IJPHCS
Open Access: e-Journal

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

THE BEST WAY TO CHOOSING THE MULTIPLE COMPARISON TESTING FOR EQUAL VARIANCE AND UNEQUAL SAMPLE SIZE IN ONE WAY ANOVA

Nawi M.A.A[1*], Ahmad W.M.A.W[1], Rohim R.A.A[1],

[1]Universiti Sains Malaysia, School of Dental Sciences, 16150 Kubang Kerian Kelantan, Malaysia.

**Corresponding author:*
Mohamad Arif Awang Nawi
Universiti Sains Malaysia, School of Dental Sciences, 16150 Kubang Kerian Kelantan and
Mohamadarif@usm.my

ABSTRACT

Background: One-way ANOVA is a method for comparisons of three or more groups of continuous data. Multiple comparison analysis (MCA) is to identify significant differences between subgroups of study. In this paper were considered the comparison and choose the best multiple comparison testing for equal variance and unequal sample size in one-way ANOVA.

Materials and Methods: The most commonly used multiple comparison analysis is the Tukey, Scheffee, Bonferroni and Dunnett T analysis method. From the results, Tukey HSD, Scheffe, Bonferroni and Dunnet T procedure, showed a significant difference in BMI between definite & normotensive of blood pressure and between definite & borderline of blood pressure.

Result: The width 95% Confidence Interval of Scheffee procedure is higher than Bonferroni, Tukey HSD, and Dunnett T procedure

Conclusion: MCA tests may be necessary and researchers should think carefully about the many tests that should be used. This is because different tests can lead to different conclusions and careful consideration for appropriate testing should be given in each case.

Keywords: One-way ANOVA, Multiple comparison analysis (MCA), Equal variance, Unequal sample size and Width of 95% Confidence Interval.

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

1.0 Introduction

ANOVA is commonly used in public health, clinical research, quality control, and social sciences. It is used for the comparison of three or more groups of continuous data when the variances are homogeneous and the data are independent and normally distributed. Once an Analysis of Variance (ANOVA) test is completed, researchers may want to identify significant differences between subgroups. Subgroup differences are called "pairwise" differences. ANOVA did not provide a paired difference test. The output of the ANOVA does not provide any differences in the analysis, so how can researchers investigate the differences between the subgroups tested with the ANOVA? That is, researchers are more likely to report significant differences between some of the pairs without significant differences [1]. Performing multiple pairwise t-tests leads to other problems such as performing multiple t-tests that will lead the researcher to a higher probability of making a Type I error. Researchers may want to test the differences between one or more study groups and one set of combined studies. A paired t-test cannot do such an analysis. However, there is a multivariate set of statistics that overcomes all the limitations of the pairwise t-tests approach. This category of statistics is called multiple comparison analysis (MCA) [2].

Therefore, the choice of MCA statistics should be based on specific research questions. For example, researchers might have one group of experiments of particular interest that should be compared separately to each control group. Alternatively, researchers might want to compare one experimental group to a combination of all control groups, or only a few control groups, or even to one or more experimental groups. Many different situations occur in research that may affect the choice of multiple comparison tests [3]. For example, the groups may have unequal sample sizes. Multiple comparison analysis tests were developed specifically to deal with unequal groups. Power may be a problem in research, and some tests have greater power than others. Testing all comparisons will be important in some studies while other studies will require testing of only a combination of predefined or control groups. When special circumstances affect the intelligence analysis of a particular pair, the choice of multiple comparison analysis tests must be controlled by the specific statistical ability to address the questions of importance and type of data to be analyzed.

Ideally, an ANOVA is performed only when the assumption of homogeneity of variance holds. However, because it is a robust statistic that can be used when there is a deviation from this assumption. Based on Mital et al. [4], they stated that multiple comparison test with unequal sample size and equal variance assumed include Tukey, Scheffe, Bonferroni, Dunnett, Fisher, Sidak, Hochberg GT2 and Gabrial (Figure 1). Each of the multiple comparison analysis (MCA) tests has its particular strengths and limitations. Some will automatically test all of the pairwise comparisons, others allow the researcher to limit the tests to only pairs or subgroups of interest. Each approach has implications for alpha inflation and for the kind of answers the researcher can derive from the test.

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

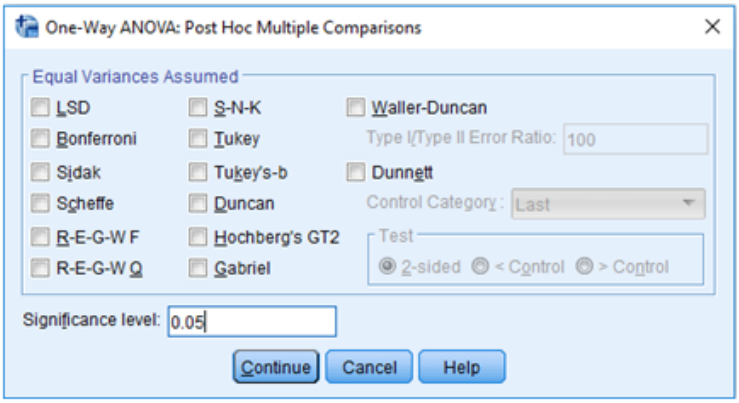


Fig 1 Post Hoc Multiple Comparisons in One-Way ANOVA for Equal Variances Assumed

When the design involves equal variances not assumed, there are several post hoc procedures as in Figure 2 below.

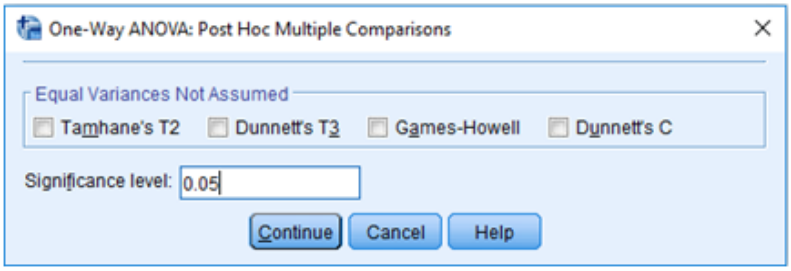


Fig 2 Post Hoc Multiple Comparisons in One-Way ANOVA for Equal Variances Not Assumed

None of the tests are exact tests, but the Tamhane T2, Dunnett T3, and Dunnett C are conservative procedures [5]. In this paper, we have considered the comparison and choose the best multiple comparison testing for equal variance and unequal sample size in one-way ANOVA.

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

2.0 Materials and Methods

The most commonly used multiple comparison analysis is the Tukey, Scheffee, Bonferroni and Dunnett T analysis method.

A. Tukey

The Tukey test is performed with one critical level, as described earlier, and the results of all pairwise comparisons are presented in one table under the section 'post-hoc test.' This test uses a pairwise post-hoc test to determine if there is a difference between the mean of all possible pairs using a studentized range distribution. This method tests each possible pair of all groups. Initially, Tukey's test was called the 'Honestly significant difference' test, or simply the 'T test,' because this method was based on t-distribution [6]. The Tukey test is a generous method to detect the difference during pairwise comparison (less conservative); to avoid this illogical result, an adequate sample size should be guaranteed, which gives rise to smaller standard errors and increases the probability of rejecting the null hypothesis [7].

B. The Scheffee method

The Scheffee method tests all possible contrasts, simple and complex. If it is known in advance that all contrasts are going to be tested, the Scheffee method is slightly more powerful than all other two methods (compare to Tukey and Bonferroni). Thus the Scheffee, like the Tukey test, is the more appropriate test to use when predicted differences are small, and the consequences of a Type II error outweigh the consequences of a Type I error. The Scheffee is a good exploratory statistic because it tests all possible comparisons. As a result, it allows the researcher to observe which groups or combinations of groups produced the significant difference found in the original ANOVA test. Using the Scheffee as a theory testing statistic, the theory is confirmed when differences predicted by the theory are found by Scheffee. When theory predicts no differences between other groups, Scheffee confirms the theory when it finds no significant differences among those groups. The Scheffee test is ideal for testing the well-developed theory because, with minimal alpha inflation, it tests all possible pairwise differences, including combinations of pairs [8].

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

The Scheffee test is also a good tool to use when theory is not sufficiently developed to confidently predict which pairs and combinations of pairs will be significantly different. The overall ANOVA can produce a significant F-test even when two or more groups within the analysis are not significantly different. It is often important to discover exactly which group differences produced the significant F-test. The Scheffee test allows the researcher to conduct a theory generation study by testing all possible contrasts to discover which are significant. Scheffé's method is not a simple pairwise comparison test. Based on F-distribution, it is a method for performing simultaneous, joint pairwise comparisons for all possible pairwise combinations of each group mean [9]. It controls family-wise error rate (FWER) after considering every possible pairwise combination, whereas the Tukey test controls the family-wise error rate (FWER) when only all pairwise comparisons are made [10]. This is why the Scheffé's method is very conservative than other methods and has small power to detect the differences. Finally, Scheffé's method enables simple or complex averaging comparisons in both balanced and unbalanced data.

C. The Bonferroni (Dunn) method

Bonferroni procedure also is a family comparison method. Besides, like the Scheffee procedure, the Bonferroni method can test complex pairs. However, Bonferroni statistics is not a tool for exploratory data analysis. It requires researchers to determine all the differences to be tested first. Researchers must have sufficient theory of the phenomenon of interest to know the differences to be determined. As a result, this is a better test to confirm the theory of experimental results than exploratory methods such as Scheffee. Because Bonferroni limits the number of tests to predetermined by researchers, it reduces the problem of alpha inflation. The major advantage of the Bonferroni method is that it reduces the probability of Type I error by limiting it to alpha inflation. However, it cannot make serendipitous findings and therefore provides less information about differences between groups because not all differences are tested [8].

D. The Dunnett method

Dunnett's method is useful for the design of test control groups. It is a very powerful statistic and therefore it can discover relatively small but significant differences among groups or combinations of groups. Dunnett's method is particularly useful when researchers want to test two or more experimental groups on a single control group. It tests each experimental group's mean against the control group. Another method tests each study group against the total number

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

of groups (i.e., grand mean). This difference in the testing approach makes the Dunnett method much more likely to find a significant difference because the grand mean includes all group means and thus mathematically it is less extreme than individual group means. The more extreme group means will produce larger mean differences than tests comparing one group mean to the grand mean. The Bonferroni method can be determined to test only experimental groups against a single control group, but by comparing them with large-value study groups, it has less power than Dunnett's method [8].

3.0 Result

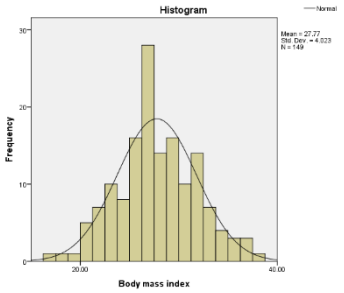
A. RESULTS

A retrospective cohort study based on secondary data of Type 2 diabetic patients who attended outpatient KRK, USM was carried out by Zainab et al, [11] with a sample size of 149 patients. The dependent variable is Body Mass Index among diabetes patients and the independent variable is blood pressure status. Blood pressure status was divided into three groups were normotensive (42 samples), borderline (24 samples) and definite (83 samples). The researcher wants to determine whether there is a significant difference between the body mass index (BMI) toward blood pressure status among type 2 diabetic patients by using IBM SPSS version 24. Thus this example addresses the following research question: Does the mean body mass index (BMI) differ depending on the blood pressure status among types 2 diabetes patients in Hospital Universiti Sains Malaysia?

This can be stated in the form of a null hypothesis: H0 = There is no difference in the body mass index (BMI) across the different blood pressure status among type 2 diabetic patients.

Before doing analysis, the researcher must have checked the Normality by histogram plot and Kolmogorov-Smirnov Test and found that observations are normally distributed.

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020



One-Sample Kolmogorov-Smirnov Test

		Body mass index
N		149
Normal Parameters[a,b]	Mean	27.7706
	Std. Deviation	4.02321
Most Extreme Differences	Absolute	.054
	Positive	.054
	Negative	-.036
Test Statistic		.054
Asymp. Sig. (2-tailed)		.200[c,d]

a. Test distribution is Normal.
b. Calculated from data.
c. Lilliefors Significance Correction.
d. This is a lower bound of the true significance.

Fig 3 Normal Distribution Result

We have also checked the variance Homogeneity by Levene statistics at 5% level of significance and found that all the populations have the same variance.

Test of Homogeneity of Variances

Body mass index

Levene Statistic	df1	df2	Sig.
.173	2	146	.841

Fig 4 Test of Homogeneity of Variences

We have used ANOVA to check whether there is a significant difference between the body mass index (BMI) toward blood pressure status among type 2 diabetic patients.

ANOVA

Body mass index

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	206.121	2	103.060	6.872	.001
Within Groups	2189.438	146	14.996		
Total	2395.558	148			

Fig 5 ANOVA Table of BMI toward Blood Pressure

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

Multiple comparisons (post hoc test) can be applied when there is a significant difference between the body mass index (BMI) toward blood pressure status among type 2 diabetic patients at 5 % level of significance and found that there is at least one significant difference.

Table 1 above, there were four procedures in the result produced such as: 1) Tukey HSD procedure, there is a significant difference in BMI between definite & normotensive of blood pressure and between definite & borderline of blood pressure. 2) By Scheffe procedure, there is a significant difference in BMI between definite & normotensive of blood pressure and between definite & borderline of blood pressure. 3) Bonferroni procedure also gives the same result as Tukey HSD and Scheffe procedure where there is a significant difference in BMI between definite & normotensive of blood pressure and between definite & borderline of blood pressure. 4) By Dunnett T test treat one group as a control and compare all other groups against it. There is a significant difference in BMI between normotensive & definite and borderline & definite. From this analysis, we can see that width 95% Confidence Interval of Scheffe procedure is higher than Bonferroni, Tukey HSD, and Dunnett T procedure.

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

Table 1: Multiple Comparison Test

Test	(I) Hypertension status	(J) Hypertension status	Mean Difference (I-J)	Std. Error	Sig.	Width of 95% CI
Tukey HSD	normotensive	borderline	0.83923	0.9909	0.674	4.6927
		definite	-2.00000*	0.7333	0.02	3.4727
	borderline	normotensive	-0.83923	0.9909	0.674	4.6927
		definite	-2.83923*	0.89751	0.005	4.2503
	definite	normotensive	2.00000*	0.7333	0.02	3.4727
		borderline	2.83923*	0.89751	0.005	4.2503
Scheffe	normotensive	borderline	0.83923	0.9909	0.699	4.9011
		definite	-2.00000*	0.7333	0.027	3.627
	borderline	normotensive	-0.83923	0.9909	0.699	4.9011
		definite	-2.83923*	0.89751	0.008	4.4392
	definite	normotensive	2.00000*	0.7333	0.027	3.627
		borderline	2.83923*	0.89751	0.008	4.4392
Bonferroni	normotensive	borderline	0.83923	0.9909	1	4.7997
		definite	-2.00000*	0.7333	0.021	3.552
	borderline	normotensive	-0.83923	0.9909	1	4.7997
		definite	-2.83923*	0.89751	0.006	4.3473
	definite	normotensive	2.00000*	0.7333	0.021	3.552
		borderline	2.83923*	0.89751	0.006	4.3473
Dunnett t	normotensive	definite	-2.00000*	0.7333	0.014	3.303
	borderline	definite	-2.83923*	0.89751	0.004	4.0427

** The mean difference is significant at the 0.05 level.*

Dependent Variable: Body mass index

4.0 Discussion

Based on the result shown in Table 1, all procedure gives the same result such as mean difference, standard error and significant value (p-value). The difference between all procedures is the width of 95% confidence interval. Neeraj Hirpara et al. [12] stated that precision is a measure of consistency and is a function of random error and confidence required. At 95% confidence interval the results are more precise. The width of confidence interval (CI) is associated with sample size. Narrow width of CI means there is a small range of effect size in the study indicates study size is quite large since the range of effect size is narrow & hence the study has reasonable certainty. Wide or diverse range of effect size & hence the estimate is not precise [12]. From this study, Dunnett procedure produce narrow width of CI compare to other procedure. Lee & Lee [8] stated that Dunnett test is a powerful statistic and, therefore, can discover relatively small but significant differences among groups or combinations of groups. A researcher can use Dunnett test in testing two or more experimental groups against a single control group only.

From the study, Chen et al [7] stated that Tukey method uses the harmonic mean of the cell size of the two comparisons and the statistical assumptions of ANOVA should be applied to the Tukey method, as well. Subsequent studies testing specific subgroup contrasts discovered through the Scheffee method should use the Bonferroni method which is more appropriate for theory testing studies. Bonferroni methods that are appropriate for theoretical test studies. It is further noted that Bonferroni methods are less sensitive to type I errors than Scheffé's method. The Bonferroni method is less susceptible to Type I errors than the Scheffee method. The Bonferroni procedure is more stringent than the Tukey procedure, which tolerates type I errors, and is more generous than the highly conservative Scheffes method

However, the Bonferroni test has its drawbacks, as it does not need to be conservative (with weak statistical power). The adjusted α is often smaller than required, especially if there are many positive correlated tests and/or test statistics. Therefore, this method often fails to detect the true difference. If the proposed study requires a type II error to be avoided and the possible effects cannot be ruled out, we cannot use Bonferroni correction. Instead, we should use more liberal methods such as Fisher's LSD, which does not control the family-wise error rate (FWER) [9]. Another alternative for Bonferroni's correction to produce too conservative results is to use a stepwise method, whereby Bonferroni-Holm and Hochberg fit, which is less conservative than the Bonferroni test [13].

In other words, Bonferroni's test is applied as a post-hoc test in many statistical procedures such as ANOVA and its variants, including analysis of covariance (ANCOVA) and multivariate ANOVA (MANOVA); multiple t-tests; and Pearson's correlation analysis. It is also used in several nonparametric tests, including the Mann-Whitney U test, Wilcoxon signed rank test, and Kruskal-Wallis test by ranks [6] and as a test for categorical data, such as Chi-squared test.

International Journal of Public Health and Clinical Sciences
e-ISSN : 2289-7577. Vol. 7:No. 4
July/August 2020

5.0 Conclusion and recommendation

There are various post hoc tests available to explain the group differences that contributed to the significance of the ANOVA test. Each test has a specific application, advantages, and disadvantages. Therefore, it is important to select the test that best fits the data, the type of information about the group comparison, and the strength of the analysis needed. It is also important to choose a test that fits the state of the theory in terms of the theory test. The consequences of poor test selection are typically related to Type 1 errors, but may also involve failure to discover important differences among groups. Multiple comparison analysis tests are important because while the ANOVA provides a lot of information, it does not provide detailed information on differences between specific study groups, and cannot provide information on complex comparisons. Secondary analyzes with these post hoc tests can provide researchers with the most important findings of the study. In general, most of the pairwise MCTs are based on balanced data. Therefore, when there is a significant difference in the number of samples, care must be taken when selecting various comparison procedures. Tukey, Scheffe, Bonferroni, and Dunnett using t-statistics do not cause problems, as there is no assumption that the sample numbers in each group are the same. The Tukey test, which uses the harmonic mean of sample numbers, can be used when the sample numbers are different.

If the data were analyzed using ANOVA, and significant F values were obtained, a more detailed analysis of the differences between treatment methods would be needed. The best option is to design a specific comparison between the treatment means before the experiment is carried out and test them using ‘contrasts’. In some cases, post-hoc tests may be necessary and researchers should think carefully about the many tests that should be used. Different tests can lead to different conclusions and careful consideration for appropriate testing should be given in each case. Confidence intervals such as p-values guide to help interpret research findings in the light effect of chance. The findings may not apply to other groups of patients (external validity). An assessment of this external validity should be made. Neither confidence interval nor p-value is of much help for this judgement.

Acknowledgement

The authors would like to express their gratitude to Universiti Sains Malaysia (USM) for providing the research funding (Short Term Grant No.304/PPSG/6315410, School of Dental Sciences, Health Campus, Kelantan, Malaysia).

Declaration

Author(s) declare that there is no conflict of interest with the publication of this article.

Author's contribution

Author 1: initiation of idea, final manuscript review and editing

Author 2: manuscript review, literature searching, drafting the manuscript and editing

Author 3: manuscript review and editing

References

[1] Lakovac V (2009). Statistical hypothesis testing and some pitfalls. Biochem Med; 19:10-6.

[2] Mary L. (2011).McHugh.Multiple comparison analysis testing in ANOVA. Biochemia Medica; 21(3):203–9.

[3] Marusteri M, Bacarea V. (2010). Comparing groups for statistical differences: How to choose the right statistical test. Biochem Med; 20:15-32.

[4] Mital C. Shingala & Arti Rajyaguru (2015). Comparison of Post Hoc Tests for Unequal Variance. International Journal of New Technologies in Science and Engineering Vol. 2, Issue 5, ISSN 2349-078.

[5] De Muth, J. (2006). Basic Statistics and Pharmaceutical Statistical Applications. Second Edition.CRC Press.2nd edition.

[6] Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol; 57:289–300.

[7] Chen SY, Feng Z, Yi X. (2017). A general introduction to adjustment for multiple comparisons. J Thorac Dis; 9: 1725-9.

[8] Lee, S. & Lee, D. K. (2018) What is the proper way to apply the multiple comparison test? Korean J Anesthesiol. 71(5): 353-360.

[9] Keselman HJ, Keselman JC, Games PA. (1991). Maximum familywise Type I error rate: The least significant difference, Newman-Keuls, and other multiple comparison procedures. Psychol Bull; 110:155-62.

[10] Cue RI. Multiple Comparisons. Department of Animal Science, McGill University, 2003.

[11] Zainab MY, Lili Husniati Y, Norul Badriah H, Saiful Bahari I, Nani D, Siti Suhaila MY.. (2016). Achievement of LDL cholesterol goal and adherence to statin by diabetes patients in Kelantan. Malays J Med Sci; 24(3):44– 50.

[12] Neeraj Hirpara, Sandesh Jain, Alpana Gupta, Soumya Dubey intern (2015). Interpreting Research Findings with Confidence Interval. Journal of Orthodontics & Endodontics. Vol. 1 No. 1:8.

[13] McHugh MM. (2008)Standard error: meaning and interpretation. Biochem Med; 18:7-13.