International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 469–483

Comparing the efficiency levels of Multiple Comparison Methods for Normal Distributed Observations

Wandee Wanishsakpong[1], Jeeraporn Thaithanan[1], Bright Emmanuel Owusu[2], Tahiru Mahama[3]

[1]Department of Statistics
Faculty of Science
Kasetsart University
Bangkok 10900, Thailand

[2]Department of Mathematics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

[3]Department of Statistics
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana

email: wandee.w@ku.th, jeeraporn.t@ku.th,
bright.owusu@knust.edu.gh, tahirum25@gmail.com

(Received September 1, 2021, Accepted September 30, 2021)

Abstract

In this study, we compare the efficiency of six multiple comparison methods: Tukey's, Duncan's new multiple range tests, Scheffe's, Bonferroni's, Hochberg's, and Sidak's methods with normality assumption and homogeneity of variance assumption for three populations. Three different sample sizes were employed and classified into three levels: small (10,10,10), medium (30,30,30), and large (50,50,50). The mean of the populations was assumed to be equal to 30 and 100 ($\mu_1 = \mu_2 = \mu_3 = 30$ and $\mu_1 = \mu_2 = \mu_3 = 100$) for calculating

Key words and phrases: Multiple comparison, power of the test, Type I error rate, Tukey's method, Duncan's new multiple range test, Scheffe's method, Bonferroni's method, Hochberg's method, Sidak's method.
AMS (MOS) Subject Classifications: 62F03.
Corresponding Author: Jeeraporn Thaithanan
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

Type I error rate. The mean of the three populations used at different levels; ($\mu_1 = 28, \mu_2 = 30, \mu_3 = 32$), ($\mu_1 = 26, \mu_2 = 30, \mu_3 = 34$), ($\mu_1 = 24, \mu_2 = 30, \mu_3 = 36$), ($\mu_1 = 98, \mu_2 = 100, \mu_3 = 102$), ($\mu_1 = 95, \mu_2 = 100, \mu_3 = 105$), and ($\mu_1 = 92, \mu_2 = 100, \mu_3 = 108$) for evaluating the power of the various multiple comparison tests. In this paper, each population variance is assumed to be equal to 10 and 30 with a significance level of 0.05. Monte Carlo simulation was performed and repeated 5000 times for each situation. Based on the results captured in this research, Tukey's method, Scheffe's method, Bonferroni's method, and Hochberg's method can control Type I error rates for all situations. Moreover, Hochberg's method provides the highest power of the test for all cases. The power of the test increases as the sample size increases and the variance decreases with the difference between population means for each group also increases. As a result, Hochberg's method tends to be the best efficient multiple comparison method based on normal distribution data.

1 Introduction

Analysis of variance (ANOVA) is the most common analytical technique in comparing the differences among the treatment means of three or more populations. Constructing ANOVA relies on the assumption that all the k populations ($k \geq 3$) are homoscedastic which refers to the k populations having equal variance (homogeneity), and the normality assumption (the data follows a normal distribution) of the data is assumed. The normality and homoscedastic assumptions can be assessed using the Kolmogorov-Smirnov method and Bartlett's test for equality of variance, respectively (Cobb [1] and Hsu [2]).

ANOVA (Parametric test) and Kruskal-Wallis (non-parametric version for ANOVA) are sufficient for testing the differences among more than 2 sample groups since they all look at how much the variation is in each subgroup (Frutos [3]). As the sum of squares treatments (between groups variation) increases, the higher the chance of resulting in a significant difference among the groups and vice versa. As the sum of squares errors (within-group variation) increases, the higher the chance of resulting in a non-significant difference among the groups and vice versa. If the null hypothesis (H_0) is rejected in the ANOVA for three or more populations, then there is an indication that a significant difference exists in the k population treatment means or at least one pair of the treatment means differs at a certain level of significance. However, the result of ANOVA does not provide detailed information regarding the differences between each pair of populations. Post Hoc Methods/Multiple Comparison Tests (MCTs) arise when the null hypothesis (H_0: All the treatment means are the same) is rejected. Therefore, further tests such as the multiple comparison methods will be required to assess which pairs of the experimental population are different.

According to Sangseok and Dong [4], there are several methods for performing multiple comparison tests. Some methods are appropriate in one situation but may not be appropriate in another. MCT methods such as Tukey's method, Duncan's new multiple range test method, Scheffe's method, Bonferroni's method, Hochberg's method, and Sidak's method can be used to find and compare the appropriate methods based on normal distribution data. In choosing and considering the appropriate test, we must maintain the balance between the statistical power of a test and Type I error.

Researchers such as Sangseok and Dong [4] discussed how to test multiple hypotheses simultaneously while limiting the Type I error rate which is caused by α inflation, and also help researchers to understand the differences among MCTs and apply them appropriately. MCTs such as Bonferroni, Dunnett, Scheffe Statistics and, Newman-Keuls Tukey have been evaluated by these researchers. A problem occurs if the error rate increases while multiple hypothesis tests are performed simultaneously. In identifying which of the treatment pairs differ, we need to control the Type I error rate (level of significance, α). When a False is accepted (i.e., must be rejected), a Type II error occurs. When a True null hypothesis (H_0) is rejected (i.e., not to be rejected), a Type I error occurs. Type I error is not likely to occur if the test is too conservative. The p-value represents the lowest value of the significance level of which the null hypothesis will be rejected. It denotes the probability that an extreme result will be observed if H_0 is true. The p-value is sometimes called the significance probability, asymptotic significance, or probability value. The p-value of any experiment is specified as a random variable in a sample space of an experiment that lies between 0 and 1. Comparison of hypotheses among groups is termed as Family. Family Wise Error (FWE) is a Type I error that occurs when each Family is compared (Kim [5]).

Type I error (α inflation) should always be considered and controlled in multiple comparisons applied in each multiple comparison method. An α

inflation increases when the number of comparisons among groups increases. Inflated $\alpha = 1-(1-\alpha)^N$, where N represents the number of hypotheses tested and α is the error rate. MCTs are classified under two categories: single-step and stepwise procedures. The single-step procedure assumes one hypothetical Type I error rate. All pairwise comparisons/multiple hypotheses are performed under this assumption using one critical value; i.e., every comparison is independent. Fisher's least significant difference (LSD) test, Bonferroni, Sidak, Scheffe, Tukey, Tukey-Kramer, Hochberg's GF2, Gabriel and Dunnett test are all examples of Single-step procedures. The stepwise procedure, on the other hand, tackles Type I error according to previously selected comparison results, or it manipulates pairwise comparisons in a predetermined manner and makes each comparison only when the former comparison result is statistically significant. This method is lucrative because it ameliorates the statistical power of the process while controlling the Type I error rate throughout the process. It identifies the most significant test among the comparison test statistics. Comparisons are successively performed when the precedent test is significant. All the remaining tests are rejected if one comparison test during the process is insignificant (Pizarro et al. [6]).

The stepwise procedure method, unlike the single-step method, which indicates the same level of significance, categorizes all relevant groups into statistically similar subgroups. Ryan Einot-Gabriel Welsch, Student-Newman-Keuls (SNK), and Duncan tests use Stepwise Procedure. All the methods mentioned previously can be used under the equal variance assumption. Other methods such as Tamhanes T2, Dunnetts T3, and Dunnets C tests (non-parametric approach) can be used if the equal variances assumption is violated in the ANOVA process. According to Sangseok and Dong [4], there are four criteria for assessing and comparing the Post Hoc methods: conservativeness, optimality, convenience, and robustness. Conservativeness is described as the statistical result of the post hoc method is significant with a certain controlled Type I error rate. Optimality refers to the smallest CI among conservative statistics. It is the smallest standard error among the conservative statistics. Conveniency is considered easy to calculate since most computer programs will tackle it. Robustness refers to the violation of the equal variance assumption in ANOVA (Sauder and DeMars [7]). Some methods are less insensitive since they do violate the equal variance assumption.

Stoline [8] asserted that LSD, Sidak, Bonferroni, and Dunnet tests do not pose any problem when using the t-statistic since there is no assumption that the number of samples in each group is the same. The Tukey test can be problematic when using the studentized range distribution since all the sample sizes are the same in the null hypothesis. The best method to use when the sample numbers are different is the Tukey-Kramar test which uses the harmonic mean of sample numbers.

Fisher's LSD procedure is not recommended because it is not conservative (cannot control the overall confidence level or the error rate). Bonferroni, Sidak, and Scheffe's tests are recommended since they are conservative. Although Tukey's HSD test is somewhat in between, it can be used when comparing all pairs of means. With a large number of groups, Tukey's HSD will be more conservative and much preferred (Pizarro et al. [6]).
This research, therefore, presents the comparison of six Post Hoc Methods. These methods are Tukey's method, Duncan's new multiple range test method, Scheffe's method, Bonferroni's method, Hochberg's method, and Sidak's method based on their efficiency levels using normal distribution data.

2 Data and methods

The purpose of this research is to compare the efficiency levels of six multiple comparison methods: Tukey's, Duncan's new multiple range test, Scheffe's, Bonferroni's, Hochberg's, and Sidak's methods. Type I error rate and power of the test analyses are performed in R programming.

2.1 Generating data through simulation study

Three populations were considered in this research. The sample sizes are classified into three levels: small (10,10,10), medium (30,30,30) and large (50,50,50). The mean of the populations is assumed to be equal to 30 and 100 ($\mu_1 = \mu_2 = \mu_3 = 30$ and $\mu_1 = \mu_2 = \mu_3 = 100$) for calculating Type I error rate. The mean of the three (3) populations were considered at different levels; ($\mu_1 = 28, \mu_2 = 30, \mu_3 = 32$), ($\mu_1 = 26, \mu_2 = 30, \mu_3 = 34$), ($\mu_1 = 24, \mu_2 = 30, \mu_3 = 36$), ($\mu_1 = 98, \mu_2 = 100, \mu_3 = 102$), ($\mu_1 = 95, \mu_2 = 100, \mu_3 = 105$) and ($\mu_1 = 92, \mu_2 = 100, \mu_3 = 108$). The data is generated from a normal distribution with the parameters mean (μ) and variance (σ^2) using simulation study in `R`. Replication is done on the data generated for each situation using Monte Carlo Simulation runs.

2.2 Analyzing the simulated data

The test statistics for the six multiple comparison methods were obtained in R software. The Type I error rate for the various methods were calculated as (see Douglas [9])

$$\text{Type I error rate} = \frac{\text{number of null hypothesis rejected}}{5000}. \tag{1}$$

Besides, the performance of the methods were then compared and evaluated by using a Type I error rate at a significance level of 0.05. The Bradley method criteria (Bradley [10]) could control the Type I error if the Type I error rate lies in the interval [0.025,0.075]. The power of the test for the six multiple comparison methods were then calculated to control the Type I error rate. The performance of the proposed methods were compared and evaluated by using the power of the test.

2.3 Test statistics for multiple comparison methods

2.3.1 Tukey's method

$$T = q_{\alpha,k,N-k}\sqrt{\frac{MSE}{n}}, \tag{2}$$

where $q_{\alpha,k,N-k}$ is the critical value from the Studentized range table, represents the Mean Squared Error (Mean Squared Error within groups) from the ANOVA Table, n is the sample size and k is the number of groups (treatments).

2.3.2 The Duncan's New Multiple Range Test

$$D = q_{\alpha,r,N-k}\sqrt{\frac{MSE}{n}}, \tag{3}$$

where $q_{\alpha,k,N-k}$ represents the critical value from Duncans New Multiple rank test tables and r is the interval between the means of each pair to be compared.

2.3.3 Scheffe's method

$$CV_d = \sqrt{(k-1)(F^*)(MSE)(\frac{2}{n})}, \tag{4}$$

where F^* is the value from the Upper Percentage of the F Distribution table with lower degrees of freedom $k-1$ and upper degrees of freedom $N-k$.

2.3.4 Bonferroni's method

$$B = t_{\frac{\alpha}{2S},N-k}\sqrt{\frac{2MSE}{n}}, \tag{5}$$

where $t_{\frac{\alpha}{2S},N-k}$ is the value from the t distribution table with a significance level α, with degrees of freedom $N-k$ and S is the number of pairs of samples to be compared. The value can also be read from the Bonferroni table.

2.3.5 Hochberg's method

$$GT2 = \frac{(\bar{y}_i - \bar{y}_j)}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}}, \tag{6}$$

where n_i and n_j are the number of repeats for each sample group i and j respectively, $\bar{y}_i$ and$\bar{y}_j$ are the means of the sample groups i and j respectively.

2.3.6 Sidak's method

$$DS = t_{S,N-k}\sqrt{\frac{MSE}{n}}, \tag{7}$$

where $t_{S,N-k}$ is a value from the percentage point of Dunn-Sidaks Multiple comparison Test tables.

3 Results and Discussion

The results of the six multiple comparison methods are based on normally distributed data (Ozkaya and Ercan [11]) and the homogeneity of variance assumption for three populations. The results are summarized into sections.

3.1 Type I error rate

This section comprises Type I error rate for Multiple Comparison Methods and Bradley criteria to control Type I error rate when the significance is equal to 0.05.

From Tables 1 and 2, an asterisk attached to the Type I error rate indicates that the method can control the Type I error rate reasonably well. It is evident from both tables that Tukey's, Scheffe's, Bonferroni's, and Hochberg's methods perform well. These methods could control the Type I error rate for all situations (when the population means are equal to 30

		Sample size (n_1, n_2, n_3)		
Variance	Methods	Small	Medium	Large
		(10,10,10)	(30,30,30)	(50,50,50)
10	Tukey	0.0488*	0.0494*	0.0542*
	Duncan	0.0988	0.0976	0.0978
	Scheffe	0.0374*	0.0376*	0.0422*
	Bonferroni	0.0412*	0.0418*	0.0494*
	Hochberg	0.0418*	0.0422*	0.0496*
	Sidak	0.0822	0.0802	0.0898
30	Tukey	0.0524*	0.0498*	0.0518*
	Duncan	0.1032	0.1002	0.0970
	Scheffe	0.0388*	0.0400*	0.0390*
	Bonferroni	0.0426*	0.0448*	0.0450*
	Hochberg	0.0432*	0.0450*	0.0462*
	Sidak	0.0866	0.0860	0.0854

Table 1: Type I error rate when $\mu_1 = \mu_2 = \mu_3 = 30$. (* can control Type I error rate)

		Sample size (n_1, n_2, n_3)		
Variance	Methods	Small	Medium	Large
		(10,10,10)	(30,30,30)	(50,50,50)
10	Tukey	0.0510*	0.0504*	0.0462*
	Duncan	0.0992	0.0968	0.0966
	Scheffe	0.0388*	0.0370*	0.0338*
	Bonferroni	0.0422*	0.0428*	0.0396*
	Hochberg	0.0430*	0.0438*	0.0406*
	Sidak	0.0836	0.0880	0.0842
30	Tukey	0.0484*	0.0464*	0.0526*
	Duncan	0.0920	0.0914	0.0970
	Scheffe	0.0392*	0.0364*	0.0406*
	Bonferroni	0.0422*	0.0408*	0.0460*
	Hochberg	0.0432*	0.0414*	0.0464*
	Sidak	0.0802	0.0796	0.0852

Table 2: Type I error rate when $\mu_1 = \mu_2 = \mu_3 = 100$. (* can control Type I error rate)

and 100 with different variances at specific sample size). On the other hand, Duncan's new multiple range test and Sidak's methods could not control the Type error rate for all situations.

3.2 Power of the test

The power of the test for the multiple comparison methods are given in Tables 3–6. Analysis of the results revealed that Hochberg's method provides the highest power of the test for all situations. The power of the test increases as the sample size increases with a decreasing variance or the difference between populations means in each group increases. The efficiency level of Hochberg's method was the greatest followed by Tukey's, Bonferroni's, and Scheffe's methods. Duncan's and Sidak's method yielded no efficiency level because they could not control the Type I error rate for all the situations. Relative to the efficiency of the methods, Hochberg's method tends to be the best multiple comparison method based on normal distribution data.

3.3 MCT results

The MCT which could control the Type I error rate was also examined and the results have been given in Table 7. The results show that Tukey's (T), Scheffe's (S), Bonferroni's (B) and Hochberg's (GT) methods could control the Type I error rate for all situations while Duncan's new multiple range test and Sidak's methods could not control the Type I error rate efficiently.

The power of the test for MCT methods was also investigated and the results are given in Table 8. Analysis of the results revealed that Hochberg's method (GT) provided the highest power of the test for all situations and conditions (Table 8). The power of the test increases as the difference between population means for each group increases or the variance decreases. The results revealed by Tukey's, Scheffe's and Hochberg's methods are consistent with results presented by Sangseok and Dong [3] and Pizarro et al. [4]. These methods are recommended since they are conservative. Even though Sidak's method was recommended by the study of Sangseok and Dong [3], we do not recommend it for normally distributed data since it could not control the Type I error rate efficiently.

		Sample size (n_1, n_2, n_3)		
Variance	Methods	Small (10,10,10)	Medium (30,30,30)	Large (50,50,50)
$\mu_1 = 28$ $\mu_2 = 30$ $\mu_3 = 32$	Tukey	0.0084	0.2008	0.5802
	Duncan	-	-	-
	Scheffe	0.0052	0.1602	0.5218
	Bonferroni	0.0060	0.1778	0.5502
	Hochberg	0.0324*	0.3940*	0.7620*
	Sidak	-	-	-
$\mu_1 = 26$ $\mu_2 = 30$ $\mu_3 = 34$	Tukey	0.3614	0.9848	1.0000*
	Duncan	-	-	-
	Scheffe	0.3120	0.9806	0.9998
	Bonferroni	0.3306	0.9824	1.0000*
	Hochberg	0.5732*	0.9962*	1.0000*
	Sidak	-	-	-
$\mu_1 = 24$ $\mu_2 = 30$ $\mu_3 = 36$	Tukey	0.9152	1.0000*	1.0000*
	Duncan	-	-	-
	Scheffe	0.8878	1.0000*	1.0000*
	Bonferroni	0.8962	1.0000*	1.0000*
	Hochberg	0.9676*	1.0000*	1.0000*
	Sidak	-	-	-

Table 3: Power of the test when variance is equal to 10 and distance of means are equal 2,4 and 6 respectively.

		Sample size (n_1, n_2, n_3)		
Variance	Methods	Small	Medium	Large
		(10,10,10)	(30,30,30)	(50,50,50)
$\mu_1 = 28$ $\mu_2 = 30$ $\mu_3 = 32$	Tukey	0.0004	0.0068	0.0346
	Duncan	-	-	-
	Scheffe	0.0002	0.0044	0.0222
	Bonferroni	0.0004	0.0052	0.0278
	Hochberg	0.0036*	0.0316*	0.1118*
	Sidak	-	-	-
$\mu_1 = 26$ $\mu_2 = 30$ $\mu_3 = 34$	Tukey	0.0218	0.3932	0.7982
	Duncan	-	-	-
	Scheffe	0.0140	0.3456	0.7522
	Bonferroni	0.0154	0.369	0.7762
	Hochberg	0.0766*	0.6028*	0.9082*
	Sidak	-	-	-
$\mu_1 = 24$ $\mu_2 = 30$ $\mu_3 = 36$	Tukey	0.1968	0.9342	0.9984
	Duncan	-	-	-
	Scheffe	0.1624	0.9166	0.9972
	Bonferroni	0.1718	0.9234	0.9976
	Hochberg	0.3840*	0.9756*	0.9996*
	Sidak	-	-	-

Table 4: Power of the test when variance is equal to 30 and the mean distance is equal to 2,4 and 6, respectively.

Variance	Methods	Sample size (n_1, n_2, n_3)		
		Small (10,10,10)	Medium (30,30,30)	Large (50,50,50)
$\mu_1 = 98$ $\mu_2 = 100$ $\mu_3 = 102$	Tukey	0.0068	0.2072	0.5956
	Duncan	-	-	-
	Scheffe	0.0050	0.1632	0.5350
	Bonferroni	0.0052	0.1828	0.5642
	Hochberg	0.0344*	0.4124*	0.7740*
	Sidak	-	-	-
$\mu_1 = 95$ $\mu_2 = 100$ $\mu_3 = 105$	Tukey	0.7156	0.9998	1.0000*
	Duncan	-	-	-
	Scheffe	0.6686	0.9998	1.0000*
	Bonferroni	0.6840	0.9998	1.0000*
	Hochberg	0.8622*	1.0000*	1.0000*
	Sidak	-	-	-
$\mu_1 = 92$ $\mu_2 = 100$ $\mu_3 = 108$	Tukey	0.9976	1.0000*	1.0000*
	Duncan	-	-	-
	Scheffe	0.9974	1.0000*	1.0000*
	Bonferroni	0.9974	1.0000*	1.0000*
	Hochberg	0.9994*	1.0000*	1.0000*
	Sidak	-	-	-

Table 5: Power of the test when variance is equal to 10 and the mean distance is equal to 2, 5 and 8, respectively.

Variance	Methods	Sample size (n_1, n_2, n_3)		
		Small (10,10,10)	Medium (30,30,30)	Large (50,50,50)
$\mu_1 = 98$ $\mu_2 = 100$ $\mu_3 = 102$	Tukey	0.0004	0.0068	0.0346
	Duncan	-	-	-
	Scheffe	0.0002	0.0044	0.0222
	Bonferroni	0.0004	0.0052	0.0278
	Hochberg	0.0036*	0.0316*	0.1118*
	Sidak	-	-	-
$\mu_1 = 95$ $\mu_2 = 100$ $\mu_3 = 105$	Tukey	0.0804	0.7358	0.9696
	Duncan	-	-	-
	Scheffe	0.0586	0.6906	0.9624
	Bonferroni	0.0660	0.7124	0.9662
	Hochberg	0.1926*	0.8704*	0.9892*
	Sidak	-	-	-
$\mu_1 = 92$ $\mu_2 = 100$ $\mu_3 = 108$	Tukey	0.5906	0.9994	1.0000*
	Duncan	-	-	-
	Scheffe	0.5332	0.9988	1.0000*
	Bonferroni	0.5514	0.9994	1.0000*
	Hochberg	0.7826*	0.9998*	1.0000*
	Sidak	-	-	-

Table 6: Power of the test when variance is equal to 30 and the mean distance is equal to 2, 5 and 8, respectively. (**Note:** - cannot control Type error rates by using Bradley criteria and * is the highest power of the test in this situation.)

$(\mu_1 = \mu_2 = \mu_3 = 30)$	Population size (n_1, n_2, n_3)		
Variance	Small (10,10,10)	Medium (30,30,30)	Large (50,50,50)
10	T, S, B, GT	T, S, B, GT	T, S, B, GT
30	T, S, B, GT	T, S, B, GT	T, S, B, GT
$(\mu_1 = \mu_2 = \mu_3 = 100)$	Population size (n_1, n_2, n_3)		
Variance	Small (10,10,10)	Medium (30,30,30)	Large (50,50,50)
10	T, S, B, GT	T, S, B, GT	T, S, B, GT
30	T, S, B, GT	T, S, B, GT	T, S, B, GT

Table 7: MCT Methods which can control Type I error rate when $\mu_1 = \mu_2 = \mu_3 = 30$ and $\mu_1 = \mu_2 = \mu_3 = 100$

		Population size (n_1, n_2, n_3)		
Variance	Difference of mean	Small	Medium	Large
		(10,10,10)	(30,30,30)	(50,50,50)
10	2	GT	GT	GT
	4	GT	GT	T, B, GT
	6	GT	T, S, B, GT	T, S, B, GT
30	2	GT	GT	GT
	4	GT	GT	GT
	6	GT	GT	GT
10	2	GT	GT	GT
	5	GT	GT	T, S, B, GT
	8	GT	T, S, B, GT	T, S, B, GT
30	2	GT	GT	GT
	5	GT	GT	GT
	8	GT	GT	T, S, B, GT

Table 8: Highest power of the MCT Methods.

4 Conclusions

We recommend Tukey's, Scheffe's, Bonferroni's, and Hochberg's methods for Multiple Comparison Test on normal distribution data since these methods can control Type I error rate for all situations. Duncan's new multiple range test and Sidak's method are not recommended for the Multiple Comparison Test since they could not control the Type I error rate for all situations.

Acknowledgment. The authors would like to thank the International SciKU Branding (ISB), Faculty of Science, Kasetsart University, Thailand for supporting this research.

References

[1] G. W. Cobb, An Introduction to Design and Analysis of Experiments, Springer-Verlag, New York, 1998.

[2] J. C. Hsu, Multiple Comparisons: Theory and Methods. Chapman & Hall, New York, 1996.

[3] I. P. Frutos, Controlling the Type I error rate by using the nonparametric bootstrap when comparing means, The British Journal of Mathematical and Statistical Psychology, **67,** (2014), 117–132.

[4] L. Sangseok, K. L. Dong, What is the proper way to apply the multiple comparison test?, Korean Journal of Anesthesiology, **71,** (2018), 354–360.

[5] T. K. Kim, Understanding one-way ANOVA using conceptual figures, Korean Journal of Anesthesiology, **70,** (2017), 22–26.

[6] J. Pizarro, E. Guerrero, P. L. Galindo, Multiple comparison procedures applied to model selection, Neurocomputing, **48,** (2002), 155–173.

[7] D. C. Sauder, C. E. DeMars, An Updated Recommendation for Multiple Comparisons, Advances in Methods and Practices in Psychological Science, **2,** (2019), 26–44.

[8] M. R. Stoline, The Status of Multiple Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs, The American Statistician, **35,** (1981), 134–141.

[9] C. E. Douglas, Multiple Comparisons: Philosophies and Illustrations, Amer. J. Physiol. Regulatory Integrative Comp. Physiol., **279,** (2000), R1–R8.

[10] J. V. Bradley, Robustness, Journal of Mathematical and Statistical Psychology, **31,** (1978), 144–151.

[11] G. Ozkaya, I. Ercan, Examining Multiple Comparison Procedures According to Error Rate, Power Type and False Discovery Rate, Journal of Modern Applied Statistical Methods, **11,** (2012), 348–360.