Menemui Matematik
(Discovering Mathematics)

journal homepage: https://myjms.mohe.gov.my/index.php/dismath/

Robust Post-Hoc Multiple Comparisons: A Study on Tukey Test Using Winsorized Mean

Nur Aniesa Amiera Ahmad Wahizan[1] and Nazihah Mohamed Ali[2*]
[1, 2] *Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia,*
43400 UPM Serdang, Selangor

[1]207748@student.upm.edu.my, [2]nazihanma@upm.edu.my
*Corresponding author

Received: 10 August 2024
Accepted: 29 October 2024

ABSTRACT

Post-hoc multiple comparisons methods play a crucial role in determining differences in group means following one-way Analysis of Variances (ANOVA) performance. Traditionally relying on assumptions of normality and homogeneity of variance that is common in statistical analyses, this method faces challenges in datasets with prevalent heterogeneity of variances. Therefore, robust central tendency methods, such as the Winsorized mean, become significant in statistical analyses. This research introduces the O'Brien test, a standard tool for testing homogeneity of variances, incorporating Winsorized mean in scenarios where the assumptions of one-way ANOVA are violated. Consequently, the interpretation results of one-way ANOVA and Tukey's test become more reliable. Simulation studies show that the proposed methodology is more preferable when dealing with non-normal distributions and moderately unequal variances. The comparison of power values of the test statistics leads to the conclusion that the modified approaches using robust estimators are more powerful than the classical approach.

Keywords: Post-Hoc Multiple Comparisons, Robust One Way ANOVA, Winsorized Mean.

INTRODUCTION

The heart of statistical analysis techniques for such comparison lies within the necessity to compare across multiple groups, which is best exemplified through post-host multiple comparison after establishing the existence of significant differences by conducting an analysis of variance (ANOVA). The simplicity and adequacy of ANOVA have assisted and evaluated the diversity of interest in helping the decision-making through testing hypotheses on different outcomes. However, this notion fails to describe and explain such differences. The comprehension of performing post-hoc multiple comparisons is much more detailed as the test was conducted to specify the sources of the group means that significantly differ among the groups.

The essence of post-hoc multiple comparison testing in reaching optimised utilisation and obtaining reliable statistical inference is adhering to the one-way ANOVA assumptions which is that the data must satisfy the normality, homogeneity variances and independence. Conventionally, the classical post-hoc multiple comparison approach needs the execution of the error term obtained from the omnibus test within the interest group. Hence, the assumption of homogeneity variances must be met in order to pairwise comparisons to yield valid results. Accordingly, the focus on this study fixates on modification of the test for homogeneity variances, allowing it to play a crucial role in robustness onto post-hoc multiple comparison. Various statistical tests establish by many statisticians in regards to evaluate the homogeneity of variances, encompassing diverse conditions (Bartlett, 1937; Cochran, 1941; Levy, 1975a, 1975b; O'Brien, 1978; Gupta & Rathie, 1983; Tang & Gupta, 1987; Nelson, 2000; Wilcox, 2002; Gupta et al., 2004).

The interest of this study is to do a transformation of the data by adopting Winsorized mean to the O'Brien (1978) test due to its versatility and compatibility with ANOVA model. The rationale behind the modification of O'Brien test is to seek and deem a flexible and robust method that is able to handle violations of variance homogeneity. The nuance of Winsorized mean offers the minimization of the extreme values' influence on the central tendency. Therefore, it provides more accurate identification of significant differences across the mean groups.

There are various statistical methods that are centred around the post-hoc multiple comparison and the specified test in this study is Tukey's Honest Significant Difference (HSD) test. The robustness of Tukey's HSD test with the adoption of modified Winsorized mean on the O'Brien test can be assessed on its risk of committing Type I error and power through simulation of different manipulations to gauge its performance. Petrinovich and Hardyck (1969) conducted research on finding the robustness of the Tukey and Scheffe test under differences such as different shapes, population variances, numbers of treatment levels, and sample sizes. The study's findings have concluded that both multiple comparison procedures lacked robustness in scenarios with uneven variances. Hence, this study proposes to delve deeper into the practical implications of utilising the modified O'Brien test with Winsorized mean as a robust for Tukey's HSD test in preserving the integrity of statistical inference, emphasising its violation of normality and homogeneity of variances. The methodology used in this study is presented in a systematic manner in the hope of enhancing the comprehension of conducting robust post-hoc multiple comparisons as a whole.

METHODOLOGY

O'brien Test with Winsorized Mean

The methodological studies pivot around a comprehensive analysis of one-way ANOVA as a whole, encompassing the satisfaction of the three assumptions, followed subsequently the calculations of one-way ANOVA, and finally, conducting the post-hoc multiple comparisons.

The O'Brien test is specifically designed to evaluate whether variances across groups are homogeneous, supporting valid ANOVA assumptions. However, real-world data often violate these assumptions, with non-normality and heterogeneous variances frequently present (Bishop, 1976, as cited in Wright, 2009). Such violations increase the risk of Type I errors, leading to incorrect rejections of the null hypothesis.

The Winsorized mean, which replaces the arithmetic mean, is introduced in the O'Brien test to address these issues by replacing extreme values with less extreme ones. This reduces sensitivity to outliers and improves the robustness of the measure of central tendency. This modification provides a more robust measure of central tendency, enhances the stability of variance estimates, and allows for a more reliable test of the null hypothesis, such that:

$$H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_k \tag{1}$$

This approach reduces the likelihood of Type I errors, thus improving the reliability of post-hoc comparisons.

Cochran (1941) claimed that the presence of heterogeneity in the sample should not be overlooked prior to conducting one-way ANOVA. When the violation ANOVA assumptions of normality and heterogeneity of variances occurred, transforming data using Winsorized mean in the O'Brien test poses as a robust central tendency, such that,

$$r_{ij}(0.5) = \frac{(n_i - 0.5 - 1)(n_i)\left(\mathrm{y_{ij}} - \bar{y}_w\right)^2 - 0.5S_i^2(n_i - 1)}{(n_i - 1)(n_i - 2)} \tag{2}$$

where n_i is the sample size, $\bar{y}_w$ is the 10% Winsorization mean and ${S_i}^2$ is the variances of the i^{th} subgroups. The 10% Winsorization means that the most extreme 10% of values are replaced with the nearest values that fall within the remaining 90%. The failure to reject the null hypothesis in equation (1) at the level of significance, α, must be achieved indicating that the variances are homogeneous.

The Winsorized mean stabilizes variance estimates across subgroups and enhances the accuracy of statistical inference. By mitigating the influence of extreme values, it reduces potential biases in variance estimation and increases the reliability of ANOVA results.

The Mechanics of Modified One-Way ANOVA

The utilisation and calculation of the modified one-way ANOVA can be performed confidently after ensuring that all three ANOVA assumptions are fully satisfied by adopting 10% Winsorized mean in the O'Brien test.

The mathematical model of one-way ANOVA follows the general linear regression model which describes the quantification of the relationship between the independent variable (treatment) and the dependent variable (response). The one-way ANOVA model can be expressed as,

$$y_{ij} = \mu + \tau + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, t;\ j = 1, 2, \ldots, r \tag{3}$$

where y_{ij} is the ij^{th} response, μ is the overall mean, τ is the treatment effect and ε_{ij} is the independent and identically distributed error terms.

One-way ANOVA is developed to determine whether multiple population means possess the same means by comparing the variation within the sample means. Thus, the null hypothesis testing can be stated as

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{4}$$

The decomposition of the modified F statistic for one-way ANOVA when the 10% Winsorized mean adopted can be defined such that

$$\mathrm{F} = \frac{MSTr_w}{MSE_w} \tag{5}$$

where $MSTr_w$ is the Winsorized mean square treatment and MSE_w is the Winsorized mean square error. The null hypothesis in Eqn. (4) is rejected when the F statistic in Eqn. (5) falls in the rejection region of the F tabulated with degrees of freedom $t - 1$ and $N - 1$. The rejection of the null hypothesis signifies that there is enough evidence to indicate that at least two populations differ significantly in mean from one another.

Robust Post-hoc Multiple Comparisons

The further statistical procedure preliminary to the assurance of the existence of differences in mean (reject H_0) in the performance of one-way ANOVA is post-hoc multiple comparisons. This study uses the Tukey Honest Significant Difference (HSD) test, also known as the Tukey's test, for its optimal balance between statistical power and having the ability to control Type I error superlatively (Sawyer, 2009). Tukey's test plays a crucial role in enhancing the depth of the specific and significant differences between pairs of treatment groups. Hence, the null hypothesis testing can be expressed as follows

H_0: There are no significant differences between the means of all possible pairs.

Despite its widespread use and practicability, the omnibus Tukey's test still underlies limitations that should be recognized to ensure the reliability of statistical inference. Tukey's test requires an equal number of observations for each group to optimize its efficiency. Additionally, the studentized range distribution, which follows the normal distribution, poses challenges by obligating the fulfilment of the homogeneity variance assumption based on the same number of observations (Rodger & Roberts, 2013). Recognizing the assumptions inherent in the test, it highlights that the problem of studentized range distribution in Tukey's test may restrict its efficiency if the violations of homogeneity are breached, thereby alleviating the severity of the robustness analysis.

In light of these challenges, the implementation of the Winsorized mean in the O'Brien test serves as a robust measure of central tendency for post-hoc multiple comparisons. This modification allows for a more reliable comparison of treatment means since the mean square error value has been adjusted to reduce the influence of outliers, particularly in cases where the assumptions of Tukey's test may be violated. Thus, the modified Tukey's test can be defined as follows:

$$\mathrm{HSD} = q_{\alpha,k,N-k}\sqrt{\frac{MSE_w}{n_k}} \tag{6}$$

where q is the studentized range. The conclusion of the existing pair treatment groups that have a significant difference in mean (reject H_0) in Tukey's test when $|\bar{r}_i - \bar{r}_j|$ is greater than or equal to the critical value from equation (6).

SIMULATION STUDY

The comparison of this study was made based on the robustness value obtained in Type I error between the classical O'Brien test, One-way ANOVA and Tukey's test, along with their respective modified tests in equation (2), (5) and (6). The experimental conditions of this study were manipulated in terms of the following:

Table 1: Simulation conditions.

Levels of sample sizes	Small (N_1), $n_1 = n_2 = n_3 = 8$ Moderate (N_2), $n_1 = n_2 = n_3 = 20$ Large (N_3), $n_1 = n_2 = n_3 = 40$
Degree of variance heterogeneity	Equal, (1:1:1) Moderately unequal, (1:1:4)
Types of distributions	Normal distribution Chi-square distribution Lognormal distribution

The consideration of committing a Type I error arises if it fails to remain within the desired alpha value or nominal level. The chosen nominal value in this study is 0.05 because it is commonly used by many statistical practitioners. Bradley (1978) suggested the criterion of empirical Type I error to detect robustness as being yielded between alpha values of 0.025 and 0.075.

The variance heterogeneity differs in the degrees ranging from heterogeneity equal to moderately and extremely unequal variances. The underlying concept of comparing the extreme degree across variance heterogeneity in the subgroup is expected to perform well in various degrees when there is less extreme variance of heterogeneity that are commonly encountered by researchers (Syed Yahaya, 2005).

Normal Distribution Simulation

Based on the findings presented in Table 1, under the equal degree of variances, both the classical and modified O'Brien tests demonstrate robustness across varying sample sizes, encompassing small, moderate, and large samples.

In the case of one-way ANOVA, the classical approach consistently exhibits robust characteristics regardless of different levels of sample sizes. In contrast, the modified approach has exhibited liberal outcomes in the context of small sample size.

As for the Tukey's test, the classical approach exhibits robust characteristics only when dealing with moderate sample size, while demonstrating conservative results for the remaining level of sample sizes. In contrast, the modified Tukey's test is robust regardless of sample sizes.

Table 2: Type I error values for classical and modified tests under normal distribution

		Classical			Modified		
Degree of Variances	Sample Size	O'Brien	One-way ANOVA	Tukey	2	5	6
1:1:1	N_1	0.0359	0.0501	0.0200	0.0277	0.0976	0.0415
	N_2	0.0390	0.0499	0.0283	0.0339	0.0741	0.0331
	N_3	0.0476	0.0513	0.0181	0.0436	0.0612	0.0257

Chi-Square Distribution Simulation

Based on the findings presented in Table 2, under moderately unequal degree of variance, both the classical and modified O'Brien tests exhibit liberal results across an array of sample sizes. In the context of one-way ANOVA, the classical approach consistently exhibits liberal characteristics, regardless of different levels of sample sizes. In contrast, the modified approach has achieved robust outcomes across the entire range of sample sizes.

As for the Tukey's test, the classical approach displays conservative characteristics, while the modified Tukey's test shows robustness, regardless of sample size considerations.

Table 3: Type I error values for classical and modified tests under chi-square distribution

		Classical			Modified		
Degree of Variances	Sample Size	O'Brien	One-way ANOVA	Tukey	2	5	6
1:1:4	N_1	0.1709	0.2174	0.0036	0.1008	0.0391	0.0245
	N_2	0.3572	0.5798	0.0023	0.2388	0.0741	0.0280
	N_3	0.6489	0.9052	0.0019	0.5447	0.0461	0.0260

Lognormal Distribution Simulation

Based on the findings presented in Table 3, under moderately unequal degree of variance, the classical O'Brien tests show inconsistency in outcomes across varying sample sizes, encompassing small, moderate, and large samples. Robustness is only achieved when dealing with moderate sample sizes. Contrarily, the modified O'Brien approach appears to exhibit consistent robustness regardless of any sample size.

Following the one-way ANOVA, the classical approach consistently exhibits liberal outcomes regardless of different levels of sample sizes. In contrast, the modified approach has exhibited robust characteristics when dealing with moderate and large sample sizes.

As for the Tukey's test, classical approaches exhibit inconsistent results when dealing with all sample sizes, with robustness characteristics that only be depicted when dealing with moderate sample sizes. Conversely, the modified approach has shown a great consistency in achieving robustness across an array of sample sizes.

Table 4: Type I error values for classical and modified tests under lognormal distribution

Degree of Variances	Sample Size	Classical			Modified		
		O'Brien	One-way ANOVA	Tukey	2	5	6
1:1:4	N_1	0.1187	0.1543	0.1187	0.0378	0.0817	0.0395
	N_2	0.0725	0.2059	0.0725	0.0290	0.0699	0.0293
	N_3	0.0799	0.2917	0.0799	0.0496	0.0567	0.0496

Power of Test

The evaluation of the scale of power in the test can be depicted through Sharma and Kibris (2013) assertion that 70% power is desirable, as it is acceptable power. This benchmark serves as a criterion for gauging the adequacy of the test's ability to detect significant effects. The simulation results indicated that both the original and modified approaches have equally produced acceptable power when underlying varied conditions. Despite this similarity, it is noteworthy that the modified approach outperformed the original approach by consistently exhibiting higher power, particularly in scenarios of moderately unequal variances, the modified O'Brien test achieved a power of 0.9369 compared to Tukey's test at 0.9295 under the chi-square distribution. This implies that the O'Brien test with Winsorized mean has enhanced their overall statistical power, which is crucial for addressing the challenges inherent in one-way ANOVA and subsequent Tukey's test.

The findings highlight the efficacy of the modified O'Brien tests in maintaining reliable statistical power, reinforcing their robustness in post-hoc multiple comparisons. The enhanced power values observed in the simulation study have significant implications for environmental science. For instance, analyzing the penguin dataset from the study by Dr. Kristen Gorman, Tony Williams, and William Fraser (2014) facilitates meaningful comparisons of means across different geographical locations, each characterized by unique ecological conditions and sample sizes. This increased statistical power can lead to more robust conclusions about differences in body mass among penguin species, allowing for better understanding of how environmental factors like habitat availability and food sources impact these species. Ultimately, this knowledge can inform policy-making and resource allocation, significantly contributing to conservation efforts for these Antarctic inhabitants based on solid statistical evidence.

CONCLUSION

The notion of this study has delved into a comprehensive insight into the performance of both the classical and modified O'Brien tests under the nuance of violating one-way ANOVA assumptions across diverse datasets, paving the path for a deeper understanding of the statistical robustness in one-way ANOVA and post-hoc multiple comparisons, particularly with respect to Tukey's test. The relevance of the modification to the O'Brien test by adopting a robust central tendency, namely the Winsorized mean, is to develop a valuable tool for Tukey's test in determining more reliable results, given its heavy reliance and dependability on the fulfilment of the three ANOVA assumptions which are normality, independence, and homogeneity of variances.

The evaluation of performance conducted sequentially starting from the O'Brien test and progressing through one-way ANOVA, and concluding with Tukey's test for both classical and modified approaches, involves assessing Type I error, illuminating behaviours exhibited across different types of distribution, degrees of variance heterogeneity, and sample sizes.

The sequential analysis obtained through a simulation study, leads to the conclusion such that in regard to normal distribution, both the classical and modified O'Brien test with Winsorized mean have performed significantly well in producing Type I error under conditions of equal degree of variances (1:1:1). However, the O'Brien test with Winsorized mean is more preferable since the achieved robustness value is slightly higher than the classical method. This preference extends to the modified one-way ANOVA followed by Tukey's test.

Meanwhile, under the non-normal distribution (chi-square and lognormal), the classical O'Brien test has performed well in producing Type I error under circumstances of equal variances. However, the O'Brien test with Winsorized mean has exhibited exceptional robustness, particularly when dealing with moderately unequal variances (1:1:4). The modified one-way ANOVA followed by Tukey's test is preferred under both distributions, as it produces greater robustness values.

REFERENCES

Ali, N. M., Kamal, M., Rahmin, N. A. A. (2022). Comparing O'Brien Test Using Mean, Median, Symmetric and Asymmetric Trimmed Mean. *Menemui Matematik (Discovering Mathematics)*, **44(1)**: 53-59.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, **160(901)**: 268-282.

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

Celik, N. (2022). Robust Post-Hoc Multiple Comparisons: Skew t Distributed Error Terms. *Revista Colombiana de Estadística*, **45(2)**: 363-372.

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*, **11(1)**: 47-52.

Demarest, M., & Kim, J. (2022). Investigating the Factors Affecting Video Conference Fatigue among Students Attending Online Classes. In Proceedings of the 5th European International Conference on Industrial Engineering and Operations Management (pp. 1-8). Rome, Italy: IEOM Society International.

Gorman, K. B., Williams, T. D., & Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (Genus Pygoscelis). *PLoS ONE,* **9(3)**: e90081. https://doi.org/10.1371/journal.pone.0090081

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, *74*(368), 877-880.

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, **89(3)**: 570.

Petrinovich, L. F., & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin*, **71(1)**: 43.

Rodger, R. S., & Roberts, M. (2013). Comparison of power for multiple comparison procedures. *Journal of Methods and Measurement in the Social Sciences*, **4(1)**: 20-47.

Sharma, D., & Kibria, B. G. (2013). On some test statistics for testing homogeneity of variances: a comparative study. *Journal of Statistical Computation and Simulation*, **83(10)**: 1944-1963.

Sawyer, S. F. (2009). Analysis of variance: the fundamental concepts. *Journal of Manual & Manipulative Therapy*, **17(2)**: 27E-38E.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99-114.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, **32(7)**: 771-780.

Wilcox, R. R. (1997). A bootstrap modification of the Alexander-Govern ANOVA method, plus comments on comparing trimmed means. *Educational and Psychological Measurement*, **57(4)**: 655-665.

Wright, D. C. S. (2009). Testing Homogeneity of Variance. *Available at SSRN 2953900.*

Zuo, Y. (2023). Non-asymptotic analysis and inference for an outlyingness induced winsorized mean. *Statistical Papers*, **64(5)**: 1465-1481.