Journal of Modern Applied Statistical Methods

Volume 11 | Issue 2 Article 7

11-1-2012

Examining Multiple Comparison Procedures According to Error Rate, Power Type and False Discovery Rate

Guven Ozkaya
Uludag University, Gorukle/Bursa, Turkey

Ilker Ercan
Uludag University, Görükle/Bursa, Turkey

Journal of Modern Applied Statistical Methods
November 2012, Vol. 11, No. 2, 348-360

1538 – 9472/12/$95.00

Examining Multiple Comparison Procedures According to Error Rate, Power Type and False Discovery Rate

Guven Ozkaya Ilker Ercan
Uludag University,
Gorukle/Bursa, Turkey

Examining pairwise differences between means is a common practice of applied researchers, and the selection of an appropriate multiple comparison procedure (MCP) is important for analyzing pairwise comparisons. This study examines the performance of MCPs under the assumption of homogeneity of variances for various numbers of groups with equal and unequal sample sizes via a simulation study. MCPs are compared according to type I error rate, power type and false discovery rate (FDR). Results show that the LSD and Duncan procedures have high error rates and Scheffe's procedure has low power; no remarkable differences between the other procedures considered were identified.

Key words: Multiple comparison procedures, pairwise comparison, error rates, power, false discovery rate.

Introduction

Multiple comparison procedures (MCPs) are used to test differences between the means of three or more groups after performing variance analysis. Although MCPs are used often, many are not used correctly (Lowry,1992; Hsu, 1996). Homogeneity of variances, normality and independence of data are assumptions made for variance analysis; these assumptions should also hold when performing MCPs. In addition, sample size also affects MCP performance and should be considered. Some MCPs are purported to apply when the assumptions hold, and some are proposed for the cases in which some assumptions are violated (Demirhan, 2010). Selecting an appropriate MCP is important, it is necessary to choose a method that is best given the research situation and data. This study examines the performance of MCPs under the assumption of homogeneity of variances for various numbers of groups with equal and unequal sample sizes via a simulation study. MCPs are compared according to type I error rate, power type and false discovery rate (FDR).

Guven Ozkaya is a Research Assistant in the Department of Biostatistics, Faculty of Medicine. His research interests are statistical shape analysis and experimental design. Email him at: guvenozkaya@gmail.com. Ilker Ercan is a Professor in the Department of Biostatistics, Faculty of Medicine. His research interests are statistical shape analysis, reliability and validity analyses. Email him at: ercan@uludag.edu.tr.

General Information

Many MCPs rely on contrasts; a comparison of *k* groups comprises a comparison of two groups and a comparison of a single group with the remaining groups. This definition is symbolized as

$$\psi = c_1\mu_1 + c_2\mu_2 + ... + c_k\mu_k = \sum_{j=1}^{k} c_k\mu_k$$

where the contrast constants c_1, c_2, …, c_k sum to zero. MCPs can be categorized according to contrast type as: pairwise comparison, complex comparison, comparison with control and comparison with the best. Generally, the method of contrasts is useful for preplanned, or *a priori*, comparisons, that is, the contrasts are specified prior to conducting the experiment and examining the data. The rationale behind *a priori* contrasts is that, if comparisons are

selected after examining data, many experimenters would construct tests that correspond to large observed differences in means (Montgomery, 2001).

This study focuses on MCPs for examining all possible pairwise comparisons. The LSD, Bonferroni, Dunn-Sidak, Scheffe, REGW-F and Q, Student Neuman Keuls (SNK), Tukey's a and b, Duncan, Hochberg's GT2 and Gabriel procedures are examined for various numbers of groups, variances and sample sizes according to error rate, power type and false discovery rate (FDR).

Measures Used to Evaluate Procedures

The statistical problem that arises from the use of MCPs is that subsequent hypothesis tests will be performed on the outcome with the same data on which the global test was performed: this can result in an uncontrolled type I error rate (Cabral, 2008). However, determining how to control type I errors is much more difficult when multiple significance tests are computed (Jaccard, 2002a, 2002b). This difficulty arises because the decision regarding control of type I errors when MCPs of significance are computed can affect whether the effects are statistically significant (Keselman, 2004).

Choosing from among the various strategies available to control Type I errors could be based on the multiplicity of testing issue. The multiplicity problem in statistical inference refers to the selection of statistically significant findings from a large set of findings (tests) to either support or refute a research hypothesis. Selecting statistically significant findings from a larger pool of results, which also contains non-significant findings, is problematic because when multiple tests of significance are computed the probability that at least one will be significant by chance alone increases with the number of tests examined (Keselman, 2004).

Testing many variables with univariate analysis is typically the first choice for various hypotheses (Tatlidil, 2002), however, due to error rate inflation, this solution is not convenient. There has been much debate concerning the necessity of statistical adjustment for multiplicity (Kemp, 1975; Bender, 2001). One argument suggests controlling the probability that at least one type I error will occur in the set of pairwise comparison tests by setting that probability equal to alpha (Kemp, 1975; Ludbrook, 1998; Cabral, 2008). This type of control is referred to in the literature as experimentwise or familywise control (Kemp, 1975; Klockars, 1986; Toothaker, 1993; Ludbrook, 1998; Keselman, 2004; Cabral, 2008). In the opposing argument, this type of adjustment is not necessary: instead, each comparison is dealt with separately (O'Neill, 1971; O'Brien, 1983; Perry, 1986; Rothman, 1990). This type of control has been referred to as the comparisonwise error rate (Kemp, 1975; Klockars, 1986; Toothaker, 1993; Keselman, 2004; Cabral, 2008).

The controversy concerning MCPs is whether to control for comparisonwise or experimentwise type I error rates. Related to this controversy is the power of the procedure (Kemp, 1975). It is widely accepted among statisticians that the goal of MCP analysis should be to control the familywise error rate (Keselman, 2004; Toothaker, 1993; Ryan, 1959; Shaffer, 1995; Roback, 2005). Another argument (Benjamini & Hochberg, 1995) is to control the false discovery rate (FDR) (Cabral, 2008; Keselman, 2004; Ludbrook, 1998). Benjamini and Hochberg (1995) developed an alternative approach to multiple hypothesis testing that controls the expected proportion of false positive findings among all rejected hypotheses.

Familywise Error Rate

Familywise error rate is the probability that at least one type I error occurs. Perfect MCPs control the familywise error rate, thus the error rate cannot exceed the α-level (Klockars, 1986; Toothaker, 1993; Ludbrook, 1998; Cabral, 2008) and the type II error rate is minimized. The simultaneous occurrence of two events is impossible (Ludbrook, 1998). MCPs that maintain the overall α-level for a set of tests are said to control the familywise error rate and effectively reduce the α-level for each post hoc test.

Comparisonwise Error Rate

MCPs that apply a separate α-level for each test are called comparisonwise error control procedures (Klockars, 1986; Toothaker, 1993; Cabral, 2008). In a study with groups A, B and C, the use of comparisonwise error control after the global null hypothesis has been rejected entails the performance of 6 individual tests (A-B, A-C, B-C, AB-C, AC-B, BC-A) and the application of an α-level of 0.05 for each test.

False Discovery Rate (FDR)

Much of the debate concerning error rates relates to familywise and comparisonwise error rates. One of the newer interesting contributions to the field of multiple hypothesis testing is an alternative conceptualization for defining errors in the multiple testing problems: the false discovery rate, or FDR, as presented by Benjamini and Hochberg (1995). The FDR is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections. Benjamini and Hochberg (1995) provided several scenarios in which the FDR control seems more reasonable than the familywise or comparisonwise control.

Consider J means, $\mu_1, \mu_2, \ldots, \mu_J$, where interest is in testing a family of $m = J(J-1)/2$ pairwise hypotheses, H_i: $\mu_j - \mu_{j'} = 0$ ($j = 1, \ldots, J$; $j' = 1, \ldots, J$; $j \neq j'$), of which m_0 are true. Let S equal the number of correctly rejected hypothesis pairs from the set of R rejections and let the number of falsely rejected pairs be V. Benjamini and Hochberg (1995) summarized the relationship between these random variables (see Table 1). In terms of random variable V, the comparisonwise error rate is E(V/m), whereas the familywise rate is given by $P(V \geq 1)$. Thus, testing each comparison at α guarantees that $E(V/m) \leq \alpha$, whereas testing each comparison at α/m guarantees $P(V \geq 1) \leq \alpha$ (Keselman, 1999). According to Benjamini and Hochberg (1995), the proportion of errors committed by falsely rejecting null hypotheses can be expressed through the random variable Q = V/(V+S), that is, the proportion of rejected hypotheses that are erroneously rejected. It is important to note that Q is defined to be zero when R = 0; that is, the error rate is zero when there are no rejections. The FDR was defined by Benjamini and Hochberg as the mean value of Q, that is,

$$\begin{aligned} E(Q) &= E\left(\frac{V}{V+S}\right) \\ &= E\left(\frac{V}{R}\right) \\ &= E\left(\frac{\text{number of false rejections}}{\text{number of total rejections}}\right). \end{aligned}$$

FDR is thus the mean value of the proportion of falsely rejected pairwise tests to the total number of pairwise tests declared significant. As Benjamini and Hochberg indicate, this error rate has a number of important properties:

a) If $\mu_1 = \mu_2 = \ldots = \mu_J$, then all m pairwise comparisons truly equal zero and, therefore, the FDR is equivalent to the familywise error rate; that is, in the case that s = 0 and v = r, if v = 0, then Q = 0, and if V > 0, then Q = 1, and thus $P(V \geq 1) = E(Q)$. Therefore, control of the FDR implies control of the familywise error (Benjamini, 1995).

b) When $m_0 < m$, the FDR is smaller than or equal to the familywise error rate; in this case, if v > 1, then $v/r \leq 1$, and if V = 0, then v/r = 0 and, thus, $P(V \geq 1) \geq E(Q)$. This result indicates that if the familywise error rate is controlled for a given procedure, then the FDR is also controlled (Keselman, 1999).

c) v/r tends to be smaller when there are fewer pairs of equal means and when the unequal pairs are more divergent, resulting in a greater difference between the FDR and the familywise value and thus a greater likelihood of increased power by adopting FDR control.

Table 1: Number of Errors Committed when Testing m Null Hypotheses
(Benjamini & Hochberg, 1995)

	Declared Non-Significant	Declared Significant	Total
True H_0	U	V	m_0
False H_0	T	S	$m-m_0$
	m–R	R	m

Statistical Power Types

The power of MCPs can be categorized for different situations. Over the years, many different conceptualizations of power for (pairwise) comparisons have appeared in the literature. For example, Einot and Gabriel (1975) considered each single subset hypothesis and summarized their findings for all subsets of a particular number of means. Einot and Gabriel (1975) provided results on pair power, triplet power, quadruplet power and quintuplet power. One of the methods to measure power is any-pair power, which is defined as the probability of at least one rejection of a false null hypothesis on a pair of means.

A second measure of power is all-pair power, which is defined as the probability of rejecting all false null hypotheses on pairs of means. If a single false hypothesis is considered, then the probability of rejecting it is called per-pair power. Another power definition considered by Ramsey (Ramsey, 1978; Horn, 2000; Ramsey, 2002; Ramsey & Ramsey, 2008) states that power types are per-pair power, any-pair power, and all-pair power. Ramsey's power types are used frequently in the literature. It should be noted that different names for these terms exist in the literature (Keselman, 2004; Ekenstierna, 2004).

Methodology

The LSD, Bonferroni, Dunn-Sidak, Scheffe, REGW-F and Q, SNK, Tukey's a and b, Duncan, Hochberg's GT2 and Gabriel procedures were examined via simulation scenarios according to error rates, power type and FDR for different numbers of groups, variances and sample sizes. The simulations used 4 cases and 27 scenarios for each case, and 250 replications were made for the 108 scenarios. Data were generated from a normal distribution using R software V.2.11.1 and analyses were performed using SPSS 17.0 for Windows. The four cases examined are:

Case I

Error rates were calculated for equal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with a mean of 40 and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were 10, 30 and 100.

Case II

Error rates were calculated for unequal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with a mean of 40 and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were chosen from 10/12/14/16/18/20/22, 30/35/40/45/50/55/60 and 100/110/120/130/140/150/160 for group numbers 3, 5 and 7, respectively.

Case III

Power-type and FDR calculation for equal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with means of 40/40/42/44/46/48/50 (the first two means are the same due to the FDR calculation) for groups and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were 10, 30 and 100

Case IV

Power-type and FDR calculation for non-equal sample sizes, different numbers of groups and different variances: The data were generated from normal distributions with means of 40/40/42/44/46/48/50 (the first two means are the same because of the FDR calculation) for groups and variances 2, 4 and 8. The numbers of compared groups (k) were 3, 5 and 7. Sample sizes were chosen from 10/12/14/16/18/20/22, 30/35/40/45/50/55/60 and 100/110/120/130/140/150/160 for group sizes 3, 5, and 7, respectively.

Results

Simulation results for error rates are shown in Tables 2-5, and the power-type and FDR results are summarized in Table 6. When the number of groups is small, for both equal and unequal sample sizes, the LSD and Duncan error rates are higher than the other MCPs; the other MCP error rates are very similar. Although the number of groups is increasing, the familywise error rates are highest with the LSD and Duncan procedures and are lowest with the Scheffe. In addition, the comparisonwise error rates of the LSD and Duncan procedures are the highest and

the Bonferroni, Dunn-Sidak and Scheffe error rates are the lowest for both equal and unequal sample sizes.

For a small number of groups and equal and unequal sample sizes, the LSD, Duncan and REGW-F procedures have the highest per-pair power and all-pair power. As the number of groups increases, the LSD, Duncan and SNK procedures have the highest per-pair power. The Scheffe per-pair power is the lowest among all the MCPs. For a large sample size, the per-pair power and all-pair power of all the MCPs are very close.

The three highest MCP any-pair powers are those for the LSD, Duncan and REGW-F procedures for small groups. As the number of groups increases, all the MCP powers reach their highest values for equal and unequal sample sizes.

For small groups and equal and unequal sample sizes, the LSD, REGW-F, REGW-Q, SNK, Tukey's b and Duncan procedures have FDR values higher than those of the other procedures. As the number of groups increases, the FDR values of all MCPs become similar. Also, as the number of comparisons increases, the FDR decreases.

Discussion

MCPs were studied in terms of the familywise error rate for equal and unequal sample sizes; the Duncan's and the LSD's familywise error rates were very high. For both procedures, when the number of groups increased, the familywise error rate also increased. For a large number of groups, the Scheffe procedure has the lowest familywise error rate of all the MCPs and was not affected by the change in the-sample sizes.

The familywise error rates of the Bonferroni, Dunn-Sidak, Gabriel and Hochberg's GT2 procedures were low and similar to each other for small numbers of groups. When the number of groups was increased, the familywise error rates of the LSD and Duncan procedures were the highest.The Scheffe procedure was the lowest of all the MCPs.

The FWE values of REGW-F, REGW-Q, SNK, Tukey's a and b, Gabriel and Hochberg's GT2 procedures were similar.

There was no significant change in the MCP familywise error rates due to an increase or decrease in number of groups, with the exceptions of the LSD and Duncan procedures. There also was no significant change in the familywise error rate for any MCP due to changes in sample size. Based on the homogeneity of variance assumption, changes in variance were not considered to have a serious an effect on familywise error rate.

According to comparisonwise error rate results for equal and unequal sample sizes, the familywise error rates of the LSD and the Duncan procedures were higher than those of other procedures. The comparisonwise error rate of the LSD procedure was not greatly affected by changes in group number. Conversely, the Duncan comparisonwise error rate significantly increased with increases in group number. For equal and unequal sample sizes, the Bonferroni, Dunn-Sidak and Scheffe procedures generally had the lowest comparisonwise error rates.

There were no significant changes in the comparisonwise error rates of these procedures due to increases in group number. The comparisonwise error rates of the REGW-F, REGW-Q, SNK, Tukey's a and b, Gabriel and Hochberg's GT2 procedures were not as low as those of the Bonferroni, Dunn-Sidak and Scheffe procedures, however, they were lower than those of the LSD and Duncan procedures. The comparisonwise REGW-Q error rate was lower than the LSD error rate. There were no significant changes in comparisonwise error rate because of the increases in group number; in addition, there was no significant change in the comparisonwise error rate of any MCP because of the changes in sample size and variance.

The per-pair power of all MCPs increased as variance decreased and as sample size increased for equal and unequal sample sizes. Furthermore, the per-pair powers of all MCPs increased with the increase of group number. The LSD, Duncan and REGW-F procedures had the highest per-pair power for a small number of groups and a small sample size.

With larger group numbers and sample sizes, the LSD, Duncan and SNK procedures had the highest per-pair powers and the Scheffe

Table 2: Familywise Error Rates Results for Equal Sample Size

	Number of Groups 3								
	n_i:10/10/10			n_i:30/30/30			n_i:100/100/100		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.116	0.108	0.120	0.136	0.096	0.088	0.080	0.132	0.128
Bonferroni	0.020	0.028	0.032	0.036	0.036	0.036	0.028	0.040	0.036
Dunn-Sidak	0.020	0.028	0.032	0.036	0.036	0.036	0.028	0.040	0.036
Scheffe	0.020	0.028	0.028	0.036	0.032	0.032	0.020	0.036	0.028
REGW - F	0.032	0.032	0.040	0.036	0.040	0.040	0.032	0.044	0.032
REGW - Q	0.024	0.032	0.048	0.036	0.036	0.040	0.040	0.044	0.048
SNK	0.028	0.032	0.048	0.036	0.036	0.040	0.040	0.044	0.048
Tukey a	0.024	0.036	0.052	0.040	0.040	0.044	0.044	0.048	0.056
Tukey b	0.028	0.032	0.048	0.036	0.036	0.040	0.040	0.044	0.048
Duncan	0.076	0.072	0.096	0.092	0.068	0.076	0.056	0.104	0.092
Hochberg's GT2	0.020	0.028	0.040	0.036	0.036	0.036	0.028	0.044	0.036
Gabriel	0.020	0.028	0.040	0.036	0.036	0.036	0.028	0.044	0.036

	Number of Groups 5								
	n_i:10/10/10/10/10			n_i:30/30/30/30/30			n_i:100/100/100/100/100		
LSD	0.240	0.196	0.196	0.212	0.216	0.188	0.244	0.184	0.236
Bonferroni	0.032	0.028	0.028	0.028	0.044	0.032	0.032	0.024	0.044
Dunn-Sidak	0.032	0.028	0.028	0.028	0.044	0.032	0.032	0.032	0.048
Scheffe	0.024	0.016	0.012	0.004	0.020	0.008	0.008	0.012	0.016
REGW - F	0.040	0.040	0.032	0.060	0.064	0.048	0.048	0.032	0.052
REGW - Q	0.040	0.044	0.032	0.032	0.052	0.052	0.044	0.032	0.048
SNK	0.040	0.044	0.032	0.032	0.052	0.056	0.044	0.032	0.048
Tukey a	0.040	0.044	0.032	0.032	0.052	0.052	0.044	0.032	0.048
Tukey b	0.040	0.044	0.032	0.032	0.052	0.056	0.044	0.032	0.048
Duncan	0.168	0.132	0.136	0.160	0.148	0.140	0.168	0.132	0.160
Hochberg's GT2	0.044	0.040	0.036	0.032	0.044	0.032	0.040	0.036	0.048
Gabriel	0.032	0.032	0.028	0.028	0.044	0.032	0.032	0.032	0.048

	Number of Groups 7								
	n_i:10/10/10/10/10/10/10			n_i:30/30/30/30/30/30/30			n_i:100/100/100/100/100/100/100		
LSD	0.436	0.436	0.464	0.424	0.420	0.408	0.436	0.432	0.400
Bonferroni	0.036	0.024	0.040	0.044	0.032	0.064	0.044	0.024	0.052
Dunn-Sidak	0.036	0.028	0.040	0.044	0.032	0.068	0.044	0.028	0.052
Scheffe	0.008	0.004	0.016	0.016	0.012	0.016	0.004	0.004	0.008
REGW - F	0.048	0.028	0.044	0.048	0.032	0.048	0.052	0.036	0.044
REGW - Q	0.048	0.020	0.036	0.048	0.040	0.068	0.044	0.028	0.048
SNK	0.048	0.020	0.036	0.048	0.040	0.068	0.044	0.028	0.048
Tukey a	0.052	0.040	0.056	0.052	0.052	0.076	0.052	0.040	0.060
Tukey b	0.048	0.020	0.036	0.048	0.044	0.068	0.044	0.028	0.048
Duncan	0.220	0.184	0.248	0.232	0.216	0.208	0.248	0.200	0.196
Hochberg's GT2	0.036	0.028	0.040	0.044	0.032	0.072	0.044	0.028	0.052
Gabriel	0.036	0.028	0.040	0.044	0.036	0.072	0.044	0.028	0.052

Table 3: Familywise Error Rates Results for Unequal Sample Size

	Number of Groups 3								
	n_i:10/12/14			n_i:30/35/40			n_i:100/110/120		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.132	0.124	0.084	0.124	0.172	0.140	0.116	0.080	0.128
Bonferroni	0.024	0.036	0.020	0.044	0.056	0.048	0.052	0.024	0.024
Dunn-Sidak	0.028	0.036	0.020	0.044	0.056	0.048	0.052	0.024	0.024
Scheffe	0.020	0.032	0.008	0.036	0.056	0.040	0.044	0.020	0.024
REGW - F	0.028	0.048	0.024	0.056	0.068	0.044	0.044	0.036	0.028
REGW - Q	0.024	0.032	0.016	0.036	0.052	0.044	0.044	0.020	0.024
SNK	0.036	0.032	0.020	0.056	0.068	0.052	0.052	0.032	0.024
Tukey a	0.044	0.040	0.028	0.060	0.076	0.048	0.060	0.036	0.028
Tukey b	0.040	0.032	0.020	0.056	0.068	0.052	0.052	0.032	0.024
Duncan	0.092	0.088	0.056	0.092	0.128	0.108	0.080	0.060	0.084
Hochberg's GT2	0.032	0.032	0.020	0.044	0.064	0.044	0.048	0.024	0.024
Gabriel	0.032	0.032	0.020	0.044	0.064	0.044	0.048	0.024	0.024

	Number of Groups 5								
	n_i:10/12/14/16/18			n_i:30/35/40/45/50			n_i:100/110/120/130/140		
LSD	0.240	0.188	0.200	0.188	0.240	0.280	0.224	0.224	0.240
Bonferroni	0.032	0.040	0.024	0.028	0.032	0.036	0.024	0.044	0.032
Dunn-Sidak	0.032	0.040	0.024	0.028	0.032	0.040	0.024	0.044	0.032
Scheffe	0.012	0.024	0.012	0.016	0.000	0.016	0.012	0.024	0.012
REGW - F	0.032	0.036	0.028	0.032	0.036	0.040	0.028	0.044	0.040
REGW - Q	0.020	0.032	0.016	0.024	0.020	0.020	0.020	0.044	0.024
SNK	0.028	0.032	0.020	0.032	0.036	0.056	0.032	0.056	0.032
Tukey a	0.036	0.036	0.024	0.036	0.028	0.052	0.032	0.056	0.032
Tukey b	0.028	0.032	0.020	0.032	0.036	0.056	0.032	0.056	0.032
Duncan	0.136	0.128	0.128	0.128	0.152	0.204	0.140	0.164	0.152
Hochberg's GT2	0.036	0.028	0.024	0.028	0.032	0.044	0.028	0.044	0.036
Gabriel	0.032	0.028	0.024	0.028	0.028	0.040	0.028	0.044	0.032

	Number of Groups 7								
	n_i:10/12/14/16/18/20/22			n_i:30/35/40/45/50/55/60			n_i:100/110/120/130/140/150/160		
LSD	0.396	0.444	0.488	0.436	0.408	0.432	0.444	0.408	0.468
Bonferroni	0.024	0.044	0.040	0.040	0.012	0.032	0.024	0.028	0.036
Dunn-Sidak	0.024	0.044	0.044	0.040	0.016	0.032	0.024	0.028	0.036
Scheffe	0.012	0.020	0.012	0.012	0.004	0.000	0.000	0.012	0.004
REGW - F	0.048	0.048	0.044	0.032	0.028	0.044	0.032	0.032	0.036
REGW - Q	0.020	0.032	0.012	0.028	0.020	0.024	0.012	0.020	0.024
SNK	0.040	0.052	0.068	0.052	0.020	0.036	0.028	0.028	0.044
Tukey a	0.036	0.056	0.100	0.068	0.024	0.036	0.044	0.036	0.052
Tukey b	0.040	0.052	0.068	0.052	0.020	0.036	0.028	0.028	0.044
Duncan	0.204	0.244	0.232	0.228	0.216	0.224	0.196	0.204	0.212
Hochberg's GT2	0.032	0.044	0.084	0.044	0.016	0.032	0.024	0.028	0.040
Gabriel	0.032	0.044	0.080	0.044	0.016	0.032	0.024	0.028	0.040

Table 4: Comparisonwise Error Rates Results for Equal Sample Size

	Number of Groups 3								
	n_i:10/10/10			n_i:30/30/30			n_i:100/100/100		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.048	0.048	0.055	0.057	0.048	0.039	0.040	0.063	0.061
Bonferroni	0.011	0.011	0.011	0.016	0.021	0.016	0.015	0.015	0.017
Dunn-Sidak	0.011	0.011	0.011	0.016	0.021	0.016	0.015	0.015	0.017
Scheffe	0.011	0.011	0.009	0.015	0.019	0.015	0.009	0.012	0.013
REGW - F	0.013	0.016	0.020	0.020	0.024	0.019	0.016	0.024	0.019
REGW - Q	0.009	0.016	0.023	0.019	0.020	0.017	0.019	0.023	0.024
SNK	0.012	0.016	0.023	0.019	0.020	0.017	0.019	0.023	0.024
Tukey a	0.012	0.015	0.019	0.020	0.023	0.019	0.020	0.017	0.025
Tukey b	0.011	0.016	0.020	0.017	0.019	0.017	0.017	0.019	0.021
Duncan	0.029	0.029	0.039	0.040	0.033	0.031	0.024	0.045	0.039
Hochberg's GT2	0.011	0.011	0.013	0.016	0.021	0.016	0.015	0.015	0.017
Gabriel	0.011	0.011	0.013	0.016	0.021	0.016	0.015	0.015	0.017
	Number of Groups 5								
	n_i:10/10/10/10/10			n_i:30/30/30/30/30			n_i:100/100/100/100/100		
LSD	0.058	0.050	0.048	0.053	0.048	0.044	0.048	0.042	0.055
Bonferroni	0.004	0.006	0.007	0.005	0.007	0.006	0.004	0.004	0.008
Dunn-Sidak	0.004	0.006	0.007	0.005	0.007	0.006	0.004	0.004	0.008
Scheffe	0.005	0.004	0.006	0.002	0.008	0.002	0.002	0.002	0.003
REGW - F	0.013	0.010	0.008	0.022	0.031	0.018	0.016	0.007	0.014
REGW - Q	0.009	0.012	0.008	0.016	0.027	0.020	0.013	0.007	0.012
SNK	0.012	0.012	0.008	0.016	0.027	0.022	0.014	0.007	0.013
Tukey a	0.010	0.013	0.012	0.017	0.021	0.020	0.012	0.008	0.013
Tukey b	0.009	0.011	0.008	0.015	0.027	0.020	0.013	0.007	0.012
Duncan	0.060	0.054	0.042	0.070	0.069	0.057	0.065	0.044	0.059
Hochberg's GT2	0.008	0.011	0.012	0.016	0.018	0.012	0.011	0.009	0.013
Gabriel	0.006	0.010	0.010	0.014	0.018	0.012	0.009	0.007	0.012
	Number of Groups 7								
	n_i:10/10/10/10/10/10/10			n_i:30/30/30/30/30/30/30			n_i:100/100/100/100/100/100/100		
LSD	0.052	0.047	0.057	0.054	0.045	0.050	0.053	0.049	0.050
Bonferroni	0.002	0.001	0.003	0.003	0.002	0.004	0.003	0.001	0.004
Dunn-Sidak	0.002	0.001	0.003	0.003	0.002	0.004	0.003	0.002	0.005
Scheffe	0.003	0.003	0.002	0.003	0.004	0.004	0.000	0.000	0.005
REGW - F	0.014	0.010	0.014	0.018	0.012	0.020	0.022	0.012	0.023
REGW - Q	0.013	0.007	0.010	0.013	0.014	0.022	0.018	0.005	0.024
SNK	0.016	0.010	0.010	0.013	0.016	0.026	0.018	0.005	0.028
Tukey a	0.013	0.008	0.011	0.013	0.030	0.018	0.012	0.006	0.022
Tukey b	0.013	0.007	0.009	0.013	0.018	0.022	0.017	0.005	0.024
Duncan	0.106	0.074	0.100	0.114	0.095	0.083	0.097	0.083	0.099
Hochberg's GT2	0.012	0.007	0.007	0.011	0.011	0.018	0.012	0.004	0.016
Gabriel	0.012	0.007	0.007	0.011	0.016	0.018	0.012	0.004	0.016

Table 5: Comparisonwise Error Rates Results for Unequal Sample Size

	Number of Groups 3								
	n_i: 10/12/14			n_i: 30/35/40			n_i: 100/110/120		
	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 8$
LSD	0.051	0.053	0.033	0.049	0.075	0.075	0.052	0.045	0.052
Bonferroni	0.011	0.017	0.007	0.015	0.024	0.024	0.019	0.011	0.008
Dunn-Sidak	0.012	0.017	0.007	0.016	0.024	0.025	0.019	0.011	0.008
Scheffe	0.009	0.016	0.003	0.012	□ v Ĥ	0.021	0.016	0.009	0.008
REGW - F	0.013	0.023	0.012	0.025	0.033	0.024	0.024	0.019	0.012
REGW - Q	0.011	0.015	0.008	0.013	0.020	0.024	0.024	0.009	0.009
SNK	0.015	0.015	0.009	0.025	0.028	0.028	0.027	0.016	0.011
Tukey a	0.017	0.019	0.011	0.021	0.031	0.027	0.021	0.015	0.009
Tukey b	0.016	0.013	0.008	0.021	0.024	0.024	0.023	0.011	0.009
Duncan	0.035	0.037	0.023	0.039	0.052	0.051	0.037	0.027	0.032
Hochberg's GT2	0.013	0.016	0.007	0.016	0.027	0.024	0.017	0.011	0.008
Gabriel	0.013	0.016	0.007	0.016	0.027	0.024	0.017	0.011	0.008

	Number of Groups 5								
	n_i: 10/12/14/16/18			n_i: 30/35/40/45/50			n_i: 100/110/120/130/140		
LSD	0.053	0.047	0.043	0.047	0.054	0.057	0.050	0.050	0.054
Bonferroni	0.006	0.006	0.004	0.005	0.006	0.006	0.003	0.007	0.006
Dunn-Sidak	0.006	0.006	0.004	0.005	0.006	0.006	0.003	0.007	0.006
Scheffe	0.004	0.005	0.005	0.007	0.001	0.006	0.004	0.006	0.003
REGW - F	0.013	0.016	0.010	0.011	0.009	0.016	0.010	0.013	0.012
REGW - Q	0.006	0.010	0.006	0.010	0.005	0.008	0.008	0.009	0.010
SNK	0.012	0.016	0.012	0.014	0.010	0.017	0.012	0.012	0.011
Tukey a	0.014	0.010	0.010	0.014	0.008	0.015	0.010	0.013	0.009
Tukey b	0.011	0.012	0.010	0.012	0.010	0.017	0.012	0.012	0.011
Duncan	0.054	0.057	0.043	0.059	0.053	0.065	0.062	0.055	0.048
Hochberg's GT2	0.014	0.008	0.010	0.013	0.009	0.014	0.010	0.012	0.011
Gabriel	0.013	0.007	0.009	0.012	0.007	0.012	0.009	0.010	0.009

	Number of Groups 7								
	n_i: 10/12/14/16/18/20/22			n_i: 30/35/40/45/50/55/60			n_i: 100/110/120/130/140/150/160		
LSD	0.050	0.052	0.063	0.049	0.040	0.046	0.049	0.046	0.055
Bonferroni	0.002	0.002	0.003	0.003	0.001	0.002	0.002	0.002	0.002
Dunn-Sidak	0.002	0.002	0.003	0.004	0.001	0.002	0.002	0.002	0.002
Scheffe	0.005	0.003	0.002	0.002	0.000	0.000	0.000	0.005	0.000
REGW - F	0.017	0.013	0.007	0.006	0.006	0.013	0.013	0.012	0.012
REGW - Q	0.008	0.008	0.002	0.005	0.005	0.011	0.002	0.010	0.005
SNK	0.024	0.021	0.018	0.011	0.005	0.013	0.005	0.018	0.008
Tukey a	0.016	0.014	0.016	0.011	0.004	0.012	0.006	0.014	0.007
Tukey b	0.021	0.017	0.016	0.010	0.004	0.013	0.005	0.014	0.008
Duncan	0.115	0.097	0.084	0.083	0.071	0.083	0.093	0.078	0.086
Hochberg's GT2	0.012	0.012	0.014	0.008	0.003	0.010	0.002	0.010	0.005
Gabriel	0.012	0.012	0.014	0.008	0.003	0.010	0.002	0.010	0.005

Table 6: Power and FDR Results of MCPs According to Variance, Number of Groups and Sample Size*

n	σ^2	Per-Pair Power			Any-Pair Power		
		k=3	k=5	k=7	k=3	k=5	k=7
$n_i = \ldots = n_j = 10$	2	.712-.854	.769-.942	.795-.961	.864-.932	1.00-1.00	1.00-1.00
	4	.366-.566	.598-.817	.674-.879	.492-.696	1.00-1.00	1.00-1.00
	8	.180-.330	.397-.666	.504-.771	.284-.476	1.00-1.00	1.00-1.00
$n_i = \ldots = n_j = 30$	2	.996-.996	.996-1.00	.991-1.00	.996-.996	1.00-1.00	1.00-1.00
	4	.898-.960	.889-.987	.881-.993	.968-.996	1.00-1.00	1.00-1.00
	8	.614-.774	.702-.894	.753-.932	.752-.876	1.00-1.00	1.00-1.00
$n_i = \ldots = n_j = 100$	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	8	.992-1.00	.987-.999	.978-1.00	1.00-1.00	1.00-1.00	1.00-1.00
$n_i \neq \ldots \neq n_j$ 10/12/14/16/ 18/20/22	2	.858-.932	.873-.977	.892-.993	.948-.980	1.00-1.00	1.00-1.00
	4	.444-.664	.669-.886	.762-.939	.620-.820	1.00-1.00	1.00-1.00
	8	.246-.432	.519-.746	.624-.841	.380-.588	.996-1.00	1.00-1.00
$n_i \neq \ldots \neq n_j$ 30/35/40/45/ 50/55/60	2	1.00-1.00	.999-1.00	.999-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	.968-.990	.946-.996	.953-.999	1.00-1.00	1.00-1.00	1.00-1.00
	8	.732-.866	.789-.949	.821-.970	.864-.948	1.00-1.00	1.00-1.00
$n_i \neq \ldots \neq n_j$ 100/110/120/ 130/140/150/160	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	8	.996-1.00	.996-1.00	.993-1.00	1.00-1.00	1.00-1.00	1.00-1.00

*Results are summarized with minimum and maximum values (min-max)

procedure has the lowest per-pair power. The per-pair power of the REGW-Q procedure was not as powerful as the LSD or Duncan procedures; also, the SNK per-pair power was similar to the LSD and Duncan procedures. The power of the Bonferroni, Dunn-Sidak, Gabriel and Hochberg GT2 procedures were very close and the REGW-F, REGW-Q, and Tukey's a and b procedures are close for equal and unequal sample sizes.

Like per-pair, any-pair and all-pair powers increased as variance decreased and as sample size increased. If the group number and sample sizes were small, the LSD, Duncan and REGW-F procedures had high any-pair powers and the Scheffe procedure had the lowest any-pair power. As the number of groups increased, any-pair power reached its highest level.

All-pair-power decreased as the number of groups and the variance both increased. When sample size and number of groups were small,

Table 6 (continued): Power and FDR Results of MCPs According to Variance, Number of Groups and Sample Size*

n	σ^2	All-Pair-Power			False Discovery Rate		
		k=3	k=5	k=3	k=5	k=3	k=5
$n_i = \ldots = n_j = 10$	2	.560-.776	.044-.564	.560-.776	.044-.564	.560-.776	.044-.564
	4	.240-.436	.000-.092	.240-.436	.000-.092	.240-.436	.000-.092
	8	.076-.196	.000-.004	.076-.196	.000-.004	.076-.196	.000-.004
$n_i = \ldots = n_j = 30$	2	.996-.996	.968-1.00	.996-.996	.968-1.00	.996-.996	.968-1.00
	4	.828-.924	.276-.884	.828-.924	.276-.884	.828-.924	.276-.884
	8	.476-.672	.000-.292	.476-.672	.000-.292	.476-.672	.000-.292
$n_i = \ldots = n_j = 100$	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	.996-1.00	1.00-1.00	.996-1.00	1.00-1.00	.996-1.00
	8	.984-1.00	.888-.992	.984-1.00	.888-.992	.984-1.00	.888-.992
$n_i \neq \ldots \neq n_j$ 10/12/14/16/ 18/20/22	2	.768-.884	.232-.812	.768-.884	.232-.812	.768-.884	.232-.812
	4	.264-.508	.000-.284	.264-.508	.000-.284	.264-.508	.000-.284
	8	.112-.296	.000-.012	.112-.296	.000-.012	.112-.296	.000-.012
$n_i \neq \ldots \neq n_j$ 30/35/40/45/ 50/55/60	2	1.00-1.00	.992-1.00	1.00-1.00	.992-1.00	1.00-1.00	.992-1.00
	4	.936-.980	.616-.964	.936-.980	.616-.964	.936-.980	.616-.964
	8	.592-.784	.044-.592	.592-.784	.044-.592	.592-.784	.044-.592
$n_i \neq \ldots \neq n_j$ 100/110/120/ 130/140/150/160	2	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	4	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
	8	.992-1.00	.968-.100	.992-1.00	.968-.100	.992-1.00	.968-.100

*Results are summarized with minimum and maximum values (min-max)

the LSD, Duncan and REGW-F procedures had the highest power and the Scheffe procedure had the lowest power. As the number of groups increased, the LSD, Duncan and SNK procedures had the highest all-pair powers and the Scheffe procedure has the lowest all-pair power for both equal and unequal sample sizes.

For small groups and equal and unequal sample sizes, the LSD, REGW-F, REGW-Q, SNK, Tukey's b and Duncan FDR values were higher than the other values. As the number of groups increased all MCPs become closer and, as the number of comparisons increased, the FDR gets smaller.

The familywise error rates of the SNK and REGW-Q procedures were not as high as those of the LSD and Duncan procedures. Similarly, for the comparisonwise error rate, the LSD and Duncan procedures had the highest rates, whereas the Scheffe, Bonferroni and Dunn-Sidak procedures had the smallest rates. The LSD, Duncan and SNK procedures had the

highest powers, whereas the Scheffe procedure had the lowest power.

The any-pair powers of the LSD and Duncan procedures were high, but the Scheffe power was low due to small sample size and number of groups; power reached its maximum value as the number of groups increased. The all-pair powers of the LSD, Duncan, REGW-F, REGW-Q, SNK and Tukey's a and b procedures were the highest, but the Scheffe power was the lowest due to small sample size and number of groups. As the number of groups increased, the LSD, Duncan and SNK procedures had the highest power and the Scheffe procedure had the lowest power. FDR values of the LSD, REGW-F, REGW-Q, SNK and Duncan procedures were higher than those of the other procedures for low number of groups.

Conclusion

Findings from this study that differed from the literature were: (1) the SNK procedure is as robust as the LSD and Duncan procedures for controlling the error rate (Bernhardson, 1975; Curran-Everett, 2000; Maxwell, 2004), and (2) the REGW-Q procedure is as robust as the LSD for CWE (Menéndez De La Fuente, 1999).

Based on study results, the LSD and Duncan procedures are not recommended due to high error rates. The Scheffe procedure is not recommended due to its low power. There were no remarkable differences between the other procedures, thus, it is not possible to recommend one specific pairwise MCP for all situations that applied researchers may encounter.

References

Bender, R., & Lange, S. (2001). Adjusting for multiple testing-when and how? *Journal of Clinical Epidemiology*, *54*(*4*), 343-349.

Benjamini, Y., & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(*1*), 289-300.

Bernhardson, C. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, *31*, 229-232.

Cabral, H. J. (2008). Multiple comparisons procedures. *Circulation*, *117*(*5*), 698-701.

Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *279*(*1*), R1-R8.

Dayton, C. M. (1998). Information criteria for the paired-comparisons problem. *The American Statistician*, *52*(*2*), 144-151.

Demirhan, H., Dolgun, N. A., Demirhan, Y. P., & Dolgun, M. O. (2010). Performance of some multiple comparison tests under heteroscedasticity and dependency. *Journal of Statistical Computation and Simulation*, *80*(*10*), 1083-1100.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*(*351*), 574-583.

Ekenstierna, M. (2004). *Multiple comparison procedures based on marginal p-values*. Uppsala University Department of Mathematics Project Report No: 2004(12), 1-26. http://www2.math.uu.se/research/pub/Ekenstierna.pdf.

Horn, M., & Vollandt, R. (2000). A survey of sample size formulas for pairwise and many one multiple comparisons in the parametric, nonparametric and binomial Case. *Biometrical Journal*, *42*(*1*), 27-44.

Hsu, J. C. (1996). *Multiple comparisons: theory and methods*. London, England: Chapman & Hall/CRC.

Jaccard, J., & Guilamo-Ramos, V. (2002a). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*, 130-146.

Jaccard, J., & Guilamo-Ramos V. (2002b). Analysis of variance frameworks in clinical chìld and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child and Adolescent Psychology*, *31*, 278-294.

Kemp, K. (1975). Multiple comparisons: Comparisonwise versuse experimentwise type I error rates and their relationship to power. *Journal of Dairy Science*, *58*(*9*), 1374-1378.

Keselman, H., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods*, *4*, 58-69.

Keselman, H., Cribbie, R. A., & Holland, B. (2004). Pairwise multiple comparison test procedures: an update for clinical child and adolescent psychologists. *Journal of Clinical Child & Adolescent Psychology*, *33*(*3*), 623-645.

Keselman, H., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(*3*), 350.

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons*. Newbury Park, CA: Sage Publications.

Lowry, S. (1992). Use and misuse of multiple comparisons in animal experiments. *Journal of Animal Science*, *70*(*6*), 1971-1977.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, *25*(*12*), 1032-1037.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. London, England: Lawrence Erlbaum Associates.

Menéndez De La Fuente, I., Vallejo Seco, G., & Fernández García P. (1999). Multiple comparison procedures for simple one-way ANOVA with dependent data. *Spanish journal of psychology*, *2*(*1*), 55-63.

Montgomery, D. C.(2001). *Design and analysis of experiments*. New York, NY: John Wiley&Sons.

O'Brien, P. (1983). The appropriateness of analysis of variance and multiple-comparison procedures. *Biometrics*, *39*(*3*), 787-788.

O'Neill, R., & Wetherill, G. (1971). The present state of multiple comparison methods. *Journal of the Royal Statistical Society*, *Series B (Methodological)*, *33*(*2*), 218-250.

Perry, J. (1986). Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology*, *79*(*5*), 1149-1155.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, *73*(*363*), 479-485.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods*, *7*(*4*), 504-523.

Ramsey, P. H., & Ramsey P. P. (2008). Power of pairwise comparisons in the equal variance and unequal sample size case. *British Journal of Mathematical and Statistical Psychology*, *61*(*1*), 115-131.

Roback, P. J., & Askins, R. A. (2005). Judicious use of multiple hypothesis tests. *Conservation Biology*, *19*(*1*), 261-267.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*(*1*), 43-46.

Ryan ,T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin*, *56*(*1*), 26-47.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561-584.

Tatlidil, H. (2002). Applied Multivariate Statistical Analysis, Ziraat Matbaacılık, Ankara.(in Turkish).

Toothaker, L. E. (1993*). Multiple comparison procedures*. Newbury Park, CA: Sage Publications.