METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Post-Hoc Tests in One-Way ANOVA: The Case for Normal Distribution

Joel Juarros-Basterretxea[1], Gema Aonso-Diego[2], Álvaro Postigo[2], Pelayo Montes-Álvarez[2], Álvaro Menéndez-Aller[2], Eduardo García-Cueto[2]

[1] *Departamento de Psicología y Sociología, Universidad de Zaragoza, Zaragoza, Spain.* [2] *Departamento de Psicología, Universidad de Oviedo, Oviedo, Spain.*

Methodology, 2024, Vol. 20(2), 84–99, https://doi.org/10.5964/meth.11721

Received: 2023-04-10 • **Accepted:** 2024-03-27 • **Published (VoR):** 2024-06-28

Handling Editor: Van Deun Katrijn, Tilburg University, Tilburg, the Netherlands

Corresponding Author: Joel Juarros-Basterretxea, Faculty of Education, University of Zaragoza, Pedro Cerbuna Street 12, Zaragoza, 50009, Spain. E-mail: joeljuarros@unizar.es

Abstract

When one-way ANOVA is statistically significant, a multiple comparison problem arises, hence post-hoc tests are needed to elucidate between which groups significant differences are found. Different post-hoc tests have been proposed for each situation regarding heteroscedasticity and sample size groups. This study aims to compare the Type I error (α) rate of 10 post-hoc tests in four different conditions based on heteroscedasticity and balance between-group sample size. A Montecarlo simulation study was carried out on a total of 28 data sets, with 10,000 resamples in each, distributed through four conditions. One-way ANOVA tests and post-hoc tests were conducted to estimate the α rate at a 95% confidence level. The percentage of times the null hypothesis was falsely refused is used to compare the tests. Three out of four conditions demonstrated considerable variability among sample sizes. However, the best post-hoc test in the second condition (heteroscedastic and balance group) did not depend on simple size. In some cases, inappropriate post-hoc tests were more accurate. Homoscedasticity and balance between-group sample size should be considered for appropriate post-hoc test selection.

Keywords

One-way ANOVA, post-hoc tests, homoscedasticity, heteroscedasticity, sample size

The analysis of variance (ANOVA) is a standard statistical method in many scientific disciplines and one of the most used techniques in social and health research (Counsell & Harlow, 2017; Howell, 2010). The wide use of ANOVA has been attributed to its usefulness to answer to experimental method´s and general research problems. The objective of this technique is using a *F*-statistic to test the null hypothesis (H0) of equality of group means when more than two groups are compared. ANOVA enables not only the individual effect of each independent variable separately, but also the interacting effects of the *k* independent variables.

ANOVA is a parametric test that depends on three distributional assumptions: (a) study groups scores must be independent; (b) distribution of each group scores must be normal (normality); (c) the variances of group scores must be equal or constant (homoscedasticity). The violation of these assumptions is affected the Type I error rate (α; when the null hypothesis is falsely rejected) (Ruscio & Roche, 2012; Sharma & Kibria, 2013; Zimmerman, 2004). Unfortunately, it is difficult to obtain normally distributed and homoscedastic samples, especially in social sciences. This usual limitation has increased the interest in the robustness of ANOVA when these assumptions are violated (Bhat et al., 2002; Keselman et al., 1998). Accordingly, some authors point out that ANOVA is robust to departures from normality even group sample sizes are different (unbalanced groups) (Blanca et al., 2017; Sarstedt & Mooi, 2019). Nevertheless, the robustness of ANOVA is questionable for violations of homoscedasticity, so it must be tested. This violation is particularly relevant when groups are unbalanced, and sample sizes are lower than 30 per group (Sarstedt & Mooi, 2019). Therefore, the imbalance between group sample sizes should be considered in addition to homoscedasticity/heteroscedasticity.

ANOVA detects the presence or the absence of a global effect of the independent variable on the dependent variable. When the null hypothesis is rejected ($p \leq .05$), it informs the researcher that there is at least one comparison that is statistically significant, but it does not inform the researcher about which pair of means are significantly different. Thus, post-hoc tests (also called *a posteriori analysis* or *multiple comparison analysis tests*) must be used to determine which levels of independent variable means differ significantly from other levels (Kim, 2015; McHugh, 2011; Meyers et al., 2016).

In order to overcome the multiple-comparison problem and the potential limitations joined to this analysis (e.g., familywise error) (see Wilcox, 2023 for a further review), different available post-hoc tests have been proposed (Cramer et al., 2016; McHugh, 2011). In SPSS these tests are encompassed into two groups, depending on whether equal variances are assumed or not. Some post-hoc tests present in SPSS are briefly reviewed.

Post-hoc tests from SPSS only consider the difference in homoscedasticity, whereas other conditions, such as sample size and imbalance between group sizes, are also influencing the capability to detect potential differences amongst groups. Tests included in this category are Fisher's Least Significant Difference, Bonferroni-Dunn's test, Šidák's test, Scheffé's test, Tukey's Honestly Significant Differences, Hochberg's GT2, and Gabriel's.

Fisher's Least Significant Difference (LSD) (Fisher, 1935) is one of the first multiple comparisons procedures (Hayter, 1986). Fisher's LSD is a set of individual *t*-tests, so it does not make any correction for multiple comparisons (Ato & Vallejo, 2015), although some authors have indicated its appropriateness under homoscedasticity (Meyers et al., 2016). The lack of multiple comparison correction makes this test the most liberal, thus, it has high statistical power, but greater likelihood of committing a Type I error (Meyers et al., 2016).

Bonferroni-Dunn's test (Dunn, 1961) is appropriate when groups' sample sizes are equal (Day & Quinn, 1989; McHugh, 2011; Sarstedt & Mooi, 2019). This post-hoc test is characterized by higher control of Type I error when the assumptions are violated compared with other tests, even though it loses test power (Abdi & Williams, 2010). The higher control of Type I error is based on the Bonferroni correction. Briefly, the Bonferroni correction is applied to every single test to maintain the p critical (α) level at .05 to all tests dividing it by the number of tests carried out (m) (Sedgwick, 2012). The Bonferroni's test is intended for the confirmation of the hypothesis basing on the planned comparison (McHugh, 2011).

Šidák's test (Šidák, 1967), as Bonferroni's test, is appropriate when groups' sizes are equal (Day & Quinn, 1989; Lee & Lee, 2020). This test is usually considered better than Bonferroni's test because of its smaller Type I error (Abdi, 2007; Abdi & Williams, 2010; Ruxton & Beauchamp, 2008). Nevertheless, this higher control of the Type I error is achieved at the expense of losing test power (Ruxton & Beauchamp, 2008). It makes the test very conservative when the number of comparisons increases and the groups are not independent (Abdi, 2007).

Scheffé's test (Scheffé, 1953) is the best option when no previous results or theoretical framework guides the analysis because it is the only test entirely consistent with ANOVA results (McHugh, 2011; Ruxton & Beauchamp, 2008). It is a particularly good estimator when the groups are balanced (Day & Quinn, 1989; McHugh, 2011; Ruxton & Beauchamp, 2008; Shaffer, 1995). However, it can be used even when groups are unbalanced due to similar robustness to normality and homoscedasticity violations such as ANOVA (Ruxton & Beauchamp, 2008). Despite it has been considered less sensitive than other tests (e.g., Tukey or Bonferroni) for pairwise comparisons (Brown, 2005; Keppel & Wickens, 2004; Lee & Lee, 2020), Abdi and Williams (2010) have pointed out that Scheffé's test is less conservative than Bonferroni's when group means are equal.

Tukey's Honestly Significant Differences (HSD) (Tukey, 1953) is the most common used post-hoc test, considered the most adequate in several situations (Brown, 2005). It is recommended when balanced groups are compared (Abdi & Williams, 2010; Day & Quinn, 1989; Lee & Lee, 2020; Sarstedt & Mooi, 2019), but it is considered robust for imbalance between groups' sizes under homoscedasticity (Brown, 2005; Jaccard et al., 1984; McHugh, 2011), especially if the unbalance is not severe (Spjøtvoll & Stoline, 1973). Regardless of whether it is relatively conservative (Brown, 2005), it is more liberal, and thus, more powerful alternative to the Bonferroni's test (Abdi & Williams, 2010) and, therefore, than the Šidák's and Scheffé's tests (Ruxton & Beauchamp, 2008).

Hochberg's GT2 (Hochberg, 1974) is a modification of Tukey's method appropriate to test differences between unbalanced groups (Day & Quinn, 1989; Hochberg, 1974; Ruxton & Beauchamp, 2008), although it could also be used for comparison of imbalance groups even under heteroscedasticity (Hochberg, 1974) due to its conservative nature (Hochberg, 1975; Tamhane, 1979). Nevertheless, of note is that the robustness of the GT2 test is at the expense of losing power. It has been suggested as an appropriate alternative when a correction is required, but Bonferroni's method is considered too conservative (Armstrong, 2014).

Gabriel's test (Gabriel, 1978) is a more liberal variation of Hochberg's GT2 (Meyers et al., 2016). Thus, it is also recommended for unbalanced groups comparison (Day & Quinn, 1989). Its liberal nature makes this test generally more powerful than Hochberg's test among other (e.g., Keppel & Wickens, 2004). Concerning post-hoc tests encompassed in the SPSS category of "equal variances are not assumed", they are presumably accurate when comparison groups' variances are unequal (heteroscedastic). As in the case of previous tests, the capability to detect potential differences between groups could be affected by other conditions as sample size and unbalanced sizes between groups. The Games-Howell's test, Tamhane's T2, and Dunnett's T3 are some tests included in this category.

Games-Howell's test (Games & Howell, 1976; Howell & Games, 1974) is an appropriate option for pairwise comparisons when group sample-sizes are different (unbalanced groups) (Day & Quinn, 1989; Sarstedt & Mooi, 2019). However, it can produce Type I error slightly above the significance level in small sizes: Thus, it is recommended when test power is more important than significance (Day & Quinn, 1989; Dunnett, 1980; Jaccard et al., 1984; Tamhane, 1979).

Tamhane's T2 (Tamhane, 1977, 1979) is a modified version of the Games-Howell test. Like the previous test, it is appropriate when sample sizes are different (Day & Quinn, 1989). Nevertheless, it is more conservative than Games-Howell's test, which is somewhat liberal under conditions of heteroscedasticity and unbalances between-groups sample sizes (Jaccard et al., 1984; Tamhane, 1979).

Dunnett's T3 (Dunnett, 1980) is a variation of Tamhane's T2, more conservative and appropriate to a comparison of balanced groups (Day & Quinn, 1989; Ruxton & Beauchamp, 2008). It is not robust for unequal sample sizes and has more power in small samples ($n < 15$) (Day & Quinn, 1989).

In this background, choosing the accurate test in each situation would facilitate the correct assessment of the differences between groups, and thus obtain results that are more precise to make conclusions (Grinde et al., 2017). To our knowledge, simulation research has not been conducted to determine the suitable post-hoc test for one-way ANOVA. Considering the characteristics of different post-hoc tests for one-way ANOVA, the potential strengths and limitations of each of them, the present study aimed to compare the Type I error rate of ten post-hoc tests in four different conditions based on homoscedasticity and balance between group sample size when positive pairing between group sample-size and group variance, as well as big variance ratio, was maintained.

Method

Data Simulation

A Montecarlo simulation study was conducted to generate 280,000 data distributed in seven sample sizes in each of the four conditions. Hence, a total of 28 data sets were generated with 10,000 resamples in each, to guarantee the stability of estimations. Normal distribution condition was imposed for all simulated data. Open software R (Version 3.5.1) using the *R commander* (Rcmdr) package (Fox, 2005) was utilized to carry out the simulations.

Procedure

In each of the 28 sets, data were distributed among three groups. Group means were maintained constant whereas standard deviations and group sample sizes were manipulated to create four different conditions. In the first condition, groups were homoscedastic and balanced, in the second condition groups were homoscedastic and unbalanced, in the third condition groups were heteroscedastic and balanced, and, finally, in the fourth condition groups were heteroscedastic and unbalanced. Seven sample sizes (*N*) were used: 30, 90, 150, 300, 750, 1,500, and 3,000. In those conditions with balanced groups, group sizes (*n*) were identical in the three groups (e.g., $N = 30$ condition implicate $n1 = 10$, $n2 = 10$, and $n3 = 10$). In conditions with unbalanced groups, groups-sizes were manipulated to get imbalance (e.g., $N = 30$ condition implicate $n1 = 5$, $n2 = 10$, $n3 = 15$, following the same rule in all cases; imbalance ratio = 3). Regarding the heterogeneity of variances, in homoscedastic groups, the standard deviation was 10 in each group, whereas in heteroscedastic groups, the standard deviation was manipulated to get imbalances ($\sigma x1 = 1$, $\sigma x2 = 5$, $\sigma x3 = 10$; variance ratio = 100), to guarantee big heteroscedasticity. In this research, only positive pairing was analyzed. In other words, the group with the biggest sample size always had the biggest variance and the smallest group always had the smallest variance.

Data Analysis

One-way ANOVA and ten (seven for equal variances and three for unequal variances) post-hoc tests were conducted in each condition and for all sample sizes to estimate the *F* statistic Type I error rate at a 95% confidence level and assume a normal distribution of group scores. Estimators' power to detect statistical differences among groups was better when Type I error was near 5% (see Pedrosa et al., 2015 for a similar approach) based on Bradley's liberal criterion, according to which Type I error rate higher than 5.25 is considered conservative and lower than 4.75 liberal (Bradley, 1978). IBM SPSS Statistics software was used to calculate one-way ANOVAs (IBM, 2016).

Results

Condition 1: Homoscedastic and Balanced Groups

In the first condition (homoscedastic and balanced groups) some consistency was observed across the different group sizes (see Table 1). In general, Bonferroni's (α_{rate} = 4.76–5.17), Šidák's (α_{rate} = 4.86–5.20), Hochberg's GT2 (α_{rate} = 4.95–5.20), and Gabriel's (α_{rate} = 4.89–5.20) tests were the most accurate for the majority of sample sizes (excepting N = 150 and N = 750). For sample sizes of N = 90, N = 300, N = 1500 and N = 3000, also Tamhane's T2 (α_{rate} = 4.93–5.11) and Dunnett's T3 (α_{rate} = 4.95–5.22) tests were accurate. The Games-Howell test was the unique test accurate for the sample size of N = 150. None of the tests was accurate for N = 750 sample size, being nearest to desirable accuracy Bonferroni´s (α_{rate} = 5.26), Tamhane´s T2 (α_{rate} = 5.27), and Dunnett´s T3 (α_{rate} = 5.28) tests respectively.

Table 1

Percentage of Times Null Hypothesis Is Refused When Is True (Type I Error) When Groups Are Homoscedastic and Balanced

	Sample size (*N*)						
Test	**30**	**90**	**150**	**300**	**750**	**1500**	**3000**
Equal variances are assumed							
LSD	14.88	15.19	13.75	15.07	15.38	16.71	14.20
Bonferroni	4.76*	4.85*	4.59	4.91*	5.26	5.13*	5.17*
Šidák	4.85*	4.94*	4.65	4.96*	5.33	5.20*	4.94*
Scheffé	4.28	4.39	3.93	4.28	4.46	4.43	4.14
HSD	5.78	5.65	5.26	5.59	5.94	5.84	5.57
GT2	4.95*	4.96*	4.65	4.96*	5.33	5.20*	4.95*
Gabriel	4.95*	4.96*	4.65	4.96*	5.33	5.20*	4.89*
Equal variances are not assumed							
T2	4.46	5.11*	4.40	4.95*	5.27	5.22*	4.93*
T3	4.60	5.13*	4.48	4.95*	5.28	5.22*	5*
Games-Howell	5.34	5.73	5.07*	5.50	5.91	5.86	5.69

Note. * indicates the accurate tests for the given sample size.

Condition 2: Heteroscedastic and Balanced Groups

All tests assuming equal variances were excessively liberal for this condition, whereas tests that assume unequal variances performed systematically better when groups were heteroscedastic and balanced, but none of the tests assuming heteroscedasticity was accurate for all sample sizes (see Table 2). Particularly, Tamhane's T2 (α_{rate} = 5.04–5.23) and Dunnett's T3 (α_{rate} = 5.10–5.24) were the most accurate tests for N = 150, N = 750, and N = 3000. Nevertheless, no test was enough accurate for sample sizes of N = 30 and N = 90 (the most accurate was Tamhane's T2, α_{rate} = 5.32 and α_{rate} = 5.70 respectively), N = 300 (Tamanhe's T2 α_{rate} = 5.26 and Dunnett's T3 α_{rate} = 5.27), and N = 1500 (Tamanhe's T2 and Dunnet's T3 α_{rate} = 5.37).

Table 2

Percentage of Times Null Hypothesis Is Refused When Is True (Type I Error) when Groups Are Heteroscedastic and Balanced

	Sample size (*N*)						
Test	**30**	**90**	**150**	**300**	**750**	**1500**	**3000**
Equal variances are assumed							
LSD	21.14	19.12	16.84	18.27	17.84	19.13	18.32
Bonferroni	10.14	8.89	7.56	7.84	7.76	8.30	7.62
Šidák	9.13	8.97	8.57	7.92	7.83	8.74	7.72
Scheffé	9.62	8.13	6.71	7.04	7.23	7.94	6.83
HSD	10.77	9.80	8.51	8.67	8.63	9.50	8.31
GT2	9.74	8.99	7.71	7.93	7.84	8.74	7.72
Gabriel	9.74	8.99	7.70	7.93	7.84	8.74	7.72
Equal variances are not assumed							
T2	5.32	5.70	5.04*	5.26	5.13*	5.37	5.23*
T3	5.58	5.75	5.10*	5.27	5.13*	5.37	5.24*
Games-Howell	6.51	6.43	5.41	5.96	5.83	5.94	6.01

Note. * indicates the accurate tests for the given sample size.

Condition 3: Homoscedastic and Unbalanced Groups

In the third condition (homoscedastic and unbalanced), more variation than in previous conditions was observed (Table 3). For sample size of N = 30, Bonferroni's (α_{rate} = 5), Šidák's (α_{rate} = 5.08) and Hochberg's GT2 (α_{rate} = 5.17) were the most accurate. As for N = 30, for N = 90 sample size, Bonferroni's (α_{rate} = 4.82), Šidák (αrate = 4.92) and Hochberg's GT2 (α_{rate} = 4.94), but also Tamanhe's T2 (αrate = 5.20) were the most accurate. For N = 300, Tamanhe's T2 (α_{rate} = 4.75) and Dunnett's T3 (α_{rate} = 4.79) showed greater accuracy. For sample sizes of N = 750 and N = 3000, Bonferroni's (α_{rate} = 5.05 and 5.06), Šidák's (α_{rate} = 5.08 and 5.17), Tamanhe's T2 (α_{rate} = 5.13 and 5.16) and Dunnett's T3 (α_{rate} = 5.16

Table 3

Percentage of Times Null Hypothesis Is Refused When Is True (Type I Error) When Groups Are Homoscedastic and Imbalanced

	Sample size (*N*)						
Test	**30**	**90**	**150**	**300**	**750**	**1500**	**3000**
Equal variances are assumed							
LSD	14.90	15.41	15.85	13.62	14.57	14.21	14.86
Bonferroni	5*	4.82*	5.41	4.55	5.05*	4.71	5.06*
Šidák	5.08*	4.92*	5.48	4.69	5.08*	4.87*	5.17*
Scheffé	4.69	4.36	4.60	3.92	4.47	4.16	4.42
HSD	6.05	5.68	6.27	5.28	5.84	5.48	5.72
GT2	5.17*	4.94*	5.50	4.69	5.09*	4.85*	5.17*
Gabriel	5.87	5.61	6.29	5.30	5.74	5.51	5.76
Equal variances are not assumed							
T2	5.81	5.20*	5.48	4.75*	5.13*	4.89*	5.16*
T3	6	5.27	5.55	4.79*	5.16*	4.91*	5.16*
Games-Howell	6.75	6.08	6.28	5.36	5.85	5.52	5.80

Note. * indicates the accurate tests for the given sample size.

in both cases) tests were the most accurate. Finally, Šidák's (α_{rate} = 4.87) Tamanhe's T2 (α_{rate} = 4.89) and Dunnett's T3 (α_{rate} = 4.91) tests were the most accurate for the sample size of N = 1500. According to the used criteria, no test was accurate enough for N = 150 sample size, being the Scheffé test (α_{rate} = 4.60) the most nearest to the desirable accuracy.

Condition 4: Heteroscedastic and Unbalanced Group

In the fourth condition (heteroscedastic and unbalanced groups), tests assuming equal variances were generally conservative, excepting Fisher's LSD (Table 4). Specifically, LSD test showed appropriate accuracy for all sample sizes (α_{rate} = 4.74–5.21) excepting or N = 15000. Similarly, Tamhane's T2 was accurate for all sample sizes (α_{rate} = 5.05–5.20) excepting for N = 150 and N = 750. Finally, Dunnett's T3 test was accurate for sample sizes of N = 90, 300, 1,500 and 3,000 (αrate = 5.11–5.20).

Table 4

Percentage of Times Null Hypothesis Is Refused When Is True (Type I Error) When Groups Are Heteroscedastic and Imbalanced

	Sample size (*N*)						
Test	**30**	**90**	**150**	**300**	**750**	**1500**	**3000**
Equal variances are assumed							
LSD	5.12*	5.21*	4.88*	5.05*	5.14*	4.72	4.75*
Bonferroni	1.74	1.31	1.44	1.43	1.68	1.53	1.52
Šidák	1.76	1.33	1.50	1.44	1.69	1.56	1.54
Scheffé	1.61	1.18	1.30	1.25	1.42	1.25	1.32
HSD	2.02	1.45	1.7	1.75	2.14	1.86	1.72
GT2	1.78	1.33	1.50	1.44	1.69	1.56	1.54
Gabriel	1.85	1.36	1.56	153	1.79	1.66	1.60
Equal variances are not assumed							
T2	5.15*	5.05*	5.29	5.11*	5.56	5.20*	5.10*
T3	5.33	5.11*	5.30	5.13*	5.56	5.20*	5.11*
Games-Howell	6.06	5.78	5.99	5.76	6.20	5.77	5.83

Note. * indicates the accurate tests for the given sample size.

Discussion

The present study aimed to determine the suitability of ten post-hoc tests for one-way ANOVA in four different conditions and seven sample sizes in 28 different data sets for the best case scenario of normal distributions. Then, samples were normally distributed and group means were equal, but group sample sizes (balanced vs. unbalanced) and group variances (homoscedastic vs. heteroscedastic) were manipulated. Furthermore, in all these conditions positive pairing between group sample size and group variance as well as big variance ratio was maintained.

Following the obtained results, two different conclusions can be highlighted. Firstly, regarding the appropriateness of post-hoc tests, big variability was observed through different sample sizes, except in the second condition (heteroscedastic and balanced). Second, some incongruences were detected regarding the nature of the test (e.g., the test which assumes equal variances) and the condition in which the test is more accurate (e.g., in heteroscedastic condition). Some post-hoc tests were designed for heteroscedastic or homoscedastic conditions and were more accurate in homoscedastic or heteroscedastic conditions, respectively.

Regarding the variability of best estimators in different conditions, only in the second condition (heteroscedastic and balanced groups), the same post-hoc test was the best estimator in all sample sizes and should be selected under these conditions: Tamhane's

T2. As expected, when between-group sizes are the same (balanced groups), and under the violation of the homoscedasticity assumption, those tests which do not assume equal variances were systematically better estimators. Nevertheless, following Bradley's liberal criterion (Bradley, 1978), only Tamanhe's T2 and Dunnett's T3 tests were accurate and only for *N* = 150, *N* = 750 and *N* = 3,000, being out of the ± .25 range for considering one test accurate. Contrary to expectations, Dunnett's T3, which is designed for this condition, was not the most accurate nor more conservative than Tamhane's T2 (Day & Quinn, 1989; Ruxton & Beauchamp, 2008). In general, Dunnett's T3 was more liberal and its performance worse than Tamanhe's T2. On the contrary, accordingly to Tamhane (1979), Games-Howell's test was the most liberal of these three post-hoc tests. Regarding the tests assuming equal variances, they were systematically more liberal, demonstrating that the use of these post-hoc tests under heteroscedasticity increased the Type I error when positive pairing and normal distribution even when the groups are balanced. Likewise, it is important to note that all post-hoc tests used in the present research were liberal in this condition.

Contrary to the results obtained in the second condition, variability is remarkable in the first, third, and fourth conditions. In the first condition (homoscedastic and balanced), and according to previous research, Scheffé's test (more conservative) and Tukey's HSD (more liberal) would theoretically be the better estimators for comparison of balanced groups under homoscedasticity (Day & Quinn, 1989; Shaffer, 1995). Nevertheless, findings indicated that Scheffé's test was systematically too conservative, while Tukey's HSD test was systematically too liberal.

Considering the conditions of the study, these results are in line with previous research. At first, it was expected to obtain results that are more conservative with Scheffé's test than with Bonferroni's, Šidák's, or Tukey's HSD´s tests (Brown, 2005; Keppel & Wickens, 2004; Lee & Lee, 2020). In addition, the excessively conservative results of Scheffé's test are in line with recent research. As recent literature has shown when ANOVA is carried out under certain conditions according to current research (i.e., positive pairing and big variance ratio), the probability of committing Type I error tends to decrease under positive pairing condition, which makes conservative (Blanca, et al., 2018), and Scheffé's test is the unique test entirely consistent with ANOVA results (McHugh, 2011; Ruxton & Beauchamp, 2008). On the other hand, more liberal results with Tukey's HSD was observed compared to Bonferroni's, Šidák's, or Scheffé's tests (Abdi & Williams, 2010; Ruxton & Beauchamp, 2008). In line with these results, Bonferroni's, Šidák's, Hochberg's GT2 and Gabriel's tests achieved the ideal accuracy for this condition and they are considered the preferable alternatives. These four tests are characterized as for being more conservative than Tukey's HSD test and more liberal than Scheffé's test. Bonferroni's and Šidák's tests were expected to perform well in this first condition considering that they are made for comparing balanced groups under homoscedasticity (Day & Quinn, 1989; McHugh, 2011; Sarstedt & Mooi, 2019). Hochberg's

GT2 and Gabriel’s tests yielded an expected performance. Based on the literature, these tests are a more conservative alternative to Tukey’s HSD recommended for unbalanced group comparison (see Day & Quinn, 1989; Ruxton & Beauchamp, 2008). Hochberg’s GT2 findings were the expected: more conservative than Tukey’s HSD, but more liberal than Bonferroni’s test, so than Scheffé’s test (Armstrong, 2014).

Maybe more surprising are the results obtained with Gabriel’s test, which is a more liberal variation of Hochberg’s GT2 (Meyers et al., 2016) but it was indistinguishable from it. In spite of Tamanhe’s T2 and Dunnett’s T3 tests showing appropriate accuracy for some sample sizes, considering the overall results and conditions, the previously mentioned four tests seem to be the best option.

Regarding the third condition (homoscedastic and unbalanced), higher variability than in the first condition was observed, but some similarities too. As in the first condition, Scheffé’s test was systematically too conservative and Šidák’s test was systematically too liberal when groups are unbalanced contradicting previous literature (see Brown, 2005; Ruxton & Beauchamp, 2008). Regarding Scheffé’s test, the same as in the first condition could be argued: ANOVA tends to be conservative when positive pairing and big variance ratio (Blanca et al., 2018) occur and the Scheffé’s test is the only test entirely consistent with ANOVA results. In respect to Tukey’s HSD test, some authors have pointed out that it is robust enough when the group sample size imbalance is not severe (Spjøtvoll & Stoline, 1973), but as in the first condition, excessively liberal results were obtained. Similar to the first condition, other tests assuming equal variances were more liberal than Scheffé’s test and more conservative than Tukey’s HSD, showing the highest accuracy, but little differences were observed. For example, Hochberg’s GT2 and its more liberal version Gabriel’s test should be the most accurate tests for this condition (see Day & Quinn, 1989; Ruxton & Beauchamp, 2008), but contrary to what was expected, the second one was excessively liberal for all sample sizes. On the contrary, Bonferroni’s, Šidák’s and Hochberg’s GT2 tests were generally accurate except for medium sample sizes. These results support the idea of the appropriateness of Hochberg’s GT2, but also the robustness of Bonferroni’s and Šidák’s test to imbalance for this specific condition. Furthermore, obtained results also show, as in the first condition, the tendency to higher liberality of Hochberg’s GT2 compared with Bonferroni’s test as indicated in previous literature (Armstrong, 2014). Finally, Tamanhe’s T2 and Dunnett’s T3 also showed general accuracy, despite the fact that they were designed for heteroscedasticity. Games-Howell´s test was liberal when groups were unbalanced without the necessity of heteroscedasticity (Jaccard et al., 1984; Tamhane, 1979), but the more conservative alternative, Tamanhe’s T2, show accuracy not only for medium to big sample sizes but also for $N = 90$. In addition, Dunnett’s T3 showed good accuracy for medium to big sample sizes, but in line with other studies (Day & Quinn, 1989), it was not accurate for small sample sizes.

Regarding the fourth condition (heteroscedastic and unbalanced), three estimators were predominantly better, making them the most appropriate alternatives: Dunnett’s T3, Tamhane’s T2, and Fisher’s LSD. As pointed out by other authors (Day & Quinn, 1989; Dunnett, 1980; Jaccard et al., 1984; Tamhane, 1979), Games-Howell’s test was systematically liberal in this condition, although it is theoretically appropriate for an unbalanced group comparison under heteroscedasticity.

On the contrary, Tamanhe’s T2 and Dunnett’s T3 were more conservative but also were generally accurate. Even though Dunnett’s T3 has been considered more conservative than Tamanhe’s T2, in the present research it was similar but slightly more liberal, probably because it is more appropriate for balanced groups (Day & Quinn, 1989; Ruxton & Beauchamp, 2008). Likewise, these differences were minimal except for the smallest sample size (*N* = 30). In this case, Dunnett´s T3 was excessively liberal, demonstrating that is not robust for unequal and small sample sizes (Day & Quinn, 1989). Finally, the post-hoc tests that assume equal variances were excessively conservative in all cases, except Fisher’s LSD. This test has been criticized due to its liberal nature, cause of its lack of correction for multiple comparisons. According to expected, this liberality was manifest in the first, second and third conditions (Meyers et al., 2016). Contrarily, while all tests for the heteroscedastic condition were excessively conservative, Fisher’s LSD was accurate for six of seven sample sizes.

Despite the irrefutable importance of the accuracy of the test, other theoretical issues must be considered also. In this sense, the relative importance of the Type I (and Type II) error rate will be dependent on the specific hypothesis. As Armstrong (2014) pointed out, in certain situations it is preferable to avoid the missing possible effects and then more liberal tests are more appropriate, while in other conditions the objective is to be sure (to extend the possible) of particular effects and then more conservative options are more desirable. For example, following this premise, in an exploratory study for the third condition (homoscedastic and unbalanced), Šidák’s and Hochberg’s GT2 tests are more appropriate, while on the contrary situation, where specific effects interest, Bonferroni’s test is the better option.

Strengths and Limitations

The present research contributes to a deeper knowledge of the appropriateness of different post-hoc tests in particular conditions. More specifically, this work provides researchers with a guide for post-hoc test selection in different four conditions when data scores are normally distributed and group-variance pairing is positive. The obtained results demonstrate the relevance of sample sizes, between-group sizes (balanced vs. unbalanced), and the violation of the homoscedasticity assumption (homoscedastic vs. heteroscedastic) in the choice of post-hoc tests for one-way ANOVA. Contrary to the idea that some tests (e.g., Tukey’s HSD) are appropriate in a variety of situations (Brown, 2005), significant variability was detected through different group sizes in each condition, except for the second condition (heteroscedastic and balanced). Nevertheless, it is important to note that the conclusions yielded from this research are not applicable to every condition, then the researchers should be cautious with the generalization of these results.

These finding must be interpreted under several limitations. First, our data simulation was focused on a particular condition of positive paring between group size and group variance (the smallest group had the smallest variance and the biggest group had the biggest variance). Second, heteroscedasticity conditions had a big ratio variance. Third, the case of the best scenario of normal distribution was analyzed. Fourth, only ten post-hoc tests' accuracy was studied here. In line with these limitations, future research should conduct simulation studies with the same conditions, but with a negative pairing between group size and variance (the smallest group have the biggest variance and the biggest group has the smallest variance) as well as with different variance ratios (see e.g., Blanca et al., 2018; Kirk, 2013). It would be also crucial for future research to test other scenarios (heavy-tailed distributions, outliers, and difference in skewness). Furthermore, future research should study the accuracy of other post-hoc tests different from those analyzed here (e.g., those aimed at controlling familywise error rate without first rejecting the global hypothesis) (Wilcox, 2023) under the same conditions as well as in new study conditions.

Funding: The authors have no funding to report.

Acknowledgments: We would like to express our gratitude to Dr. Juan Ramón Barrada (University of Zaragoza) for his support, comments and suggestions.

Competing Interests: The authors have declared that no competing interests exist.

References

Abdi, H. (2007). The Bonferroni and Sidak corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. SAGE.

Abdi, H., & Williams, L. J. (2010). Tukey's Honestly Significant Difference (HSD) test. In N. Salkind (Ed.), *Encyclopedia of research design*. SAGE.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic & Physiological Optics, 34*(5), 502–508. https://doi.org/10.1111/opo.12131

Ato, M., & Vallejo, G. (2015). *Diseños experimentales en psicología* (2nd ed.). Pirámide.

Bhat, B. R., Badade, M. N., & Aruna Rao, K. (2002). A new test for equality of variances for k Normal Populations. *Communications in Statistics. Simulation and Computation, 31*(4), 567–587. https://doi.org/10.1081/SAC-120004313

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods, 50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology, 31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, A. M. (2005). A new software for carrying out one-way ANOVA post hoc tests. *Computer Methods and Programs in Biomedicine, 79*(1), 89–95. https://doi.org/10.1016/j.cmpb.2005.02.007

Counsell, A., & Harlow, L. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology, 58*(2), 140–147. https://doi.org/10.1037/cap0000074

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L. J., & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23*, 640–647. https://doi.org/10.3758/s13423-015-0913-5

Day, R. W., & Quinn, G. P. (1989). Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs, 59*(4), 433–463. https://doi.org/10.2307/1943075

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association, 75*(372), 796–800. https://doi.org/10.1080/01621459.1980.10477552

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*(1), 39–82. https://doi.org/10.2307/2342435

Fox, J. (2005). The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software, 14*(9), 1–42. https://doi.org/10.18637/jss.v014.i09

Gabriel, K. R. (1978). A simple method of multiple comparisons of means. *Journal of the American Statistical Association, 73*(364), 724–729. https://doi.org/10.1080/01621459.1978.10480084

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1*(2), 113–125. https://doi.org/10.3102/10769986001002113

Grinde, K. E., Arbet, J., Green, A., O'Connell, M., Valcarcel, A., Westra, J., & Tintle, N. (2017). Illustrating, quantifying, and correcting for bias in post-hoc analysis of gene-based rare variant tests of association. *Frontiers in Genetics, 8*, Article 117. https://doi.org/10.3389/fgene.2017.00117

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's Least Significant Difference Test. *Journal of the American Statistical Association, 81*(396), 1000–1004. https://doi.org/10.1080/01621459.1986.10478364

Hochberg, Y. (1974). Some generalizations of the T-method in simultaneous inference. *Journal of Multivariate Analysis, 4*(2), 224–234. https://doi.org/10.1016/0047-259X(74)90015-3

Hochberg, Y. (1975). An extension of the T-method to general unbalanced models of fixed effects. *Journal of the Royal Statistical Society. Series A (General), 37*(3), 426–433. https://doi.org/10.1111/j.2517-6161.1975.tb01557.x

Howell, D. C. (2010). *Statistical methods for psychology.* Wadsworth Cengage Learning.

Howell, J. F., & Games, P. A. (1974). The effects of variance heterogeneity on simultaneous multiple-comparison procedures with unequal sample size. *British Journal of Mathematical & Statistical Psychology, 27*(1), 72–81. https://doi.org/10.1111/j.2044-8317.1974.tb00529.x

IBM. (2016). *IBM SPSS Statistics for Windows, Version 24.0* [Computer software]. IBM.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 96*(3), 589–596. https://doi.org/10.1037/0033-2909.96.3.589

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook.* Pearson Prentice-Hall.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350–386. https://doi.org/10.3102/00346543068003350

Kim, H. Y. (2015). Statistical notes for clinical researchers: Post-hoc multiple comparisons. *Restorative Dentistry & Endodontics, 40*(2), 172–176. https://doi.org/10.5395/rde.2015.40.2.172

Kirk, R. E. (2013). *Experimental design. Procedures for the behavioral sciences* (4th ed.). SAGE.

Lee, S., & Lee, D. K. (2020). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology, 71*(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

McHugh, M. L. (2011). Multiple comparison analysis testing ANOVA. *Biochemia Medica, 21*(3), 203–209. https://doi.org/10.11613/BM.2011.029

Meyers, L. S., Gamst, G., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation* (3rd ed.). SAGE.

Pedrosa, I., Juarros-Basterretxea, J., Robles-Fernández, A., Basteiro, J., & García- Cueto, E. (2015). Goodness of fit tests for symmetric distributions, which statistical should I use? *Universitas Psychologica, 14*(1), 15–24.

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*(1), 1–11. https://doi.org/10.1027/1614-2241/a000034

Ruxton, G. D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral Ecology, 19*(3), 690–693. https://doi.org/10.1093/beheco/arn020

Sarstedt, M., & Mooi, E. (2019). Hypothesis testing and ANOVA. In M. Sarstedt & E. Mooi (Eds.), *A concise guide to market research: The process, data, and methods using IBM SPSS statistics* (pp. 151–208). Springer. https://doi.org/10.1007/978-3-662-56707-4_6

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40*(1–2), 87–110. https://doi.org/10.1093/biomet/40.1-2.87

Sedgwick, P. (2012). Multiple significance tests: The Bonferroni correction. *BMJ (Clinical Research Ed.), 344*, Article e509. https://doi.org/10.1136/bmj.e509

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*(1), 561–584. https://doi.org/10.1146/annurev.ps.46.020195.003021

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. https://doi.org/10.1080/00949655.2012.675336

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*(318), 626–633. https://doi.org/10.2307/2283989

Spjøtvoll, E., & Stoline, M. R. (1973). An extension of the T-method of multiple comparison to include the cases with unequal sample sizes. *Journal of the American Statistical Association, 68*(344), 975–978. https://doi.org/10.1080/01621459.1973.10481458

Tamhane, A. C. (1977). Multiple-comparisons in Model I one-way ANOVA with unequal variances. *Communications in Statistics, 6*(1), 15–32. https://doi.org/10.1080/03610927708827466

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association, 74*(366), 471–480. https://doi.org/10.2307/2286358

Tukey, J. W. (1953). *The problem of multiple comparisons,* Unpublished manuscript, Princeton University.

Wilcox, R. R. (2023). *A guide to robust statistical methods.* Springer. https://doi.org/10.1007/978-3-031-41713-9

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical & Statistical Psychology, 57*(1), 173–181. https://doi.org/10.1348/000711004849222

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.