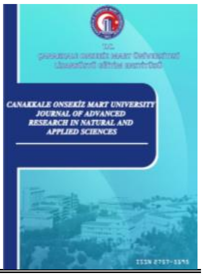
Çanakkale Onsekiz Mart University

Journal of Advanced Research in Natural and Applied Sciences

dergipark.org.tr/tr/pub/jarnas

Open Access



e-ISSN: 2757-5195 | 2024, Vol. 10, Issue 3, Pages: 493-505 | doi.org/10.28979/jarnas.1429315

A Comprehensive Monte Carlo Simulation Study on Multiple Comparison Methods after ANOVA

Hanife Avcı[1], Osman Dağ[2,*]

[1,2]Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Türkiye

Article Info
Received: 31 Jan 2024
Accepted: 05 May 2024
Published: 30 Sep 2024
Research Article

Abstract − Multiple comparison methods are applied to control the type I error rate at the nominal level. In this study, we investigate the performance of multiple comparison methods after analysis of variance (ANOVA) is implemented under different conditions. We include Bonferroni, Holm, Hochberg, Hommel, Benjamini-Hochberg (BH), and Benjamini-Yekutieli (BY) correction methods. Monte-Carlo simulation study is applied to assess their performances under different patterns, including sample size and group number combinations. Wide inferences are drawn on considered methods, and suggestions are provided for selecting appropriate methods. Moreover, the methods are implemented on three different types of real-life data sets to emphasize the importance of these correction methods in the research.

Keywords *– ANOVA, Monte Carlo simulation, pairwise comparison, type I error rate, variance homogeneity*

1. Introduction

Testing equality of means is a general process in many data applications. In this study, we compare multiple comparison methods after ANOVA concerning controlling the type I error rate (α) at its nominal level. There exist many multiple comparison tests in the literature [1]. We include Bonferroni [2], Holm [3], Hommel [4], Hochberg [5], Benjamini-Hochberg (BH) [6], and Benjamini-Yekutieli (BY) [7] comparison methods in this study. These methods adjust the estimated type I error rate using the number of hypotheses tested. The number of hypotheses is specified by the number of groups compared in ANOVA. For instance, the total number of hypotheses becomes $k(k-1)/2$ when the number of groups compared within ANOVA is k. Each pair in the comparison for testing the hypotheses is called a family. The type I error rate is called the family-wise error rate (FWER) since each family is compared separately [8].

Multiple comparison methods are used in various fields. Bender and Lange [9] discussed adjustment methods for multiple comparisons in medical and epidemiological studies. Westphal and Troendle [10] proposed multiple comparison methods based on resampling that control FWER. They noted that the methods could be applied specifically to gene expression data but more generally to multivariate and multigroup data. Blakesley et al. [11] proposed the Hommel and Hochberg methods for mildly correlated measures to increase power while controlling type I errors in neuropsychological data sets. Felix and Menezes [12] showed that BH correction provided the best type I error rate and the second strongest correction by varying the sample size, sample distribution, and degree of variability. Staffa and Zurakowski [13] suggested that surgeons use multiple comparison methods, including Bonferroni, Tukey, Scheffe, Holm, and Dunnett while planning clinical or

[1]hanife.avci@hacettepe.edu.tr; [2]osman.dag@hacettepe.edu.tr (Corresponding Author)

research studies. Their study discussed the results of each approach for controlling FWER on pediatric surgical research data.

In a study conducted by Dimitriev et al. [14], the Bonferroni method was used to investigate the effects of stress on heart rate and heart rate variability (HRV) measurements. Participants were divided into three groups according to baseline HRV (<25th percentile, 25th -75th percentile, and >75th percentile). Stress-related changes were compared in HRV groups. Kharola et al. [15] criticized the work done by Dimitriev et al. [14] concerning using multiple comparison methods. They stated that the Holm method was more powerful than the Bonferroni method. In our simulation study, this design fits the one extreme pattern with three groups and a sample size between 50 and 100. The findings state that the Holm method has a type I error rate larger than its nominal level, while Bonferroni holds the nominal level. That is, the Holm method may find a difference when the difference is not real.

Musicus et al. [16] assessed the effects of text-based alerts, health picture alerts, sugar picture alerts, and control on parents' beverage choices for their children in a virtual convenience store. They used the Holm method for multiple comparisons after ANOVA. In our simulation study, this study design approximately fits the equal N pattern with four groups and a large sample size. Our findings reveal that the Holm method has a type I error rate larger than its nominal level, while Bonferroni holds the nominal level. In other words, the Holm technique might detect a difference when no one exists.

Generally, the researchers are unaware of choosing an appropriate method when performing multiple comparisons. However, this situation leads to misleading serious consequences. Therefore, the researchers should consider the number of groups, the sample size, and the pattern of sample sizes to reduce the risk of misleading consequences while conducting multiple comparison methods. No study compares multiple comparison tests' type I error rates in different scenarios, such as sample size, number of groups, and other sample size designs. Therefore, this study is essential to give recommendations for the applied researchers on selecting appropriate multiple comparison methods under various scenarios.

This study compares multiple comparison methods' type I error rates at different sample sizes, sample size patterns (under equal N, progressive N, and one extreme scenario), and the number of groups after ANOVA. Therefore, a comprehensive Monte-Carlo simulation study is conducted to compare multiple comparison tests' type I error rates.

This study has some limitations since ANOVA is conducted under certain assumptions. These assumptions of ANOVA are independence of observations within and between groups, normality (i.e., k samples are drawn randomly from a normal distribution), and variance homogeneity (i.e., k populations have identical variances).

The sections of the paper are organized as follows: Section 2 introduces the multiple comparison methods in chronological order. Section 3 provides the Monte-Carlo simulation study and its results. Section 4 presented the implementation of the pairwise comparison methods following ANOVA in R. Finally, the paper concludes with a summary of the main findings and a discussion of the results of similar studies.

2. Materials and Methods

This section includes six multiple comparison methods commonly used in the literature.

2.1. Bonferroni Correction

Bonferroni correction is one of the most widely used approaches for multiple comparisons [2]. The Bonferroni multiple comparison test is conservative when the large sample size and the number of pairwise comparisons increases [3]. With a pre-specified significance level (α) and the number of hypotheses tested (m), the adjusted significance level is calculated as follows:

$$\alpha'_{(i)} = \frac{\alpha}{m} \tag{2.1}$$

2.2. Holm Correction

The less conservative Holm method is proposed after the Bonferroni correction [3]. Holm correction is more powerful than Bonferroni correction [17]. Holm method controls the false discovery rate (FDR). In the Holm method, the p-value for each hypothesis is arranged from least to greatest. For the i^{th} ordered hypotheses $H_{(i)}$, the formula is given in (2.2) for the type I error rate:

$$\alpha'_{(i)} = \frac{\alpha}{m-i+1} \tag{2.2}$$

In this part, we include an example to clarify Holm's correction. For instance, we have three groups to compare with ANOVA and obtain statistically significant results. After that, we apply multiple comparison methods to adjust estimated type I error rates. We obtain p-values of 0.002, 0.010, and 0.035 (given in order from smallest to largest) for hypotheses I-III, respectively. For the smallest p-value (0.002), we obtain the nominal type I error rate for the comparison as follows: $\alpha'_{(i)} = \frac{\alpha}{m-i+1} = \frac{0.05}{3-1+1} \cong 0.0167$. We compare the smallest p-value (0.002) with its nominal rate (0.0167). The hypotheses are rejected since the p-value of 0.002 is smaller than 0.0167. The following steps are similar to the other hypotheses and α''s are estimated as 0.0167, 0.025, and 0.05, respectively.

2.3. Hochberg Correction

Like the Holm method, Hochberg correction uses the same formula to calculate the associated significance levels [5]. Hochberg's multiple comparison method, controlling FDR, is more powerful than the Holm method [18]. Both multiple comparison methods compare ordinal p-values with the same set of critical values. The algorithm of the method is the same as within the Holm method, except that the p-values are ordered from largest to smallest [18]. Thus, the Hochberg method rejects a hypothesis that results in all hypotheses being rejected (2.2).

2.4. Hommel Correction

Simes revised the Bonferroni method and proposed a new multiple comparison method by combining all m hypotheses [19]. However, Hommel extended this new method to test each hypothesis (Hommel 1988) since Simes method cannot be used to evaluate each hypothesis alone. The decisions for the individual hypotheses can be performed in the following simpler way: $j = \max\left\{i \in \{1,2,\dots,m\} : p_{(m-i+k)} > \frac{k\alpha}{i} \text{ for } k \in \{1,2,\dots,i\}\right\}$. If the maximum value does not exist, reject all H_i ($i \in \{1,2,\dots,m\}$), otherwise, reject all H_i with $p_i \leq \alpha/j$. It is not easy to calculate adjusted p-values with the Hommel method. It can be easily calculated with the pair comp function in the oneway tests *R* package.

2.5. Benjamini-Hochberg (BH) Correction

Benjamini and Hochberg [6] developed a method called BH (known as false discovery rate, FDR) to control FDR. It is often preferred when the number of hypotheses is large. BH multiple comparison method is preferred since it is simpler than other methods. First, m hypotheses are arranged in order from largest to smallest according to their p-values ($i \in \{1,2,\dots,m\}$). The q value is the upper bound of FDR (e.g., q = 0.05). BH critical value of each p-value is calculated using the $\frac{i}{m}q$ formula. p_i is the p-value related to the H_i hypotheses, and k is the largest I:

$$k = \max\left\{p_i \leq \frac{i}{m}q\right\} \tag{2.4}$$

The p values are compared with the critical BH value. The largest p-value that is less than the critical BH value is found. All p-values under this largest value are considered statistically significant.

2.6. Benjamini and Yekutieli (BY) Correction

Benjamini and Yekutieli proposed a more conservative multiple comparison method than BH to control FDR [7]. BY method is similar to the BH method. However, the way of finding k is different from the BH method.

$$k = \max\left\{i: p_{(i)} \leq \frac{i}{m}\tilde{q}, \tilde{q} = \frac{q}{\sum_{i=1}^{m}\frac{1}{i}}\right\} \tag{2.5}$$

where the q value is an upper bound of FDR. Unlike other multiple comparison methods, the BY correction method considers the dependency of hypotheses. The difference between BH and BY is the dependency structure. Benjamini and Yekutieli proved that the estimator is valid under some forms of the dependency structure.

3. Results

This study compares the type-I error rates of six correction methods with the Monte Carlo simulation study. Moreover, add the results obtained without correction to emphasize the importance of multiple comparison methods.

3.1. Simulation Design

Monte Carlo simulation is applied to show the performances of these tests for scenarios where normality and homogeneity of variance assumptions are met. The algorithm of the simulation study is planned as follows.

i. Random samples are generated from a normal distribution with a mean of 0 and a standard deviation of 1 as much as the number of groups.

ii. The number of groups is set to 3 - 4 (small), 5 - 6 (medium), and 7 - 10 (large).

iii. The sample sizes for each group are 10, 20, 30, 40, 50, and 100.

iv. The sample size patterns are specified as equal N, progressive N, and one extreme scenario is given in Table 1.

All steps are repeated 10,000 runs, and type I errors are calculated. After calculating the type I error for each pairwise comparison, we combine the type I error rates as given in Equation 6.

$$\hat{\alpha}_n = 1 - \prod_{i=1}^{h}\left(1 - \hat{\alpha}_{i,n}\right) \tag{3.1}$$

where $h = \binom{k}{2}$ and k is the number of groups.

3.2. Simulation Results

This study compares the performances of multiple comparison tests concerning sample size designs of equal, progressive, and one-extreme cases. The results are presented in Tables 2-4. We outline the results according to the number of groups for selecting the appropriate methods for practical use.

When k is small: The type I error of the Holm method is closest to the nominal level when the sample size is equal and smaller than 40 under the equal N scenario. For instance, the type I error rates of the Holm method vary between 0.049-0.051 when k = 3. In the same scenario, the type I error rates of the Bonferroni method range from 0.045 to 0.048, while those of the Hommel method range from 0.051 to 0.054. On the other hand, the Bonferroni method holds the nominal level of type I error when the number of observations exceeds 40. For example, the type I error rates of the Bonferroni method vary between 0.049-0.050 when k = 3. Under progressive N and one extreme scenario, the Bonferroni method holds the nominal level of type I error rate in all observation numbers.

When k is medium: Results obtained under equal N, progressive N, and one extreme scenario show similar patterns. Type I error of the Hommel method holds the nominal level when the number of observations is equal and lower than 30. The type I error rate of the Bonferroni method is very close to the nominal level when the number of observations exceeds 30.

When k is large**:** Under equal N, progressive N, and one extreme scenario, Bonferroni, Holm, Hommel, and Hochberg methods hold the nominal level of type I error rate regardless of the number of observations.

When the multiple comparison methods are not applied, the type I error rates increase excessively as the number of groups and comparisons increase. For example, when the number of groups is 9, and the number of observations in each group is 10 under the equal N scenario, multiple comparison methods give results ranging from 0.018 to 0.078. In the same scenario, the type I error reaches 0.834 without any corrections.

Table 1. Simulation study the sample size patterns

	Progressive N	**Equal N**	**One extreme**
k=3	8	10	8
	10	10	8
	12	10	14
Average N	10	10	10
k=4	7	10	8
	9	10	8
	11	10	8
	13	10	16
Average N	10	10	10
k=5	6	10	8
	8	10	8
	10	10	8
	12	10	8
	14	10	18
Average N	10	10	10
k=6	5	10	8
	7	10	8
	9	10	8
	11	10	8
	13	10	8
	15	10	20
Average N	10	10	10
k=7	7	10	8
	8	10	8
	9	10	8
	10	10	8
	11	10	8
	12	10	8
	13	10	22
Average N	10	10	10
k=8	6	10	8
	7	10	8
	8	10	8
	9	10	8
	11	10	8
	12	10	8
	13	10	8
	14	10	24
Average N	10	10	10
k=9	6	10	8
	7	10	8
	8	10	8
	9	10	8
	10	10	8
	11	10	8
	12	10	8
	13	10	8
	14	10	26
Average N	10	10	10

Table 1. (Continued) Simulation study sample size patterns

	Progressive N	Equal N	One extreme
k=10	5	10	8
	6	10	8
	7	10	8
	8	10	8
	9	10	8
	11	10	8
	12	10	8
	13	10	8
	14	10	8
	15	10	28
Average N	10	10	10

Table 2. Type I error for equal

Group	Method	$\hat{\alpha}_{10}$	$\hat{\alpha}_{20}$	$\hat{\alpha}_{30}$	$\hat{\alpha}_{40}$	$\hat{\alpha}_{50}$	$\hat{\alpha}_{100}$
3	Bonferroni	0.048	0.045	0.047	0.047	**0.049**	**0.050**
	Holm	**0.051**	**0.049**	**0.051**	**0.050**	0.053	0.055
	Hommel	0.054	0.051	0.054	0.052	0.057	0.057
	Hochberg	0.053	0.050	0.052	0.051	0.055	0.055
	BH	0.060	0.058	0.059	0.057	0.062	0.065
	BY	0.031	0.030	0.031	0.031	0.030	0.035
	None	0.141	0.140	0.140	0.135	0.142	0.143
4	Bonferroni	0.046	0.047	0.048	0.050	**0.051**	**0.054**
	Holm	**0.048**	**0.049**	**0.051**	**0.053**	0.054	0.057
	Hommel	0.051	0.051	0.053	0.054	0.056	0.059
	Hochberg	0.048	0.050	0.051	0.053	0.055	0.057
	BH	0.064	0.065	0.068	0.069	0.070	0.075
	BY	0.024	0.025	0.024	0.026	0.028	0.030
	None	0.261	0.263	0.266	0.256	0.267	0.274
5	Bonferroni	0.047	0.042	0.047	**0.051**	**0.051**	**0.054**
	Holm	0.049	0.043	0.048	0.052	0.053	0.056
	Hommel	**0.050**	**0.045**	**0.050**	0.054	0.055	0.057
	Hochberg	0.049	0.044	0.049	0.053	0.054	0.056
	BH	0.069	0.066	0.074	0.076	0.076	0.080
	BY	0.020	0.017	0.020	0.024	0.024	0.027
	None	0.388	0.393	0.401	0.392	0.397	0.408
6	Bonferroni	0.045	0.044	0.047	**0.051**	**0.053**	**0.052**
	Holm	0.046	0.045	0.048	0.053	0.054	0.054
	Hommel	**0.047**	**0.046**	**0.049**	0.055	0.055	0.055
	Hochberg	0.046	0.045	0.048	0.053	0.054	0.054
	BH	0.071	0.071	0.077	0.089	0.086	0.088
	BY	0.019	0.017	0.018	0.024	0.025	0.024
	None	0.529	0.531	0.541	0.532	0.536	0.548
7	Bonferroni	**0.045**	**0.044**	**0.046**	**0.051**	**0.051**	**0.050**
	Holm	**0.046**	**0.046**	**0.047**	**0.053**	**0.052**	**0.051**
	Hommel	**0.047**	**0.047**	**0.047**	**0.053**	**0.054**	**0.052**
	Hochberg	**0.046**	**0.046**	**0.047**	**0.053**	**0.052**	**0.051**
	BH	0.073	0.079	0.079	0.087	0.089	0.086
	BY	0.016	0.017	0.018	0.023	0.023	0.020
	None	0.646	0.651	0.663	0.656	0.652	0.664
8	Bonferroni	**0.043**	**0.044**	**0.049**	**0.053**	**0.048**	**0.051**
	Holm	**0.044**	**0.044**	**0.049**	**0.054**	**0.050**	**0.051**
	Hommel	**0.045**	**0.045**	**0.050**	**0.054**	**0.050**	**0.052**
	Hochberg	**0.044**	**0.044**	**0.049**	**0.054**	**0.050**	**0.051**
	BH	0.077	0.079	0.086	0.090	0.087	0.087
	BY	0.016	0.016	0.017	0.019	0.019	0.019
	None	0.756	0.758	0.768	0.760	0.760	0.770
9	Bonferroni	**0.043**	**0.048**	**0.049**	**0.054**	**0.050**	**0.052**
	Holm	**0.044**	**0.048**	**0.050**	**0.056**	**0.051**	**0.053**
	Hommel	**0.044**	**0.049**	**0.050**	**0.056**	**0.051**	**0.053**
	Hochberg	**0.044**	**0.048**	**0.050**	**0.056**	**0.051**	**0.053**
	BH	0.078	0.085	0.091	0.096	0.090	0.096
	BY	0.018	0.016	0.016	0.020	0.018	0.019
	None	0.834	0.839	0.844	0.841	0.841	0.847
10	Bonferroni	**0.042**	**0.045**	**0.046**	**0.050**	**0.052**	**0.052**
	Holm	**0.042**	**0.046**	**0.046**	**0.051**	**0.053**	**0.052**
	Hommel	**0.042**	**0.046**	**0.047**	**0.052**	**0.053**	**0.053**
	Hochberg	**0.042**	**0.046**	**0.046**	**0.051**	**0.053**	**0.052**
	BH	0.077	0.084	0.092	0.097	0.102	0.096
	BY	0.013	0.016	0.016	0.016	0.017	0.017
	None	0.893	0.899	0.904	0.900	0.899	0.904

Boldfaced values indicate the closest type I error rates to the nominal ones.

Table 3. Type I error for progressive

Group	Method	$\hat{\alpha}_{10}$	$\hat{\alpha}_{20}$	$\hat{\alpha}_{30}$	$\hat{\alpha}_{40}$	$\hat{\alpha}_{50}$	$\hat{\alpha}_{100}$
3	Bonferroni	**0.049**	**0.047**	**0.049**	**0.048**	**0.052**	**0.049**
	Holm	0.053	0.051	0.053	0.052	0.056	0.054
	Hommel	0.056	0.053	0.056	0.056	0.057	0.057
	Hochberg	0.055	0.052	0.055	0.054	0.056	0.055
	BH	0.061	0.059	0.062	0.061	0.063	0.063
	BY	0.030	0.030	0.035	0.030	0.036	0.034
	None	0.140	0.140	0.143	0.144	0.146	0.150
4	Bonferroni	**0.046**	**0.052**	**0.049**	**0.048**	**0.049**	**0.050**
	Holm	0.050	0.055	0.052	0.051	0.051	0.052
	Hommel	0.052	0.057	0.054	0.052	0.054	0.053
	Hochberg	0.051	0.056	0.052	0.051	0.052	0.052
	BH	0.066	0.069	0.068	0.068	0.068	0.069
	BY	0.024	0.026	0.026	0.027	0.028	0.027
	None	0.267	0.266	0.269	0.266	0.260	0.263
5	Bonferroni	0.045	0.043	0.046	**0.049**	**0.048**	**0.050**
	Holm	0.046	0.045	0.048	0.051	0.049	0.052
	Hommel	**0.048**	**0.046**	**0.050**	0.052	0.050	0.054
	Hochberg	0.047	0.045	0.049	0.051	0.049	0.052
	BH	0.069	0.070	0.071	0.075	0.069	0.079
	BY	0.020	0.018	0.022	0.023	0.022	0.024
	None	0.394	0.389	0.396	0.409	0.391	0.405
6	Bonferroni	0.050	0.047	0.047	**0.052**	**0.052**	**0.047**
	Holm	0.051	0.049	0.049	0.054	0.053	0.048
	Hommel	**0.052**	**0.050**	**0.050**	0.056	0.054	0.050
	Hochberg	0.051	0.049	0.049	0.054	0.053	0.048
	BH	0.083	0.077	0.080	0.090	0.081	0.081
	BY	0.020	0.021	0.020	0.023	0.024	0.019
	None	0.536	0.533	0.544	0.544	0.536	0.547
7	Bonferroni	**0.050**	**0.050**	**0.053**	**0.051**	**0.052**	**0.052**
	Holm	**0.051**	**0.050**	**0.054**	**0.052**	**0.053**	**0.054**
	Hommel	**0.052**	**0.051**	**0.055**	**0.053**	**0.054**	**0.055**
	Hochberg	**0.051**	**0.050**	**0.054**	**0.052**	**0.053**	**0.054**
	BH	0.081	0.087	0.095	0.083	0.087	0.088
	BY	0.020	0.017	0.023	0.018	0.018	0.022
	None	0.656	0.662	0.666	0.668	0.672	0.658
8	Bonferroni	**0.042**	**0.046**	**0.049**	**0.053**	**0.050**	**0.047**
	Holm	**0.043**	**0.047**	**0.050**	**0.054**	**0.051**	**0.048**
	Hommel	**0.043**	**0.048**	**0.050**	**0.055**	**0.052**	**0.049**
	Hochberg	**0.043**	**0.047**	**0.050**	**0.054**	**0.051**	**0.048**
	BH	0.074	0.082	0.088	0.093	0.089	0.086
	BY	0.014	0.020	0.019	0.021	0.017	0.019
	None	0.755	0.757	0.765	0.769	0.769	0.771
9	Bonferroni	**0.049**	**0.048**	**0.052**	**0.054**	**0.052**	**0.052**
	Holm	**0.050**	**0.049**	**0.053**	**0.055**	**0.052**	**0.053**
	Hommel	**0.051**	**0.050**	**0.053**	**0.056**	**0.053**	**0.053**
	Hochberg	**0.050**	**0.049**	**0.053**	**0.055**	**0.053**	**0.053**
	BH	0.089	0.094	0.094	0.097	0.098	0.100
	BY	0.016	0.016	0.020	0.022	0.020	0.020
	None	0.838	0.848	0.846	0.844	0.840	0.845
10	Bonferroni	**0.045**	**0.046**	**0.053**	**0.052**	**0.054**	**0.054**
	Holm	**0.045**	**0.046**	**0.054**	**0.052**	**0.055**	**0.055**
	Hommel	**0.046**	**0.047**	**0.054**	**0.053**	**0.055**	**0.055**
	Hochberg	**0.046**	**0.046**	**0.054**	**0.052**	**0.055**	**0.055**
	BH	0.080	0.091	0.103	0.102	0.104	0.104
	BY	0.013	0.015	0.019	0.017	0.022	0.020
	None	0.898	0.904	0.901	0.902	0.904	0.902

Boldfaced values indicate the closest type I error rates to the nominal ones.

Table 4. Type I error for one extreme

Group	Method	$\hat{\alpha}_{10}$	$\hat{\alpha}_{20}$	$\hat{\alpha}_{30}$	$\hat{\alpha}_{40}$	$\hat{\alpha}_{50}$	$\hat{\alpha}_{100}$
3	**Bonferroni**	**0.050**	**0.046**	**0.049**	**0.053**	**0.050**	**0.050**
	Holm	0.053	0.050	0.053	0.056	0.053	0.054
	Hommel	0.056	0.052	0.056	0.059	0.056	0.057
	Hochberg	0.055	0.051	0.055	0.057	0.054	0.056
	BH	0.062	0.057	0.061	0.065	0.061	0.063
	BY	0.033	0.029	0.033	0.032	0.032	0.032
	None	0.137	0.137	0.142	0.142	0.144	0.147
4	**Bonferroni**	**0.047**	**0.049**	**0.050**	**0.050**	**0.051**	**0.049**
	Holm	0.050	0.053	0.052	0.054	0.053	0.052
	Hommel	0.053	0.054	0.054	0.057	0.055	0.054
	Hochberg	0.051	0.053	0.053	0.054	0.053	0.052
	BH	0.066	0.069	0.069	0.071	0.069	0.068
	BY	0.026	0.028	0.026	0.027	0.028	0.027
	None	0.261	0.266	0.262	0.271	0.263	0.262
5	**Bonferroni**	0.044	0.046	0.049	**0.052**	**0.052**	**0.050**
	Holm	0.046	0.048	0.051	0.053	0.055	0.052
	Hommel	**0.047**	**0.049**	**0.052**	0.055	0.056	0.053
	Hochberg	0.046	0.048	0.051	0.053	0.055	0.052
	BH	0.068	0.071	0.074	0.080	0.075	0.077
	BY	0.021	0.020	0.024	0.023	0.025	0.024
	None	0.396	0.397	0.393	0.413	0.404	0.406
6	**Bonferroni**	0.046	0.045	0.051	**0.051**	**0.053**	**0.052**
	Holm	0.047	0.047	0.053	0.053	0.054	0.054
	Hommel	**0.048**	**0.048**	**0.055**	0.054	0.056	0.055
	Hochberg	0.047	0.047	0.053	0.053	0.054	0.054
	BH	0.073	0.076	0.082	0.086	0.088	0.085
	BY	0.021	0.022	0.024	0.020	0.025	0.022
	None	0.531	0.531	0.539	0.539	0.539	0.542
7	**Bonferroni**	**0.054**	**0.049**	**0.048**	**0.053**	**0.050**	**0.051**
	Holm	**0.055**	**0.050**	**0.049**	**0.054**	**0.052**	**0.052**
	Hommel	**0.056**	**0.051**	**0.051**	**0.055**	**0.053**	**0.053**
	Hochberg	**0.055**	**0.050**	**0.050**	**0.054**	**0.052**	**0.052**
	BH	0.088	0.083	0.087	0.091	0.091	0.091
	BY	0.024	0.019	0.020	0.019	0.021	0.021
	None	0.657	0.654	0.662	0.667	0.665	0.659
8	**Bonferroni**	**0.049**	**0.049**	**0.051**	**0.049**	**0.053**	**0.051**
	Holm	**0.051**	**0.050**	**0.052**	**0.050**	**0.054**	**0.051**
	Hommel	**0.052**	**0.051**	**0.052**	**0.051**	**0.055**	**0.052**
	Hochberg	**0.051**	**0.050**	**0.052**	**0.050**	**0.054**	**0.051**
	BH	0.088	0.095	0.095	0.095	0.093	0.091
	BY	0.017	0.019	0.022	0.019	0.021	0.018
	None	0.763	0.763	0.767	0.771	0.766	0.767
9	**Bonferroni**	**0.045**	**0.047**	**0.051**	**0.049**	**0.056**	**0.053**
	Holm	**0.045**	**0.048**	**0.052**	**0.049**	**0.057**	**0.054**
	Hommel	**0.046**	**0.048**	**0.053**	**0.050**	**0.058**	**0.055**
	Hochberg	**0.045**	**0.048**	**0.052**	**0.049**	**0.057**	**0.054**
	BH	0.080	0.091	0.093	0.097	0.106	0.105
	BY	0.017	0.015	0.019	0.018	0.024	0.020
	None	0.845	0.848	0.836	0.852	0.847	0.847
10	**Bonferroni**	**0.044**	**0.047**	**0.052**	**0.049**	**0.053**	**0.047**
	Holm	**0.044**	**0.048**	**0.053**	**0.050**	**0.054**	**0.048**
	Hommel	**0.045**	**0.049**	**0.053**	**0.051**	**0.054**	**0.049**
	Hochberg	**0.044**	**0.048**	**0.053**	**0.050**	**0.054**	**0.048**
	BH	0.082	0.090	0.097	0.098	0.102	0.098
	BY	0.013	0.015	0.018	0.017	0.020	0.017
	None	0.897	0.904	0.902	0.905	0.903	0.901

Boldfaced values indicate the closest type I error rates to the nominal ones.

3.3. Demonstration of the Pairwise Comparison Methods

This section demonstrates how to apply multiple comparison methods to three real data applications. ANOVA and multiple comparison methods are conducted on cholesterol data using aov test and pair comp functions, respectively, available in the oneway tests R package [20]. The associated p-values of all multiple comparison methods are placed in Table 5.

3.3.1. Cholesterol Data

In this part, we work with the cholesterol data set collected by Westfall, available in a multcomp R package [21]. A clinical study assesses the effect of three formulations of the same drug on reducing cholesterol. The formulations are 20 mg at once (1 time), 10 mg twice a day (2 times), and 5 mg four times a day (4 times). In addition, two competing drugs are used as control groups (drug D and drug E). The study aims to find which formulations, if any, are efficacious and how these formulations compare with the existing drugs. This data set has 50 observations (10 observations for each treatment). The descriptive statistics (mean ± standard deviation) are 5.78±2.88, 9.22±3.48, 12.37±2.92, 15.36±3.45, and 20.95±3.35 for the groups, 1time, 2time, 4times, drugD, and drugE, respectively.

In this part, we apply ANOVA for the comparison of the groups. Before ANOVA, we assess the normality of the reduced amount in cholesterol values for five treatment groups. All groups satisfy the assumption of normality (e.g., Shapiro-Wilk normality test: all p-values=$0.4541 - 0.9696$). Moreover, Levene’s homogeneity test suggests that the variances of the five treatment methods are homogeneous (p-value = 0.9875).

```
library(oneway tests)
model <- aov.test(response ~ trt, data = cholesterol)
  One-Way Analysis of Variance (alpha = 0.05)
-------------------------------------------------------------
  data: response and trt
  statistic: 32.43283
  num df: 4
  denom df: 45
  p-value: 9.818516e-13
  Result: The difference is statistically significant.
-------------------------------------------------------------
```

Since the p-value obtained as a result of ANOVA is smaller than 0.05, there is a statistically significant difference between the treatment groups ($\mathrm{F} = 32.43283$, $\mathrm{df}_{num} = 4$, $\mathrm{df}_{denom} = 45$, $\mathrm{p-value} = 9.818516 \times 10^{-13}$). After obtaining statistically significant results in ANOVA, we need to investigate the groups which create the difference. In this part, we make pairwise comparisons with the Hommel method since our simulation results suggest that the Hommel method holds a nominal level of type I error rate when the number of groups is 5. The number of observations in each group is 10.

```
paircomp(model, adjust.method = "hommel")
  Hommel Correction (alpha = 0.05)
-----------------------------------------------------
     Level (a)  Level (b)   p.value        No difference
1     1time     2times   5.139629e-02     Not reject
2     1time     4times   4.664132e-04     Reject
3     1time     drugD   2.066193e-05      Reject
4     1time     drugE   2.444640e-08      Reject
5    2times    4times   5.139629e-02     Not reject
6    2times     drugD   4.343976e-03     Reject
7    2times     drugE   3.951905e-06      Reject
8    4times     drugD   5.139629e-02     Not reject
9    4times     drugE   6.398084e-05      Reject
10  drugD     drugE   6.950361e-03      Reject
-----------------------------------------------------
```

According to the result obtained adjusting with the Hommel method, statistical differences in cholesterol reduction between all other pairs are significant, except for three pairs: 1time-2times, 2times-4 times, and 4times-drugD.

3.3.2. Diet and Weight Loss Data

In this part, we work with the diet data set in the WRS2 R package [22]. Weight loss is studied for three different types of diets. There are 24 observations in diet group A, 25 in diet group B, and 27 in diet group C, with a total of 76 observations. The descriptive statistics (mean ± standard deviation) are 3.30±2.24, 3.27±2.46, and 5.15±2.40 for the diet types A, B, and C, respectively.

In this section, we perform ANOVA for the comparison of diet groups. Before ANOVA, we check the normality of the weight loss in diet values for three different diet groups. The normality assumption is met for all groups (e.g., Shapiro-Wilk normality test: all p-values = $0.0774 - 0.8721$). Further, Levene's homogeneity test states that the variances of the three temperature methods are homogeneous (p-value = 0.6122).

Since the p-value obtained as a result of ANOVA is smaller than 0.05, there is a statistically significant difference between the diet groups ($\mathrm{F} = 5.383104$, $\mathrm{df}_{num} = 2$, $\mathrm{df}_{denom} = 73$, and $\mathrm{p-value} = 0.006595853$). After obtaining statistically significant results in ANOVA, we need to investigate the groups which create the difference. In this part, we make pairwise comparisons with the Bonferroni method since our simulation results suggest that the Bonferroni method holds a nominal level of type I error rate when the number of groups is three. The sample size pattern is progressive N. The result obtained adjusting with Bonferroni method points out that, diet type C leads to statistically greater weight loss than the others

3.3.3. Pottery Data

In this part, we work with the pottery data set in the carData R package [23]. The data give the chemical composition of ancient pottery found at three sites in Great Britain: AshleyRails, IsleThorns, and Llanedyrn. There are 5 observations in the AshleyRails group, 5 in the IsleThorns group, and 14 in the Llanedyrn group, totaling 24 observations. The descriptive statistics (mean ± standard deviation) are 17.32±1.66, 18.18±1.77, and 12.56±1.38 for AshleyRails, IsleThorns, and Llanedyrn sites, respectively. The Caldicot site was not included in this part since there were two ancient potteries.

In this section, we perform ANOVA for the comparison of site groups. Before ANOVA, we check the normality of the Aluminum values for three different groups. The normality assumption is met for all groups (e.g., Shapiro-Wilk normality test: all p-values = $0.780 - 0.967$). Furthermore, Levene's homogeneity test states that the variances of three sites are homogeneous (p-value = 0.950).

Since the p-value obtained as a result of ANOVA is less than 0.05, there is a statistically significant difference between the three sites ($\mathrm{F} = 34.52644$, $\mathrm{df}_{num} = 2$, $\mathrm{df}_{denom} = 21$, $\mathrm{p-value} = 2.296561 \times 10^{-7}$). After obtaining a statistically significant result in ANOVA, we need to investigate the sites making the difference. In this section, we make pairwise comparisons with the Bonferroni method since our simulation results illustrate that the Bonferroni method holds a nominal type I error rate in extreme cases where the number of groups is three. The sample size design is one extreme.

The results obtained by correcting with the Bonferroni method show that the Llanedyrn group leads to statistically smaller aluminum than the AshleyRails and IsleThorns sites.

In comparisons of two or more groups, multiple comparison tests are employed to investigate the difference if there is a statistically significant difference among the groups. At this point, multiple comparison tests are performed to control Type I error rates. Although studies are published in the literature to control type I error rates, it is important to choose an appropriate multiple comparison test under different conditions. Therefore, it will be important to provide researchers with an overview of multiple comparison tests under various scenarios.

Table 5. The p-values of all multiple comparison methods on cholesterol, diet, and pottery data sets

Data sets	Level (a)	Level (b)	Bonferroni	Holm	Hommel	Hochberg	BH	BY	None
Cholesterol	1time	2times	0.269	0.081	**0.051**	0.051	0.034	0.098	0.027
	1time	4times	0.001	0.000	**0.000**	0.000	0.000	0.000	0.000
	1time	drugD	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
	1time	drugE	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
	2times	4times	0.419	0.084	**0.051**	0.051	0.047	0.136	0.042
	2times	drugD	0.009	0.005	**0.004**	0.005	0.002	0.005	0.001
	2times	drugE	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
	4times	drugD	0.514	0.084	**0.051**	0.051	0.051	0.151	0.051
	4times	drugE	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
	drugD	drugE	0.017	0.007	**0.007**	0.007	0.002	0.007	0.002
Diet	A	B	**1.000**	0.962	0.962	0.962	0.962	1.000	0.962
	A	C	**0.020**	0.020	0.013	0.015	0.011	0.021	0.007
	B	C	**0.022**	0.020	0.015	0.015	0.011	0.021	0.007
Pottery	AshleyRails	IsleThorns	**1.000**	0.451	0.451	0.451	0.451	0.827	0.451
	AshleyRails	Llanedyrn	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000
	IsleThorns	Llanedyrn	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000

All p-values are rounded to three decimals. Associated p-values of suggested methods are written in bold.

4. Discussion

In this paper, we compare the type I error rates of Bonferroni, Holm, Hommel, Hochberg, BH (FDR), and BY multiple comparison methods under different conditions with Monte Carlo simulation and three real data applications. Blakesley et al. [11] introduced the Hommel and Hochberg techniques tailored for moderately correlated measures to enhance statistical power while maintaining control over type I errors within neuropsychological datasets. Felix and Menezes [12] showed that BH correction provides a smaller type I error rate by varying the sample size and sample distribution. Staffa and Zurakowski [13] proposed that surgeons use multiple comparison methods, including Bonferroni, Tukey, Scheffe, Holm, and Dunnett. They provide guidance on strategies for how to handle multiplicity and multiple significance testing in surgical research studies. Kharola et al. [15] criticized the work done by Dimitriev et al. [14] concerning using multiple comparison methods. They stated that the Holm method was more powerful than the Bonferroni method.

5. Conclusion

The criteria for selecting multiple comparison tests in the literature are not clearly defined, leaving uncertainty about the appropriateness of their application in various scenarios. Furthermore, selecting these multiple comparison methods without considering factors like group sizes, sample sizes and sample size designs could produce misleading results. For 3-4 groups, we recommend Bonferroni correction when the number of observations is equal and greater than 50 and Holm correction when the number of observations is less than 50 under the equal N scenario. We recommend the Bonferroni method under progressive N and one extreme scenario regardless of the number of observations. For 5-6 groups, we suggest the Bonferroni method when the number of observations is equal and more than 40 and the Hommel correction when the number of observations is less than 40 under all sample size patterns. For 7 or more groups, we propose Bonferroni, Holm, Hommel, and Hochberg methods under all sample size patterns regardless of the number of observations. In future studies, simulation studies are planned on which post-hoc test should be used under non-normality and/or heterogeneity.

Author Contributions

All the authors equally contributed to this work. They all read and approved the final version of the paper.

Conflicts of Interest

All the authors declare no conflict of interest.

Ethical Review and Approval

No approval from the Board of Ethics is required.

References

[1] S. K. Sarkar, C. K. Chang, *The Simes method for multiple hypotheses testing with positively dependent test statistics*, Journal of the American Statistical Association 92 (440) (2012) 1601–1608.

[2] O. J. Dunn, *Multiple comparisons among means*, Journal of the American Statistical Association 56 (293) (1961) 52–64.

[3] S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics 6 (2) (1979) 65–70.

[4] G. Hommel, *A stagewise rejective multiple test procedure based on a modified Bonferroni test*, Biometrika 75 (2) (1988) 383–386.

[5] Y. Hochberg, *A sharper Bonferroni procedure for multiple tests of significance*, Biometrika 75 (4) (1988) 800–802.

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

[7] Y. Benjamini, D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, The Annals of Statistics 29 (4) (2001) 1165–1188.

[8] S. Lee, D. K. Lee, *What is the proper way to apply the multiple comparison test?*, Korean Journal of Anesthesiology 71 (5) (2018) 353–360.

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

[10] P. H. Westfall, J. F. Troendle, *Multiple testing with minimal assumptions*, Biometrical Journal 50 (5) (2008) 745–755.

[11] R. E. Blakesley, S. Mazumdar, M. A. Dew, P. R. Houck, G. Tang, C. F. Reynolds, M. A. Butters, *Comparisons of methods for multiple hypotheses testing in neuropsychological research*, Neuropsychology 23 (2) (2009) 255–264.

[12] V. B. Felix, A. F. B. Menezes, *Comparisons of ten corrections methods for t-test in multiple comparisons via Monte Carlo study*, Electronic Journal of Applied Statistical Analysis 11 (01) (2018) 74–91.

[13] S. J. Staffa, D. Zurakowski, *Strategies in adjusting for multiple comparisons: A primer for pediatric surgeons*, Journal of Pediatric Surgery 55 (9) (2020) 1699–1705.

[14] D. A. Dimitriev, E. V. Saperova, O. S. Indeykina, A. D. Dimitriev, *Heart rate variability in mental stress: The data reveal regression to the mean*, Data in Brief 22 (2019) 245–250.

[15] S. S. Kharola, D. Gupta, A. Agrawal, *Heart rate variability in mental stress: The data reveal regression to the mean*, Indian Statistical Institute Bangalore Centre (2023) 19 pages.

[16] A. A. Musicus, L. A. Gibson, S. L. Bellamy, J. A. Orr, D. Hammond, K. Glanz, K. G. Volpp, M. B. Schwartz, A. Bleakley, A. A. Strasser, C. A. Roberto, *Effects of sugary beverage text and pictorial warnings: A randomized trial*, American Journal of Preventive Medicine 64 (5) (2023) 716–727.

[17] M. Giacalone, Z. Agata, P.C. Cozzucoli, A. Alibrandi, *Bonferroni-Holm and permutation tests to compare health data: Methodological and applicative issues*, BMC Medical Research Methodology 18 (81) (2018) 1–9.

[18] S. Chen, Z. Feng, X. Yi, *A general introduction to adjustment for multiple comparisons*, Journal of Thoracic Disease 9 (6) (2017) 1725–1729.

[19] R. J. Simes, *An improved Bonferroni procedure for multiple tests of significance*, Biometrika 73 (1986) 751–754.

[20] O. Dag, N. A. B. Dolgun, N. M. Konar, *Onewaytests: An R package for one-way tests in independent groups designs*, The R Journal 10 (1) (2018) 175–199.

[21] T. Hothorn, F. Bretz, P. Westfall, *Simultaneous inference in general parametric models*, Biometrical Journal 50 (3) (2008) 346–363.

[22] P. Mair, R. Wilcox, *Robust statistical methods in R using the WRS2 package*, Behavior Research Methods 52 (2020) 464–488.

[23] J. Fox, S. Weisberg, B. Price, carData: Companion to Applied Regression Data Sets, 2022.