The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 139-149

Type I Error Rates and Power for Omnibus Tests of Repeated Measures Means in the Split-Plot Design : F test, $\tilde{\varepsilon}$-adjusted F test, and CIGA test

Hyunchul Kim[1)]

Abstract

For split plot designs exact univariate F tests of the within-subjects main effect are based on the assumption of multisample sphericity. Type I error rates and power are reported for the F test and two tests designed for use when multisample sphericity is violated: the $\tilde{\varepsilon}$-adjusted test and the Corrected Improved General Approximation (CIGA) test. The results indicate that even though the F test and the $\tilde{\varepsilon}$-adjusted test have better power than the CIGA test in some conditions, the F test and the $\tilde{\varepsilon}$-adjusted test do not control Type I error rates when the design is unbalanced and the F test dose not have a good control of Type I error rates when sphericity assumption is severely violated.

1. Introduction

Consider a design with one between-subjects factor with J levels and one within-subjects factor with K levels and let nj (j = 1, ..., J) denotes the number of subjects in level j and N = n1 + n2 + ... + nJ. Each subject is measured at each of K levels of a within-subjects factor. This design is called a split plot design. In the split plot design, the hypothesis of no within-subjects effect is tested by the ratio

$$F_K = \frac{MS_B}{MS_{B \cdot S/A}} = \frac{\frac{SS_B}{K-1}}{\frac{SS_{B \cdot S/A}}{(N-J)(K-1)}}, \tag{1.1}$$

where MSB is the mean square due to between-subjects factor and $MS_{B \cdot S/A}$ is the mean square for error. Let $x_j \sim N(\mu_j, \Sigma_j)$ denotes a (K x 1) random vector that generates the scores in the jth group and assume score vectors are independently distributed both within and

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

between groups. Let 1K denotes the (K x 1) unit vector and A denotes a [K x (K-1)] matrix of full column rank such that A'1K = 0 and A'A = IK-1. Huynh and Feldt (1970) have shown that when $A'\Sigma 1A = \cdots = A'\Sigma JA = \sigma 2$ IK-1, a condition referred to as multisample sphericity (Huynh, 1976), FK~F[(K-1), (N-J)(K-1)], where FK denotes the traditional F ratio for the within-subjects main effect in equation (1.1) and F[df1, df2] denotes the F distribution.

Repeated observations seldom satisfy the sphericity assumption (Rogan, Keselman, and Mendoza, 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 ε(K-1) and ε(N-J)(K-1) degrees of freedom in a split plot design. When the sphericity assumption is satisfied ε=1. Huynh and Feldt (1976) suggested $\tilde{\varepsilon}$ as an estimate of ε,

$$\tilde{\varepsilon} = \frac{[N(K-1)\hat{\varepsilon}-2]}{(K-1)[N-J-(K-1)\hat{\varepsilon}]}, \qquad (1.2)$$

where $\hat{\varepsilon}$ is the estimate of ε by Greenhouse and Geisser (1959). Lecoutre (1991) corrected a minor error in the formula for $\tilde{\varepsilon}$:

$$\tilde{\varepsilon}^* = \frac{(N-J+1)(K-1)\hat{\varepsilon}-2}{(K-1)[N-J-(K-1)\hat{\varepsilon}]}. \qquad (1.3)$$

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 variance-covariance (v-c) matrices into account. Algina (1994) developed the Corrected Improved General Approximation (CIGA) test based on Lecoutre's (1991) results.

Keselman and Keselman (1990) investigated the effect of heterogeneous v-c matrices on the Type I error rates in balanced and unbalanced split plot designs using three estimates of ε to adjust critical values. Three ε-adjusted tests were $\hat{\varepsilon}$-adjusted test (Greenhouse and Geisser, 1959), $\tilde{\varepsilon}$-adjusted test (Huynh and Feldt, 1976), and $\bar{\varepsilon}$-adjusted test (Maxwell and Arvey, 1982). Let $\mu = \Sigma n_j\mu_j / N$, where μ_j (j = 1, ..., J) is the (K x 1) mean vector for group j. They tested main effect null hypothesis H01 : $C'\mu = 0$, where C is a K x (K - 1) matrix with full row rank and rows that are contrast vectors. They concluded that heterogeneity has little effect on Type I error rate of the within-subjects main effect when group sizes are equal or not, and even though the $\tilde{\varepsilon}$-adjusted test was liberal in some conditions, Type I error rates for the $\tilde{\varepsilon}$-adjusted test were closest to nominal significance level.

Algina and Oshima (1995) investigated Type I error rates of the within-subjects main effect for the $\tilde{\varepsilon}$-adjusted, GA, IGA, and CIGA tests. Let $\mu^* = \Sigma \mu_j/J$, where μ_j (j = 1, ..., J) is the (K x 1) mean vector for group j. They tested main effect null hypothesis H02 : $C'\mu^* = 0$. Conditions with data sampled from multivariate normal distributions and from multivariate lognormal distributions were included. The multivariate lognormal distribution has skewed marginal distribution. Conditions that heterogeneous v-c matrices and unbalanced sample sizes are positively paired ($\Sigma1:\Sigma2 \neq 1:1$, n1<n2) and negatively paired ($\Sigma1:\Sigma2 \neq 1:1$, n1>n2) were included. Algina and Oshima (1995) reported that the $\tilde{\varepsilon}$-adjusted, IGA, and CIGA test provided adequate control of the Type I error rate in the conditions in their study.

In this study Type I error rates and power for the F test, $\tilde{\varepsilon}$-adjusted, and CIGA tests were investigated. Main effect null hypothesis of H01 : $C'\mu = 0$ using weighted means $\mu = \Sigma n_j\mu_j / N$ was tested. In contrast to Algina and Oshima (1995) only symmetric distributions with wide range of kurtosis were included. A configuration with equal means and three configurations with unequal means were included.

2. Method

In all conditions J = 2 and K = 4. In addition to conditions with n1 = n2, conditions with n1/n2 = 3/7 and 7/3 were included. Although more extreme sample size ratios may be used in non-experimental studies, it seems unlikely that sample size ratios more extreme than 3/7 and 7/3 will occur in experimental studies. Total sample size had two levels (n = 40 and 60). Thus nj ranged from 12 to 42, a range that covers small to moderate cell frequencies. In addition to conditions with equal dispersion matrices, two levels of dispersion matrix inequality were included ($\Sigma1:\Sigma2$ = 1:2 and 1:5). Three levels of ε were employed (ε = .96, .75 and .40). The dispersion matrices for ε = .96, .75 and .40 were taken from Keselman and Keselman (1990).

Both multivariate normal and multivariate nonnormal distribution were included. However, to keep the study to a manageable size only symmetric distributions were included. Nonnormal data were generated using the g-and-h distribution suggested by Tukey (1977) and developed by Hoaglin (1985). Data sampled from long-tailed distributions (h = .109 and .35) and from a short-tailed distribution (h = -.244) were included. The distribution with h = .109 has the same kurtosis (6.00) as a double-exponential distribution. The distribution with h = .35 has larger kurtosis than the distribution with h = .109. Thus our long-tailed distributions ranged from being very long-tailed to extremely long-tailed. The distribution with h = -.244 has the same kurtosis (-1.2) as an uniform distribution.

A configuration with equal means and three configurations with unequal means were included. The configurations with unequal means were minimum range for the means, maximum range for the means, and equally spaced means. For each combinations of n1 and n2, the means for each configuration are presented in Table 1. The means were calculated from the noncentrality parameter necessary to produce omnibus power of .50 when α = .05, ε = 1.0, and the dispersion matrices are equal.

<Table 1> Means of Repeated Measures for each Sample Size Arrangement (S) to get 50 percent Power in Omnibus F Test when all Assumptions for Omnibus F Test are satisfied

configuration of the nonnull means	S	μ1	μ2	μ3	μ4
minimum range means	a,c	-0.470918	-0.470918	0.470918	0.470918
	b	-0.431603	-0.431603	0.431603	0.431603
	d,f	-0.382238	-0.382238	0.382238	0.382238
	e	-0.350327	-0.350327	0.350327	0.350327
maximum range means	a,c	-0.665978	0	0	0.665978
	b	-0.610379	0	0	0.610379
	d,f	-0.540567	0	0	0.540567
	e	-0.495438	0	0	0.495438
equal space means	a,c	-0.631802	-0.210601	0.210601	0.631802
	b	-0.579056	-0.193019	0.193019	0.579056
	d,f	-0.512827	-0.170942	0.170942	0.512827
	e	-0.470013	-0.156671	0.156671	0.470013

Note. a: nj = 12 and 28; b: nj = 20 and 20; c: nj = 28 and 12; d: nj = 18 and 42; e: nj = 30 and 30; f: nj = 42 and 18.

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 Zj, an $nj \times 4$ 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 Zj was transformed to $Xj = \mu + djZjU'$, where μ is an $nj \times 4$ matrix of means selected to simulate the required configuration of means, dj 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 data from nonnormal distributions were generated using the g-and-h distribution by replacing the second step for generating normally distributed variables by the following steps:

1. An nj×4 matrix Xj* was constructed by applying, Xij* = Zij · exp(h Zij2/2).
2. The nj×4 matrix Xj* was transformed to Xj = μ + djXj*U′, where μ, dj, 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, μ, 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

3.1 Type I Error Rates

The distribution of Type I error rates is summarized in Table 2. A 4 (Distribution) x 3 (ε) x 3 (V-C Heteroscedasticity) x 3 (n1/n2) 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 (MSeffect - MSerror) / 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 test x n1/n2 x Σ1:Σ2 interaction (.1719) and the test x ε interaction (.0508) were the highest order interaction for the factors in the interactions. Shown in Table 3 are the effects that accounted for more than 1% of the total variance.

<Table 2> Distributions of Type I Error Rates at α = .05

Test	Min	10	25	50	75	90	Max
F	0.0046	0.0198	0.0506	0.0636	0.1034	0.1964	0.2310
$\tilde{\varepsilon}$	0.0042	0.0126	0.0374	0.0504	0.0572	0.1376	0.2104
CIGA	0.0292	0.0360	0.0440	0.0496	0.0526	0.0564	0.0660

Note. F = F test; $\tilde{\varepsilon}$ = $\tilde{\varepsilon}$-adjusted F test; CIGA = CIGA test.

<Table 3> 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	0.2400
T x n_1/n_2 x $\Sigma_1:\Sigma_2$	0.1719
T	0.1712
n_1/n_2	0.1659
n_1/n_2 x $\Sigma_1:\Sigma_2$	0.1188
T x ε	0.0508
T x $\Sigma_1:\Sigma_2$	0.0257
$\Sigma_1:\Sigma_2$	0.0161

Note. T = Test ; n1/n2 = Sample Size Arrangement ; Σ1:Σ2 = V-C Heteroscedasticity ; ε = Sphericity.

Means for interpreting the test x n1/n2 x Σ1:Σ2 interaction are presented in Table 4. Results in the table indicate that when n1 = n2, heterogeneity of the v-c matrices has little effect on the estimated actual Type I error rate ($\hat{\tau}$). When n1 = n2, the $\tilde{\varepsilon}$-adjusted test and CIGA test maintain $\hat{\tau}$ close to α, but F test has inflated Type I error rate. When Σ1:Σ2 = 1:1, n1/n2 appears to have relatively little effect on $\hat{\tau}$. However, when heterogeneous v-c matrices are positively paired with unequal group sizes, Type I error rates are strongly deflated for the F test and the $\tilde{\varepsilon}$-adjusted test. The CIGA test shows a minor degree of deflation. When heterogeneous v-c matrices are negatively paired with unequal group sizes, Type I error rates are strongly inflated for the F test and the $\tilde{\varepsilon}$-adjusted test. The CIGA test shows a minor degree of inflation. If heterogeneity of v-c matrices is expected, the F test and the $\tilde{\varepsilon}$-adjusted test are not good choices. Means for interpreting test x ε interaction are presented in Table 5. Type I error rates for the CIGA test shows minor inflation for the severe violation of the sphericity. Type I error rates are strongly inflated for the F test as the degree of violation increases. The degree of the inflation for the $\tilde{\varepsilon}$-adjusted F test decreases as the degree of violation increases. Only CIGA test controls Type I error rates close to the nominal Type I error rate.

<Table 4> 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	0.0712	0.0698	0.0723
	$\tilde{\varepsilon}$	0.0492	0.0479	0.0497
	CIGA	0.0486	0.0468	0.0494
1:2	F	0.0357	0.0702	0.1270
	$\tilde{\varepsilon}$	0.0213	0.0487	0.0961
	CIGA	0.0472	0.0475	0.0493
1:5	F	0.0150	0.0710	0.2117
	$\tilde{\varepsilon}$	0.0076	0.0488	0.1696
	CIGA	0.0462	0.0462	0.0512

Note. See Table 2 for abbreviations.

<Table 5> Means of Type I Error Rates of Main Effect as a Function of Test and Sphericity at α = .05

Test	ε=.96	ε=.75	ε=.40
F	0.0668	0.0760	0.1051
$\tilde{\varepsilon}$	0.0619	0.0604	0.0573
CIGA	0.0459	0.0475	0.0508

Note. See Table 2 for abbreviations.

3.2 Power

The results of a 3 (Configuration of Means) x 4 (Distribution) x 3 (ε) x 3 (V-C Heteroscedasticity) x 3 (n1/n2) x 2 (N) x 3 (Test) ANOVA with repeated measures on the test factor was used to analyze the power. The distribution of power is summarized in Table 6. Shown in Table 7 are the effects that accounted for more than 1% of the total variance.

<Table 6> Distributions of Power at α = .05

Test	Min	10	25	50	75	90	Max
F	0.0430	0.1534	0.2496	0.3950	0.5900	0.7418	0.8820
$\tilde{\varepsilon}$	0.0326	0.1224	0.1860	0.3370	0.4984	0.7010	0.8780
CIGA	0.0974	0.1214	0.1736	0.3378	0.4864	0.6874	0.9024

Note. See Table 2 for abbreviations.

<Table 7> Percent of Variance for Power of Main Effect for Effects that Accounted for at least 1% of the Variance at α = .05

Effect	$\hat{\omega}^2$
D	0.6584
T x n_1/n_2	0.1027
T x n_1/n_2 x $\Sigma_1:\Sigma_2$	0.0700
ε	0.0406
T	0.0381
T x ε	0.0209
D x ε	0.0206
n_1/n_2	0.0153

Note. See Table 3 for abbreviations.

Means of power for interpreting the D effects are presented in Table 8. Results in the table indicate that all three tests have less power when the data are generated from long-tailed distributions. Means for interpreting the test x n1/n2 x Σ1:Σ2 interaction are presented in Table 9. When Σ1:Σ2 = 1:1, n1/n2 appears to have relatively little effect on power. The F test has better power than the other tests, but the $\tilde{\varepsilon}$-adjusted test and the CIGA test appear to have almost same power. When heterogeneous v-c matrices are positively paired with unequal group sizes, power rates for the F test and the $\tilde{\varepsilon}$-adjusted test decline; power rates for the CIGA test increase. When heterogeneous v-c matrices are negatively paired with unequal group sizes, power rates for the CIGA test decline; power rates for the F test and the $\tilde{\varepsilon}$-adjusted test increase. For all tests the degree of deflation or inflation increases with increases in the degree of v-c matrices heterogeneity. The $\tilde{\varepsilon}$-adjusted test has very low power in positive pairing. When n1 = n2, v-c matrix heterogeneity appears to have relatively little effect on power.

<Table 8> Mean of Power of Main Effect as a Function of Distribution at α = .05

h=-.244	h=0	h=.109	h=.35
0.6610	0.4203	0.3146	0.1463

<Table 9> Estimated Power of Test for Degree of V-C Matrices Heteroscedasticity and Sample Size Arrangement Combinations at α = .05

Σ_1:Σ_2	Test	$n_1<n_2$	$n_1=n_2$	$n_1>n_2$
1:1	F	0.4178	0.4180	0.4201
	$\tilde{\varepsilon}$	0.3649	0.3651	0.3672
	CIGA	0.3628	0.3617	0.3645
1:2	F	0.3508	0.4196	0.4923
	$\tilde{\varepsilon}$	0.2976	0.3657	0.4394
	CIGA	0.4036	0.3612	0.3297
1:5	F	0.2867	0.4215	0.5726
	$\tilde{\varepsilon}$	0.2340	0.3660	0.5182
	CIGA	0.4565	0.3577	0.2943

Note. See Table 2 for abbreviations.

4. Conclusions

The standard error of these estimated Type I error rates is [τ(1-τ)/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 H0 : α = .05 is .055 at a .05 significance level.

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. The results reported in this paper indicate that the CIGA test provides good control over the Type I error for the conditions used in the study. In contrast to Algina and Oshima (1995) the $\tilde{\varepsilon}$-adjusted test did not perform well for testing within-subjects main effect. Type I error rates for the F test and the $\tilde{\varepsilon}$-adjusted test are below the lower limit of Bradley's liberal criterion (.025) when heterogeneous v-c matrices are positively paired with unequal group sizes. When heterogeneous v-c matrices are negatively paired with unequal group sizes, Type I error rates for the tests are above the upper limit (.075). But, when sample sizes are equal the $\tilde{\varepsilon}$-adjusted test provides good control of the Type I error rate even when the v-c matrices are heterogeneous. The F test does not control the Type I error rates under the nominal Type I error rate when the sphericity condition is severely violated (ε = .40).

All three tests have less power when the data are generated from long-tailed distributions and higher power when the data are generated from short-tailed distribution. When $\Sigma1{:}\Sigma2$ = 1:1, n1/n2 appears to have relatively little effect on power and the F test has better power than the other tests. When n1 = n2, and when heterogeneous v-c matrices are negatively paired with unequal group sizes, the F test and the $\tilde{\varepsilon}$- adjusted test have better power than the CIGA test. But, they will not be meaningful to someone who cannot tolerate the liberal tendency of the F test and the $\tilde{\varepsilon}$-adjusted test in some conditions.

References

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

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

[6] 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.

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

[8] Huynh, H. and 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.

[9] Huynh, H. and 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.

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

[11] 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.

[12] Maxwell, S. E. and Arvey, R. D. (1982). Small sample profile analysis with many variables. *Psychological Bulletin*, 92, 778-785.

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

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

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

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