or multisample sphericity assumption are violated in a split plot design. However, results in Keselman, Keselman, and Shaffer (1991), Yuen (1974), and Wilcox (1992) suggest a Welch-type test using trimmed means and Winsorized variances may be robust for long-tailed distributions. Studies show that long-tailed distributions are common in psychometric and educational measures (Micceri, 1989; Wilcox, 1990). Consequently, the development and evaluation of multiple comparison procedures using trimmed means and Winsorized variances is of interest in connection with testing contrasts on levels of the within-subjects factor.

Kim (1997) investigated four pairwise multiple comparison procedures in a split plot design with one between- and one within-subjects factor. Both Type I error rates and power for procedures were estimated when the assumptions for the procedures are violated. In comparing Bonferroni procedure (B), Welsch's step-up procedure (W), Shaffer's procedure

1) Researcher, Korean Educational Development Institute, 92-6 Umyeon-Dong, Seocho-Gu, Seoul 137-791, Korea.

following the $\tilde{\varepsilon}$-adjusted F test ($S(\tilde{\varepsilon})$), and Shaffer's procedure following the corrected Improved General Approximation (CIGA) test (S(C)), Kim concluded that Welsch's (1977a) procedure provides adequate control of Type I error rates and is typically more powerful than the other procedures. Conditions in which dispersion matrices were equal were included as were conditions in which dispersion matrices were unequal. Conditions with nonsphericity and symmetrically distributed multivariate nonnormal data were included.

The classical statistics such as the sample mean and variance are sensitive to outliers. Trimming and Winsorization refer to the removal and modification of the extreme values of a sample. Let $y_{(i)}$ denote the ith order observation in a random sample of size N with $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$. The α-trimmed mean for $\alpha = g/N$ or the g-times trimmed mean is

$$\bar{y}_{tg} = \frac{1}{(N-2g)} \sum_{i=g+1}^{N-g} y_{(i)}. \tag{1.1}$$

The g-times Winsorized mean and sum of squared deviations are, respectively,

$$\bar{y}_{wg} = \frac{1}{N}[g y_{(g+1)} + \sum_{i=g+1}^{N-g} y_{(i)} + g y_{(N-g)}], \tag{1.2}$$

and,

$$SSD_{wg} = g(y_{(g+1)} - \bar{y}_{wg})^2 + \sum_{i=g+1}^{N-g} (y_{(i)} - \bar{y}_{wg})^2 + g(y_{(N-g)} - \bar{y}_{wg})^2. \tag{1.3}$$

Tukey and McLaughlin (1963) found that the Winsorized sample variance is a suitable estimate of the variance of the trimmed mean.

Three variations of the Bonferroni procedures were included: (a) the Bonferroni (B), (b) the raw score trimmed Bonferroni (B_{RT}), and (c) the trimmed difference score Bonferroni (B_{DT}). Six variations of the Shaffer's modified sequentially rejective Bonferroni (MSRB) procedures were included: (a) Shaffer's MSRB procedure using $\tilde{\varepsilon}$-adjusted test for omnibus test ($S(\tilde{\varepsilon})$), (b) Shaffer's MSRB procedure using $\tilde{\varepsilon}$-adjusted test for omnibus test with the raw score trimmed data ($S(\tilde{\varepsilon})_{RT}$), (c) Shaffer's MSRB procedure using $\tilde{\varepsilon}$-adjusted test for omnibus test with the trimmed difference score ($S(\tilde{\varepsilon})_{DT}$), (d) Shaffer's MSRB procedure using the CIGA test for omnibus test (S(C)), (e) Shaffer's raw score trimmed MSRB procedure using the trimmed CIGA test for omnibus test ($S(C)_{RT}$), and (f) Shaffer's trimmed difference score MSRB procedure using the trimmed CIGA test for omnibus test ($S(C)_{DT}$). Two variations of Welsch's step-up procedures were included: (a) Welsch's step-up procedures (W) and (b) raw score trimmed Welsch procedure (W_{RT}). To produce a trimmed version 20% trimming was used.

The Bonferroni procedures were included primarily because multiple comparison procedures using the Bonferroni critical value are well known and commonly employed. The potential advantage of $S(\tilde{\varepsilon})$ over B is that the critical values for the $S(\tilde{\varepsilon})$ procedure are uniformly

less extreme than the critical values for B. The potential disadvantage is that a lower ranked test statistics can be declared nonsignificant because of a nonsignificant test in a prior step. Nevertheless, MSRB procedures tend to be more powerful than Bonferroni procedures (Keselman, 1994). The potential advantage of Welsch's step-up procedures is better power than Shaffer's procedures because the step-up procedures declare significance by implication; in contrast Shaffer's procedures declare nonsignificance by implication.

2. Method

2.1 Multiple Comparison Procedures

1. Bonferroni procedures

Three variations of the Bonferroni procedure were investigated. In the first procedure, the test statistic was

$$\frac{\sum_j [\bar{y}_{jk} - \bar{y}_{jk'}]/J}{\sqrt{\sum_j [(S_{jk}^2 + S_{jk'}^2 - 2S_{jkk'})/n_j]/J^2}}. \tag{2.1}$$

In equation (2.1) $\bar{y}_{jk}$ and S_{jk}^2 are the mean and variance of y_k at the jth level of the between-subjects factor; $S_{jkk'}$ is the covariance between y_k and $y_{k'}$ at the jth level of the between-subjects factor. The critical value was $\pm t_{\alpha_{pc}/2, \nu_1}$, where $\alpha_{pc} = 2\,\alpha_{fw} / [K(K-1)]$ is the per comparison error rate and α_{fw} is the familywise error rate and

$$\nu_1 = \frac{[\sum_j (S_{jk}^2 + S_{jk'}^2 - 2S_{jkk'})/n_j)]^2}{\sum_j \frac{[S_{jk}^2 + S_{jk'}^2 - 2S_{jkk'})/n_j]^2}{n_j - 1}}. \tag{2.2}$$

This procedure is developed by Keselman, Keselman, and Shaffer (1991). Subsequently, this procedure is referred to as the Bonferroni (B) procedure.

In the second procedure $\bar{y}_{jk}$ and $\bar{y}_{jk'}$ were replaced by g-times trimmed means and S_{jk}^2, $S_{jk'}^2$, and $S_{jkk'}$ were replaced by Winsorized variances and covariances, respectively. The Winsorized covariances is defined as

$$S_{jkk'}^{*} = \sum_j \frac{(y_{ijk}^{*} - \bar{y}_{jk}^{*})(y_{ijk'}^{*} - \bar{y}_{jk'}^{*})}{(n_j - 2g - 1)}, \tag{2.3}$$

where y_{ijk}^{*} is a Winsorized data point and $\bar{y}_{jk}^{*}$, $\bar{y}_{jk'}^{*}$ are Winsorized means. The critical value was $\pm t_{\alpha_{pc}/2, \nu_2}$. The degrees of freedom υ_2 are obtained by replacing S_{jk}^2 and $S_{jkk'}$ by

Winsorized variances and covariances, respectively; the quantities n_j-1 were replaced by n_j-2g-1. The second procedure is referred to as the raw score trimmed Bonferroni procedure (B_{RT}).

Let $d_{kk'} = y_k - y_{k'}$. In the third procedure the test statistic was

$$\frac{\sum_j \bar{d}_{jkk'}/J}{\sqrt{\sum_j [\,S_{jkk'}^{*2}/n_j]/J^2}}, \tag{2.4}$$

where $\bar{d}_{jkk'}^{*}$ is the trimmed mean of $d_{jkk'} = y_{jk} - y_{jk'}$ and $S_{jkk'}^{*}$ is the Winsorized standard deviation of $d_{jkk'} = y_{jk} - y_{jk'}$. The critical value was $\pm t_{\alpha_{FW}/2,\,\nu_3}$, where

$$\nu_3 = \frac{[\sum_j S_{jkk'}^{*2}/(n_j-2g)]^2}{\sum_j \frac{[\,S_{jkk'}^{*2}/(n_j-2g)]^2}{n_j-2g-1}}. \tag{2.5}$$

The third procedure is called the trimmed difference score Bonferroni procedure (B_{DT}).

2. Shaffer's MSRB procedures

The MSRB procedure is a modification of the Bonferroni procedure which is implemented in several steps. The first step is to test H_0: $\bar{\mu}_1 = \cdots = \bar{\mu}_K$. In the first MSRB procedure, H_0 was tested by using the $\tilde{\varepsilon}$-adjusted test. If H_0 is not rejected, all pairwise comparisons are declared nonsignificant. If H_0 is rejected, test statistics for the K(K-1)/2 pairwise comparisons are calculated and ranked by their p values from smallest to largest. In the first MSRB procedure the test statistic was calculated by using equation (2.1) and its degrees of freedom were calculated by using equation (2.2). The critical value for the test statistic with rank i (i = 1, ..., K(K-1)/2) is $\pm t_{\alpha_{FW}/2c_i,\,\nu_1}$. In this study, K = 4; for K = 4, K(K-1)/2 = 6 and there are six c_i: $c_1 = c_2 = c_3 = c_4 = 3$; $c_5 = 2$; and $c_6 = 1$. Testing is conducted in steps. If the test statistic in the *i*th step is not significant, all subsequent steps are declared nonsignificant. The first MSRB procedure is denoted by $S(\tilde{\varepsilon})$.

In the second procedure MSRB procedure, H_0 is tested by using a trimmed $\tilde{\varepsilon}$-adjusted test. That is, in the calculation of F and $\tilde{\varepsilon}$, trimmed means replace means and Winsorized variances and covariances replace variances and covariances. In subsequent steps the test statistics and degrees of freedom employed in B_{RT} are used. The second MSRB procedure is denoted by $S(\tilde{\varepsilon})_{RT}$. The third MSRB procedures, denoted by $S(\tilde{\varepsilon})_{DT}$, used the trimmed $\tilde{\varepsilon}$ -adjusted test in the first step; in subsequent steps the test statistics and degrees of freedom used in B_{DT} are used.

The fourth MSRB procedure is the same as the first, except the CIGA test developed by

Algina (1994) based on Lecoutre's (1991) results replaces the $\tilde{\varepsilon}$-adjusted test. The fourth MSRB test is denoted by S(C). In the fifth MSRB test, denoted by $S(C)_{RT}$, a raw score trimmed CIGA test is used to H_0. In this test, the means, variances and covariances used in the CIGA test were replaced by trimmed means and Winsorized variances and covariances, respectively. The test statistics and degrees of freedom employed in the subsequent steps are the same as in $S(\tilde{\varepsilon})_{RT}$. In the sixth MSRB procedure, denoted by $S(C)_{DT}$, the trimmed CIGA test is again used in the first step. The test statistics and degrees of freedom employed in the subsequent steps are the same as in $S(\tilde{\varepsilon})_{DT}$.

3. Welsch's step-up procedures

In Welsch's step-up procedure the means for all K levels of the within-subjects factor are ordered from smallest to largest. The procedure begins by testing all 2-ranges, $\overline{Y}_{(k+1)}^{*} - \overline{Y}_{(k)}^{*}$ $(1 \leq k \leq K)$, of the ordered treatment means of each level of the within-subjects factor, $\overline{Y}_{(1)}^{*} \leq \overline{Y}_{(2)}^{*} \leq \cdots \leq \overline{Y}_{(K)}^{*}$. If the 2-range is significant, the members of the corresponding pair of treatments are declared different. Also declare the p-ranges containing that 2-range significant and all sets of treatments which contain that significant subset as heterogeneous by implication without further tests. If at least one 2-range is not significant, then proceed to test 3-ranges. In general, test a p-range $\overline{Y}_{(k+p-1)}^{*} - \overline{Y}_{(k)}^{*}$ $(1 \leq k \leq K-p+1,\ 2 \leq p \leq K)$, if that p-range is not declared significant by implication at earlier steps. If $\overline{Y}_{(k+p-1)}^{*} - \overline{Y}_{(k)}^{*}$ is significant, then declare the pair of treatments corresponding to $\overline{Y}_{(k+p-1)}^{*}$ and $\overline{Y}_{(k)}^{*}$ as different. Also declare all sets of treatments containing that subset of p treatments as heterogeneous and the corresponding q-ranges containing that p-range for all $q > p$ as significant by implication without further tests. Welsch proposed using $\alpha_p = p\alpha / K$ for $2 \leq p \leq K-2$, and $\alpha_{K-1} = \alpha_K = \alpha$. Welsch (1977b) tabulated the critical values $\xi_{\alpha_p, \nu}$ $(2 \leq p \leq K)$ based on Studentized range statistics for k = 2, 3, ..., 10, v = 5, 6, ..., 20, 24, 30, 40, 60, 120, ∞, and α = .05 which he obtained using Monte Carlo simulations. Critical values in this study were obtained by interpolation in the table.

In the first variation of Welsch's step-up procedure, denoted by W, the test statistic was calculated using (2.1). The critical values, denoted by $\xi_{\alpha_p, \nu_1}/\sqrt{2}$ are presented in Keselman (1994) for integer values of v_1. In the second variation of Welsch's procedure the ordered means used in the first variation are replaced by $\sum_j \overline{y}_{tjk}/J$, where $\overline{y}_{tjk}$ is the trimmed mean for the *jk*th cell of the design. The test statistic for comparing means is the test statistic used in B_{RT}. The critical value is $\xi_{\alpha_p, \nu_2}/\sqrt{2}$. This procedure is referred to as the raw score

trimmed Welsch procedure (W_{RT}). In both procedures the critical values were forced to follow a monotone sequence. That is ξ_{α_p, ν_i} was set equal to $\xi_{\alpha_{p-1}, \nu_i}$ whenever $\xi_{\alpha_p, \nu_i} < \xi_{\alpha_{p-1}, \nu_i}$ ($3 \leq p \leq K$ and $s = 1, 2, 3$).

2.2 Design

The conditions included in this study were based on those in Kim (1997). In all conditions $J = 2$ and $K = 4$. Four distribution types ($g = 0$ and $h = -.244$, $g = 0$ and $h = 0$, $g = 0$ and $h = .109$, and $g = 0$ and $h = .35$), three levels of sphericity of the common v-c matrix ($\varepsilon = .96, .75$, and $.40$), three levels of the degree of the heterogeneity of the v-c matrices ($\Sigma_1:\Sigma_2 = 1:1, 1:2$, and $1:5$), two levels of total sample size ($N = 40$ and 60), and three levels of sample size ratio (n_1, n_2) = (28, 12), (20, 20), and (12, 28) for $N = 40$, and (42, 18), (30, 30), and (18, 42) for $N = 60$ combine to give 216 experimental conditions.

2.3 Simulation Procedure

The data for each condition that involved multivariate normal data were generated by using the following steps:

1. For the jth level of the between-subjects factor Z_j, an $n_j \times 4$ matrix of independent normally distributed variates was generated. The NORMAL function in SAS (SAS Institute Inc., 1989) was used to generate all variates.
2. The matrix $\mathbf{Z}_j$ was transformed to $\mathbf{X}_j = \boldsymbol{\mu} + d_j \mathbf{Z}_j \mathbf{U}'$, where $\boldsymbol{\mu}$ is an $n_j \times 4$ matrix of means selected to simulate the required configuration of means, d_j is a constant selected to simulate the required degree of heteroscedasticity, and $\mathbf{U}$ is a lower triangular matrix satisfying the equality $\boldsymbol{\Sigma}_1 = \mathbf{U}\mathbf{U}'$.

The data from nonnormal distributions were generated using the g-and-h distribution suggested by Tukey (1977) and developed by Hoaglin (1985). This family of distributions is attractive in simulation studies because those distribution shapes are determined by a small number of parameters, a wide spectrum of distributions can be approximated, and simulated observations can be easily generated from independent and identically distributed normal deviates. The nonnomal data were generated by replacing the second step for generating normally distributed variables by the following steps:

1. An $n_j \times 4$ matrix $\mathbf{X}_j^*$ was constructed by applying, $X_{ij}^* = Z_{ij} \cdot \exp(h\, Z_{ij}^2/2)$.
2. The $n_j \times 4$ matrix $\mathbf{X}_j^*$ was transformed to $\mathbf{X}_j = \boldsymbol{\mu} + d_j \mathbf{X}_j^* \mathbf{U}'$, where $\boldsymbol{\mu}$, d_j, and $\mathbf{U}$ are defined as in the second step of the procedure for generating multivariate normal data.

Type I error rates were obtained under conditions where the population mean vector, $\boldsymbol{\mu}$, was the null vector. For each condition, 5000 replications were performed.

3. Results

The distribution of Type I error rates is summarized in Table 1. The standard error of these estimated Type I error rates is $[\tau(1-\tau)/5000]^{1/2}$, where τ is the actual Type I error rate. If τ were .05, the standard error would be .0031, so that the rejection region for an upper-tailed z test of H_0: α = .05 is .055 at a .05 significance level. As would be expected from the theoretical developments underlying the B and S(C) tests, using the upper-tailed z test as a criterion, these tests do not result in Type I error rates above .05. The Welsch tests appear to have a more liberal tendency, presumably because the critical value for the Welsch test has no theoretical support when multisample sphericity is violated. By Bradley's (1978) liberal criterion a test is robust if $.5\alpha \leq \tau \leq 1.5\alpha$, where α is the nominal significance level. By this criterion only the Welsch tests were liberal in some conditions. However, $\hat{\tau}$ for W was larger than .075 in only two conditions. For all of the conditions in which $\hat{\tau} >$ .075, the data were short tailed and the relationship between the dispersion matrices and the sample sizes was negative.

<Table 1> Distributions of Type I Error Rates for the Eleven Tests at α = .05

Test	Min	10	25	50	75	90	Max
B	0.0108	0.0186	0.0236	0.0306	0.0408	0.0448	0.0516
B_{RT}	0.0196	0.0230	0.0258	0.0306	0.0354	0.0394	0.0510
B_{DT}	0.0192	0.0266	0.0310	0.0420	0.0492	0.0560	0.0720
W	0.0350	0.0440	0.0496	0.0548	0.0598	0.0646	0.0800
W_{RT}	0.0360	0.0438	0.0480	0.0534	0.0590	0.0634	0.0806
S($\tilde{\varepsilon}$)	0.0030	0.0106	0.0200	0.0314	0.0410	0.0508	0.0830
S($\tilde{\varepsilon}$)$_{RT}$	0.0038	0.0140	0.0242	0.0324	0.0376	0.0484	0.0714
S($\tilde{\varepsilon}$)$_{DT}$	0.0034	0.0120	0.0202	0.0272	0.0332	0.0476	0.0678
S(C)	0.0158	0.0228	0.0280	0.0342	0.0396	0.0446	0.0500
S(C)$_{RT}$	0.0176	0.0264	0.0296	0.0332	0.0362	0.0402	0.0456
S(C)$_{DT}$	0.0154	0.0214	0.0240	0.0272	0.0302	0.0338	0.0462

Note. B=Bonferroni procedure; B_{RT}=raw score trimmed Bonferroni; B_{DT}=trimmed difference score Bonferroni; W=Welsch's step-up procedure; W_{RT}=raw score trimmed Welsch step-up procedure; S($\tilde{\varepsilon}$)=Shaffer's Modified Sequentially Rejective Bonferroni (MSRB) procedure using $\tilde{\varepsilon}$-adjusted F test; S($\tilde{\varepsilon}$)$_{RT}$=Shaffer's MSRB procedure using $\tilde{\varepsilon}$-adjusted F test; S($\tilde{\varepsilon}$)$_{DT}$= Shaffer's MSRB procedure using $\tilde{\varepsilon}$-adjusted F test;

S(C)=Shaffer's MSRB procedure using a CIGA omnibus test; $S(C)_{RT}$=Shaffer's raw score trimmed MSRB procedure using a CIGA omnibus test; $S(C)_{DT}$=Shaffer's trimmed difference score MSRB using a CIGA omnibus test.

A 4 (Distribution) × 3 (ε) × 3 (V-C Heteroscedasticity) × 3 (n_1/n_2) × 2 (N) × 11 (Test) ANOVA with repeated measures on the test factor was used to analyze the Type I error rates. Because many of the factors that affect Type I error rates were included in the study, the ANOVA was expected to yield a substantial number of significant effects. To compare the relative size of the effects, the effect component of each mean square was obtained by using $(MS_{effect} - MS_{error}) / T$, where T is the product of the numbers of levels of the factors not involved in the effect. Defining total variance as the sum of the mean square components plus the sum of the two error variances, the proportion of total variance, $\hat{\omega}^2$, associated with each effect was calculated (Myers, 1979). Only effects for which $\hat{\omega}^2$ was larger than .05 were selected for interpretation. The effects which have larger than .05 $\hat{\omega}^2$ were the main effect of test (.6222), test × n_1/n_2 interaction (.1186), test × ε (.0959) and test × n_1/n_2 × Σ_1:Σ_2 (.0678). Shown in Table 2 are the effects that accounted for more than 1% of the total variance.

<Table 2> Percent of Variance for Type I Error Rate for Effects that Accounted for at least 1% of the variance at α = .05

Effect	$\hat{\omega}^2$
T	0.6222
T× n_1/n_2	0.1186
T× ε	0.0959
T× n_1/n_2 × Σ_1:Σ_2	0.0678
T× n_1/n_2 × ε	0.0158
T× D	0.0143
D	0.0132

Note. T = Test; n_1/n_2 = Sample Size Arrangement; Σ_1:Σ_2 = V-C Heteroscedasticity; D = Distribution; ε = Sphericity.

Means for interpreting the Test × ε interaction are presented in Table 3. The B_{DT}, W, and W_{RT} had the Type I error rates close to α for ε = .96 and .75, but W had liberal tendency in the conditions. There was no difference between S($\tilde{\varepsilon}$) and S(C) procedures.

Type I error rate decreases as ε decreases in all procedures except W, W_{RT}, $S(\tilde{\varepsilon})_{DT}$, and $S(C)_{DT}$. There was small impact of violation of sphericity assumption on Type I error rate for the W, $S(\tilde{\varepsilon})_{RT}$, $S(\tilde{\varepsilon})_{DT}$, $S(C)_{RT}$, and $S(C)_{DT}$ procedures. Type I error rates are most strongly deflated for Bonferroni procedures. The degree of deflation increases with increases in the degree of the violation of sphericity assumption. The W_{RT} shows a minor degree of inflation. When ε = .40, the W and W_{RT} procedures better performed than the other procedures.

<Table 3> Mean of Type I Error Rates as a Function of Test and Sphericity at α = .05

Test	ε=.40	ε=.75	ε=.96
B	0.021	0.036	0.038
B_{RT}	0.026	0.033	0.034
B_{DT}	0.029	0.047	0.049
W	0.052	0.059	0.054
W_{RT}	0.058	0.054	0.049
$S(\tilde{\varepsilon})$	0.026	0.033	0.037
$S(\tilde{\varepsilon})_{RT}$	0.029	0.032	0.034
$S(\tilde{\varepsilon})_{DT}$	0.029	0.029	0.028
S(C)	0.028	0.034	0.039
$S(C)_{RT}$	0.031	0.033	0.035
$S(C)_{DT}$	0.030	0.027	0.026

Note. See Table 1 for abbreviations.

Means for interpreting the Test × n_1/n_2 × $\Sigma_1:\Sigma_2$ interaction are presented in Table 4. Results in the table indicate that when $n_1 = n_2$ heterogeneity of the v-c matrices has little effect on the Type I error rate. When $\Sigma_1 = \Sigma_2$, n_1/n_2 appears to have relatively little effect on Type I error rate. When $n_1 = n_2$ or $\Sigma_1 = \Sigma_2$, the best procedures appear to be W and W_{RT} because they maintain Type I error rate nearest to α. When heterogeneous v-c matrices are positively paired with unequal group sizes ($\Sigma_1:\Sigma_2 \neq 1:1$, $n_1 < n_2$), Type I error rates are strongly deflated for $S(\hat{\varepsilon})$. The Bonferroni tests show a minor degree of deflation. The W and W_{RT} shad good control of Type I error rates. When heterogeneous v-c matrices are negatively paired with unequal group sizes ($\Sigma_1:\Sigma_2 \neq 1:1$, $n_1 > n_2$), Type I error rates are strongly inflated for $S(\hat{\varepsilon})$. The Bonferroni and Welsch procedures exhibit moderate inflation.

The B_{DT}, W, and W_{RT} performed better than the others in the conditions. However, Welsch procedures appear to have a liberal tendency. If large degree of v-c matrices heterogeneity is expected, B_{DT} might be a best choice.

<Table 4> Mean Type I Error Rates as a Function of V-C Heterogeneity, Test, and Sample Size Arrangement at α = .05

Test	$\Sigma_1:\Sigma_2$ = 1:1			$\Sigma_1:\Sigma_2$ = 1:2			$\Sigma_1:\Sigma_2$ = 1:5		
	$n_1<n_2$	$n_1=n_2$	$n_1>n_2$	$n_1<n_2$	$n_1=n_2$	$n_1>n_2$	$n_1<n_2$	$n_1=n_2$	$n_1>n_2$
B	0.033	0.031	0.032	0.031	0.031	0.033	0.030	0.031	0.034
B_{RT}	0.031	0.030	0.032	0.030	0.029	0.033	0.029	0.032	0.034
B_{DT}	0.043	0.038	0.042	0.039	0.039	0.046	0.037	0.042	0.049
W	0.056	0.053	0.056	0.053	0.053	0.057	0.052	0.054	0.060
W_{RT}	0.054	0.051	0.055	0.052	0.051	0.058	0.050	0.054	0.060
$S(\tilde{\varepsilon})$	0.031	0.034	0.031	0.017	0.035	0.044	0.007	0.034	0.054
$S(\tilde{\varepsilon})_{RT}$	0.031	0.034	0.032	0.018	0.033	0.043	0.009	0.034	0.052
$S(\tilde{\varepsilon})_{DT}$	0.028	0.026	0.028	0.015	0.027	0.041	0.008	0.027	0.055
S(C)	0.034	0.034	0.034	0.034	0.034	0.037	0.033	0.033	0.034
$S(C)_{RT}$	0.033	0.033	0.034	0.033	0.033	0.033	0.033	0.033	0.032
$S(C)_{DT}$	0.028	0.026	0.028	0.027	0.026	0.029	0.026	0.026	0.030

Note. See Table 1 for abbreviations.

4. Conclusion

All eleven procedures controlled Type I error rates reasonably well. Two tests emerged as the best tests: the Welsch test on untrimmed data (W) and the Welsch test on trimmed data (W_{RT}). The W and W_{RT} tests have Type I error rates closest to the nominal significance level. Type I error rates for Bonferroni procedures were strongly deflated when ε decreases. Type I error rates for $S(\tilde{\varepsilon})$ procedures were strongly deflated when heterogeneous v-c matrices and unequal group sizes are positively paired. However, when the heterogeneous v-c matrices are positively paired with unbalanced group sizes, the Welsch tests had a slight liberal tendency. Larger $\hat{\tau}$s occurred in our study because we included a short-tailed distribution. Nevertheless, in the vast majority of conditions W and W_{RT} had $\hat{\tau} < .075$. Our results indicate that, when large degree of v-c matrices heterogeneity is expected, BDT can be better than W and WRT because of the liberal tendency of the Welsch tests.

References

[1] Algina, J. (1994). Some alternative approximate tests for a split plot design. *Multivariate Behavioral Research*, 29, 365-384.

[2] Bradley, J. C. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.

[3] Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables, trends, and shapes*. New York: Wiley.

[4] Keselman, H. J. (1994). Stepwise and simultaneous multiple comparisons of repeated measures means. *Journal of Educational Statistics*, 19, 127-162.

[5] Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychological Bulletin*, 110(1), 162-170.

[6] Kim, H. (1997). *Pairwise multiple comparison procedures: Type I error rates and power*. Manuscript submitted for publication.

[7] Lecoutre, B. (1991). A correction for the ε approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics*, 16, 371-372.

[8] Micceri, T. (1991). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.

[9] Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn and Bacon.

[10] SAS Institute Inc. (1989). *SAS/IML Software: Usage and Reference, Version 6*, 1st ed. Cary, NC: Author.

[11] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA, Addison-Wesley.

[12] Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankyha A*, 25, 331-352.

[13] Welsch, R. E. (1977a). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, 72, 566-575.

[14] Welsch, R. E. (1977b). *Tables for stepwise multiple comparison procedures* (Working paper No. 949-97). Cambridge: Massachusetts Institute of Technology.

[15] Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, 32, 771-780.

[16] Wilcox, R. R. (1992). Comparing one-step M-estimators of location corresponding to two independent groups. *Psychometrika*, 57(1), 141-154.

[17] Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika*, 61, 165-170.