Brazilian Journal of Biometrics, **42**, 20–49 (2024)
doi:10.28951/bjb.v42i1.641

ARTICLE

Ferbat's test: A Monte Carlo multiple comparison procedure with a control

Ben Dêivide de Oliveira Batista*[,1] and Daniel Furtado Ferreira[2]

[1]Statistics, Phisycs and Mathematics Department, Federal University of São João del-Rei, Campus Alto Paraopeba, Ouro Branco - MG, Brazil.
[2]Statistics Department, Institute of Exact and Technological Sciences, Federal University of Lavras, Lavras - MG, Brazil.
*Corresponding author. Email: ben.deivide@gmail.com

(Received: October 20, 2022; Revised: June 5, 2023; Accepted: June 30, 2023; Published: March 15, 2024)

Abstract

The present work presented an alternative MCC to the Dunnett's test, called Ferbat's test. The test replaced the root of the mean square of the residue used within Dunnett's test with another non-biased σ estimator. The distribution of the test statistics was determined by simulation using the Monte Carlo method. Comparing the performance evaluation of these two tests, the Ferbat's test performed better in some scenarios, such as control of the experimentwise error rate for all simulated situations and higher power when the number of treatments was small and when the number of replications increases. In the other evaluation situations, the tests presented equivalent performance.

Keywords: Experimentwise error rate; Power; Simulation.

1. Introduction

Many experiments in applied sciences aim at comparing the treatments to a control or standard treatment also referred to as many-to-one comparisons. The control treatment represents a reference in the literature for the studied factor. For example, in the pharmaceutical industry there is a drug which is always used to combat a headache due to its efficiency. As a market competition, there are companies that develop new drugs and they want to test them against the standard treatment to know the relative efficiency of these products, i.e., if the performance of these drugs are equal to or greater than the drug already used in the market (control).

The method used to answer to these questions in statistics is the multiple comparison procedure (MCP). The advantage of using a MCP is that its construction is based on the control of the global significance level for the simultaneous inferences. In particular, to compare new drugs with a control

is used the multiple comparison with a control (MCC) (Hsu, 1996). This is a particular case, since the general case of the MCPs a total of $n(n-1)/2$ comparisons are made, where n is the number of treatment levels, and this produce results in confidence intervals which are wider than necessary and, also,significance small significance levels in the tests. The MCC performed only n comparisons, that is, it performs the comparison (intervals or tests) between all new treatments with the control. This tends to present tests with more accurate results, and as asserted by Shaffer (1977), the MCCs tend to produce more powerful tests for this cases.

In this article, we will be interested in developing a new MCC. Many attempts are made in the literature in search of an ideal test, that shows an appropriate control of the type I error and high power. We can highlight the Dunnett's test (Dunnett, 1955), in cases where samples are random and independent, from random variables with normal distribution. Other MCCs can be presented in Dunnett (1994), Hsu (1996), Benjamini *et al.* (2004), Dmitrienko *et al.* (2010), Bretz *et al.* (2011), Westfall *et al.* (2011), among others.

We will show the idea behind of the two-sided Dunnett's test for the case of balanced data with normal distribution and homoscedasticity in the sequence. Of course, this test extends to unbalanced data and heteroscedasticity conditions and we will show this restriction for this test, as these conditions will be the basis for developing the multiple comparison procedure created in this paper.

Let be a random sample $Y_{11}, Y_{12}, \ldots, Y_{1r}, Y_{21}, Y_{22}, \ldots, Y_{2r}, \ldots, Y_{i1}, Y_{i2}, \ldots, Y_{ij}, \ldots, Y_{ir}, \ldots, Y_{n1}, Y_{n2}, \ldots, Y_{nr}, Y_{(n+1)1}, Y_{(n+1)2}, \ldots, Y_{(n+1)r}$, where Y_{ij} is the random observation of the ith treatment in the jth replication, $i = 1, 2, \ldots, (n+1)$ and $j = 1, 2, \ldots, r$. The $(n+1)$th treatment is the standard (control) treatment. The sample mean of the ith treatment is given by

$$\bar{Y}_{i.} = \frac{\sum_{j=1}^{r} Y_{ij}}{r} = \frac{Y_{i.}}{r}. \tag{1}$$

Without loss of generality, we will consider that this sample was submitted to an analysis of variance, according to the model given by

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}, \tag{2}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\mu_i = \mu + \tau_i$ is the mean of the ith treatment. Thus, the mean square of the error (*MSE*), estimator of error variance σ^2, is given by

$$MSE = \frac{\sum_{i=1}^{n+1}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^2}{(n+1)(r-1)} = \frac{\sum_{i=1}^{n+1}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^2}{\nu}, \tag{3}$$

where $\nu = (n+1)(r-1)$ is the error degrees of freedom.

The two-sided Dunnett's test is used to determine which treatments differ of a control treatment. For balanced experiments, this test provides the following confidence set of the set of simultaneous range of $100(1-\alpha)\%$ confidence for the all differences $\mu_k - \mu_{n+1}$, $k \neq n+1$, between the mean, μ_k, of each treatment in test and the mean of the control treatment, μ_{n+1}, given by

$$\bar{Y}_k - \bar{Y}_{n+1} \pm |d| \sqrt{MSE \times \left(\frac{2}{r}\right)}, \quad k = 1, 2, \ldots, n, \tag{4}$$

where $|d|$ is the solution given by

$$\int_0^{\infty} \int_{-\infty}^{\infty} \left[\Phi(z + \sqrt{2}|d|s) - \Phi(z - \sqrt{2}|d|s)\right]^n \phi(z) dz f_S(s; \nu) ds = 1 - \alpha, \tag{5}$$

$\Phi(.)$ and $\phi(.)$ are, respectively, the distribution and density functions from the standard normal distribution and $f_S(s;\nu)$ is the density function of $S = \frac{\sqrt{MSE}}{\sigma}$, given by

$$f_S(s;\nu) = \frac{\nu^{\nu/2}}{\Gamma(\nu/2)2^{\nu/2-1}} s^{\nu-1} e^{-\nu s^2/2}, \quad s \geq 0. \tag{6}$$

It can be seen that $|d|$ is the upper limit $100\alpha\%$ of the maximum module of the multivariate t distribution with common correlation $\rho = 0.5$ and ν degrees of freedom (Dean *et al.,* 2017) (studentized maximum module distribution). To determine other cases to obtain d distribution see Dunnett (1964). A more complete work on the non-central multivariate t distribution can be found in Broch & Ferreira (2013b) and Broch & Ferreira (2013a), where these same authors made available in CRAN a package R called nCDunnett (Broch & Ferreira, 2015) by computing all these results, including determining values of d.

The test statistic for applying the two-sided Dunnett's test of size α, under the null hypothesis $H_0 : \mu_k - \mu_{n+1} = 0 \; \forall \; k = 1, 2, \ldots, n$, is given by

$$|D_k| = \frac{|\bar{Y}_k - \bar{Y}_{n+1}|}{\sqrt{\frac{2MSE}{r}}}, \quad k = 1, 2, \ldots, n. \tag{7}$$

If $|D_k| > |d|$, where $|d|$ is the critical level of the studentized maximum module distribution at $100\alpha\%$, the null hypothesis should be rejected. The Dunnett's test performs n simultaneous comparisons with the control of the overall significance level, as can be observed in Hochberg & Tamhane (1987), Hsu (1996), Miller (1981), and Dickhaus (2014), among others. This control since an exact distribution is used for the statistic in the equation (7), which takes into account the multiplicity effect in the simultaneous comparisons.

A pioneering study of Daly (1946) proposed the use of the sample range in place of root-mean-square as an estimator of σ, in an analogue of Student's t-test. The test was called of u-test. Lord (1947) and Patnaik (1950) went further by using the sample range or mean range as a σ estimator rather than root-mean-square determined from the sample. They claim that the efficiency of range estimates of standard deviation is, of course, always less than the root-mean-square estimates. But Davies & Pearson (1934) and Pearson & Haines (1935) indicated that this efficiency is not accentuated for samples that not greatly in excess of about 10.

Lord (1950) and Daly (1946) evaluated the power of u-test in some experimental scenarios. In addition, it has been shown that in spite of some loss of precision, this did not influence in practical applications, being even compensated by the greater ease of calculation when compared with the t-test. Daly (1946) stated that the small sample size, the power of u-test was equivalent to that of the t-test. However, Lord (1950) stated that, in general, the power of the u-test shows a slightly lower performance than t-test. One of the explanations for the low performance of the u-test, may be because the distribution of the sample mean range is not obtained accurately, except in special cases, as observed in Patnaik (1950). Another point cited by the same author is that this statistic is less efficient than the root-mean-square estimate.

Of course, as already well documented in the literature, the t-test does not control the overall significance level in simultaneous comparisons. With the results presented by the authors already mentioned, it suggests that the u-test also does not control.

Therefore, the main ideas about the u-test is outline below. The Statistic of u-test is given by

$$U = U(i,j) = \frac{Y}{\bar{W}/d_r}, \tag{8}$$

where Y is random variable distributed normally about a mean zero and standard deviation σ, without loss of generality. The d_r is some constant to eliminated the bias of a new estimator of the standard deviation that will be defined latter. The $\bar{W}$ is the mean of n ranges W, obtained from n independent samples or subgroups, each containing r observations. We define the sample range of group i by $W_i = \max_j(Y_{ij}) - \min_j(Y_{ij})$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, r$, and mean range $\bar{W} = \sum_{i=1}^{n} W_i/n$. The probability density function of W from the normal population with mean μ and standard deviation σ is given by

$$f_W(w) = \int_{-\infty}^{\infty} r(r-1)\phi_\sigma(y)\phi_\sigma(w+y)\left[\Phi_\sigma(w+y) - \Phi_\sigma(y)\right]^{r-2} dy, \tag{9}$$

as shown by Ahsumullah *et al.* (2013), where ϕ_σ e Φ_σ are the density and distribution functions, respectively, from the normal population with mean 0 and standard deviation σ.

Ahsumullah *et al.* (2013) also presents the distributions of the maximum $(Y_{(r)})$ and minimum $(Y_{(1)})$ that are given, respectively, by

$$f_{Y_{(r)}}(y) = r\phi_\sigma(y)[\Phi_\sigma(y)]^{r-1}, \tag{10}$$

and

$$f_{Y_{(1)}}(y) = r\phi_\sigma(y)[1 - \Phi_\sigma(y)]^{r-1}. \tag{11}$$

We present (10) and (11) according to the parental distribution of interest in this specific case of normality. The expected value of W can be expressed by

$$\begin{aligned} E[W] &= E[Y_{(r)}] - E[Y_{(1)}] \\ &= \int_{-\infty}^{\infty} y f_{Y_{(r)}}(y)dy - \int_{-\infty}^{\infty} y f_{Y_{(1)}}(y)dy. \end{aligned} \tag{12}$$

By the symmetry about 0 of $\phi_\sigma(.)$, the $E[W]$ reduce to

$$E[W] = 2\int_{-\infty}^{\infty} y f_{Y_{(r)}}(y)dy. \tag{13}$$

Finally, doing the transformation $Z = Y/\sigma$ and rewriting $E[W]$ by μ_W, the expected value of W is given by

$$\mu_W = \sigma d_r, \tag{14}$$

where $d_r = 2r\int_{-\infty}^{\infty} z\phi(z)[\Phi(z)]^{r-1}dz$, $\phi(z)$ and $\Phi(z)$ are density and distribution functions from the standard normal population. The equation (14) shows that $\bar{W}/d_r$ is an unbiased estimator for σ.

To determine the distribution of U, this statistic will be rewrite by

$$Q = \frac{U}{d_r} = \frac{Y}{\bar{W}}, \tag{15}$$

and then multiplying by the corresponding value of d_r to obtain the percentage points of the U in the resulting distribution (Lord, 1947).

Daly (1946) proved that Y and W are statistically independent. Also Patnaik (1950) stated that if it is known that the sample mean is statistically independent of the range in the sample, then Y

is, therefore, also independent of the mean range $\bar{W}$. Thus the distribution of Q, equation (15), is given by

$$f_Q(q) = \int_{\infty}^{\infty} \int_{0}^{\infty} f_{\bar{W}}(\bar{w}) f_Y(y) d\bar{w} dy. \tag{16}$$

The analytical form of the U distribution was obtained only for some cases. In the others, Lord (1947) and Patnaik (1950) used approximations using Gaussian quadrature.

Thus, restricted to the case of comparisons of treatments with a control, this work will develop a two-sided test based on the Dunnett's test, using the same ideas that were proposed by Daly (1946) and Lord (1947) to modify the t-test. The initial idea to construct a two-sided test is that it can be extended to encompass all contrasts without additional assumptions, as stated by Shaffer (1977) for the two-sided Dunnett's test. For this, due to the problems already presented previously, the new multiple comparison procedure will be done by Monte Carlo simulation.

Therefore, the objectives of this work are to develop an MCC test, called Ferbat, and evaluate the Ferbat's test performance against Dunnett's test performance. The latter will also be evaluated in the present work, using the literature results as well as the results found in our simulations study.

2. Matherials and Methods

2.1 Ferbat's test

To develop the Ferbat's test, a Monte Carlo two-sided test, we consider the family of the n simultaneous hypotheses defined by expression 17,

$$\begin{cases} H_0 : \mu_k = \mu_{n+1}, & k = 1, 2, \ldots, n, \\ H_1 : \mu_k \neq \mu_{n+1}, & \text{for some } k. \end{cases} \tag{17}$$

For testing this hypothesis let be a random sample $Y_{11}, Y_{12}, \ldots, Y_{1r}, Y_{21}, Y_{22}, \ldots, Y_{2r}, \ldots, Y_{i1}, Y_{i2}, \ldots, Y_{ij}, \ldots, Y_{ir}, \ldots, Y_{n1}, Y_{n2}, \ldots, Y_{nr}, Y_{(n+1)1}, Y_{(n+1)2}, \ldots, Y_{(n+1)r}$, where Y_{ij} is the random observation of the ith treatment in the jth replication, $i = 1, 2, \ldots, (n + 1)$ and $j = 1, 2, \ldots, r$. The treatment $n + 1$ is the standard (control) treatment. The parental distribution was considered the normal distribution $N(\mu_i, \sigma^2)$, $\forall\ i = 1, 2, \ldots, n + 1$. The sample mean of the ith treatment is given by

$$\bar{Y}_{i.} = \frac{\sum_{j=1}^{r} Y_{ij}}{r} = \frac{Y_{i.}}{r}. \tag{18}$$

Under the null H_0 hypothesis, $\mu_i = \mu$, $\forall\ i = 1, 2, \ldots, n + 1$, the statistic for application of Ferbat's two-sided test is given by

$$FB^* = \frac{|\bar{Y}_{k.} - \bar{Y}_{(n+1).}|}{\sqrt{\frac{2\hat{\sigma}}{r}}}, \quad k = 1, 2, \ldots, n, \tag{19}$$

where $\hat{\sigma} = (\bar{W}/d_r^*)^2$, $\bar{W}$ is the mean range and d_r^* is the modified constant given by

$$d_r^* = \begin{cases} 2(r + 0.10r) \int_{-\infty}^{\infty} z\phi(z)[\Phi(z)]^{(r+0.10r)-1} dz, & \text{if } r \leq 10 \\ 2(r + 0.23r) \int_{-\infty}^{\infty} z\phi(z)[\Phi(z)]^{(r+0.23r)-1} dz, & \text{if } r > 10. \end{cases} \tag{20}$$

The expression 20 was a result of trial and error simulation to get a more powerful test. Thus, assuming normality, the algorithm for the application of the test follows the steps below:

1. A Monte Carlo sample of a completely randomized experimental design with $n + 1$ treatments and r replications generate from $N(0, 1)$ distribution. The null hypothesis was imposed by considering all treatments means equal to the same value, in the case, equal to 0. The variance σ^2 was choose as 1. Since the test statistic is an ancillary statistic there is no loss of generality to chose the common mean as 0 and the variance as 1. Consider the Monte Carlo sample is given by $X_{11}, X_{12}, \ldots, X_{1r}, X_{21}, X_{22}, \ldots, X_{2r}, \ldots, X_{i1}, X_{i2}, \ldots, X_{ij}, \ldots, X_{ir}, \ldots, X_{n1}, X_{n2}, \ldots, X_{nr}$, $X_{(n+1)1}, X_{(n+1)2}, \ldots, X_{(n+1)r}$, where X_{ij} is the random observation of the ith treatment in the jth replication, $i = 1, 2, \ldots, (n + 1)$ and $j = 1, 2, \ldots, r$ generate from the $N(0, 1)$ distribution. The sample mean of the ith treatment in the Monte Carlo simulation is given by

$$\bar{X}_{i.} = \frac{\sum_{j=1}^{r} X_{ij}}{r} = \frac{X_{i.}}{r}. \tag{21}$$

2. Compute the range for each treatment, that is, $W_i = \max_i(\bar{X}_{i.}) - \min_i(\bar{X}_{i.})$, $i = 1, 2, \ldots, n, n + 1$;
3. Compute the mean range of the experiment, given by $\bar{W} = \sum_{i=1}^{n+1} W_i/(n + 1)$;
4. Compute the Monte Carlo test statistic FB given by

$$FB = \frac{|\max_k(\bar{X}_{k.} - \bar{X}_{(n+1).})|}{\sqrt{\frac{2\hat{\sigma}}{r}}}, \quad k = 1, 2, \ldots, n, \tag{22}$$

where $\hat{\sigma} = (\bar{W}/d_r)^2$, $\bar{W}$ is the mean range and d_r is the constant given in expression (14).

5. Repeat the steps from 1 to 4 B times. Store the computed value of the Monte Carlo test statistic computed in (22) in each step of the simulations along with the values previously obtained, if any. In this paper we used $B = 50,000$.
6. Compute the upper quantile α of Monte Carlo distribution of statistic in (22), defined by $fb_{(\alpha,n,r)}$, based in the step 5.
7. Calculate the statistic of Ferbat's test, equation (19), to evaluate each k simultaneous hypothesis. Make a decision to reject the null hypothesis if $FB^* \geq fb_{(\alpha,n,r)}$ at the nominal significance level of α.

The implementation of d_r, given by equation (14), has no closed solution. For this, we will use 64 points of Gauss-Legendre quadrature by using a change of the variable given by $z = t/(1 - t^2)$, to arrive at the following approximation expressed by

$$d_r \approx 2r \sum_{i=1}^{64} w_i \left(\frac{t_i}{1 - t_i^2}\right) \Phi\left(\frac{t_i}{1 - t_i^2}\right) \left[\Phi\left(\frac{t_i}{1 - t_i^2}\right)\right]^{r-1} \frac{1 + t_i^2}{(1 + t_i^2)^2}, \tag{23}$$

where t_i e w_i are the node and weight of Gauss-Legendre quadrature. The procedure for computing the modified constant (d_r^*) was performed similarly.

2.2 Scenarios for the evaluation of Ferbat's test

Two strategies were considered in this work. The first was to evaluate the experimentwise error rates (EER or $\hat{\alpha}$) and the second was to evaluate the power of Ferbat and Dunnett tests. In both cases, Monte Carlo simulation was used for this purpose.

In each simulation the Ferbat and Dunnett tests were applied at a pre-established nominal significance level α checking whether or not the null hypothesis was rejected. This process was repeated $N = 2000$ times and the proportion of experiments with at least one incorrect decision in the first

case refers to the empirical EER and in the second case the proportion of rejections that are correct refers to empirical power.

To verify the effect of Monte Carlo simulation error in the EER, the exact binomial test with a confidence coefficient of 99% was used to test the hypotheses $H_0 : \alpha = 5\%$ versus $H_1 : \alpha \neq 5\%$ or $H_0 : \alpha = 1\%$ versus $H_1 : \alpha \neq 1\%$. If the null hypothesis is rejected and the empirical EER is considered to be significantly ($p - value < 0,01$) lower than the nominal significance level α, the test will be considered conservative. If the empirical EER is considered significantly ($p - value < 0.01$) higher than the nominal level, the test will be considered liberal. If the observed value of the empirical EER is non-significantly ($p - value > 0.01$) different of the nominal significance level, the test will be considered exact (Oliveira & Ferreira, 2010).

Considering γ the number of null hypotheses rejected in 2000 Monte Carlo simulations, for a nominal significance level of α, and, also considering the relationship between the F and binomial distributions (Leemis & Trivedi, 1996), with probability of success $p = \alpha$, the statistic of test is given by

$$F = \left(\frac{\gamma + 1}{N - \gamma}\right)\left(\frac{1 - \alpha}{\alpha}\right), \tag{24}$$

that under H_0 has an F distribution with $\nu_1 = 2(N - \gamma)$ and $\nu_2 = 2(\gamma + 1)$ degrees of freedom. If $F < F_{0.005}$ or $F > F_{0.995}$, the null hypothesis must be rejected at the significance level of 1% probability, where $F_{0.005}$ and $F_{9.995}$ are, respectively, the 0.005 and 0.995 quantiles of the F distribution with ν_1 and ν_2 degrees of freedom (Oliveira & Ferreira, 2010).

In both steps the data were simulated according to the statistical model expressed in (2), where μ is a general constant settled in 100 for all cases, without loss of generality, τ_i is the effect of the ith treatment and ϵ_{ij} is the effect of a identical and independently distributed random error with mean 0 and common variance σ^2. Also, it was assumed that $\sigma^2 = 100$, without loss of generality. Yet $i = 1, 2, \ldots, n, n + 1$ and $j = 1, 2, \ldots, r$, where r is the number of replications.

In the first step of evaluating the EER, the treatment effects τ_i were considered equal to 0 for all i, $i = 1, 2, \ldots, n, n+1$. Thus, the data were generated under a complete null hypothesis, that is, with all treatments having the same parametric mean. The probability of empirical EER ($\hat{\alpha}$) was estimated by the proportion of experiments with at least one difference detected incorrectly in relation to the total of simulated experiments given by

$$\hat{\alpha} = \frac{\sum_{k=1}^{N} I(E_k = 1)}{N}, \tag{25}$$

where E_k is a binary variable that assumes the value 1 if at least one type I error occurred in the kth experiment and 0, otherwise, for $k = 1, 2, \ldots, N$ and $I(E_k = 1)$ is the indicator function that returns 1 if the conditions is true and 0, otherwise.

In the second step the power was evaluated. Therefore, the treatment effects were simulated in two cases. The first is called of complete alternative hypotheses H_1 and the second of partial null hypotheses H_{0_p}. Besides that, each corresponds to a homogeneous and heterogeneous scenario, respectively.

Thus, the power in the first case, that is, under complete alternative hypotheses H_1 in homogeneous scenario, the control treatment effect was considered equal to 0, $\tau_{n+1} = 0$. The others treatments was fixed and presented by

$$\tau_{(n+1)-i} = \tau_{n+1} + \delta\frac{\sigma}{\sqrt{r}}, \quad \text{for } i = 1, 2, \ldots, n, \tag{26}$$

where δ = 1, 2, 4, 8, 16 and 32, representing the number of standard errors of difference between the mean of a specific treatment effect in the homogeneous group and the mean of the control treatment. Thus, the power was computed by the ratio of rejections among the mean of the control treatment and the means of the other treatments involving multiples of δ, relative to the total number of comparisons involving this difference. We have n comparisons per experiment and a total of $n \times N$ comparisons. The ratio between the total number of rejection and the total number of comparisons corresponds to the estimated power to detect δ standard errors of the difference between the mean of the control treatment and the means of the other treatments.

In the second case, that is, the power under the complete alternative hypotheses (H_1) in heterogenous scenario, the control treatment effect was considered equal to 0, $\tau_{n+1} = 0$. The others was fixed and set by

$$\tau_{(n-i)+1} = \tau_{(n-i)+2} + \delta\frac{\sigma}{\sqrt{r}}, \quad \text{for } i = 1, 2, \ldots, n, \tag{27}$$

where δ = 1, 2, 4, 8, 16 or 32, representing the number of standard errors of the difference between means to specify the consecutive treatments effect compared with the control treatment. Thus, the power was computed by the ratio of rejections among means involving multiples of δ, relative to the total number of comparisons involving this difference. We have n comparisons per experiment between the mean of the control treatment and the means of the other treatments in each experiment. The ratio between the total of rejection for fixed difference and the number of simulations N corresponds to the power to detect that multiple of δ standard errors of the difference between the mean of the control treatment and the mean of that specific treatment.

The second option for the study of power under the H_{0_p} partial null hypothesis involved a simulation of two groups of means, with $k_1 = \lfloor (n+1)/2 \rfloor$ e $k_2 = (n+1) - k_1$ means in each, where $\lfloor x \rfloor$ refers to the largest integer less than or equal to x. The control treatment is allocated to the first group. The means of the first group were all the same, for which the effects were set to $\tau_i = 0$, $i = 1, 2, 3, \ldots, k_1$, without loss of generality. The second group, with k_2 treatments, had its effects also specified by equation (26) (homogeneous scenario) and equation (27) (heterogenous scenario), with the variation of i replaced by $i = 1, 2, \ldots, k_2$. In Table 1, we present a summary table of the scenarios in which the tests will be evaluated.

Table 1. Scenarios of simulation for performance evaluation of Ferbat and Dunnett tests

Scenario	Simulation
1	Under complete null hypothesis (H_0)
2	Under partial hypothesis (H_{0_p}) in a homogeneous scenario
3	Under partial hypothesis (H_{0_p}) in a heterogeneous scenario
4	Under complete hypothesis (H_1, alternative hypothesis) in a homogeneous scenario
5	Under complete hypothesis (H_1, alternative hypothesis) in a heterogeneous scenario

2.3 Justifications for the use of the modified constant d_r^* on the Ferbat test

After implementing the Ferbat's test, a few initial results showed that the empirical type I error was statistically less than or equal to the adopted overall nominal significance level α. This sometimes provided a conservative test. In this way, the test tended to have less power. These results showed that in terms of performance the Dunnett's test was superior, since the Dunnett's test has an exact distribution for its statistics, and the evaluation results evidenciated this, as will be shown later.

Thus, as a way of trying to increase the power of the Ferbat's test and to control the global nominal significance level, we modify the constant d_r of the statistic as showed in (20). However, to generate the critical points of Ferbat's test we use the Monte Carlo distribution of the statistic without this change. This is common in multiple comparison procedures. Tukey (1953) developed the Tukey's test with has empirical type I error rates less than the overall significance level, that means it is a conservative test. After that work, Duncan (1955) tried to do some change in the test to improve his performance, as also Keuls (1952). These changes led to the Duncan and SNK tests, respectively.

Another argument for modifying the d_r constant is the fact that the standard deviation estimator $\bar{W}/d_r$ is less efficient than the root-mean-square of the residue to estimate the population standard deviation. Given these considerations, we will present below the performance evaluation of the Ferbat's test and compare it with the Dunnett's test. The performance evaluation of this last test was also simulated in this work. Besides that, we also will use results presented in the literature on this test.

3. Results and Discussion

3.0.1 Performance of tests in scenario 1

The results of the empirical EER results for the performance of the tests obtained by simulation in scenario 1 (Table 1) are shown in Tables 2 and 3. The exact binomial test at the nominal significance level of 1% was applied in each simulated configurations. The EER receive the identification (++) for liberal test and receive the identification (—), when it was considered conservative. The former situation is undesirable because it shows that the test shows type error rates greater than a the nominal significance level.

We observed that the two tests controlled appropriately the EER at the nominal significance level in all Monte Carlo simulation cases. We did not achieve any non standard pattern in the behaviour of EER when the number of treatment or the number of replications change. From now on, we will represent the number of treatments as $m = n + 1$, which represents the number of treatments in the experiment including the control treatment.

Unlike what was presented by Sousa *et al.* (2012), the Dunnett's test controls the overall significance level in the scenario 1 (Table 3), as can be confirmed by Carmer & Walker (1985), Dunnett (1994), Dunnett (1964). Sousa *et al.* (2012) states that the Dunnett's test was liberal, considering 200 Monte Carlo simulations only with 32 treatments and 4 replicates under complete null hypothesis (H_0). They found the experimentwise error rate of Dunnett's test of 0.092^{++}, a value that exceeds the critical point of the exact binomial test at a significance level of 1% of probability. One of the explanations for this high EER value is the low number of simulations, since the α value may be overinflated due to the Monte Carlo error. Despite the widespread reference of Dunnett's test control to experimentwise error rate, few studies in the literature present the results as presented in Table 3. Conagin *et al.* (2008) evaluated several MCPs, one of which was the one-sided Dunnett's test. The EER of this test was 4.3% in an experiment with 8 treatments, 4 replications and a nominal significance level of 5%, repeated 400 times. Although the evaluation methodology of this study was slightly different from this work, the results are equivalent.

Table 2. Experimentwise type I error rates of Ferbat's test as a function of the number of treatments (m), the number of replications (r) and the nominal significance level α under a complete null hypothesis H_0, evaluated by the exact binomial test with a confidence coefficient of 99% of probability

	α					
	0.01			0.05		
m \ r	4	10	20	4	10	20
5	0.0140	0.0165	0.0155	0.0544	0.0440	0.0575
10	0.0115	0.0110	0.0135	0.0510	0.0555	0.0580
20	0.0155	0.0130	0.0150	0.0570	0.0435	0.0550
40	0.0135	0.0145	0.0125	0.0610	0.0500	0.0610
100	0.0155	0.0115	0.0160	0.0470	0.0370	0.0410

* The symbol "- -" indicates that EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The symbol "++" indicates that EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

Table 3. Experimentwise type I error rates of Dunnett's test as a function of the number of treatments (m), the number of replications (r) and the nominal significance level α under a complete null hypothesis H_0, evaluated by the exact binomial test with a confidence coefficient of 99% of probability

	α					
	0.01			0.05		
m \ r	4	10	20	4	10	20
5	0.0105	0.0080	0.0065	0.0475	0.0545	0.0530
10	0.0095	0.0090	0.0090	0.0570	0.0545	0.0435
20	0.0080	0.0105	0.0145	0.0500	0.0550	0.0535
40	0.0105	0.0080	0.0075	0.0490	0.0460	0.0555
100	0.0120	0.0110	0.0095	0.0460	0.0480	0.0445

* The symbol "- -" indicates that EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The symbol "++" indicates that EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

3.0.2 Performance of tests in scenario 2

We did not find complete studies in the literature for performance of two-sided Dunnett's test in scenarios 2 to 5. Table 1 shows results of the condition of homogeneous and heterogeneous environments in the simulation under partial null hypothesis (H_{0_p}). This condition, under the partial null hypothesis the scenarios, is more realistic for practical applications.

What we actually have are statements that this test controls the experimentwise error rate. However, the results are not presented in full. In Tables 4 and 5 we present the EER of Ferbat and Dunnett tests with m treatments, r replications, δ standards error and significance level α, to the scenario 2. In general, the tests controlled the nominal level, and are, in certain circumstances, conservative.

The Ferbat's test is conservative for $\alpha = 0.05$ and for $\alpha = 0.01$ when $m = 5$. For $\alpha = 0.01$, in the other simulation settings ($m > 5$) the test shows exact size controlling the nominal significance level (Table 4). The Dunnett's test was conservative in almost all scenario 2 for both nominal significance levels. Only when $m = 100$ and $\alpha = 0.01$ the Dunnett's test has exact size with control of the EER at nominal significance level (Table 5). This result is interesting to show that the Dunnett's test was conservative under partial null hypothesis (H_{0_p}), even though the results found in the literature always state that it controls the nominal significance level.

One explanation of why the Dunnett's test was conservative in most cases of partial null hypothesis in scenario 2, is that the distribution of the test statistic is under the overall null hypothesis. This implies that the distribution of the test statistic assumes that the m means are equal in its formulation, when in reality there is a smaller number than this. Thus, the upper quantiles of the distribution are larger and more difficult to overcome by the treatments that are equal to the control. In the case of Ferbat's test this situation are mitigated due to the modified constant d_r^* in its statistic.

Comparing the results of Tables 4 and 5 with the results in Tables 7 and 9, we noticed that the tests had lower EER in the homogeneous circumstance (scenario 2) than when evaluated in the heterogeneous case (scenario 3). However, the Ferbat's test when evaluated for an $\alpha = 0.01$ showed better control of the nominal significance level in both cases.

We can also observe in Tables 4 and 5 that only the nominal significance level influenced the EER of Dunnett and Ferbat tests, that is, the performance of tests was not influenced by the number of treatments, the number of replications and the number of standard errors. Another assessment in scenario 2 was the power. In Figure 1, we show the power in an experiment with 4 replications. Note that the power of the tests tends to decrease with increasing the number of treatments, being the Ferbat's test reaching higher levels of power when m is small and equivalent at the power of Dunnett's test as m increases (Figure 2). But when the number of standard errors increases, the power also increases. The tests practically reach power equal to 1 when the difference between means equals 8 standard errors.

Table 4. Experimentwise type I error rates of Ferbat's test under the partial null hypothesis (H_{0_p}) as a function of number of replications (r), number of treatment m, at the significance level α and with δ standard errors, in scenario 2

		α					
		0.01			0.05		
m	r / δ	4	10	20	4	10	20
5	1	0.0020—	0.0040—	0.0060	0.0215—	0.0185—	0.0125—
	2	0.0040—	0.0030—	0.0035—	0.0190—	0.0180—	0.0200—
	4	0.0050—	0.0045—	0.0005—	0.0235—	0.0125—	0.0240—
	8	0.0055	0.0035—	0.0045—	0.0130—	0.0180—	0.0185—
	16	0.0035—	0.0055	0.0035—	0.0205—	0.0235—	0.0235—
	32	0.0070	0.0040—	0.0045—	0.0165—	0.0100—	0.0240—
10	1	0.0100	0.0090	0.0070	0.0295—	0.0415	0.0340—
	2	0.0090	0.0070	0.0070	0.0400	0.0290—	0.0265—
	4	0.0085	0.0105	0.0100	0.0345—	0.0295—	0.0340—
	8	0.0120	0.0090	0.0055	0.0400	0.0230—	0.0320—
	16	0.0080	0.0040—	0.0090	0.0275—	0.0385	0.0365—
	32	0.0090	0.0080	0.0100	0.0345—	0.0315—	0.0280—
20	1	0.0060	0.0080	0.0120	0.0400	0.0310—	0.0350—
	2	0.0090	0.0080	0.0100	0.0370—	0.0295—	0.0370—
	4	0.0110	0.0110	0.0090	0.0320—	0.0385	0.0355—
	8	0.0095	0.0065	0.0080	0.0330—	0.0310—	0.0395
	16	0.0090	0.0090	0.0110	0.0450	0.0285—	0.0370—
	32	0.0115	0.0095	0.0110	0.0375—	0.0340—	0.0300—
40	1	0.0070	0.0065	0.0055	0.0385	0.0280—	0.0345—
	2	0.0095	0.0050—	0.0070	0.0445	0.0365—	0.0315—
	4	0.0125	0.0070	0.0060	0.0390	0.0295—	0.0320—
	8	0.0120	0.0070	0.0080	0.0375—	0.0280—	0.0430
	16	0.0125	0.0130	0.0110	0.0410	0.0305—	0.0370—
	32	0.0090	0.0105	0.0090	0.0410	0.0395	0.0355—
100	1	0.0135	0.0065	0.0090	0.0360—	0.0320—	0.0330—
	2	0.0145	0.0090	0.0110	0.0375—	0.0325—	0.0360—
	4	0.0115	0.0085	0.0105	0.0385	0.0330—	0.0355—
	8	0.0125	0.0065	0.0130	0.0480	0.0330—	0.0310—
	16	0.0145	0.0100	0.0090	0.0425	0.0295—	0.0375—
	32	0.0150	0.0115	0.0070	0.0390	0.0315—	0.0370—

* The symbol "- -" indicates that EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The symbol "++" indicates that EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

Table 5. Experimentwise type I error rates of Dunnett's test under the partial null hypothesis (H_{0_p}) as a function of number of replications r, number of treatments m, at the significance level α and with δ standard errors, in scenario 2

		α					
		0.01			0.05		
m	r / δ	4	10	20	4	10	20
5	1	0.0015—	0.0040—	0.0040—	0.0180—	0.0120—	0.0180—
	2	0.0020—	0.0040—	0.0030—	0.0180—	0.0160—	0.0145—
	4	0.0025—	0.0010—	0.0025—	0.0190—	0.0190—	0.0155—
	8	0.0050—	0.0035—	0.0025—	0.0195—	0.0175—	0.0175—
	16	0.0025—	0.0045—	0.0030—	0.0165—	0.0105—	0.0125—
	32	0.0025—	0.0025—	0.0020—	0.0120—	0.0125—	0.0170—
10	1	0.0070	0.0040—	0.0030—	0.0270—	0.0275—	0.0315—
	2	0.0050—	0.0050—	0.0055	0.0250—	0.0315—	0.0255—
	4	0.0050—	0.0065	0.0055	0.0305—	0.0280—	0.0235—
	8	0.0030—	0.0035—	0.0040—	0.0205—	0.0210—	0.0250—
	16	0.0045—	0.0050—	0.0045—	0.0230—	0.0280—	0.0265—
	32	0.0045	0.0055	0.0050	0.0285—	0.0245—	0.0280—
20	1	0.0050—	0.0030—	0.0040—	0.0280—	0.0250—	0.0285—
	2	0.0080	0.0050—	0.0050—	0.0280—	0.0235—	0.0315—
	4	0.0020—	0.0045—	0.0055	0.0240—	0.0235—	0.0235—
	8	0.0030—	0.0050—	0.0040—	0.0305—	0.0325—	0.0275—
	16	0.0050—	0.0045—	0.0045—	0.0220—	0.0240—	0.0330—
	32	0.0045—	0.0070	0.0065	0.0280—	0.0295—	0.0255—
40	1	0.0045—	0.0035—	0.0035—	0.0275—	0.0270—	0.0295—
	2	0.0050—	0.0050—	0.0055—	0.0265—	0.0245—	0.0280—
	4	0.0055	0.0050—	0.0040—	0.0275—	0.0285—	0.0275—
	8	0.0070	0.0060	0.0070	0.0390	0.0365—	0.0355—
	16	0.0060	0.0050—	0.0055	0.0360—	0.0290—	0.0295—
	32	0.0055	0.0060	0.0030—	0.0275—	0.0370—	0.0340—
100	1	0.0060	0.0060	0.0065	0.0270—	0.0305—	0.0270—
	2	0.0025—	0.0055	0.0035—	0.0300—	0.0270—	0.0240—
	4	0.0060	0.0090	0.0085	0.0350—	0.0295—	0.0295—
	8	0.0070	0.0050—	0.0045—	0.0285—	0.0325—	0.0270—
	16	0.0055	0.0060	0.0070	0.0320—	0.0405	0.0365—
	32	0.0060	0.0045—	0.0110	0.0300—	0.0335—	0.0345—

* The symbol "- -" indicates that EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The symbol "++" indicates that EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

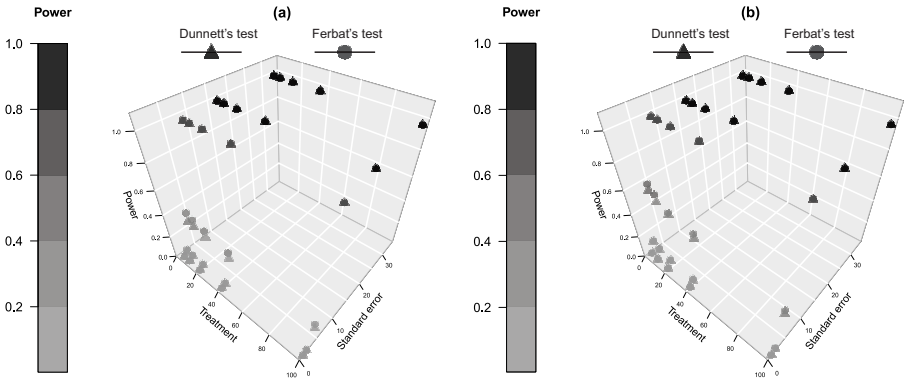

Figure 1. Power of Ferbat and Dunnett tests, under the partial null hypothesis (H_{0_p}) as a function of number of treatments m, standard errors δ, with $r = 4$ replications, at the significance level α: (a) $\alpha = 0.01$ and (b) $\alpha = 0.05$, in scenario 2.

When the number of standard errors is small (the most real situation in practical terms), the Ferbat's test tends to have a higher power than the Dunnett's test (Figure 2). As the difference between the true means increases, the tests show power very close to each other. In Figure 2, we observed that the number of replications has no major influence on the power of the tests when the true difference is fixed as number of true standard error of the mean δ. However, as well as the experimentwise error rate, the nominal significance level also influenced the power of the tests as expected theoretically. The tests had greater power at the nominal level of 0.05, with the power of Ferbat's test being greater than the power of Dunnett's test. The Tukey and Scheffé tests, for example, control the EER under the partial null hypothesis (Carmer & Swanson, 1973). However, when the number of treatments increases, the EER of these tests tends to 0 when the difference between means is at least 2 standard errors (Perecin & Malheiros, 1989). As a result, these tests have very low powers.

We can see from these results that Ferbat and Dunnett tests have more power than the MCPs[1] when the problem of the experiment focuses on comparing the treatments with a control, as stated by Shaffer (1977). As an example, under this same simulation scenario, we applied the Tukey test to compare with the power of Ferbat and Dunnett tests (Table 6). The results confirm what is observed in the literature. An interesting result, however, is that the Ferbat's test is more powerful than the Dunnett's test and may be a test alternative for multiple comparison procedures with a control. The *t* and Duncan tests, for example, are more powerful than the Ferbat and Dunnett tests when looking at the results found by Carmer & Swanson (1973). However, this is due to the high experimentwise error rates, since they show higher test sizes than the nominal significance levels.

[1]Of course we are restricted to MCPs that control the experimentwise error rate.

Table 6. Power of Tukey, Dunnett and Ferbat tests to detect a difference between means of $\delta = 4$ standard errors, under the partial null hypothesis (H_{0_p}) as a function of number of replications, treatments m at the significance level $\alpha = 0.05$, in scenario 2

		Power of tests		
Treatment	**Replication**	*Tukey*	*Dunnett*	*Ferbat*
	4	0.4348	0.5522	0.6040
5	10	0.5025	0.6080	0.6502
	20	0.5207	0.6402	0.6568
	4	0.3046	0.4857	0.5497
10	10	0.3487	0.5315	0.5584
	20	0.3595	0.5564	0.5878
	4	0.2069	0.4469	0.4661
20	10	0.2295	0.4680	0.4834
	20	0.2321	0.4770	0.5071
	4	0.1312	0.3936	0.4326
40	10	0.1416	0.3897	0.3983
	20	0.1422	0.4122	0,4265
	4	0.0665	0.3010	0.3361
100	10	0.0690	0.3209	0.3186
	20	0.0688	0.3170	0.3416

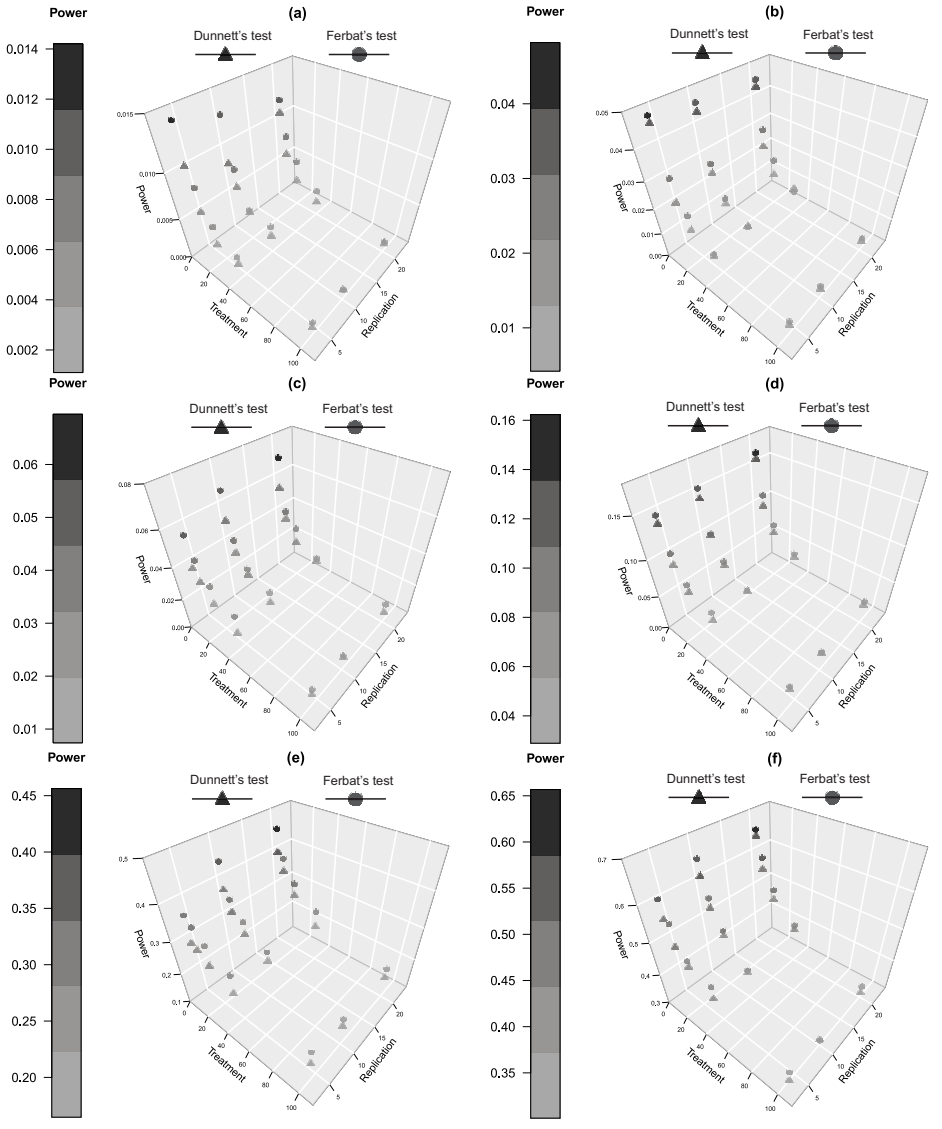

Figure 2. Power of Ferbat and Dunnett tests under the partial null hypothesis (H_{0_p}) as a function of number of treatments m, replications r, with (a) $\alpha = 0.01$ and $\delta = 1$, (b) $\alpha = 0.05$ and $\delta = 1$, (c) $\alpha = 0.01$ and $\delta = 2$, (d) $\alpha = 0.05$ and $\delta = 2$, (e) $\alpha = 0.01$ and $\delta = 4$, (f) $\alpha = 0.05$ and $\delta = 4$, in scenario 2.

3.0.3 Performance of tests in scenario 3

Only a few MCPs in the literature control the EER at nominal significance level in this scenario. It should be noticed that the relationship between experimentwise (EER) and comparisonwise (CER) error rates of MCPs shows that a test that controls the first type of error rate will control the second, but the reverse is not true (Carmer & Swanson, 1973; Boardman & Moffitt, 1971). These authors mentioned, for example, the Duncan, Waller-Duncan, t tests among others, that control the CER but they do not control the EER under the complete null hypothesis H_0. Consequently, under the partial null hypothesis (H_{0_p}), these tests will hardly control the EER, as can be observed also in the results of Bernhardson (1975) and Perecin & Malheiros (1989). Conagin (1999), Carmer & Swanson

(1973) and Silva *et al.* (1999) mentioned the Tukey, Scheffé, Dunnett, SNK and Scott-Knott tests that control the EER at the nominal significance level under complete null hypothesis. However, these authors showed that the last two tests do not control the EER under partial null hypothesis, but Tukey, Scheffé and Dunnett tests control the ERR in a conservative way.

In Tables 7 and 9 show the EER of Ferbat and Dunnett tests with the number of treatments of 40, to the scenario 3. For other values of m, the results were similar. None of the EER is significant high than the nominal levels in the exact binomial test. In general, the tests controlled the experimentwise type I error rates or they were conservative. This is a characteristic of the MCPs that control EER under this condition of scenario 3, when they also control the experimentwise error rate under complete null hypothesis.

Under H_{0_p} the tests were conservative at significance level $\alpha = 0.05$, that is, the EER was below of the nominal significance level, in all cases. When $\alpha = 0.01$, the tests preserved the overall significance level in most cases. However, in some simulations, the tests were conservative. For example, in Table 7, the Ferbat's test was conservative for $r = 4$ and $k \times \delta = 1 \times 16 = 16$ standard errors, with $\hat{\alpha} = 0.0033$. By the exact binomial test, at the 1% significance level, the test was considered conservative, and therefore the $\hat{\alpha}$ received the superscript (—). However, for this same simulation setup, the Dunnett's test controlled the overall significance level (Table 9). But when we look at tables 4 and 5 as well as tables 7 and 9, the Dunnett's test shows several cases in which it was conservative compared to the Ferbat's test. This shows that for $\alpha = 0.01$, the Ferbat's test better controlled the nominal significance level. The justification for the tests being conservative in these scenarios is the same as in scenario 2.

Another interesting evaluation when looking at Tables 7 and 9 are the initial gaps (k) in terms of standard errors, being represented by the constant k, between the control treatment and the other treatments. For the same difference between the control treatment and the other treatments, the EER were equivalent. Note in Table 8 that both the EER and the power of the tests for the same $k\delta = 4$ standard errors are the same no matter the value of k, that is, from the initial gap. Small variations in values occur due to Monte Carlo simulation error.

Table 7. Experimentwise type I error rates of Ferbat's test under the partial hypothesis H_{0_p} as a function of number of replications (r), with $m = 40$ treatments, at the significance level α and $k\delta$ standard errors, in scenario 3

		α					
		0.01			0.05		
k	δ \ r	4	10	20	4	10	20
1	1	0.0120	0.0070	0.0075	0.0365^{-}	0.0255^{-}	0.0360^{-}
	2	0.0070	0.0070	0.0055	0.0305^{-}	0.0255^{-}	0.0280^{-}
	4	0.0060	0.0070	0.0070	0.0275^{-}	0.0280^{-}	0.0315^{-}
	8	0.0120	0.0065	0.0055	0.0325^{-}	0.0235^{-}	0.0320^{-}
	16	0.0075	0.0035^{-}	0.0055	0.0255^{-}	0.0245^{-}	0.0380^{-}
	32	0.0110	0.0055	0.0095	0.0335^{-}	0.0210^{-}	0.0315^{-}
2	1	0.0035^{-}	0.0065	0.0055	0.0335^{-}	0.0270^{-}	0.0265^{-}
	2	0.0105	0.0055	0.0065	0.0285^{-}	0.0255^{-}	0.0355^{-}
	4	0.0065	0.0050^{-}	0.0110	0.0275^{-}	0.0280^{-}	0.0275^{-}
	8	0.0080	0.0055	0.0070	0.0310^{-}	0.0240^{-}	0.0225^{-}
	16	0.0065	0.0060	0.0055	0.0350^{-}	0.0245^{-}	0.0310^{-}
	32	0.0105	0.0050^{-}	0.0065	0.0260^{-}	0.0235^{-}	0.0305^{-}
4	1	0.0060	0.0070	0.0065	0.0310^{-}	0.0270^{-}	0.0220^{-}
	2	0.0100	0.0045^{-}	0.0060	0.0305^{-}	0.0210^{-}	0.0245^{-}
	4	0.0050^{-}	0.0045^{-}	0.0070	0.0285^{-}	0.0225^{-}	0.0225^{-}
	8	0.0090	0.0065	0.0040^{-}	0.0370^{-}	0.0320^{-}	0.0325^{-}
	16	0.0110	0.0090	0.0065	0.0290^{-}	0.0245^{-}	0.0270^{-}
	32	0.0055	0.0040^{-}	0.0085	0.0315^{-}	0.0255^{-}	0.0335^{-}
8	1	0.0090	0.0055	0.0085	0.0280^{-}	0.0230^{-}	0.0255^{-}
	2	0.0055	0.0085	0.0060	0.0320^{-}	0.0305^{-}	0.0340^{-}
	4	0.0090	0.0045^{-}	0.0075	0.0375^{-}	0.0220^{-}	0.0245^{-}
	8	0.0110	0.0020^{-}	0.0095	0.0270^{-}	0.0235^{-}	0.0330^{-}
	16	0.0060	0.0045^{-}	0.0090	0.0315^{-}	0.0295^{-}	0.0330^{-}
	32	0.0100	0.0060	0.0040^{-}	0.0345^{-}	0.0210^{-}	0.0265^{-}
16	1	0.0095	0.0115	0.0075	0.0385	0.0300^{-}	0.0275^{-}
	2	0.0070	0.0080	0.0050^{-}	0.0310^{-}	0.0275^{-}	0.0315^{-}
	4	0.0080	0.0050^{-}	0.0110	0.0295^{-}	0.0220^{-}	0.0240^{-}
	8	0.0070	0.0055	0.0130	0.0335^{-}	0.0295^{-}	0.0235^{-}
	16	0.0095	0.0050^{-}	0.0100	0.0375^{-}	0.0300^{-}	0.0295^{-}
	32	0.0110	0.0075	0.0085	0.0290^{-}	0.0285^{-}	0.0290^{-}

* The symbol "- -" indicates that EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The symbol "++" indicates that EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

Table 8. Experimentwise type I error rates and Power of Ferbat and Dunnett tests under the partial hypothesis (H_{0_p}) as a function of number of replications (r), with $n = 40$ treatments, at the significance level α and $k\delta = 4$ standard errors, in scenario 3

		α					
		0,01			0,05		
k	r / δ	4	10	20	4	10	20
Experimentwise error rate							
				Ferbat's test			
1	4	0.0060	0.0070	0.0070	0.0275	0.0280	0.0315
2	2	0.0105	0.0055	0.0065	0.0285	0.0255	0.0355
4	1	0.0060	0.0070	0.0065	0.0310	0.0270	0.0220
				Dunnett's test			
1	4	0.0055	0.0045	0.0055	0.0345	0.0295	0.0300
2	2	0.0055	0.0060	0.0040	0.0300	0.0280	0.0365
4	1	0.0050	0.0040	0.0055	0.0245	0.0275	0.0325
Power							
				Ferbat's test			
1	4	0.2440	0.2250	0.2545	0.4175	0.3705	0.3845
2	2	0.2425	0.2085	0.2575	0.3765	0.3840	0.4130
4	1	0.2240	0.2265	0.2375	0.3805	0.3630	0.3630
				Dunnett's test			
1	4	0.2060	0.2195	0.2335	0.3725	0.3815	0.3880
2	2	0.1950	0.2370	0.2315	0.3865	0.4035	0.3900
4	1	0.2075	0.2000	0.2250	0.3930	0.3995	0.4000

The experimentwise error rate and power of tests in detecting a $k\delta$ standard error are equivalents, regarding the tests were applied in experiments with greater dispersion or not. The justification of this result is that in the simulation pattern the true difference between means is always fixed in terms of standard error. The power of the same configuration from Tables 7 and 8 is shown in Figure 3. The power of the tests were close, and it is confirmed once again that it increases as δ and the nominal significance level also increase. Apparently, in Figure 3 the number of replications does not influence the power of the tests. However, when we look at Figure 4, we notice when m is small, the power increases with increasing number of replications r. As the number of treatments m increases, the number of replications r does not influence the power of the tests.

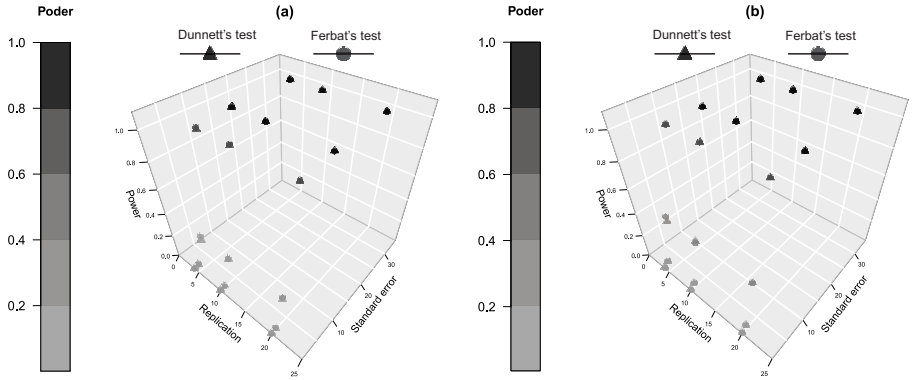

Figure 3. Power of Ferbat and Dunnett tests under the partial null hypothesis H_{0_p}, with $m = 40$ treatments, standard errors δ, replications r, at the significance level α (a) $\alpha = 0.01$ and b) $\alpha = 0.05$), in scenario 3.

In Figure 4 can be noticed that once again the Ferbat's test has greater power when m is small. However, when m increases the Dunnett's test has a slightly higher power than the Ferbat's test, which did not occur in scenario 2.

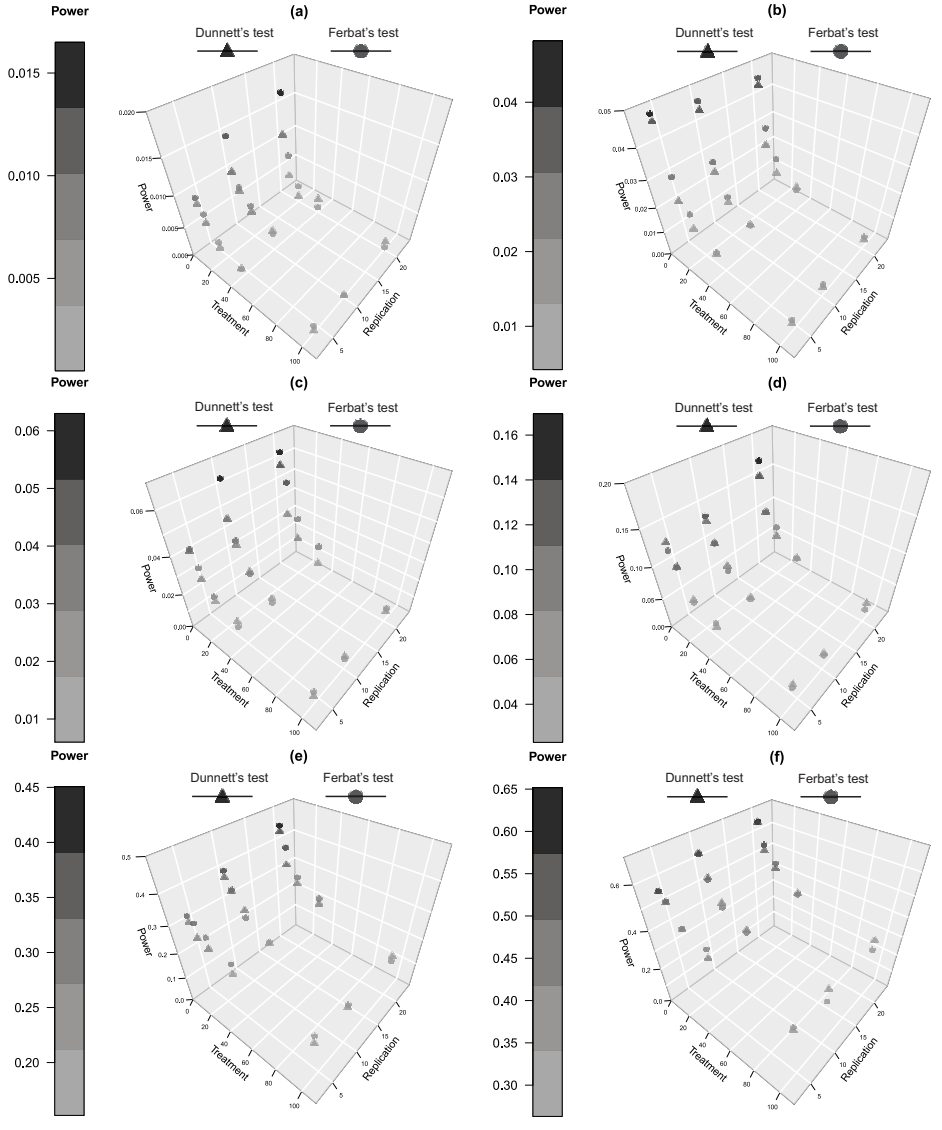

Figure 4. Power of Ferbat and Dunnett tests under the partial null hypothesis H_{0_p} as a function of number of treatments m, the number of replications r, with (a) $\alpha = 0.01$ and $\delta = 1$, (b) $\alpha = 0.05$ and $\delta = 1$, (c) $\alpha = 0.01$ and $\delta = 2$, (d) $\alpha = 0.05$ and $\delta = 2$, (e) $\alpha = 0.01$ and $\delta = 4$, (f) $\alpha = 0.05$ and $\delta = 4$, in scenario 3.

Table 9. Experimentwise type I error rates of Dunnett's test under the partial hypothesis H_{0_p} as a function of number of replications (r), with $m = 40$ treatments, at the significance level α and $k\delta$ standard errors, in an experiment on the scenario 3

		α					
		0,01			0,05		
k	r / δ	4	10	20	4	10	20
	1	0.0060	0.0035—	0.0080—	0.0275—	0.0260—	0.0280—
	2	0.0050—	0.0055	0.0065	0.0360—	0.0240—	0.0335—
1	4	0.0055	0.0045—	0.0055	0.0345—	0.0295—	0.0300—
	8	0.0045—	0.0060	0.0095	0.0210—	0.0275—	0.0300—
	16	0.0050—	0.0065	0.0030—	0.0305—	0.0295—	0.0310—
	32	0.0065	0.0070	0.0070	0.0285—	0.0285—	0.0260—
	1	0.0050—	0.0045—	0.0065	0.0310—	0.0275—	0.0265—
	2	0.0050—	0.0060	0.0040—	0.0300—	0.0280—	0.0365—
2	4	0.0045—	0.0055	0.0035—	0.0265—	0.0275—	0.0285—
	8	0.0045—	0.0050—	0.0020—	0.0320—	0.0330—	0.0350—
	16	0.0045—	0.0065	0.0065	0.0355—	0.0300—	0.0305—
	32	0.0030—	0.0055	0.0065	0.0335—	0.0195—	0.0295—
	1	0.0050—	0.0040—	0.0055	0.0245—	0.0275—	0.0325—
	2	0.0060	0.0045—	0.0040—	0.0320—	0.0235—	0.0300—
4	4	0.0060	0.0045—	0.0060	0.0340—	0.0325—	0.0275—
	8	0.0085	0.0055	0.0055	0.0295—	0.0285—	0.0370—
	16	0.0050—	0.0055	0.0045—	0.0270—	0.0275—	0.0285—
	32	0.0050—	0.0080	0.0040—	0.0290—	0.0235—	0.0235—
	1	0.0035—	0.0035—	0.0065	0.0345—	0.0275—	0.0340—
	2	0.0045—	0.0040—	0.0030—	0.0310—	0.0295—	0.0300—
8	4	0.0060	0.0045—	0.0060	0.0325—	0.0250—	0.0290—
	8	0.0050—	0.0045—	0.0055	0.0345—	0.0295—	0.0320—
	16	0.0075	0.0055	0.0075	0.0320—	0.0310—	0.0245—
	32	0.0060	0.0035—	0.0055	0.0290—	0.0315—	0.0295—
	1	0.0055	0.0010—	0.0070	0.0295—	0.0350—	0.0235—
	2	0.0040—	0.0045—	0.0075	0.0335—	0.0290—	0.0270—
16	4	0.0060	0.0060	0.0070	0.0280—	0.0290—	0.0255—
	8	0.0060	0.0065	0.0070	0.0355—	0.0280—	0.0285—
	16	0.0055	0.0075	0.0065	0.0240—	0.0255—	0.0300—
	32	0.0070	0.0065	0.0050—	0.0270—	0.0270—	0.0270—

* The symbol "- -" indicates that EER was rejected by the exact binomial test, such that $F \leq F_{0,005}$. The symbol "++" indicates that EER was rejected by the exact binomial test, such that $F \geq F_{0,995}$.

3.0.4 Performance of tests in scenario 4

In the scenario 4, we evaluated the power of the tests under the complete alternative hypothesis H_1. Both the significance level and the difference between means (δ) have effect in the power of the tests. The higher these values the greater the power of the tests (Figure 5).

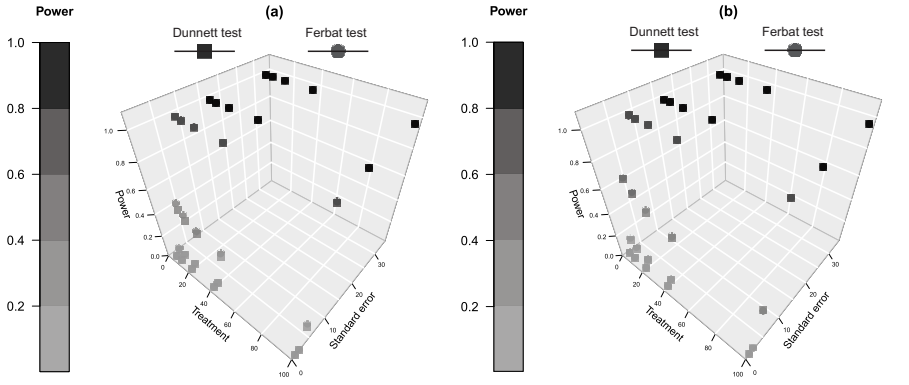

Figure 5. Power of Ferbat and Dunnett tests under the alternative hypothesis (H_1) as a function of number of treatments m, standard errors δ, with $r = 10$ replications, at the significance level (α) (a) $\alpha = 0.01$ and b) $\alpha = 0.05$), in scenario 4.

As the number of treatments increases the power of the tests decreases. When $\alpha = 0.01$ and $m = 100$, the power of the tests is close to zero (Figure 6). This means that under this condition, the probability of detecting a real difference of 1 standard error between treatments and control is very low. This is very common issue in the tests like Tukey and Scheffé, among others, that are based on the control of the experimental error rates.

Figure 6 shows how the number of replications influences the power of the tests for several number of treatments and number of replications. When m increases, regardless of the number of replications, the power of the tests does not vary greatly.

In this scenario, we realize that the Ferbat's test may be recommended as an option of a MCC when compared to the Dunnett's test, since it has greater power when the number of treatments and especially of replications is higher. However, the power of tests when m increases is practically equivalent, and once again, the Ferbat's test may be an alternative to MCC.

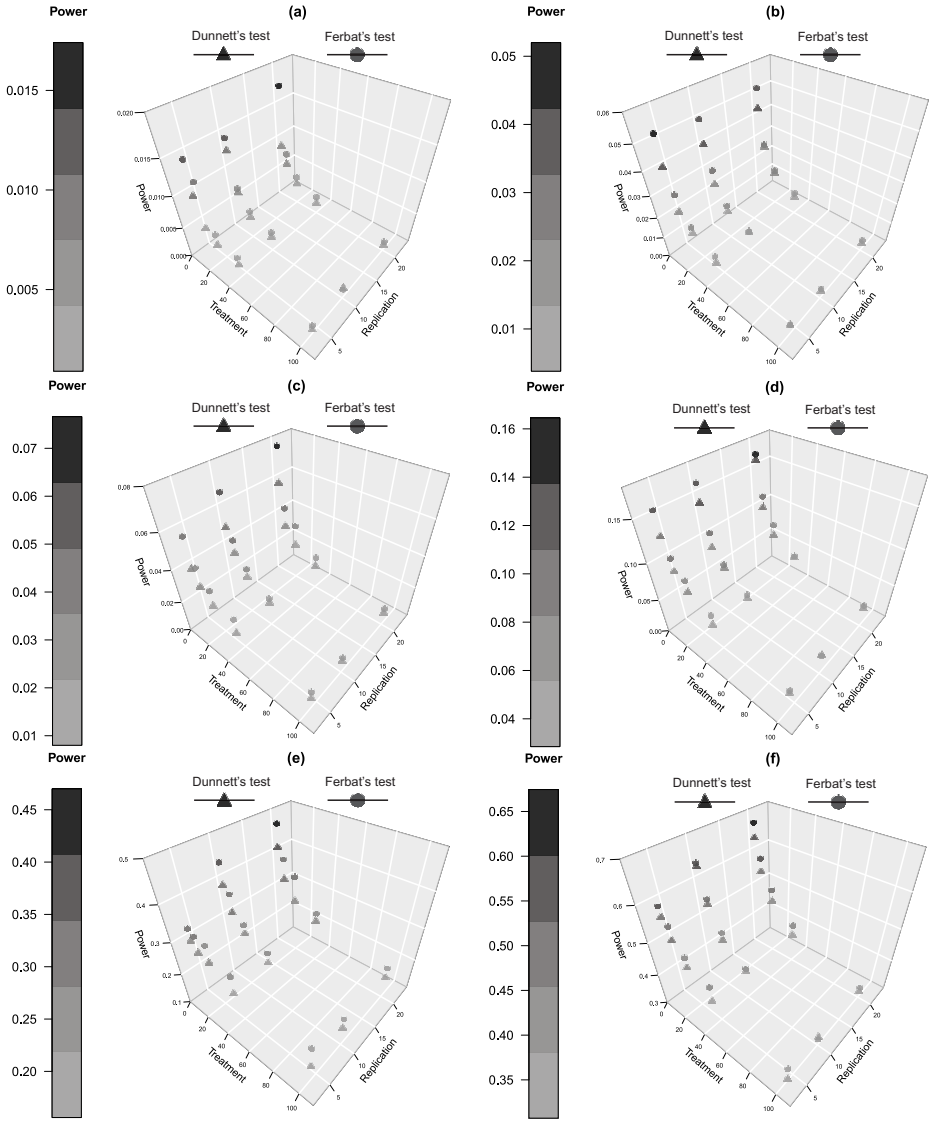

Figure 6. Power of Ferbat and Dunnett tests under the alternative hypothesis (H_1) as a function of number of treatments m, number of replications r, with (a) $\alpha = 0.01$ and $\delta = 1$, (b) $\alpha = 0.05$ and $\delta = 1$, (c) $\alpha = 0.01$ and $\delta = 2$, (d) $\alpha = 0.05$ and $\delta = 2$, (e) $\alpha = 0.01$ and $\delta = 4$, (f) $\alpha = 0.05$ and $\delta = 4$, in scenario 4.

3.0.5 Performance of tests in scenario 5

In the last evaluation scenario, scenario 5, the power of the tests are shown in Table 10 and 11, for a setting of $m = 10$, δ standard errors, replications r, and significance level α. The results were summarized, because for the other values of m they have the same pattern. As mentioned earlier in scenario 3, the power of the tests was not influenced by the initial gaps (k) between means. The convergence of the power to 1 occurs from 8 standard errors of difference. For the Ferbat and Dunnett tests, the power increases with increasing the significance level α and δ and decreasing the number of treatments m. The number of replications has certain influence with a small number of treatments, that is, as the number of replications increases the power of the tests increases.

Table 10. Power of Ferbat's test to detect a difference between means of $k\delta$ standard errors under alternative hypothesis H_1 as a function of number of replications r, with $m = 10$ treatments, at the significance level α, in scenario 5

		α					
		0,01			0,05		
k	r / δ	4	10	20	4	10	20
	1	0.0045	0.0075	0.0090	0.0210	0.0220	0.0250
	2	0.0245	0.0255	0.0375	0.1060	0.0980	0.1035
	4	0.2850	0.3205	0.3580	0.5080	0.5420	0.5700
	8	0.9840	0.9890	0.9925	0.9955	0.9945	0.9985
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0275	0.0435	0.0430	0.0970	0.0900	0.1000
	2	0.2950	0.3460	0.3895	0.5190	0.5315	0.5770
	4	0.9825	0.9905	0.9920	0.9960	0.9990	0.9980
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.3135	0.3315	0.3710	0.5205	0.5340	0.5615
	2	0.9805	0.9895	0.9930	0.9970	1.0000	0.9985
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.9820	0.9925	0.9950	0.9950	0.9970	0.9985
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 11. Power of Dunnett's test to detect a difference between means of $k\delta$ standard errors under alternative hypothesis H_1 as a function of number of replications r, with $n = 10$ treatments, at the significance level α, in scenario 5

		α					
		0,01			0,05		
k	r / δ	4	10	20	4	10	20
	1	0.0065	0.0045	0.0050	0.0250	0.0290	0.0240
	2	0.0185	0.0295	0.0365	0.0930	0.0875	0.0970
	4	0.2690	0.3370	0.3265	0.5095	0.5410	0.5460
	8	0.9755	0.9895	0.9920	0.9955	0.9975	0.9980
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.0305	0.0375	0.0370	0.0920	0.0980	0.0965
	2	0.2880	0.3215	0.3270	0.4975	0.5385	0.5550
	4	0.9805	0.9900	0.9935	0.9960	0.9985	0.9975
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.2495	0.3020	0.3325	0.4960	0.5295	0.5460
	2	0.9795	0.9890	0.9930	0.9970	0.9960	0.9990
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.9745	0.9885	0.9935	0.9965	0.9980	0.9970
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

An interesting result found in this study was the influence of the heterogeneity on the power of the Ferbat and Dunnett tests (homogeneous and heterogeneous cases), previously described. In scenarios 2 and 3, with $m = 40$, $\delta = 4$, $\alpha = 0.05$, 4, 10, and 20 replications, the power values of the Ferbat's test (homogeneous scenario) were 0.4326, 0.3989 and 0.4265, respectively. In this same configuration, but in the heterogeneous scenario, the power values of this test were 0.4175, 0.3705 and 0.3845, respectively. For the Dunnett's test the power values in this setting were 0.3936, 0.3897, 0.4122 (homogeneous scenario) and 0.3725, 0.3815, 0.3880 (heterogeneous scenario). We realized that the power of the two tests in the homogeneous environment was superior to the power in the heterogeneous environment.

In scenario 4 and 5, in the simulated configuration with $m = 5$, $\delta = 2$, $\alpha = 0.05$, with 4, 10 and 20 replications, the powers of the Ferbat's test in the homogeneous case were 0.1463, 0.1530 and 0.1623, respectively. For the heterogeneous scenario, the powers were 0.1170, 0.1345 and 0.1695, respectively. Considering the Dunnett's test, the powers for these situations were 0.1372, 0.1417 and 0.1545 (homogeneous scenario) and 0.1280, 0.1270 and 0.1496 (heterogeneous scenario), respectively. Only the power of the Ferbat's test for $r = 20$ in the homogeneous case was lower than the power in the heterogeneous environment. However, this difference was very small.

Again, in scenarios 4 and 5, the power of the test in the homogeneous environment was superior to the power in the heterogeneous case in almost all circumstances. One explanation for these results in all scenarios refers to the high entropy that occurs between treatments in the heterogeneous case, which leads to a loss of precision in the mean square of the residue for Dunnett's test and in the mean of the ranges for the Ferbat's test.

4. Applications

We adapted the example in Dunnett (1955), section II in item (b), and we considered the example for balanced data. Thus, for all treatments we consider four replications. The following data are blood count measurements on three groups of animals, one of which served as a control while the other two were treated with two drugs. The data are presented in the Table 12.

Table 12. Blood count measurements on three groups of animals

	Blood Counts (millions of cells per cubic millemeter)		
	Controls	*Drug A*	*Drug B*
	7.40	9.76	12.80
	8.50	8.80	9.68
	7.20	7.68	12.16
	8.24	9.36	9.20
Sums:	31.34	35.60	43.84
r:	4	4	4
Means:	7.84	8.90	10.96
Range:	1.30	2.08	3.60

According to the algorithm presented in subsection 2.1, the Ferbat's test statistic between the control treatment and the Drug A treatment is

$$FB^* = \frac{|7.84 - 8.90|}{0.78} = 1.36, \tag{28}$$

since $\bar{W}$ = 2.33, d_r^* = 2.10 ($r \leq 10$) and r = 4. The critical point of the test, at the significance level of 5% probability, according to step 6 (subsection 2.1) for B = 100.000, is 2.53. Therefore, as $|1.36| < 2.53$, the drug A has an effect equivalent to that of the control treatment. The same procedure is done between the control treatment and drug B, and the test statistic was 4.0. As the critical point is the same as the previous one, we observed statistical differences between the effect of drug B and the effect of the control treatment.

5. Conclusions

We note that the Dunnett's test is very complete for comparing treatments with a control because it controls the experimentwise error rate, it has high power and it can be used for unbalanced data. The test can also be applied as an one-sided and two-sided test. However, we observed in the results presented that a proposal for a two-sided test for a MCC, the Ferbat's test presented a performance similar to the Dunnett's test, and in some situations this test was superior, in particular, in the experiments with lower number of treatments and higher number of replications. Another advantage of the Ferbat test is that no information will be needed on the results of the variance analysis, such as the root-mean-square of the residue. The test is based on the calculation of mean ranges, making it easier to calculate.

The Ferbat's test performed well due to the replacement of the population standard deviation estimator, which instead of using the root-mean-square of the residue, we used $\bar{W}/d_r$ for the Ferbat's test statistic. And yet, with the modified constant $d_r{}^*$, we get more power than the Dunnett's test for some cases. However, its limitations of being used only in balanced data and being applied as a two-sided test will be overcome in a future work. Also, with the exact distribution of statistics we can achieve better performance for this test.

It is noteworthy that this paper was not intended to replace the Dunnett's test with the Ferbat's test, but rather to present a MCC alternative to the multiple comparison procedures with one control.

Acknowledgments

We would like to thank CNPq and CAPES for their financial support.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Ahsumullah, M., Nevzorov, V. B. & Shakil, M. *An introduction to order statistics* 244 (Atlantis Press, Paris, 2013).
2. Benjamini, Y., Bretz, F. & Sarkar, S. *Recent Developments in Multiple Comparison Procedures* 156 (Institute of Mathematical Statistics, Ohio, 2004).
3. Bernhardson, C. S. 375: Type I Error Rates When Multiple Comparison Procedures Follow a Significant F Test of ANOVA. *Biometrics* **31,** 229–232 (1975).
4. Boardman, T. J. & Moffitt, D. R. Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedures. *Biometrics* **27,** 738–744 (1971).
5. Bretz, F., Hothorn, T. & Westfall, P. *Multiple Comparisons using R* 182 (Chapman & Hall/CRC, Boca Raton, 2011).

6. Broch, S. C. & Ferreira, D. F. Algorithm using Gaussian quadraturas to obtain probabilities of Dunnett two-sided test for balanced data. *Tendências em Matemática Aplicada e Computacional (Tema)* **14,** 209–219 (2013).
7. Broch, S. C. & Ferreira, D. F. Multivariate distributions of non-central Dunnett test statistics. *Rev. Bras. Biom.* **31,** 501–515 (2013).
8. Broch, S. C. & Ferreira, D. F. *nCDunnett: Noncentral Dunnett's Test Distribution* R Foundation for Statistical Computing (Vienna, Austria, 2015). https://CRAN.R-project.org/package=nCDunnett.
9. Carmer, S. G. & Swanson, M. R. An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods. *Journal of the American Statistical Association* **68,** 66–74 (1973).
10. Carmer, S. G. & Walker, W. M. Pairwise multiple comparisons of treatment means in agronomic research. *Journal of Agronomic Education* **14,** 19–26 (1985).
11. Conagin, A. Discriminative power of the modified Bonferroni's test under general and partial null hypotheses. *Revista de Agricultura* **73,** 31–46 (1999).
12. Conagin, A., Barbin, D. & Demétrio, C. G. B. Modifications for the tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. *Scientia Agricola* **65,** 428–432 (2008).
13. Daly, J. F. On the Use of the Sample Range in an Analogue of Student's t-Test. *The Annals of Mathematical Statistics* **17,** 71–74 (1946).
14. Davies, O. L. & Pearson, E. S. Methods of Estimating from Samples the Population Standard Deviation. *Supplement to the Journal of the Royal Statistical Society* **1,** 76–93 (1934).
15. Dean, A., Voss, D. & Draguljić, D. *Design and Analysis of Experiments* 2nd ed., 840 (Springer International Publishing, Switzerland, 2017).
16. Dickhaus, T. *Simultaneous Statistical Inference. with applications in the life sciences* 180 (Springer, New York, 2014).
17. *Multiple testing problems in pharmaceutical statistics* (eds Dmitrienko, A., Tamhane, A. C. & Bretz, F.) 289 (CRC Press, Boca Raton, 2010).
18. Duncan, D. B. Multiple range and multiple F tests. *Biometrics* **11,** 1–42 (1955).
19. Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Ass.* **50,** 1096–1121 (1955).
20. Dunnett, C. W. New tables for multiple comparisons with a control. *Biometrics,* 482–491 (1964).
21. Dunnett, C. W. in *Proceedings of the International Conference on Linear Statistical Inference LINSTAT '93* 35–46 (Springer Netherlands, Dordrecht, 1994).
22. Hochberg, Y. & Tamhane, A. C. *Multiple Comparisons Procedures* 450 (John Wiley & Sons, New York, 1987).
23. Hsu, J. C. *Multple Comparison. Theory and methods* 277 (Chapman and Hall, London, 1996).
24. Keuls, M. The use of the "studentized range" in connection with an analysis of variance. *Euphytica* **1,** 112–122 (1952).
25. Leemis, L. M. & Trivedi, K. S. A Comparison of Approximate Interval Estimators for the Bernoulli Parameter. *The American Statistician* **50,** 63–68 (1996).
26. Lord, E. Power of the Modified t-Test (u-Test) Based on Range. *Biometrika* **37,** 64–77 (1950).

27. Lord, E. The Use of Range in Place of Standard Deviation in the t-Test. *Biometrika* **34,** 41–67 (1947).

28. Miller, R. *Simultaneous statistical inference* 2nd ed., 299 (Springer-Verlag, New York, 1981).

29. Oliveira, I. R. C. & Ferreira, D. F. Multivariate extension of chi-squared univariate normality test. *Journal of Statistical Computation and Simulation* **80,** 513–526 (2010).

30. Patnaik, P. B. The Use of Mean Range as an Estimator of Variance in Statistical Tests. *Biometrika* **37,** 78–87 (1950).

31. Pearson, E. S. & Haines, J. The Use of Range in Place of Standard Deviation in Small Samples. *Supplement to the Journal of the Royal Statistical Society* **2,** 83–98 (1935).

32. Perecin, D. & Malheiros, E. B. *Uma avaliação de seis procedimentos para comparações múltiplas* in 3º *Simpósio de Estatística aplicada à Experimentação Agonômica* (Lavras, Brazil, 1989), 66.

33. Shaffer, J. P. Multiple Comparisons Emphasizing Selected Contrasts: An Extension and Generalization of Dunnett's Procedure. *Biometrics* **33,** 293–303 (1977).

34. Silva, E. C. d., Ferreira, D. F. & Bearzoti, E. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. *Ciência e Agrotecnologia* **23,** 687–696 (1999).

35. Sousa, C. A. d., Junior, M. A. L. & Ferreira, R. L. C. Evaluation of statistical tests of multiple comparisons for means. *Revista Ceres* **59,** 350–354 (2012).

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

37. Westfall, P. H., Tobias, R. D. & Wolfinger, R. D. *Multiple comparisons and multiple tests using SAS* 2nd ed., 625 (SAS institute, North Carolina, 2011).