The Korean Communications
in Statistics Vol. 4, No. 2, 1997
pp. 581 - 588

Robustness for Omnibus Tests using Trimmed Means under Violated Assumptions

Hyunchul Kim[1)]

Abstract

Univariate F test is based on the multisample sphericity assumption. Robustness for tests of a main effect of the within-subjects factor was investigated when the assumptions of the omnibus F tests are violated in a split-plot design with one between-subjects factor using untrimmed data and trimmed data. The results indicate that when sample sizes are unbalanced and dispersion matrices are heterogeneous, the CIGA and the $CIGA_T$ tests better control Type I error rates than do the F_T test and the $\tilde{\varepsilon}_T$ test.

1. Introduction

Huynh and Feldt (1970) investigated conditions on the variance-covariance (v-c) matrix required to test the hypothesis of no within-subjects effect. They showed that necessary and sufficient conditions on the v-c matrix for level j of between-subjects factor (Σ_j) for the valid F tests are (a) homogeneity of the v-c matrices for a suitable set of orthonormalized variables at all levels of the between-subjects factor, and (b) sphericity of the common v-c matrix. These conditions have been referred to as multisample sphericity. A v-c matrix is spherical if the variances for all possible difference scores are equal. With the homogeneity of the v-c matrices assumption, the variation of the interaction of the subject and the within-subject factors is homogeneous for the groups. This allows pooling over groups to calculate the sum of squares for the interaction of the within-subjects factor and subjects within groups (Brogan and Kutner, 1980).

Repeated observations seldom satisfy the sphericity assumption (Rogan, Keselman, and Mendoza, 1979). For example, when the within-subjects factor is occasion, adjacent measurements are usually more highly correlated than non-adjacent measurements in split plot and repeated measurements designs (Huynh and Feldt, 1970; Rogan, Keselman, and Mendoza,

1) Lecturer, Department of Education, Sungkyunkwan University, 3-53, Myungryun-Dong, Chongro-Ku, Seoul, Korea, 110-745.

1979). Theory in Box (1954) can be used to show that under violations of sphericity the F test of the within-subjects main effect is distributed approximately as F with $\varepsilon(K-1)$ and $\varepsilon(N-J)(K-1)$ degrees of freedom in a split plot design. Greenhouse and Geisser (1959) suggested $\hat{\varepsilon}$, and Huynh and Feldt (1976) suggested $\tilde{\varepsilon}$ as the estimates of ε. Lecoutre (1991) corrected a minor error in the formula for $\tilde{\varepsilon}$. Huynh (1978) extended the $\hat{\varepsilon}$- and $\tilde{\varepsilon}$-adjusted tests to the General Approximation (GA) test and the Improved General Approximation (IGA) test to take the heteroscedasticity of the v-c matrices into account. Algina (1994) developed the corrected Improved General Approximation (CIGA) test based on Lecoutre's (1991) results.

Algina and Oshima (1995) tested the hypothesis of within-subjects main effect by using unweighted means. They considered normal distribution and lognormal distribution. Algina and Oshima (1995) reported that the $\tilde{\varepsilon}$-adjusted test provides adequate control of the Type I error rate as IGA and CIGA. Kim (1997) investigated Type I error rates and power in a split plot design with one between- and one within-subjects factor when the assumptions of the tests are violated for three omnibus tests: the F test, the $\tilde{\varepsilon}$-adjusted F test, and the CIGA. Conditions with normal distribution, and nonnormal distributions with symmetric but with different kurtosis were considered. The long-tailed distributions were $g = 0$, $h = .109$ and $g = 0$, $h = .35$, and short-tailed distribution was $g = 0$ and $h = -.244$. In contrast to Algina and Oshima (1995), Kim (1997) reported that the $\tilde{\varepsilon}$-adjusted test does not control Type I error rates. Kim (1997) concluded that CIGA test provides adequate control of Type I error rates and is typically more powerful than the other procedures.

Wilcox (1993), studying the design with one within-subjects factor, investigated the use of trimmed means in testing for the within-subjects effect. Under the violation of the sphericity assumption and/or normal distribution assumption, Wilcox (1993) compared Type I error and power rates of usual $\tilde{\varepsilon}$-adjusted test ($\tilde{\varepsilon}$) and $\tilde{\varepsilon}$-adjusted test using trimmed means and Winsorized variances ($\tilde{\varepsilon}_T$) in repeated measures designs. In general, they show adequate control of Type I error rates in repeated measure design. Wilcox presented simulation results indicating that there is close agreement between $\tilde{\varepsilon}$ and $\tilde{\varepsilon}_T$, but the $\tilde{\varepsilon}_T$ reduces the liberal tendency of the $\tilde{\varepsilon}$. In this study Type I error rates for a test of a main effect of the within-subjects factor was investigated for the F test using trimmed data, the $\tilde{\varepsilon}$-adjusted test using trimmed data, and the CIGA tests using untrimmed and trimmed data when the assumptions of the omnibus F tests are violated in a split plot design. Since the results of the previous study using skewed distribution and the study using symmetric distributions with different kurtosis were different, it might be interesting to investigate the performance of the procedures using the trimmed data.

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.

Let $\bar{\mathbf{X}}_j$ and S_j denote the sample means vector and dispersion matrix for the jth group. The trimmed version tests were modified by substituting the (K x 1) vector of g-times trimmed means for $\bar{\mathbf{X}}_j$ and the (K x K) Winsorized dispersion matrix for S_j and $n_j - 2g$ for n_j in the test statistics and critical values. To produce a trimmed version the 20% trimming was used in this study.

2. Method

2.1 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 (ε = .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.

2.2 Simulation Procedure

The simulation procedure is same as in Kim (1997). 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 $\mathbf{Z}_j$, 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 U is a lower triangular matrix satisfying the equality $\Sigma_1 = UU'$.

The nonnormal data were generated using the g-and-h distribution suggested by Tukey (1977) and developed by Hoaglin (1985) as 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 using the NORMAL function in SAS.
2. An $n_j \times 4$ matrix $\mathbf{X}_j^*$ was constructed by applying, $X_{ij}^* = Z_{ij} \cdot \exp(h\, Z_{ij}^2/2)$.
3. 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 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. The power were obtained under conditions where the population mean vector was not the null vector. For each condition, 5000 replications were performed.

3. Results

The distribution of Type I error rates is summarized in Table 1. By Bradley's (1978) liberal criterion a test is robust if $.05\alpha \leq \tau \leq 1.5\alpha$, where α is the nominal significance level. By this criterion F_T and $\tilde{\varepsilon}_T$ tests were liberal in some conditions. The minimum and maximum values for CIGA and $CIGA_T$ tests were within Bradley's liberal criterion. The standard error of these estimated Type I error rates is .0031 from $[\tau(1-\tau)/5000]^{1/2}$, where τ is the actual Type I error rate and if τ were .05. So, the critical value for an upper-tailed z test of H_0: $\alpha = .05$ is .055 at a .05 significance level. The actual Type I error rates for all tests were larger than .055 in some conditions.

<Table 1> Distributions of Type I Error Rates at $\alpha = .05$

Test	Min	10	25	50	75	90	Max
F_T	0.0042	0.0196	0.0462	0.0582	0.0988	0.1862	0.2190
$\tilde{\varepsilon}_T$	0.0038	0.0152	0.0414	0.0494	0.0590	0.1444	0.1840
CIGA	0.0292	0.0360	0.0440	0.0496	0.0526	0.0564	0.0660
$CIGA_T$	0.0322	0.0412	0.0442	0.0478	0.0520	0.0578	0.0736

Note. F_T = F test with trimmed data; $\tilde{\varepsilon}_T$ = $\tilde{\varepsilon}$-adjusted F test with trimmed data; CIGA = CIGA test; $CIGA_T$ = CIGA test with trimmed data.

A 4 (Distribution) x 3 (ε) x 3 (V-C Heteroscedasticity) x 3 (n_1/n_2) x 2 (N) x 3 (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 test x n_1/n_2 interaction (.3185), test x n_1/n_2 x $\Sigma_1:\Sigma_2$ interaction (.2304), the test main effect (.1647), the n_1/n_2 main effect (.0945), and n_1/n_2 x $\Sigma_1:\Sigma_2$ interaction (.0644). 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 of Main Effect for Effects that Accounted for at least 1% of the variance at α = .05

Effect	$\hat{\omega}^2$
T x n_1/n_2	0.3185
T x n_1/n_2 x $\Sigma_1:\Sigma_2$	0.2304
T	0.1647
n_1/n_2	0.0945
n_1/n_2 x $\Sigma_1:\Sigma_2$	0.0644
T x ε	0.0467
T x $\Sigma_1:\Sigma_2$	0.0371

Note. T = Test; n_1/n_2 = Sample Size Arrangement;
$\Sigma_1:\Sigma_2$ = V-C Heteroscedasticity; ε = Sphericity.

Means for interpreting the test, n_1/n_2, and $\Sigma_1:\Sigma_2$ interaction are presented in Table 3. Results in the table show that when sample sizes are balanced, heterogeneity of the v-c matrices has little effect on the estimated actual Type I error rate ($\hat{\tau}$). For balanced sample sizes, all tests except the F_T test maintain $\hat{\tau}$ close to α. The Type I error rates are inflated for the F_T test. When the v-c matrices are homogeneous, n_1/n_2 appears to have relatively little effect on $\hat{\tau}$. However, Type I error rates are higher when the group sizes are unbalanced than when the group sizes are balanced for all four tests. The CIGA and $CIGA_T$ have the Type I error rates below the nominal Type I error rate when the v-c matrices are homogeneous.

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 the F_T test and the $\tilde{\varepsilon}_T$ test. The

deflation is worse for $\tilde{\varepsilon}_T$ test than F_T. The CIGA test shows a minor degree of deflation. The $CIGA_T$ test shows a moderate degree of deflation. Both CIGA and $CIGA_T$ keep $\hat{\tau}$ close to α. 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 the F_T test and the $\tilde{\varepsilon}_T$ test. The inflation is larger for F_T than $\tilde{\varepsilon}_T$. The CIGA and $CIGA_T$ tests show a minor degree of inflation. The results show that if heterogeneity of v-c matrices is expected, the F_T test and the $\tilde{\varepsilon}_T$ test should be avoided.

<Table 3> Estimated Type I Error Rate of Test for Degree of Variance-Covariance Matrices Heteroscedasticity and Sample Size Arrangement Combinations at α = .05

$\Sigma_1:\Sigma_2$	Test	$n_1<n_2$	$n_1=n_2$	$n_1>n_2$
1:1	F_T	0.0679	0.0644	0.0676
	$\tilde{\varepsilon}_T$	0.0509	0.0485	0.0512
	CIGA	0.0486	0.0468	0.0494
	$CIGA_T$	0.0496	0.0471	0.0500
1:2	F_T	0.0343	0.0642	0.1187
	$\tilde{\varepsilon}_T$	0.0232	0.0487	0.0957
	CIGA	0.0472	0.0475	0.0493
	$CIGA_T$	0.0470	0.0469	0.0518
1:5	F_T	0.0159	0.0670	0.1974
	$\tilde{\varepsilon}_T$	0.0097	0.0504	0.1658
	CIGA	0.0462	0.0462	0.0512
	$CIGA_T$	0.0451	0.0470	0.0525

Note. See Table 1 for abbreviations.

4. Conclusions

The F_T and the $\tilde{\varepsilon}_T$ did not perform well when the unbalanced sample sizes and heterogeneous v-c matrices are paired. The CIGA test was robust for the violations of the assumptions for the omnibus test of within-subjects main effect in the split plot design. It

appears that CIGA is the best procedure in terms of Type I error rate. However, both CIGA and $CIGA_T$ maintain the Type I error rates close to nominal Type I error rate. Kim (1997) reported that the F test and the $\tilde{\varepsilon}$-adjusted test did not control the Type I error rate when the design is unbalanced and v-c matrices are heterogeneous, and when the sphericity assumption is violated. The trimmed versions of F and $\tilde{\varepsilon}$-adjusted tests temper the liberal tendency of the tests, but they still did not control the Type I error rate in those conditions. So, the F tests and the $\tilde{\varepsilon}$-adjusted tests should be avoided when heterogeneous v-c matrices are paired with unbalanced sample sizes.

References

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

[2] Algina, J., & Oshima, T. C. (1995). An improved general approximation test for the main effect in a split-plot design. *British Journal of Mathematical and Statistical Psychology*, 48, 149-160.

[3] Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.

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

[5] Brogan, D. R., & Kutner, M. H. (1980). Comparative analysis of pretest-posttest research designs. *The American Statistician*, 34(4), 229-232.

[6] Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95-112.

[7] 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.

[8] Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika*, 43(2), 161-175.

[9] Huynh, H., & Feldt, L. S. (1970). Conditions under which mean squared ratios in repeated measurements designs have exact F distributions. *Journal of the American Statistical Association*, 65, 1582-1589.

[10] Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1(1), 69-82.

[11] Keselman, J. C., & Keselman, H. J. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, 43, 265-282.

[12] Kim, H. (1997). Type I error rates and power for omnibus tests of repeated measures means in the split plot design, *The Korean Communications in Statistics*, 4(1), 139-149

[13] 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.

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

[15] Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 32, 269-286.

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

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

[18] 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.

[19] Wilcox, R. R. (1993). Analyzing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology*, 46, 63-76.