Brazilian Journal of Biometrics, **41**, 361–397 (2023)
doi:10.28951/bjb.v41i4.640

ARTICLE

Means Grouping based on Studentized Midrange

Ben Dêivide de Oliveira Batista,*[,1] Daniel Furtado Ferreira,[2] Matheus Fernando Rodrigues Santos,[3] and Henrique José de Paula Alves[4]

[1]Statistics, Phisycs and Mathematics Department, Federal University of São João del-Rei, Campus Alto Paraopeba, Ouro Branco - MG, Brazil.
[2]Statistic Department, Federal University of Lavras, Lavras - MG, Brazil.
[3]Federal University of São João del-Rei, Campus Alto Paraopeba, Ouro Branco - MG, Brazil.
[4]Instituto de Pesquisa Econômica Aplicada - IPEA, Rio de Janeiro - RJ, Brazil
*Corresponding author. Email: ben.deivide@gmail.com

(Received: October 19, 2022; Revised: March 03, 2023; Accepted: April 05, 2023; Published: December 01, 2023)

Abstract

The purpose of this paper was to develop two procedures of multiple comparisons based on methods of clustering means, that is, mean grouping tests based on the midrange (MGM) and range (MGR). The first is based on the studentized midrange distribution, and the second is based on the studentized range distribution. The tests presented similar performance (evaluation of type I error and power) to the performance of the considered tests used for comparison. Like the tests presented that were based on methods of grouping averages of literature, the MGM and MGR tests did not control the experimentwise error rate for almost all evaluated scenarios. However, under the complete H_1 hypothesis, these tests showed high power, with emphasis on the MGM test. Thus, what we propose is yet another test alternative without ambiguity in its results and not a substitution for the traditional tests already present in the literature.

Keywords: Midrange; Range; Mean grouping; Monte Carlo simulation; R software.

1. Introduction

In experimental statistics, some of the existing problems are the simultaneous comparison of hypothesis tests, so that the global type I error increases as the number of comparisons between treatments increases, and this is what we call the multiplicity effect. Statistical procedures designed to adequately control for multiplicity effects are called multiple comparison procedures (MCPs).

One of the biggest problems in the study of multiple comparison procedures is the lack of transitivity (results ambiguity) when two levels of the factor had the same difference between themselves, but they do not differ from a third party, that difficult interpretation the results. As an alternative,

different methodologies were proposed to contour this situation, for example, methods based on grouping analysis. The grouping analysis uses as a separation criterion of objects, the characteristics that these objects own. The proposal is to unite groups of objects with similar characteristics. An efficient alternative to contouring the problem of MCPs ambiguity is Scott-Knott's test (Scott & Knott, 1974).

Despite the Scott-Knott's test being able to solve the MCPs results ambiguity problem, its performance presents some problems, like small deviations from rate of type I error under complete H_0 and high type I error rates under partial H_0. In addition, an interesting situation is that the criterion of partition of the groups of means of this test, in certain situations, forms groups of means with a difference between consecutive means intra-group greater than the difference between consecutive means that delimit these groups.

Many other alternative forms of grouping tests, similar to Scott-Knott's test, were presented in the specialized literature (Bhering *et al.,* 2008; Calinski & Corsten, 1985; Conrado *et al.,* 2017; Ramos & Ferreira, 2009; Ramos & Vieira, 2014; Shimokawa & Goto, 2011). None of these methods was able to solve all problems presented beforehand and, in some cases, they displayed a performance even worse than the reference Scott-Knott's test. A considerable quantity of these methods is based on the externally studentized range or in the F statistics. The search for alternative statistics to make this grouping was the mark of these works since Scott-Knott's test used the likelihood ratio. A very interesting statistic is the midrange since according to Rider (1957), it is more efficient (it's an estimator with less variance to the population average) than the arithmetic average in some populations, such as the cosine population, the parabolic population, the rectangular population and the inverted parabolic population. Based on this information, it was noticed that these statistics could be an alternative to the development of multiple comparisons methods.

In the literature, some works about the midrange were published by Gumbel (1958), David & Nagaraja (2003), among others, obtained the midrange distribution and density functions to the case of a normally distributed population. Batista & Ferreira (2017) developed the density, distribution and quantile functions for the case of the externally studentized midrange, both theoretical and numerical methods. As a consequence of these works, was created an R package, denominated `SMR` (Batista & Ferreira, 2014a), with the implementation of the algorithms published in Batista & Ferreira (2014b), that calculates the cumulative distribution function, the density function and returns the quantile values for the distribution of this statistic. Batista & Ferreira (2020) published two tests based on this same distribution, being an alternative to Tukey's test.

Considering $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$, the statistics with an order of one random sample $Y_1, Y_2, \ldots, Y_n$ with size n with normal distribution with average 0 and variance σ^2, namely, $Y_i \sim N(\mu, \sigma^2)$, defines the externally studentized *midrange* as:

$$\bar{Q} = \frac{\bar{R}}{S}, \tag{1}$$

being $\bar{R} = (Y_{(n)} + Y_{(1)})/2$ the *midrange* and S an estimator to the population standard deviation σ, associated to ν degrees of freedom and obtained independently from $\bar{R}$. The density function and the cumulative distribution function of (1), consequence of the results of Batista & Ferreira (2017), are given by

$$f_{\bar{Q}}(\bar{q}; \mu, \sigma^2, n, \nu) = \int_0^{\infty} \int_{-\infty}^{x\bar{q}} 2n(n-1)x\phi(y - \mu/\sigma)\phi(2x\bar{q} - y - \mu/\sigma) \times \\ \times [\Phi(2x\bar{q} - y - \mu/\sigma) - \Phi(y - \mu/\sigma)]^{n-2} f_X(x; \nu) dy dx, \tag{2}$$

and

$$F_{\bar{Q}}(\bar{q};\mu,\sigma^2,n,\nu) = \int_0^{\infty}\int_{-\infty}^{x\bar{q}} n\phi(y-\mu/\sigma)[\Phi(2x\bar{q}-y-\mu/\sigma)- \\ -\Phi(y-\mu/\sigma)]^{n-1}f_X(x;\nu)dydx, \tag{3}$$

respectively, wherein the probability density function $f_X(x;\nu)$ is given by

$$f_X(x;\nu) = \frac{\nu^{\nu/2}}{\Gamma(\nu/2)2^{\nu/2-1}}x^{\nu-1}e^{-\nu x^2/2}, \quad x \geq 0. \tag{4}$$

However, $\bar{Q}$ is not an ancillary statistic to μ, avoiding the test development based on this statistic. This way, Batista & Ferreira (2014a) showed that to $Y_i \sim N(0,\sigma^2)$, the probability density function and the cumulative distribution function are

$$f_{\bar{Q}}(\bar{q};n,\nu) = \int_0^{\infty}\int_{-\infty}^{x\bar{q}} 2n(n-1)x\phi(y)\phi(2x\bar{q}-y)\times \\ \times[\Phi(2x\bar{q}-y)-\Phi(y)]^{n-2}f_X(x;\nu)dydx. \tag{5}$$

and

$$F_{\bar{Q}}(\bar{q};n,\nu) = \int_0^{\infty}\int_{-\infty}^{x\bar{q}} n\phi(y)[\Phi(2x\bar{q}-y)-\Phi(y)]^{n-1}\times \\ f_X(x;\nu)dydx, \tag{6}$$

respectively.

However, to $Y_i \sim N(\mu,\sigma^2)$, the hope of $\bar{Q}$ is given by

$$E[\bar{Q}] = \frac{\mu}{\sigma}\frac{\left(\frac{\nu}{2}\right)^{1/2}\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma(\nu/2)},$$

wherein $\mu = 0$, one has that $E[\bar{Q}] = 0$. This is fundamental information to the development of the proposed tests in this work.

Another widely used statistic in the multiple comparison tests is the externally studentized range. Its distribution was widely studied by (David *et al.,* 1954; Hartley, 1942; Newman, 1939; Pearson & Haines, 1935; Pearson & Hartley, 1943, 1942; Pearson, 1926, 1932). In the 1950 decade, the main MCPs based on this statistic were proposed, such as the SNK test started by Student (1927) and Newman (1939) and adapted by Keuls (1952), Tukey test (Tukey, 1953), and the Duncan test (Duncan, 1955).

The externally studentized range is defined by the ratio between $W = Y_{(n)} - Y_{(1)}$ and S, in which S is the population standard deviation estimator σ, associated to ν degrees of freedom, being independently distributed from W, namely,

$$Q = \frac{W}{S}.$$

The distribution function and density function of the externally studentized range are given, respectively, by:

$$F_Q(q;n,\nu) = \int_0^{\infty}\int_{-\infty}^{\infty} n\phi(y)[\Phi(xq+y)-\Phi(y)]^{n-1}f_X(x;\nu)dydx,$$

and

$$f_Q(q;n,\nu) = \int_0^{\infty}\int_{-\infty}^{\infty} n(n-1)x\phi(y)\phi(xq+y)[\Phi(xq+y)-\Phi(y)]^{n-2}\times \\ \times f_X(x;\nu)dydx,$$

wherein $\phi(y)$ and $\Phi(y)$ are the density function and the distribution function of a standard normal random variable, with $y \in \mathbb{R}$ and $f_X(x;\nu)$ is the density function of X, that was expressed in (4). These results are given in the function of the standard normal distribution. Namely, independently from the parameters of the initial normal distribution, the density function and the Q distribution function will be always expressed in terms of the standard normal distribution.

Considering that both the externally range distribution and the externally studentized midrange distribution, this work has as an objective, develop two average grouping methods that don't present ambiguity and that have control upon type I error and high power. The performance of the tests is rated by Monte Carlo simulations, considering the type I error rate experimentwise and power.

2. Materials and Methods

Considering n treatments and r repetitions, for the tests proposition, the following random sample was experimentally obtained: $Y_{11}, Y_{12}, \ldots, Y_{1r}, Y_{21}, \ldots, Y_{2r}, \ldots, Y_{i1}, Y_{i2}, \ldots, Y_{ij}, \ldots, Y_{ir}, \ldots, Y_{n1}, Y_{n2}, \ldots, Y_{nr}$, wherein Y_{ij} is the random observation referenced to the ith treatment in its jth repetition, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, r$. The ith treatment average is:

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

This sample is subjected to variance analysis, adopting the following model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}, \tag{7}$$

in which $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\mu_i = \mu + \tau_i$ are the ith treatment average. Thus, the mean squared error (*MSE*) is estimated by:

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

It is well known that $\bar{Y}_{i.}$ and the *MSE* are independently distributed and that $\hat{V}(\bar{Y}_{i.}) = MSE/r$, see Graybill (1961) and Searle (1987).

Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = \mu$, the n treatments have a common average μ. In this particular case, the statistics of order $\bar{Y}_{(1).}, \bar{Y}_{(2).}, \ldots, \bar{Y}_{(n).}$ are centered on μ. Therefore, the externally studentized midrange, defined by

$$\bar{Q} = \frac{\sqrt{r}\left[(\bar{Y}_{(1).} + \bar{Y}_{(n).})/2\right]}{\sqrt{MSE}}, \tag{8}$$

has a distribution function dependent on μ (under H_0), as presented in the expressions (2) and (3).

However, μ is unknown and hardly equal to zero in real situations. Thus, to utilize the distribution of $\bar{Q}$ with unknown μ and $\mu \neq 0$ is impossible in the test proposition.

This way, we chose to utilize the midrange distribution in the especifica case where $\mu = 0$, expressions (5) and (6) and adapt the test's statistic to adjust to the sample mean. It is observed that the distribution is centered on μ, which is unknown. Therefore, to use the distribution centred on 0, there was a correction in the statistic. Since $\bar{R} = (\bar{Y}_{(1).} + \bar{Y}_{(n).})/2$ has a distribution centred on μ, the corrected statistic was $\bar{R}_n = \bar{R} - \bar{Y}_{\cdot}^*$, wherein $\bar{Y}_{\cdot}^*$ is an estimator for μ.

Initially, the overall average was thought of as an estimator of $\bar{Y}_{\cdot}^*$, i.e. $\bar{Y}_{\cdot\cdot} = \sum_{i=1}^{n} \bar{Y}_{i.}/n$. However, when the simulated data were under H_1, this mean estimated the overall mean of the parameters $\mu = \sum_{i=1}^{n} \mu_i/n$, where $\mu_i = E(\bar{Y}_{i.})$. Thus, under H_1, it was observed utilizing of the simulation that the quantity $(\mu_1 + \mu_n)/2$ was close to μ and therefore, $E(\bar{R}_n) = (\mu_1 + \mu_n)/2 - \mu \approx 0$. Thus, the performance of the power tests was very low.

Under H_0, it is clear that the expected value $E(\bar{R}_n)$ is null, which at first would support the direct use of the externally studentized midrange in the test. However, what was observed, in a preliminary evaluation via simulation, was that the type I errors experimentwise were too high. Initially, it was speculated that this resulted from the fact of the statistic is a function of $\bar{Y}_{\cdot\cdot}$, which also has a sample error associated with it. Thus, an initial (Minimum Significant Difference) MSD that would represent the standard error of $\bar{Y}_{\cdot\cdot}$ was built. However, the test started to control adequately the type I error, but presented low power.

It happened because $E(\bar{R}_n)$, although different from zero, under H_1, presented values in magnitude not so different from zero. Thus, it was sought an estimator of μ that would maximize $E(\bar{R}_n)$ under H_1 and where $E(\bar{R}_n) = 0$ under H_0. In this case, it was used $\bar{Y}_{\cdot}^*$ that would correspond to the average of one of the two potential groups to be obtained in the test. This partition would be between two ordered means of maximum range.

To obtain an estimator with a smaller standard error, it was used the average of the group with the higher number of involved averages between the two groups considered. This estimator was determined based on empiric criteria and validated through Monte Carlo simulation. Therefore, considering the partitions $\bar{Y}_{(1).}$, $\bar{Y}_{(2).}$, ..., $\bar{Y}_{(k).}$ and $\bar{Y}_{(k+1).}$, $\bar{Y}_{(k+2).}$, ..., $\bar{Y}_{(n).}$, whose point k corresponds to the value j, where

$$\max_j(\bar{Y}_{(j+1).} - \bar{Y}_{(j).})$$

happens, for $j = 1, 2, \ldots, n-1$. if there are ties with two or more values different from k, say $k_1, k_2, \ldots$, then it is formed a partition wherein $k = \max\{\min(k_1, m - k_1), \min(k_2, m - k_2), \cdots\}$.

This way, taking

$$\bar{Y}_1^* = \frac{\sum_{j=1}^{k} \bar{Y}_{(j).}}{k}$$

and

$$\bar{Y}_2^* = \frac{\sum_{j=k+1}^{n} \bar{Y}_{(j).}}{n-k},$$

the value of $\bar{Y}_{\cdot}^*$ will correspond to $\bar{Y}_1^*$ if $k \geq n - k$ or equal to $\bar{Y}_2^*$, otherwise.

Thus, the final statistic of the test is:

$$\bar{R}_n = \frac{\bar{Y}_{(1).} + \bar{Y}_{(n).}}{2} - \bar{Y}_{\cdot}^* \tag{9}$$

The MSD to reject or not the hypothesis, initially was considered

$$\Delta_n = \bar{q}_{(\alpha/2;n,\nu)}\sqrt{\frac{MSE}{r}} + \frac{1}{\sqrt{n}}\sqrt{\frac{MSE}{r}},$$

where $\bar{q}_{(\alpha/2;n,\nu)}$ is the $100\alpha/2\%$ upper tail quantile of the distribution of $\bar{Q}$, expression 6, with n treatments and ν degrees of freedom.

On the results of preliminary Monte Carlo simulations, it was observed that the type I error per experiment was way smaller than the nominal levels of significance and that the power was low. Thus, the contribution of $\bar{Y}_{\cdot}^{*}$ for the MSD, $(1/\sqrt{n}) \times \sqrt{MSE/r}$ should be reduced by a factor between 0 and 1. By trial and error in a process of Monte Carlo simulation, was found a value that converged to $\sqrt{2}/2$. Thus, the final MSD considered was

$$\Delta_n = \bar{q}_{(\alpha/2;n,\nu)}\sqrt{\frac{MSE}{r}} + \frac{1}{\sqrt{2n}}\sqrt{\frac{MSE}{r}},$$

where $\bar{q}_{(\alpha/2;n,\nu)}$ is the $100\alpha/2\%$ upper tail quantile of the distribution of $\bar{Q}$, expression 6, with n treatments and ν degrees of freedom.

2.1 Mean grouping test based on the midrange (MGM)

The MCP was proposed using a criteria of forming a partition of m oredered means in the position k, in which

$$\max_j\{\bar{Y}_{(j+1).} - \bar{Y}_{(j).}\} = \bar{Y}_{(k+1).} - \bar{Y}_{k.}$$

happens for $j = 1, 2, \ldots, m-1$, where m represents the number of means in a considered group. Initially $m = n$. If there are ties with two of more values different from k, say k_1, k_2, ..., then it is created a partition where $k = \max\{\min(k_1, m-k_1), \min(k_2, m-k_2), \cdots\}$. Thus, defined

$$\bar{Y}_1^{*} = \frac{\sum_{j=1}^{k} \bar{Y}_{(j).}}{k}$$

and

$$\bar{Y}_2^{*} = \frac{\sum_{j=k+1}^{m} \bar{Y}_{(j).}}{m-k}.$$

Therefore, $\bar{Y}_m^{*} = \bar{Y}_1^{*}$ if $k \geq m-k$ or $\bar{Y}_m^{*} = \bar{Y}_2^{*}$, otherwise. The steps for this test's application are:

1. Do $m = n$ and take the ordered means of the treatments by: $\bar{Y}_{(1).}, \bar{Y}_{(2).}, \ldots, \bar{Y}_{(m).}$;
2. Determine k and $\bar{Y}_m^{*}$ as discussed previously;
3. Determine the statistic's value by:

$$\bar{R}_m = \frac{\bar{Y}_{(1).} + \bar{Y}_{(m).}}{2} - \bar{Y}_m^{*}; \tag{10}$$

4. The MSD is

$$\Delta_m = \bar{q}_{(\alpha/2;m,\nu)}\sqrt{\frac{QME}{r}} + \underbrace{\frac{1}{\sqrt{2m}}\sqrt{\frac{QME}{r}}}_{\text{It represents the variation of } \bar{Y}^*_m}, \tag{11}$$

where $\bar{q}_{(\alpha/2;m,\nu)}$ is the $100\alpha/2\%$ upper tail quantile of the distribution of $\bar{Q}$, expression 6, with m treatments and ν degrees of freedom;

5. If $|\bar{R}_m| \leq \Delta_m$ is the stopping criterion, therefore the m means are considered not different and then the group is marked as not partitionable. Otherwise, consider the group's means $\bar{Y}_{(1).}$, $\bar{Y}_{(2).}$, ..., $\bar{Y}_{(k).}$ as different from the group's means $\bar{Y}_{(k+1).}$, $\bar{Y}_{(k+2).}$, ..., $\bar{Y}_{(m).}$ and go to the 6th step;
6. For each group obtained and marked as partitionable, consider m as the number of means of the related group. Repeat the steps from 2 to 5, with one reservation, in the 4th step, the following must be used as the minimum significant difference:

$$\Delta_m = \bar{q}_{(\alpha/2;m,\nu)}\sqrt{\frac{MSE}{r}}, \tag{12}$$

where $\bar{q}_{(\alpha/2;m,\nu)}$ is the $100\alpha/2\%$ upper tail quantile of the distribution of $\bar{Q}$, expression 6, with m treatments and ν degrees of freedom. The reason for changing the DMS in expressions (11) and (12) is to increase the power controlling the type I error rate at nominal significance level. This was verified through simulation. The step (repetition of the steps 2 to 5) is done for all the groups until no other group can be partitioned in two new ones or until all the groups contain a single mean.

2.2 Mean grouping test based on the studentized range (MGR)

A similar version of Scott-Knott's test (Scott & Knott, 1974), based on the studentized range was also proposed. The essence of the test is the same as the proposed MGM test (Section 2.1). To partition the groups, it was used a potential partition point the position of the maximum range between ordered means. Thus, for $\bar{Y}_{(1).}$, $\bar{Y}_{(2).}$, ..., $\bar{Y}_{(m).}$, the partition must be considered in the position k where is verified:

$$\max_j\{\bar{Y}_{(j+1).} - \bar{Y}_{(j).}\} = \bar{Y}_{(k+1).} - \bar{Y}_{k.},$$

for $j = 1, 2, \ldots, m-1$. Must be considered for the application of the test, the superior quantile $100\alpha\%$, $q_{(\alpha;m,\nu)}$, of the externally studentized range.

The steps for the application of the test are:

1. Do $m = n$ and consider m ordered means: $\bar{Y}_{(1).}$, $\bar{Y}_{(2).}$, ..., $\bar{Y}_{(m).}$;
2. Determine k according to what was previously discussed;
3. Calculate the statistic of the test by:

$$q_m = \bar{Y}_{(m).} - \bar{Y}_{(1).};$$

4. The MSD is

$$\Delta_m = q_{(\alpha;m,\nu)}\sqrt{\frac{MSE}{r}};$$

5. If $q_m \leq \Delta_m$, then the m means are considered not different, mark the group as not partitionable and go to the 6th step. Otherwise, consider the group means $\bar{Y}_{(1).}$, $\bar{Y}_{(2).}$, ..., $\bar{Y}_{(k).}$ as different from the group's means $\bar{Y}_{(k+1).}$, ..., $\bar{Y}_{(m).}$ and go to the 6th step.
6. For each group obtained and marked as partitionable, consider m the number of means for the related group. Repeat steps 2 to 5. This procedure is done for all groups until no other group can be partitioned into two new ones or until every group contain a single mean.

2.3 Performance evaluation of the proposed tests

Two strategies were considered in this work. The first was to evaluate the experimentwise error rate (EER) of the proposed multiple comparison tests. The second was to evaluate the power of the tests. In both cases, Monte Carlo simulation was used in the R software (R CORE TEAM, 2022). In each simulation the multiple comparison tests were applied at a pre-established nominal level of significance α, verifying whether or not the null hypothesis was rejected. This process, in each case, was repeated $N^* = 5000$ 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 correct decisions (rejections) refers to the empirical power.

To evaluate the empirical EER simulated via Monte Carlo, it was used the exact binomial test with a coefficient of 99% of probability to test the hypothesis $H_0: \alpha = 5\%$ against $H_1: \alpha \neq 5\%$ and $H_0: \alpha = 1\%$ against $H_1: \alpha \neq 1\%$. If the null hypothesis is rejected and the empirical EER is considered significant (p-value $< 0,01$) inferior to the nominal level, the test will be considered conservative. If the empirical EER is considered significantly (p-value $< 0,01$) superior to the nominal level, the test will be considered liberal. If the observed value of the EER is not significant (p-value $> 0,01$), the test will be considered exact (Oliveira & Ferreira, 2010).

Considering γ as the number of rejected null hypotesis in $N^* = 5000$ Monte Carlo simulations, for a nominal level of significance α, the test statistic using the relation between the distribution F and the binomial distribution (Leemis & Trivedi, 1996), with success rate of $p = \alpha$, is given by

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

under H_0. This statistic 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 to the significant level 1% of probability, wherein $F_{0,005}$ and $F_{0,995}$ are the quantiles of the F distribution with ν_1 and ν_2 degrees of freedom (Oliveira & Ferreira, 2010).

In both steps data were simulated according to the statistic model described in (7), where μ is the general constant fixed at 100 for all the cases, without loss of generality, τ_i is the effect of the ith treatment and ϵ_{ij} is the effect of the random error with normal distribution and independently distributed with 0 mean and common variance of σ^2, also fixed at 100, without loss of generality, being $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, r$, wherein r is the number of repetitions.

In the first step for the evaluation of the EER, the effects of the treatment τ_i were considered equal to 0 for all i, $i = 1, 2, \ldots, n$. Therefore, the data were generated under the complete null hypothesis, namely, with all the treatments having the same parametric means. The probability of the EER ($\hat{\alpha}$) was estimated by the proportion of experiments with at least one incorrectly detected difference according to the total of N^* simulated experiments, namely,

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

wherein 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 equality is verified and 0, otherwise.

In the second step of the power evaluation, the effects of the treatments were simulated with two options, to generate a simulation of complete H_1 (alternative hypothesis) and of partial H_0 (null hypothesis). Thus, in the first case, the effect of the treatment 1 was considered equal to 0, namely, $\tau_1 = 0$, and the others are fixed by

$$\tau_i = \tau_{i-1} + \delta\frac{\sigma}{\sqrt{r}},$$

for δ, $\delta = 1, 2, 4, 8, 16$ and 32, representing the number of standard errors of the difference between means to specify the effect of the consecutive treatments, considering $i = 2, 3, \ldots, n$. Thus, the power was computed by the proportion of rejections among the means involving multiples of δ, about the total number of comparisons involving this difference. Therefore, between consecutive treatments, for example, there are $n - 1$ comparisons per experiment and $N^*(n - 1)$ comparisons in total, which corresponds to the power of detecting δ standard errors of the difference between means. In the same way, for the neighbours with step 2 (first and third means, second and fourth up to the antepenultimate and last ordered means) there are $n-2$ comparisons per experiment involving 2δ standard errors of the difference of means to be detected. This procedure is done for all the cases until the first and last means are compared, i.e., $(n - 1)\delta$ standard errors to be detected in only 1 comparison per experiment and a total of N^* comparisons to all simulated experiments.

The second option for the study of the power involving a simulation under partial H_0 involved the simulation of two mean groups, with $k_1 = \lfloor n/2 \rfloor$ and $k_2 = n - k_1$ means in each, where $\lfloor x \rfloor$ refer to the biggest integer lesser or equal to x. The means of the first group were all the same, for which the effects were $\tau_i = 0$, $i = 1, 2, 3, \ldots, k_1$, without loss of generality. The second group, with k_2 means, had its effects also equal to

$$\tau_i = \tau_1 + \delta\frac{\sigma}{\sqrt{r}}, \quad i = k_1 + 1,\ k_1 + 2,\ \ldots,\ n,$$

where different values of δ were considered as $\delta = 1, 2, 4, 8, 16$. In this case, the proportion of rejections involving comparisons of different groups in the total of $N^*k_1k_2$ comparisons involving means of two groups in the N^* simulated experiments provided an estimate of the power. The intragroup comparisons allowed us to also evaluate the ratios of the EER under partial H_0. The proportion of experiments with at least one rejection of the null hypothesis of equality between the two intragroup means was an estimate of this ratio of test error. All the tests were applied to each one of the simulated scenarios, the EER (intragroup comparisons) and power (intergroup comparisons) were computed and the results were compared.

Considered some configurations in both steps with different values of n and r. Thus, were considered the cases with $n = 5$, 10, 20, 40 and 100, and $r = 4$, 10 and 20. Considered also the nominal significance level of 1% and 5%. The coefficient of variation of the experiment adopted was 10% because, in the simulated results, it was noticed that the evaluated MCPs were not influenced by the coefficient of variation, considering a normal population, for the evaluation of the performance of type I error per experiment and power, since, in the simulation, when the means differed, it was always in terms of standard errors. Besides, preliminary analyses were made with the proposed tests and this same behaviour was verified. Therefore, the simulations were fixed in a single variation coefficient.

3. Results and Discussion

The performance evaluation of the proposed tests will be compared with the results of existing tests in the literature, emphasizing the Tukey, SNK and Scott-Knott tests, the first two being a

classical reference in the literature and the last one being the reference for proposing the tests. To understand more about these tests, an updated reference can be found in Cui *et al.* (2021) and an updated recommendation for multiple comparisons can be found in Sauder & DeMars (2019). Even so, the Tukey and SNK tests were also simulated, to confirm the results that already exist on these in the literature. The results found by Ramos & Vieira (2014), Ramos & Ferreira (2009), Conrado *et al.* (2017), Bernhardson (1975), Carmer & Swanson (1973), Bhering *et al.* (2008) and Einot & Gabriel (1975) will also be used for discussion. The performance evaluation will be based on the type I error and the power of the tests. Several arrangements have been chosen for performance evaluation. The results will be discussed and presented through tables and graphs to facilitate exposure and interpretation. As the simulation was performed in several scenarios, only some will be presented due to the amount of simulated data and also taking into account that in some simulation scenarios the performance of the tests was similar. The first evaluation of the tests was based on the EER. Two scenarios were evaluated under complete H_0 and partial H_{0_p}. Tables 1 and 2 show the results of the EER under complete H_0.

Table 1. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR and MGM tests, as a function of the number of treatments and of the number of repetitions, under complete H_0, at the significance level $\alpha = 1\%$ probability, evaluated by the exact binomial test with a confidence coefficient of 99% probability

		Tests			
Replications	**Treatments**	*Tukey*	*SNK*	*MGR*	*MGM*
4	5	1.160	1.160	1.160	$0.640^{—}$
	10	1.100	1.100	1.100	0.720
	20	1.040	1.040	1.040	0.900
	30	0.820	0.820	0.820	$0.660^{—}$
	40	0.820	0.820	0.820	0.700
	100	1.060	1.060	1.060	$0.600^{—}$
10	5	0.980	0.980	0.980	$0.460^{—}$
	10	1.000	1.000	1.000	0.740
	20	0.900	0.900	0.900	0.780
	30	1.240	1.240	1.240	$0.560^{—}$
	40	0.880	0.880	0.880	0.860
	100	0.940	0.940	0.940	$0.580^{—}$
20	5	0.940	0.940	0.940	$0.500^{—}$
	10	1.080	1.080	1.080	0.880
	20	1.100	1.100	1.100	0.720
	30	1.020	1.020	1.020	$0.520^{—}$
	40	0.840	0.840	0.840	$0.600^{—}$
	100	1.020	1.020	1.020	0.700

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

It was observed that the MGM and MGR tests controlled the experimentwise error rate because none of them had the empirical EER rejected by the exact binomial test, such that $F \geq F_{0.995}$. However, in some cases, the empirical nominal levels for the MGM test were rejected by the exact binomial test, such that $F \leq F_{0.995}$, making it conservative. Confirming the results of this work (Tables 1 and 2), Carmer & Swanson (1973) and Bernhardson (1975), they also showed that the Tukey and SNK tests have the control of the EER.

Silva *et al.* (1999) and Borges & Ferreira (2003), evaluating the performance of the Scott-Knott's

test and considering the same simulation settings of the present study, except for N = 2.000 simulations, observed that some values of the EER were higher than the nominal levels of significance (α) of 1% and 5%. The values of the EER that exceeded the value of α (liberal tests) were those in which the number of treatments was 5, although they did not distance themselves much from the nominal values. Conrado *et al.* (2017) presented a version of the Scott-Knott test for unbalanced designs. The performance evaluation of the adjusted Scott-Knott's test showed results similar to those found by Silva *et al.* (1999) and Borges & Ferreira (2003). In some situations, the test exceeded the overall level of significance. This can be justified by the Monte Carlo error.

Ramos & Ferreira (2009) as well as Ramos & Vieira (2014) developed the tests created by Calinski & Corsten (1985), in the bootstrap version. They also evaluated the original tests and those that were developed. The tests developed by Calinski & Corsten (1985) are Calinski-Corsten's test based on the studentized range distribution (CCR test) and Calinski-Corsten's test based on the F distribution (CCF test). The bootstrap versions of these tests will be called CCRb and CCFb tests, respectively. All tests were considered exact under the null hypothesis and normal distribution. This is because the tests were evaluated under non-normal conditions, something that was not the objective of this work.

Table 2. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR, and MGM tests, depending on the number of treatments and the number of repetitions, under H_0 complete, at the significance level α = 5% probability, assessed by the exact binomial test with a confidence coefficient of 99% probability

		Tests			
Replication	**Treatments**	*Tukey*	*SNK*	*MGR*	*MGM*
4	5	5.240	5.240	5.240	3.680—
	10	5.660	5.660	5.660	4.720
	20	5.080	5.080	5.080	5.060
	30	4.960	4.960	4.960	4.340
	40	4.980	4.980	4.980	3.980—
	100	4.680	4.680	4.680	3.340—
10	5	4.940	4.940	4.940	3.860—
	10	5.060	5.060	5.060	4.820
	20	5.240	5.240	5.240	5.140
	30	4.840	4.840	4.840	4.160—
	40	4.620	4.620	4.620	4.020—
	100	5.140	5.140	5.140	3.700—
20	5	4.880	4.880	4.880	2.540—
	10	5.060	5.060	5.060	4.440
	20	4.940	4.940	4.940	4.760
	30	4.960	4.960	4.960	4.120—
	40	5.020	5.020	5.020	4.180—
	100	4.820	4.820	4.820	3.720—

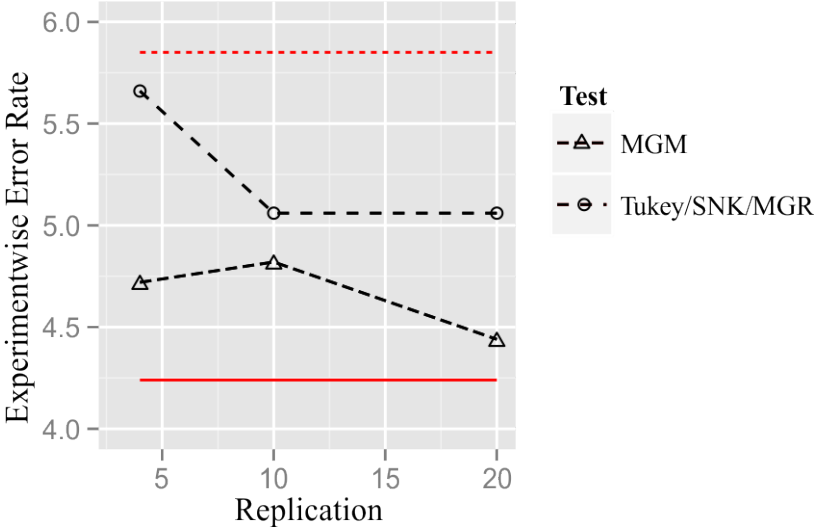
* The symbol "—" indicates that the EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The "++" symbol indicates that the EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

Regardless of the number of replications, the proposed tests (MGM and MGR) controlled the EER (Tables 1 and 2). This was also verified by Borges & Ferreira (2003) when they evaluated the performance of Tukey and SNK tests, thus confirming the results of Tables 1 and 2. They used the same methodology of simulation of the present work, concerning the number of replications and the coefficient of variation.

The reason for this is the simulation adopted. Treatment parameters are linked to the number of replications and the difference between means is always preserved in terms of standard error, which is therefore related to the coefficient of variation and the number of replications. However, considering Scott-Knott's test, these same authors observed that for a large number of replications, $r = 20$, only when the number of treatments was small, $n = 5$, the test was liberal. Conrado *et al.* (2017) also verified this for the adjusted Scott-Knott test, when $\alpha > 12\%$ in the simulated experimental conditions.

In Figures 1 and 2 that represent the evaluation of the experimentwise error rate, two red lines can be observed, one is full and the other is dashed. The first delimit the rejection where the EER values below this line were lower than the overall nominal level, i.e., the hypothesis $H_0 : \alpha = 5\%$ was rejected by the exact binomial test because $F \leq F_{0.005}$. Thus, it is a conservative test. The second delimits the region in which the EER values above this line were higher than the overall nominal level, that is, the hypothesis $H_0 : \alpha = 5\%$ was rejected by the exact binomial test because $F > F_{0.995}$. This is a liberal test.

In Figure 1, the MGM, MGR, Tukey and SNK tests controlled the EER, since none of the tests evaluated by the exact binomial test exceeded the red lines identifying the rejection of the H_0 hypothesis.

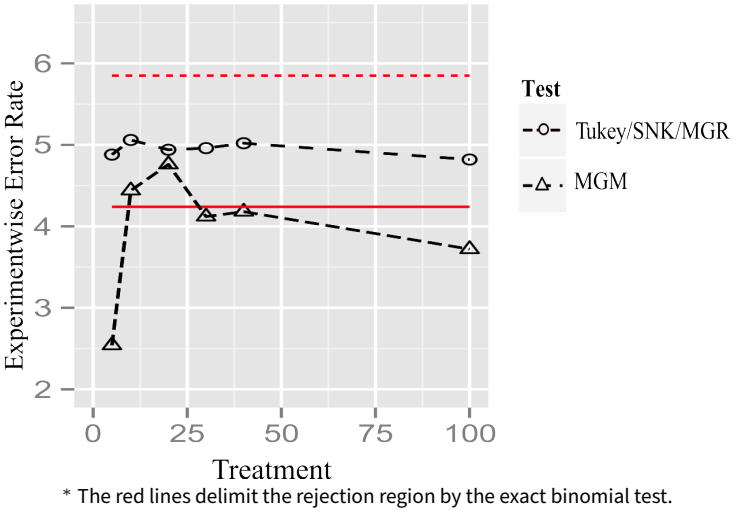


* The red lines delimit the rejection region by the exact binomial test.

Figure 1. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR and MGM tests, depending on the number of replications, hypothesis H_0 complete, $n = 10$, $\alpha = 5\%$.

Regarding the number of treatments, based on a graphical representation for the arrangement $r = 20$ and $\alpha = 5\%$ in Figure 2, it was found that the MGR and MGM tests control the EER. It was noticed that in the MGM test as the number of treatments increases, the EER decreases to the point of being conservative for both $\alpha = 1\%$ and $\alpha = 5\%$ (Tables 1 and 2). The other simulated settings can be seen in Tables 1 and 2, and the results were similar to those shown in Figure 2. Carmer & Swanson (1973) and Boardman & Moffitt (1971) verified this same behaviour for Scheffé's test, considering 4000 experiments. For $n = 20$ treatments, the EER of this test reached almost 0% type I error per experiment, a very conservative test.

For the Tukey, SNK and Scott-Knott tests, regardless of the number of treatments, considering a normal population, Borges & Ferreira (2003) showed that the experimentwise error rate remains equal to the level of overall significance. However, when considering non-normal populations, the



* The red lines delimit the rejection region by the exact binomial test.

Figure 2. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR and MGM tests, depending on the number of replications, hypothesis H_0 complete, n = 20, α = 5%.

Tukey and SNK tests showed EER in the order of 55% with 80 treatments for the log-normal distribution. These same authors also found that the Scott-Knott's test presents a certain control of type I error by experiment considering non-normal populations, since the positive bias in the control of the ERR was very small, presenting certain robustness. Ramos & Ferreira (2009) as well as Ramos & Vieira (2014) also showed robustness to the CCR and CCF tests under non-normality, the CCRb and CCFb tests being more robust, as the latter were exact.

Unlike these MCPs, Bernhardson (1975) showed that the LSD test (test based on the *t* distribution of *Student*) and Duncan's test, considering 10 treatments, present high rates of type I error, 49.0% and 36.3%, respectively. This fact was also confirmed in Boardman & Moffitt (1971) and Carmer & Swanson (1973).

Carmer & Swanson (1973) and Perecin & Malheiros (1989) evaluated the t-bayesian test proposed by Waller & Duncan (1969) and high rates of type I error were found per experiment for this test. Carmer & Swanson (1973) observed that for the numbers of treatments equal to 5, 10 and 20 and significance level α = 5%, the values of type I error rates per experiment were 15.6%, 18.4%, and 18.7%, respectively, confirming that it was a liberal test.

An interesting fact for the proposed MGR test is that it presents EER identical to the Tukey and SNK tests, regardless of the number of replications and treatments, under complete H_0. This is due to the similarity in the theoretical development of the tests. For example, the Tukey and SNK tests for the first difference between the extreme means (lowest mean and highest mean), present the same MSD, as observed by Carmer & Swanson (1973). This fact was also observed in Borges & Ferreira (2003), considering the normal distribution and variation coefficient of 10% (same conditions of the simulation of this study), in which the Tukey and SNK tests present equal EER.

However, the assumption in the null hypothesis that the treatment means are the same, can be observed in the experiments that very rarely all these means are the same. Based on this, in the following subsection, we considered the scenario in which the simulation was based on the partial null hypothesis. Thus, another way to evaluate the type I error is through simulations taking into account the partial null hypothesis (H_{0_p}), see Tables 3 to 5.

Figure 3 shows the performance evaluation of the tests about the difference in consecutive means (δ), fixing the number of treatments (n = 5, 20 and 100) and the number of replications (r =10). In

Table 3. Experimentwise, in percentage, of the Tukey, SNK, MGR, and MGM tests, depending on the number of treatments and the number of replications, under partial H_0, at the significance level $\alpha = 5\%$ probability and $\delta = 1\sigma_{\bar{Y}}$, evaluated by the exact binomial test with a confidence coefficient of 99% probability

		Evaluated tests			
Treatment	**Replication**	*Tukey*	*SNK*	*MGR*	*MGM*
5	4	2.700—	3.600—	8.560++	5.220
	10	2.820—	3.740—	9.960++	4.760
	20	2.240—	3.240—	9.420++	4.620
10	4	2.980—	3.560—	10.820++	9.400++
	10	2.880—	3.560—	12.360++	9.080++
	20	2.460—	3.000—	12.900++	8.080++
20	4	3.140—	3.500—	13.120++	8.580++
	10	2.740—	3.180—	13.440++	7.900++
	20	2.520—	3.000—	14.160++	7.640++
40	4	3.140—	3.480—	15.400++	5.600++
	10	3.400—	3.400—	16.000++	5.820
	20	2.700—	2.880—	14.980++	5.440
100	4	3.200—	3.380—	17.960++	4.560
	10	3.380—	3.460—	19.400++	4.720
	20	2.920—	3.020—	18.320++	4.200—

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

general, it was observed that the proposed tests exceed the established levels of significance, especially when the difference in groups of consecutive means is greater than $2\sigma_{\bar{Y}}$. For Ramos & Vieira (2014), the CCR and CCF tests, as well as their bootstrap versions, presented these problems from $\delta \geq 1$, making them liberal tests. Borges & Ferreira (2003) evaluated the Scott-knott's test for $\delta = 0.5$ and 4, and $\alpha = 5\%$. In almost all scenarios evaluated, the Scott-Knott test exceeds the nominal level adopted, being also a liberal test.

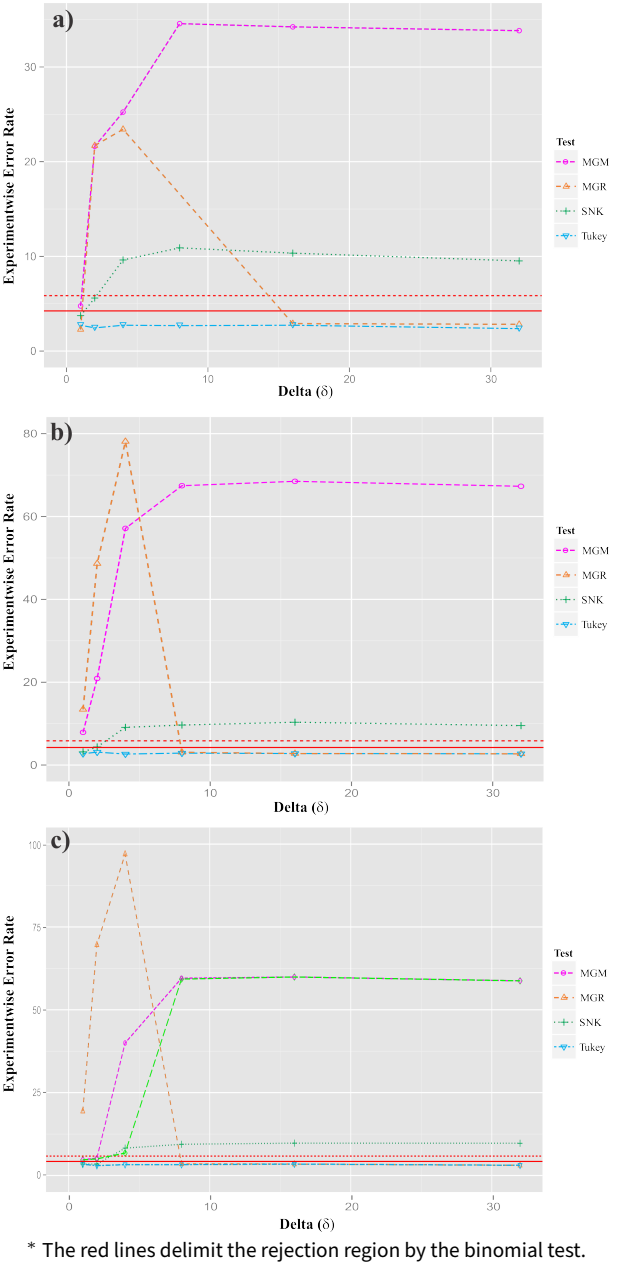
According to the simulation performed in the present study, Tukey's test is the only test that controls the EER to the level of overall significance, regardless of the configuration of the experiment. Carmer & Swanson (1973) found that the Tukey and Scheffé tests did not exceed the 3.1% EER in all configurations, considering a significance level of $\alpha = 5\%$. In this same study, it was also observed that the Duncan and t-Bayesian tests have the highest type I error rates per experiment under partial H_0.

The SNK's test showed control in the EER only for small differences in consecutive means. When the difference between groups of consecutive means was greater than $4\sigma_{\bar{Y}}$, the EER of this test exceeded the nominal level for both $\alpha = 0.01$ and $\alpha = 0.05$, which characterizes it as a liberal test being also confirmed by Carmer & Swanson (1973.

The MGR test showed control over the overall significance level when the difference between groups of consecutive means was greater than or equal to $8\sigma_{\bar{Y}}$. The MGM test controlled the level of overall significance only for a large number of treatments ($n = 100$) and $\delta \leq 2$, see Tables 3 and 4.

The tests evaluated did not suffer expressive influences, as can be seen in Figure 4, which presents the results of the EERs for the proposed tests concerning the number of replications, for $n = 10$, $\alpha = 5\%$ and the difference between consecutive group means of $4\sigma_{\bar{Y}}$.

The number of treatments had an influence on the EER at the overall significance level, under



* The red lines delimit the rejection region by the binomial test.

Figure 3. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR, and MGM tests, depending on the difference in consecutive averages (δ), under partial H_0, a) $n = 5$, b) $n = 20$ e c) $n = 100$, for $\alpha = 5\%$ and $r = 10$.

Table 4. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR, and MGM tests, depending on the number of treatments and the number of replications, under partial H_0, at the significance level probability $\alpha = 5\%$ and $\delta = 2\sigma_{\bar{Y}}$, assessed by the exact binomial test with a confidence coefficient of 99% probability

		Evaluated tests			
Treatment	**Replication**	*Tukey*	*SNK*	*MGR*	*MGM*
5	4	2.460$^{—}$	5.500	18.420^{++}	10.760^{++}
	10	2.540$^{—}$	5.600	21.700^{++}	10.140^{++}
	20	2.240$^{—}$	3.240$^{—}$	9.420^{++}	15.020^{++}
10	4	2.600$^{—}$	4.640	31.220^{++}	22.560^{++}
	10	3.140$^{—}$	5.260	38.220^{++}	21.620^{++}
	20	2.460$^{—}$	3.000$^{—}$	12.900^{++}	22.140^{++}
20	4	3.480$^{—}$	4.800	42.920^{++}	22.220^{++}
	10	3.120$^{—}$	4.380	48.640^{++}	20.900^{++}
	20	2.520$^{—}$	3.000$^{—}$	12.900^{++}	20.640^{++}
40	4	2.820$^{—}$	3.800$^{—}$	53.500^{++}	13.240^{++}
	10	2.980$^{—}$	3.860$^{—}$	58.160^{++}	11.460^{++}
	20	2.520$^{—}$	3.000$^{—}$	14.160^{++}	12.680^{++}
100	4	3.360$^{—}$	3.800$^{—}$	67.120^{++}	5.760
	10	2.980$^{—}$	3.440$^{—}$	69.600^{++}	5.260
	20	2.700$^{—}$	2.880$^{—}$	14.980^{++}	5.780

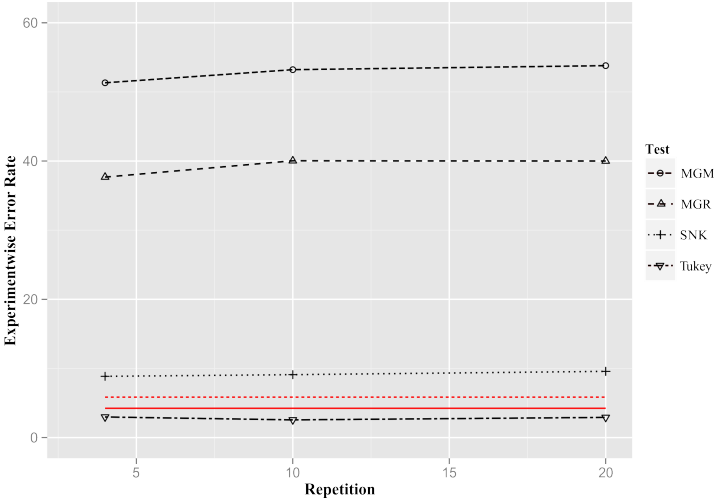
* The symbol "$^{—}$" indicates that the EER was rejected by the exact binomial test, such that $F \leq F_{0.005}$. The "++" symbol indicates that the EER was rejected by the exact binomial test, such that $F \geq F_{0.995}$.

Table 5. Experimentwise error rate, in percentage, of the Tukey, SNK, MGR, and MGM tests, depending on the number of treatments and the number of replications, under partial H_0, at the significance level $\alpha = 5\%$ probability and $\delta = 4\sigma_{\bar{Y}}$, assessed by the exact binomial test with a 99% probability confidence coefficient

		Evaluated tests			
Treatment	**Replication**	*Tukey*	*SNK*	*MGR*	*MGM*
5	4	2.560$^{—}$	9.740^{++}	21.720^{++}	24.880^{++}
	10	2.820$^{—}$	9.620^{++}	23.420^{++}	25.240^{++}
	20	2.460$^{—}$	9.360^{++}	23.100^{++}	25.500^{++}
10	4	2.980$^{—}$	8.860^{++}	37.680^{++}	51.320^{++}
	10	2.560$^{—}$	9.100^{++}	40.040^{++}	53.220^{++}
	20	2.920$^{—}$	9.580^{++}	40.000^{++}	53.800^{++}
20	4	3.000$^{—}$	8.540^{++}	57.660^{++}	64.980^{++}
	10	2.660$^{—}$	9.100^{++}	58.240^{++}	64.460^{++}
	20	3.140$^{—}$	9.500^{++}	57.480^{++}	65.380^{++}
40	4	3.540$^{—}$	8.660^{++}	78.060^{++}	58.720^{++}
	10	2.600$^{—}$	8.360^{++}	78.120^{++}	57.140^{++}
	20	2.920$^{—}$	8.840^{++}	76.840^{++}	58.000^{++}
100	4	3.000$^{—}$	7.800^{++}	97.080^{++}	32.720^{++}
	10	3.280$^{—}$	8.260^{++}	96.980^{++}	32.980^{++}
	20	2.940$^{—}$	7.720^{++}	96.440^{++}	32.720^{++}

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

partial H_0 (Figure 5), mainly for the MGM and MGR tests. But the behaviours of the tests were different from each other. Initially, Tukey's test preserved the EER at the significance level when the number of treatments varied. Compared to the performance evaluation made by Borges & Ferreira (2003), it was found that the results of the EER for Tukey's test are similar.



* The red lines delimit the rejection region by the exact binomial test.

Figure 4. Experimentwise, in percentage, of the Tukey, SNK, MGR, and MGM tests, depending on the number of replications, under hypothesis H_0 partial, for $n = 10$, $\alpha = 5\%$.

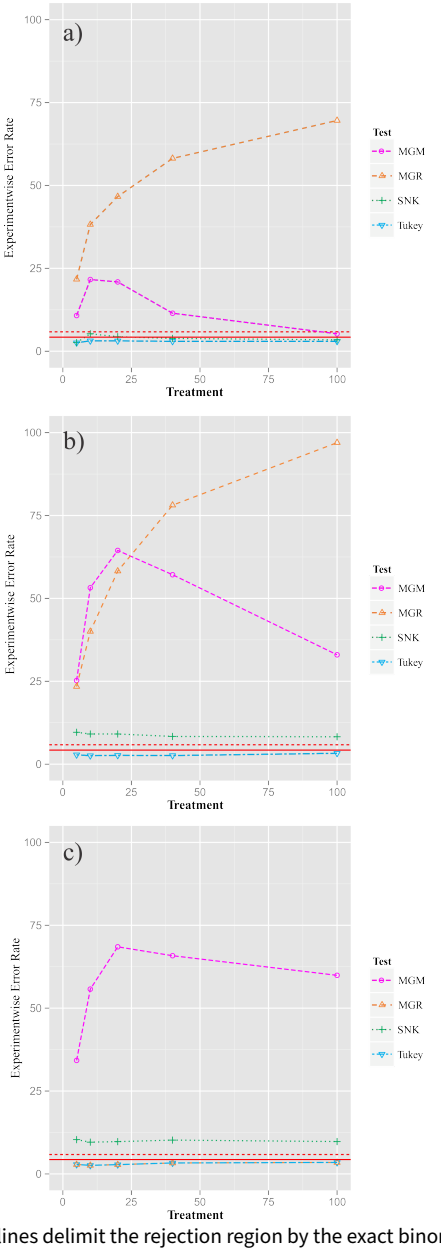
The SNK's test, for $\delta \leq 2$, presented the EER according to the level of significance adopted, regardless of the number of treatments, see Tables 3 and 4. However, when $\delta > 2$ the test presented the EER > α, becoming a liberal test. It was also observed that the MGR test, for $\delta < 8$, considerably increases the values of the EER with the increase in the number of treatments. For $\delta \geq 8$, the EER of this test stabilizes and becomes identical to the overall significance level. When $\delta = 16$, the experimentwise is identical to Tukey's test. This is due to the similarity between the structures of the two methods. So, it doesn't matter whether you use one or the other, under partial H_0 when $\delta \geq 8$. The problem is that this difference between averages is not very common in practical situations and in fact, we don't know the real difference between the means.

For the MGM test, when $\delta \geq 8$, it is observed that by increasing the number of treatments, this test also increases the EER, becoming a very liberal test, under partial H_0. When $\delta < 8$, the EERs of these tests as a function of the number of treatments (n), present a behaviour of a parable (Figure 5), having a peak in the values of the EER when the number of treatments is equal to 20.

In the second step, the tests were compared through the study of power. Several situations were considered: number of replications, number of treatments, differences between means, level of significance, and number of populations. In the latter situation, two groups were simulated that had the same means internally and differed from each other by a quantity of δ standard errors, that is, under partial H_0. The power study was also evaluated under the hypothesis H_1, in which comparisons between groups of different means were considered.

The power of the tests, under complete H_1, was influenced by the number of treatments. In Table 9, the performance evaluation of the Tukey, SNK, MGM and MGR tests can be observed, for the difference between averages of $2\sigma_{\bar{Y}}$, $r = 4$ replications and a significance level of $\alpha = 5\%$ probability.

The SNK test increased power as the number of treatments (n) increased, while the Tukey test decreased power as the number of treatments increased. Tukey's test has almost 0% power when the



* The red lines delimit the rejection region by the exact binomial test.

Figure 5. Type I error rate per experiment, in percentage, of the Tukey, SNK, MGR and MGM tests, depending on the number of treatments, under hypothesis H_0 partial, for a) $\delta = 2$, b) $\delta = 4$ and c) $\delta = 16$, with $r = 10$, $\alpha = 5\%$, assessed by the exact binomial test with a confidence coefficient of 99% probability.

Table 6. Power of the Tukey, SNK, MGR, and MGM tests, in percentage, to detect a difference between averages starting with standard error ($1\sigma_{\overline{Y}}$) to $32\sigma_{\overline{Y}}$, depending on the number of treatments and the number of replications equal to 4 (r = 4), under H_1 complete, at the significance level of 5% probability

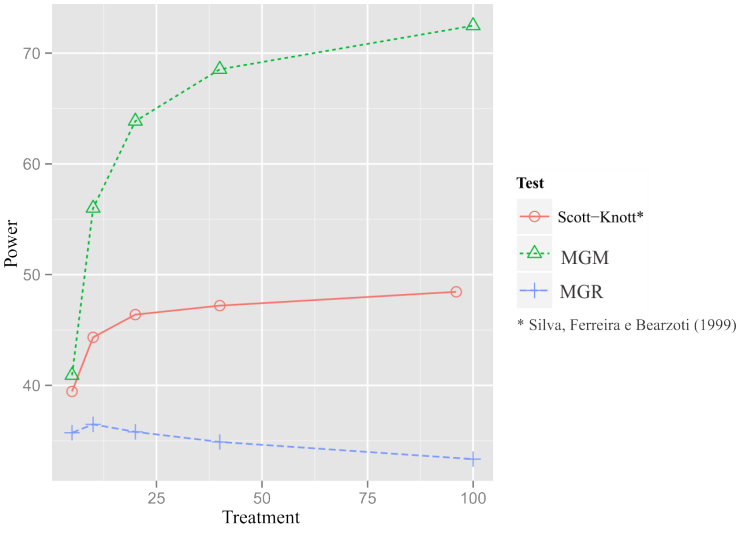
		Real difference between means					
Test	**Treatment**	$1\sigma_{\overline{Y}}$	$2\sigma_{\overline{Y}}$	$4\sigma_{\overline{Y}}$	$8\sigma_{\overline{Y}}$	$16\sigma_{\overline{Y}}$	$32\sigma_{\overline{Y}}$
	5	2.040	7.630	41.640	-	-	-
	10	0.700	3.580	30.960	98.130	-	-
Tukey	20	0.220	1.510	20.760	96.730	100.000	-
	40	0.070	0.620	13.080	94.690	100.000	100.000
	100	0.020	0.190	6.690	90.060	100.000	100.000
	5	5.020	12.570	44.980	-	-	-
	10	4.150	11.930	48.350	99.090	-	-
SNK	20	3.840	11.920	51.180	99.470	100.000	-
	40	3.590	11.760	52.590	99.620	100.000	100.000
	100	3.550	11.910	53.710	99.700	100.000	100.000
	5	15.540	25.620	50.800	-	-	-
	10	23.690	39.270	72.520	99.620	-	-
MGR	20	24.190	40.020	72.870	99.560	100.000	-
	40	23.630	39.280	71.380	99.330	100.000	100.000
	100	22.650	37.790	68.850	98.710	100.000	100.000
	5	11.970	21.210	34.940	-	-	-
	10	25.910	41.000	65.540	78.030	-	-
MGM	20	36.610	54.610	80.600	89.790	92.090	-
	40	43.360	62.150	87.280	94.480	95.910	96.480
	100	49.360	68.600	92.620	98.010	98.720	98.980

number of treatments equals 100 for δ = 1. This can also be verified for the Scheffé test, according to Carmer & Swanson (1973). This fact shows that Tukey and Scheffé tests are not recommended for comparison of two to two means with a large number of treatments. This result shows the importance of the type I error for these tests, i.e. a very conservative test, increases the type II error, and consequently decreases the power of the test.

Another power performance evaluation was performed for the MGR, MGM and Scott-Knott tests, Figure 6. This evaluation was performed based on the number of treatments, fixing the difference between means of $2\sigma_{\overline{Y}}$, r = 4 replications and α = 5% probability. The power of Scott-Knott's test was taken from the work of Silva *et al.* (1999). These authors evaluated Scott-Knott's test in the same experimental scenario as the present study.

The MGM test had an increase in power with an increase in the number of treatments. This behaviour was also verified for the Scott-Knott test. However, the MGR test practically did not change power with the variation in the number of treatments. What occurred was a small decrease in power with an increase of n. For n = 5, the power of the MGM, Scott-Knott and MGR tests was 42.99%, 39.45% and 37.59%, respectively. While for n = 100, the power of the tests was 72.64%, 48.45% and 33.51%, respectively. The MGM test showed greater power than the Scott-Knott and MGR tests. The latter showed the worst performance among the three tests.

For a broader evaluation, all the tests presented to date were compared with the performance evaluations of other tests performed by Perecin & Malheiros (1989). These authors showed that the test with the greatest power was the t-Bayesian test, followed by the t-test, Duncan test, modified Newman-Keus test and the Newman-Keuls test. All of them had power above 60% for $4\sigma_{\overline{Y}}$ standard

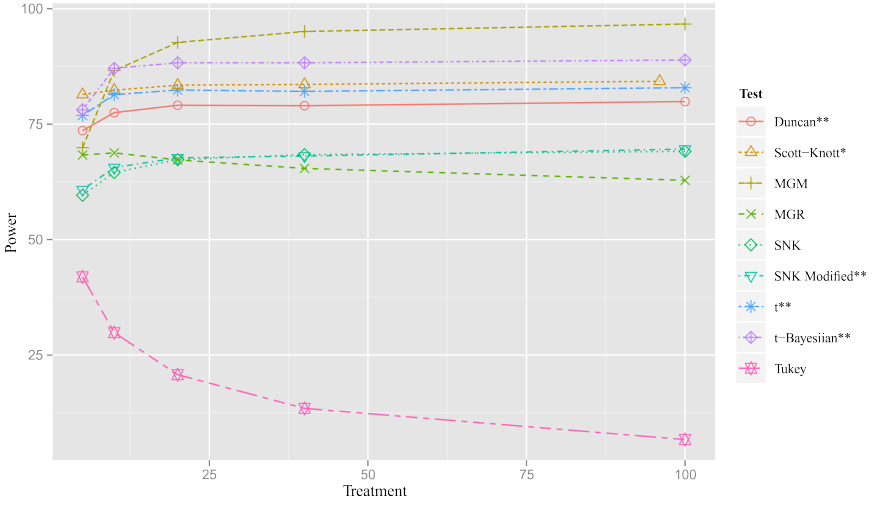


* Result of Silva *et al.* (1999).

Figure 6. Power of the Scott-Knott, MGR and MGM tests, in percentage, under H_1 complete, to detect a difference between averages of $2\sigma_{\bar{Y}}$, with $r = 4$ replications, in depending on the number of treatments, for a $\alpha = 0.05$.

errors, this effect being more expressive with the t-bayesian test, power above 78%. However, this is due to the high type I error rates per experiment, under the H_0 hypothesis.

Figure 7 shows the power of the tests in the scenario $\delta = 4$, $r = 4$ replications and $\alpha = 5\%$ probability. This scenario served as the basis for presenting the other situations since the results were equivalent.



* Results of Silva *et al.* (1999) and
** Results of Perecin & Malheiros (1989).

Figure 7. Power of Duncan, Scott-Knott, MGM, MGR, SNK, modified SNK, t, t-bayesian, and Tukey tests, in percentage, under H_1 complete, to detect a difference between averages of $4\sigma_{\bar{Y}}$, with $r = 4$ replications, depending on the number of treatments, for a $\alpha = 0.05$.

It was observed that the MGM, t-bayesian and Scott-Knott tests showed the greatest power, and the first test showed the greatest prominence. Tukey's test presented the worst performance. The other tests presented the values of intermediate power between those of the MGM test (test with greater power) and Tukey's test (test with less power).

The number of replications was another aspect that influenced the power of the tests, but not as expressive as in the case of the number of treatments. Figure 8 has presented the evaluation of power performance concerning the number of replications. This evaluation was analyzed in three scenarios: (a) 5 treatments, (b) 20 treatments and (c) 100 treatments, for a difference between the mean of $2\sigma_{\bar{Y}}$ and $\alpha = 5\%$.

For a small number of treatments, Figure 8(a), there was an increase in the power of the tests with an increase in the number of replications, mainly from 4 to 10. However, when the number of treatments increased ($n \geq 20$), Figures 8(b) and 8(c), the power of the tests hardly changed with the increase in the number of replications. This may be due to the higher accuracy of the estimation of residual variance, because of the increase in the number of treatments, regardless of the increase in the number of replications, degrees of freedom are high. However, with a small number of treatments, the degrees of freedom are also small for a small number of replications, and high for a large number of replications. Thus, we observe the greatest effect of the number of replications for the power in the latter situation, once the accuracy of the experiment was fixed.

Another evaluation of power performance was based on the difference between means. The power of the tests increased rapidly as the difference between means increased. Silva *et al.* (1999) and Perecin & Malheiros (1989) showed that when the magnitude between means was equal to or greater than $6\sigma_{\bar{Y}}$, the correct decision percentages of the tests evaluated were high.

In Figure 9 presented the power of the tests for the real differences between averages of 1 to $32\sigma_{\bar{Y}}$, with 4 replications and $\alpha = 0.05$. The scenario was divided concerning the number of treatments: (a) $n = 5$, (b) $n = 20$ and (c) $n = 100$. When $\delta \leq 6$, Figure 9(a), the MGR (for $n = 5$) and MGM (for $n = 20$ and 100) tests obtained higher power than the others. When $\delta > 6$ almost all the tests reached 100% power (Figures 9(b) and 9(c)), remembering that the proposed tests converged more slowly to this value. With 100 treatments, the power of these tests did not exceed 50%, even when the actual difference between averages was $32\sigma_{\bar{Y}}$ (Figure 9(c)). It is interesting to note that the SNK test tends to be slightly higher than the Tukey test in all configurations, as can also be seen in Borges & Ferreira (2003), and that this test has converged more quickly to 100%.

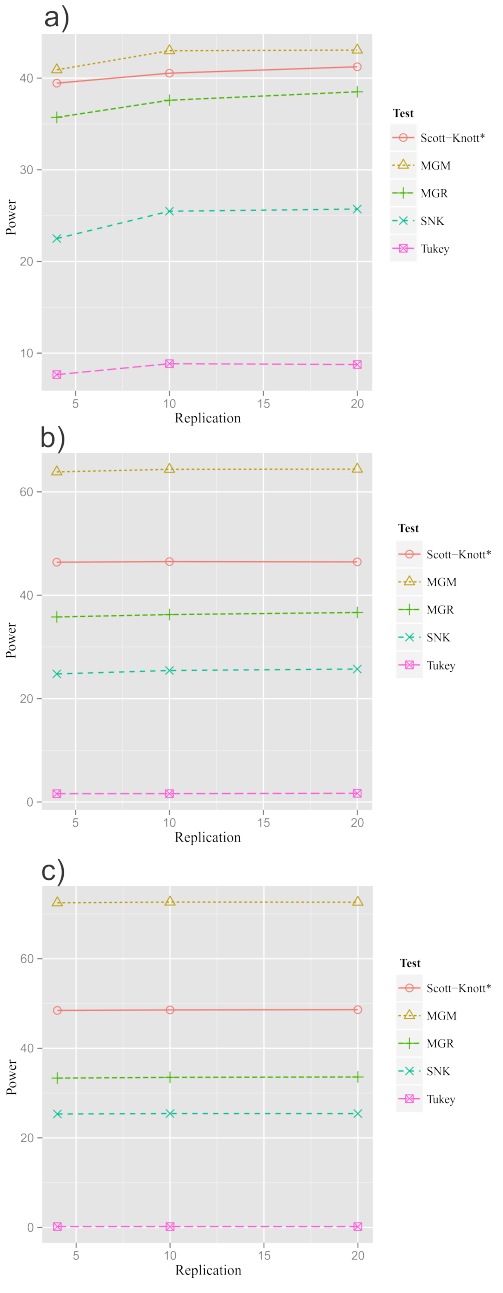
To compare the power of the proposed tests with the power of other tests found in the literature, the real difference between means was considered from 2 to $32\sigma_{\bar{Y}}$, for the number of treatments 5, 20 and 100, with 4 replications and $\alpha = 0.05$, Figure 10.

For a small real difference between means, regardless of the size of n, the MGM test obtained higher power than the others, being even more accentuated as n increased, especially from the Tukey test, a test with less power for this situation.

The t-Bayesian and Duncan tests were highlighted concerning power, as expected, because these two tests have high rates of error type I per experiment Bernhardson (1975), in the case of liberal tests. Being liberal tests, a high type I experimentwise error rate implies a small type II error rate and consequently high power. With the increase in the difference between means, these two tests converged more quickly at 100%.

Unlike the MGM test, the MGR test showed a power lower than the Scott-Knott test. However, in the performance evaluation, it was classified as having intermediate test power concerning that of the evaluated tests. The modified SNK and SNK tests also showed intermediate power, but lower than the MGR test for $n \leq 5$ and higher than the MGR test for $n > 5$, Figure 10.

When the actual difference between means increased, the power of the t-bayesian and Duncan tests increased, although the MGM test maintained high power and control of type I error per experiment.



* Results of Silva *et al.* (1999).

Figure 8. Power of the Scott-Knott, MGM, MGR, SNK, and Tukey tests, in percentage, under H_1 complete, to detect a difference between averages of $2\sigma_{\bar{Y}}$, with (a) $n = 5$, (b) $n = 20$ and (c) $n = 100$, depending on the number of replications, for a $\alpha = 0.05$.

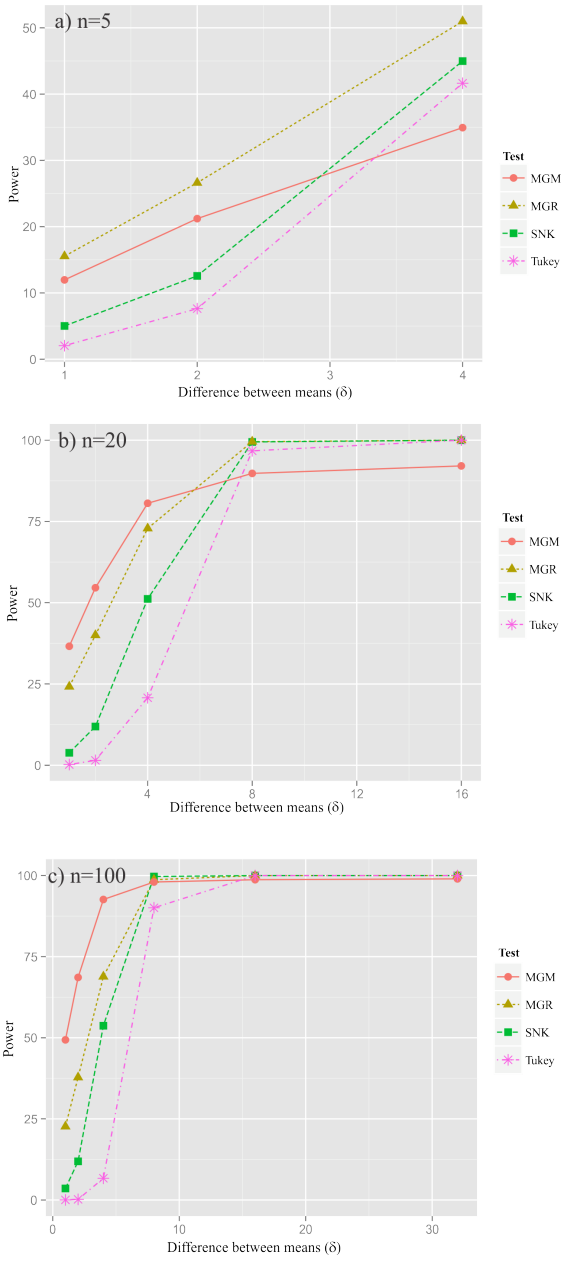


Figure 9. Power of the MGM, MGR, SNK, and Tukey tests, in percentage, under H_1 complete, to detect differences between averages from 1 to 32 $\sigma_{\bar{Y}}$, considering treatments (a) $n = 5$, (b) $n = 20$ and (c) $n = 100$, 4 replications, and a significance level of 5% probability.

The Scott-Knott and modified SNK tests obtained higher power than the MGR tests, with slightly higher values with an increase of n. Interestingly, in almost all configurations, the modified SNK and SNK tests were practically the same, except for the difference between means between 4 and $8\sigma_{\bar{Y}}$.

Comparing the power of the MGM and MGR tests with the CCR, CCF, CCRb and CCFb tests (Ramos & Ferreira, 2009; Ramos & Vieira, 2014), we noticed that the tests proposed in this work showed greater power for smaller differences between means ($\delta \leq 2$). As this difference increases, the CCRb and CCFb tests have greater power than the other tests.

A very relevant aspect in the proposed tests (MGR and MGM) was that although they may have presented a slower convergence to the maximum percentage of correct decisions (100%), for small values of δ, these tests were superior when compared to the other tests presented in this paper. In real experiments, this is the most common situation, Figure 9.

In Figure 11, we observe the setting for the real difference between means of 4 to $32\sigma_{\bar{Y}}$, for n = 5, 20 and 100 treatments, with 4 replications and α = 0.05. For this scenario, the proposed tests were compared with the Tukey and SNK tests. Regardless of the number of treatments, the MGM test obtained higher power, followed by the SNK and MGR tests. Once again, the test with the worst performance was Tukey's test. When the initial value of the difference between means was greater than $4\sigma_{\bar{Y}}$, the power of the tests quickly converged at 100%, since the real difference between means was very large.

In the present study, it was verified that the initial values of the real differences between means influenced the power of the proposed tests. This was not verified in any other study. Consider the power value of the MGM test as an example. In Table 9, the value of the difference between means (δ) for the n = 5, r = 4 and α = 0.05 scenario was 1 to 32. In Table 7, the value of δ was 2 to 32, and in Table 8, 4 to 32.

Note that the initial δ values were different. Thus, for these three scenarios, considering the same difference between averages of $4\sigma_{\bar{Y}}$, the power for the three situations was 34.94%, 69.91% and 89.58%, respectively, Figure 12.

This shows that the power of the proposed tests and the SNK test has increased as population means have become more heterogeneous. However, this did not happen with the MGR test. When the initial values of δ were 1 to 32 for 2 to 32, the power of this test increased when evaluating the same difference between means ($4\sigma_{\bar{Y}}$).

However, when the initial values of δ went from 2 to 32 to 4 to 32, the power of this test increased. Thus, what is observed is that the MGR test tends to be more powerful when it is evaluated medium populations that are more homogeneous than medium populations that are more heterogeneous. For the Tukey test, power has become constant for the same difference between increasingly more heterogeneous population means. This can be explained by the fact that the Tukey test is very conservative. Excessive control in type I error ends up influencing power, as predicted in the literature.

Another way to evaluate the tests is under the hypothesis under partial H_0. The evaluation took into account the number of treatments, number of replications, difference between means and level of significance.

The number of treatments was a point that influenced the power, under partial H_0, although the number of replications did not show as much influence. In Figure 13, it was observed that the increase in the number of treatments (n) decreases the power of the tests. However, when the real difference between means was $4\sigma_{\bar{Y}}$, Figure 13(c), the MGR test started to increase power with an increase of n, being the only test to reach power around 90% when n = 100. This test and the MGM test obtained the highest percentages of correct decisions. However, when $\delta \leq 4$, the power values did not exceed 30%. Even so, the Tukey test performed worst in almost all situations. With the increase of n, its power came close to 0%. From $\delta > 8$, almost all the tests converged to 100% power.

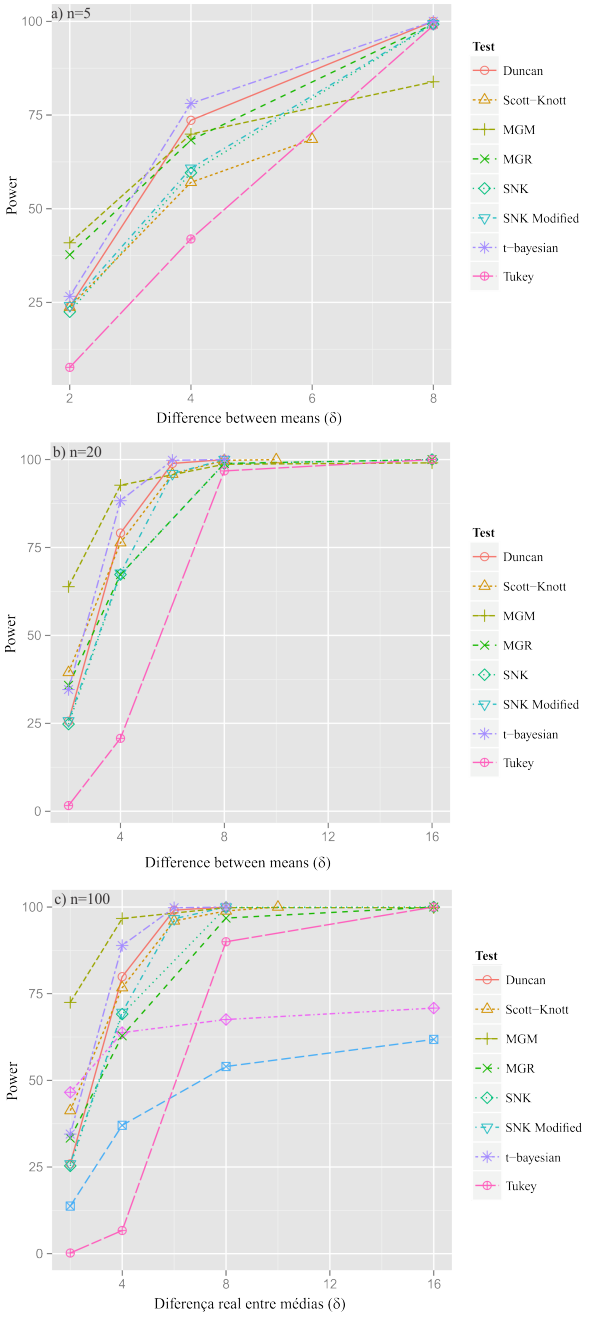


Figure 10. Power of the Scott-Knott MGM, MGR, SNK, Tukey, Duncan, modified SNK and t-Bayesian tests, in percentage, under H_1 complete, to detect differences between averages from 2 to 32 $\sigma_{\bar{Y}}$, considering the number of treatments (a) n = 5, (b) n = 20 and (c) n = 100, 4 replications, and a significance level of 5% probability.

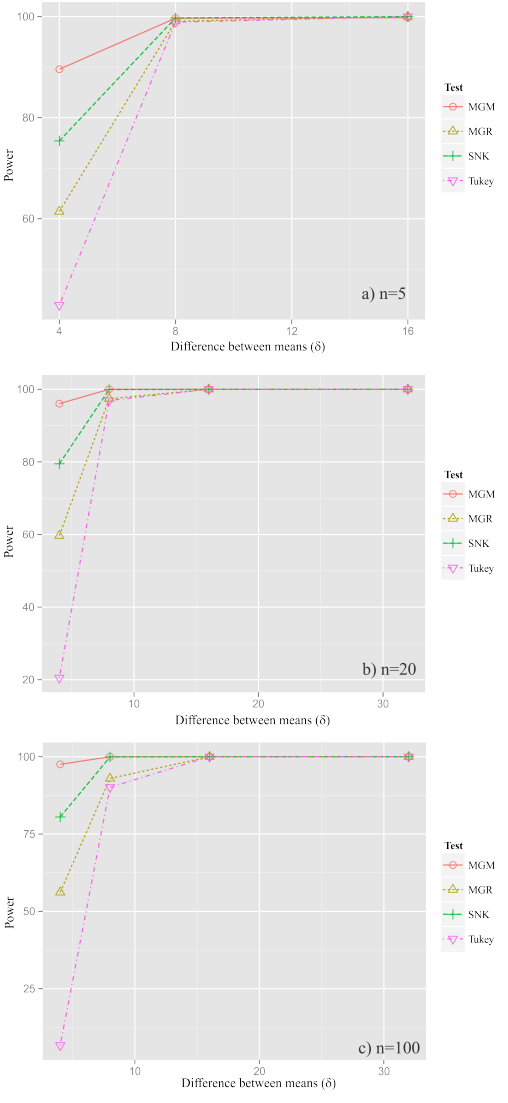


Figure 11. Power of the MGM, MGR, SNK and Tukey tests, in percentage, under H_1 complete, to detect differences between averages from 4 $\sigma_{\bar{Y}}$ to 32 $\sigma_{\bar{Y}}$, considering the number of treatments (a) $n = 5$, (b) $n = 20$ and (c) $n = 100$, 4 replications, and a significance level of 5% probability.

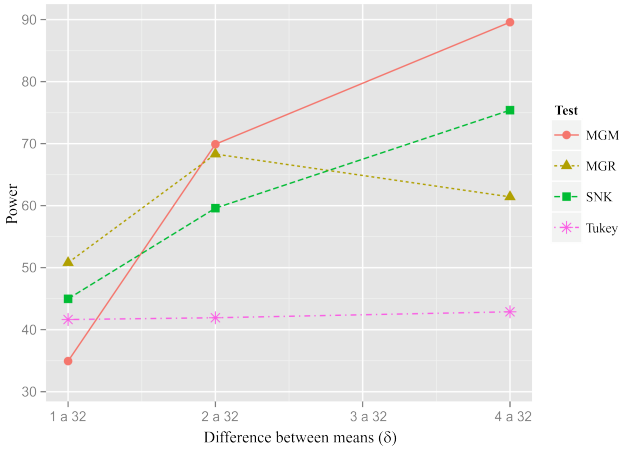


Figure 12. Power of the MGM, MGR, SNK and Tukey tests, in percentage, under H_1 complete, for the initial values of the actual differences between means for $4\sigma_{\bar{Y}}$, for n = 5, r = 4 and α = 0.05.

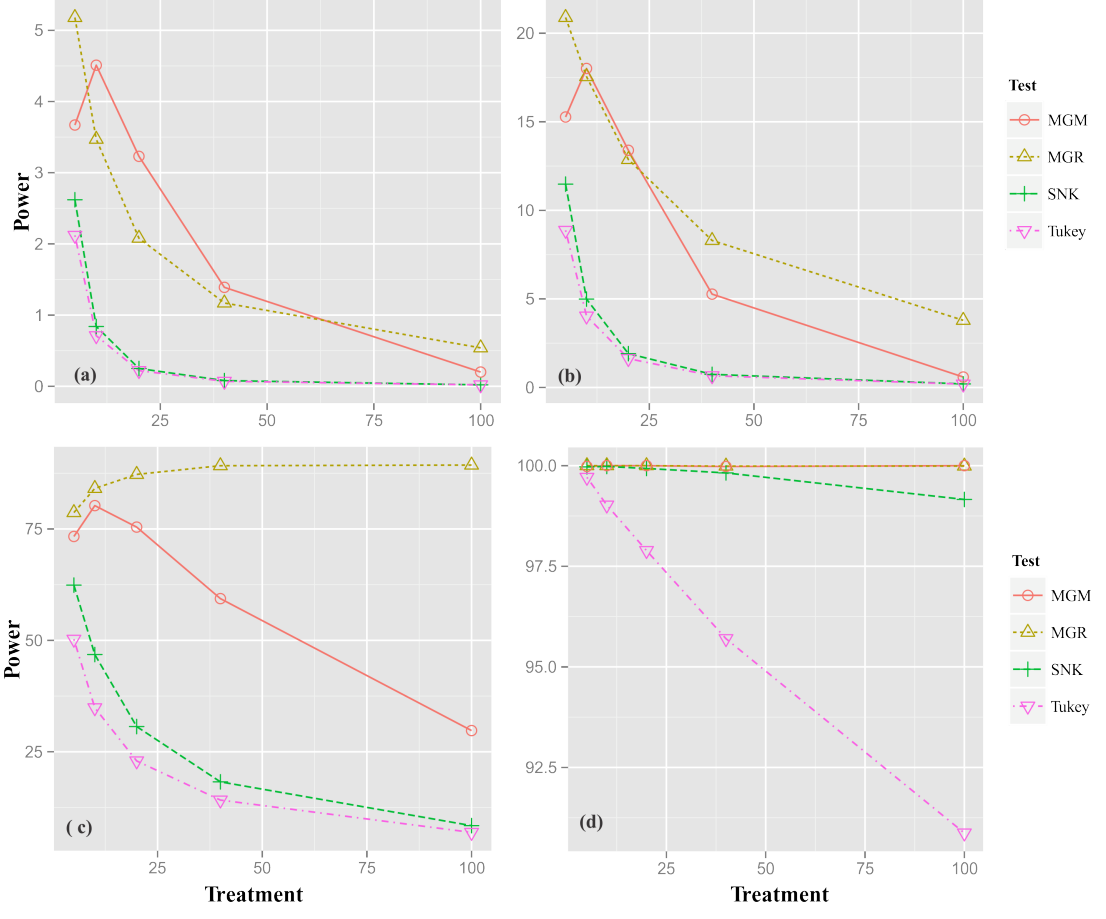


Figure 13. Power of the MGM, MGR, SNK and Tukey tests, in percentage, under partial H_0, depending on the number of treatments, to detect differences between averages of (a) 1 $\sigma_{\bar{Y}}$, (b) $2\sigma_{\bar{Y}}$, (c) $4\sigma_{\bar{Y}}$ e (d) $8\sigma_{\bar{Y}}$, with 4 replications, and a significance level of 5% probability.

Table 7. Power of the Tukey, SNK, MGR, and MGM tests, in percentage, to detect a difference between averages starting with two standard errors ($2\sigma_{\overline{Y}}$) to $32\sigma_{\overline{Y}}$, depending on the number of treatments and the number of replications equal to 4 ($r = 4$), under H_1 complete, at the significance level of 5% probability

		Real difference between means					
Test	**Treatment**	$1\sigma_{\overline{Y}}$	$2\sigma_{\overline{Y}}$	$4\sigma_{\overline{Y}}$	$8\sigma_{\overline{Y}}$	$16\sigma_{\overline{Y}}$	$32\sigma_{\overline{Y}}$
Tukey	5	-	7.660	41.930	98.880	-	-
	10	-	3.350	29.840	98.060	100.000	-
	20	-	1.610	20.740	96.770	100.000	100.000
	40	-	0.680	13.440	94.820	100.000	100.000
	100	-	0.190	6.690	89.980	100.000	100.000
SNK	5	-	22.500	59.600	99.220	-	-
	10	-	23.440	64.530	99.710	100.000	-
	20	-	24.780	67.310	99.850	100.000	100.000
	40	-	25.170	68.410	99.880	100.000	100.000
	100	-	25.330	69.110	99.890	100.000	100.000
MGR	5	-	35.720	68.320	99.320	-	-
	10	-	36.480	68.790	99.280	100.000	-
	20	-	35.800	67.260	98.900	100.000	100.000
	40	-	34.890	65.420	98.240	100.000	100.000
	100	-	33.340	62.840	96.810	100.000	100.000
MGM	5	-	40.910	69.910	83.880	-	-
	10	-	56.000	86.610	95.970	97.180	-
	20	-	63.860	92.710	98.670	99.070	99.220
	40	-	68.540	95.080	99.430	99.580	99.640
	100	-	72.480	96.690	99.820	99.820	99.890

The power of the tests evaluated had little practical significance, since the EER of all these tests was higher than the level of significance adopted, under partial H_0. Only the Tukey and Scheffé tests had EER identical to the nominal level, as verified in Carmer & Swanson (1973). However, the power of these EERs came to 0% in certain situations.

A characteristic that can be improved in the MGM and MGR tests, for the control of type I error by experiment and high power, under partial H_0, is to try to improve the contribution that the unknown population mean influences the MSD of the tests, since the distribution of the centred midrange in μ depends on the location parameter.

4. Application

We apply the tests the proposed tests, using the experiment of Figueiredo *et al.* (2015). The results will be compared with Scott-Knott's test.

Example 1 *The experiment was performed in a triple-lattice design evaluating the genotypes in 7 environments. For this study, the evaluation of the environments will not be taken into consideration, since it goes beyond the study objective of this work. The evaluation of the Scott-Knott test for the flowering period was also performed by Figueiredo* et al. *(2015) and is presented in Table 10. The additional information of this study was: an analysis of variance whose residual mean square was 6.3078 with 252 degrees of freedom. The number of replications with which the genotype means were estimated was 21.*

Table 8. Power of the Tukey, SNK, MGR, and MGM tests, in percentage, to detect a difference between averages starting with four standard errors ($4\sigma_{\bar{Y}}$) to $32\sigma_{\bar{Y}}$, depending on the number of treatments and the number of replications equal to 4 (r = 4), under H_1 complete, at the significance level of 1% probability

		Real difference between means					
Test	**Treatment**	$1\sigma_{\bar{Y}}$	$2\sigma_{\bar{Y}}$	$4\sigma_{\bar{Y}}$	$8\sigma_{\bar{Y}}$	$16\sigma_{\bar{Y}}$	$32\sigma_{\bar{Y}}$
	5	-	-	20.020	92.730	100.000	-
	10	-	-	14.430	92.290	100.000	100.000
Tukey	20	-	-	9.320	90.760	100.000	100.000
	40	-	-	5.840	87.440	100.000	100.000
	100	-	-	2.860	80.910	100.000	100.000
	5	-	-	47.790	97.530	100.000	-
	10	-	-	53.980	99.100	100.000	100.000
SNK	20	-	-	56.840	99.500	100.000	100.000
	40	-	-	58.250	99.590	100.000	100.000
	100	-	-	59.400	80.910	100.000	100.000
	5	-	-	54.120	94.850	100.000	-
	10	-	-	55.790	94.540	100.000	100.000
MGR	20	-	-	55.990	93.600	100.000	100.000
	40	-	-	54.710	91.740	100.000	100.000
	100	-	-	59.400	99.670	100.000	100.000
	5	-	-	74.340	95.050	97.500	-
	10	-	-	86.520	99.540	99.790	99.820
MGM	20	-	-	91.470	99.940	99.960	99.970
	40	-	-	93.450	99.990	100.000	100.000
	100	-	-	95.100	100.000	100.000	100.000

A data entry not very common in routines is the average of the treatments, which will be presented in this example. In this example, it will be shown that by entering only the results of the mean square of the residue, the degree of freedom and the number of replications, the `MRtest` function can perform the procedure of the four proposed tests.

In Table 11, the test results are presented for comparison, as well as the consecutive differences between the ordered means to assist in the comparison of the test results. Another aspect is the emphasis on the lines in which one of the tests separated the group of means.

The results show that the proposed tests (MGM and MGR) showed a greater separation of the groups of means than Scott-Knott's test in a more coherent way. The proposed tests showed very similar results. The means were ordered to facilitate discussion. Note the difference in test results in the first groups of means. The means of the genotypes BR507 and BR506 were considered statistically equal by the Scott-Knott test, but different by the MGM and MGR tests. Subsequently, the means of genotypes BR506 and BR508 were considered statistically equal by the proposed tests, but different by the Scott-Knott test. There is an inconsistency in the Scott-Knott test, which is very common in practice. Note the difference $\bar{Y}_{BR507} - \bar{Y}_{BR506} = 0.84$. The value of 0.84 between these two means was not enough for the Scott-Knott test to detect that they are sampled from populations with different means. However, this same test found that the difference $\bar{Y}_{BR506} - \bar{Y}_{BR508} = 0.63$ was significant, and therefore the mean effects of these genotypes are different. This is due to the philosophy of how the Scott-Knott test was developed. The separation of groups occurs by the likelihood ratio between groups. The differences between the limiting means of each group can often be smaller than the differences between consecutive means within the groups.

Table 9. Power of Tukey, SNK, MGR, and MGM tests, in percentage, to detect a difference between averages starting with standard error ($1\sigma_{\overline{Y}}$) to $32\sigma_{\overline{Y}}$, as a function of the number of treatments and the number of replications equal to 4 ($r = 4$), under H_1 complete, at the significance level of 5% probability

		Real difference between means					
Test	**Treatment**	$1\sigma_{\overline{Y}}$	$2\sigma_{\overline{Y}}$	$4\sigma_{\overline{Y}}$	$8\sigma_{\overline{Y}}$	$16\sigma_{\overline{Y}}$	$32\sigma_{\overline{Y}}$
Tukey	5	2.040	7.630	41.640	-	-	-
	10	0.700	3.580	30.960	98.130	-	-
	20	0.220	1.510	20.760	96.730	100.000	-
	40	0.070	0.620	13.080	94.690	100.000	100.000
	100	0.020	0.190	6.690	90.060	100.000	100.000
SNK	5	5.020	12.570	44.980	-	-	-
	10	4.150	11.930	48.350	99.090	-	-
	20	3.840	11.920	51.180	99.470	100.000	-
	40	3.590	11.760	52.590	99.620	100.000	100.000
	100	3.550	11.910	53.710	99.700	100.000	100.000
MGR	5	15.540	25.620	50.800	-	-	-
	10	23.690	39.270	72.520	99.620	-	-
	20	24.190	40.020	72.870	99.560	100.000	-
	40	23.630	39.280	71.380	99.330	100.000	100.000
	100	22.650	37.790	68.850	98.710	100.000	100.000
MGM	5	11.970	21.210	34.940	-	-	-
	10	25.910	41.000	65.540	78.030	-	-
	20	36.610	54.610	80.600	89.790	92.090	-
	40	43.360	62.150	87.280	94.480	95.910	96.480
	100	49.360	68.600	92.620	98.010	98.720	98.980

Unlike Scott-Knott's test, the MGM and MGR tests are more consistent in this respect. The difference between the BR506 and BR508 genotypes of 0.63 was not sufficient for the proposed tests to evaluate these two genotypes as statistically different. However, for the major difference between the BR507 and BR506 genotypes of 0.84, they were statistically different.

However, in one situation the MGR test did not get rid of this aspect either. Verifying the difference between the genotypes V82393 and V82392, which was 1.31, the MGR test did not detect a difference between these means, as it was not verified by Scott-Knott's test. This question is because the difference between the BR507 and BR506 genotypes of 0.84 was detected as a significant difference by the MGR test. For the MGM test, this does not occur; the difference for genotypes V82393 and V82392 of 1.31 was detected as statistically different genotypes. Only in one situation did none of the tests detect significance in a difference of 0.86 (difference between the genotypes CMSXS639 and CMSXS642). The smallest significant difference detected for the MGM and MGR tests was 0.84, and for the Scott-Knott test was 0.63. Thus, all tests above these values should also detect differences. It is worth remembering that for the proposed tests, the values 0.84 and 0.86 are very close, being a threshold for these tests to detect the significance of the difference between the means.

In all other situations in which Scott-Knott's test differentiated the groups of means, the MGM and MGR tests were also able to detect. Taking into account that the MGM test further refined the group separation.

More coherently, the greater group separation occurs in the MGM and MGR tests due to the development of how the tests were proposed. The separation of the groups of these tests takes into

Table 10. Selection of twenty-five sorghum genotypes based on the flowering period, evaluated by the Scott-Knott test

Genotype	Flowering period	Scott-Knott test
CMSXS643	87.51	A
CMSXS630	85.78	B
BR507	85.17	B
BR506	84.33	B
BR508	83.70	C
CMSXS629	83.27	C
BR501	83.25	C
CMSXS635	82.48	C
CMSXS644	82.35	C
BRS511	81.42	D
CMSXS648	81.12	D
CMSXS633	80.91	D
BR505	79.91	E
CMSXS637	79.59	E
XBSW80140	79.35	E
CMSXS646	78.59	E
BRS601	78.33	E
CMSXS639	78.15	E
CMSXS647	77.29	E
SUGARGRAZE	75.45	F
CMSXS636	75.43	F
V82391	75.36	F
XBSW80007	75.15	F
V82393	73.83	G
V82392	72.52	G

Table 11. Results of the MGM, MGR and Scott-Knott tests evaluating the 25 sorghum genotypes presented in Example 1

			Tests		
Genotype	Means	Difference between means	*MGM*	*MGR*	*Scott-Knott*
CMSXS643	87.51	-	g1	g1	A
CMSXS630	85.75	1.77	g2	g2	B
BR507	85.17	0.58	g2	g2	B
BR506	84.33	0.84	g3	g3	B
BR508	83.70	0.63	g3	g3	C
CMSXS629	83.27	0.43	g3	g3	C
BR501	83.25	0.01	g3	g3	C
CMSXS635	82.48	0.77	g3	g3	C
CMSXS644	82.35	0.13	g3	g3	C
BRS511	81.42	0.93	g4	g4	D
CMSXS648	81.12	0.30	g4	g4	D
CMSXS633	80.91	0.21	g4	g4	D
BR505	79.91	1.00	g5	g5	E
CMSXS637	79.59	0.32	g5	g5	E
XBSW80140	79.35	0.24	g5	g5	E
CMSXS646	78.59	0.76	g5	g5	E
BRS601	78.33	0.26	g5	g5	E
CMSXS639	78.15	0.18	g5	g5	E
CMSXS647	77.29	0.86	g5	g5	E
SUGARGRAZE	75.45	1.84	g6	g6	F
CMSXS636	75.43	0.02	g6	g6	F
V82391	75.36	0.07	g6	g6	F
XBSW80007	75.15	0.21	g6	g6	F
V82393	73.83	1.32	g7	g7	G
V82392	72.52	1.31	g8	g7	G

account the greater consecutive difference between means, and this was determinant to avoid the inconsistency that often occurs in the Scott-Knott test.

Example 2 *Calinski & Corsten (1985) proposed two grouping methods, one based on the F distribution, we will call it the CF test, and the other based on the studentized range, we will call it the CR. For these authors, the idea of the tests was to provide unambiguous results, to have a separation of small groups and that this separation would provide more homogeneous groups among any other formation of groups, that is, groups with lower intra-group variances. Thus, they exemplified the application of the two proposed tests applying them in the experiment analyzed by Duncan (1955) and then by Scott & Knott (1974). The experiment evaluated the yields (bushels per acre) of seven varieties of barley were compared in a randomized block design, which contained 6 blocks. The means of the varieties were:*

Treatments	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Means	*49,6*	*58,1*	*61,0*	*61,5*	*67,6*	*71,2*	*71,3*

We will present the results for the tests proposed by Calinski & Corsten (1985), the Scott-Knott test, and the MGM and MGR tests, in Table 12. Results that present equal letters in the means between treatments represent that they are statistically equal. The different letters represent the means of different groups. To verify

Table 12. Result of multiple comparison tests

	Test results				
Varieties	***CF***	***CR***	***Scott-Knott***	***MGM***	***MGR***
1	b	b	b	c	b
2	b	b	b	b	a
3	b	b	b	b	a
4	b	b	b	b	a
5	a	a	a	a	a
6	a	a	a	a	a
7	a	a	a	a	a

the homogeneity of the groups, we applied a weighted average of the variances of the formed groups, in which the weights were the degrees of freedom computed in each group. For example, for the CF test, we have two groups formed (1-4)(5-7). Thus, the variance of means for the first group was 30.33667, and for the second it was 4.443333. In the first group, there are four means, and therefore, 3 degrees of freedom. In the second group, there were 2 degrees of freedom. Thus, the value for the weighted mean of the variances of the groups for the CF test is $(30.33667 \times 3 + 4.443333 \times 2)/5 = 19.97933$. *For the other tests, the weighted average of the variances is given in Table 13.*

Table 13. The weighted average of the variances of the groups formed from the multiple comparison tests

Tests	**Average of variances**
CF	19.97933
CR	19.97933
Scott-Knott	19.97933
MGM	3.906667
MGR	32.13367

It can be observed that the MGM test presented more homogeneous groups and the MGR test presented the formation of more heterogeneous groups. One could think that this occurred because the MGM test formed

more groups. However, as well mentioned by Calinski & Corsten (1985), it can be observed that the difference between consecutive means of treatments 1 and 2, in the order of $2.33\sigma_{\bar{Y}}$*, was greater than the difference between treatments 4 and 5, in the order of* $1.67\sigma_{\bar{Y}}$*. These last treatments were the limited treatments for the breakdown of group formation. However, the formation of the groups by the CF, CR and Scott-Knott tests, with the inclusion of treatment 1 in the group (1-4), provided that this group had a greater sum of squares, and consequently, a greater variance for the group. This shows that the treatment included in this group differs from the other treatments and, therefore, could not be included in the group. This was verified by the MGM test, which resulted in the formation of groups (1)(2-4)(5-7).*

Something that also draws attention is that Ramos & Vieira (2014) evaluated these tests, and the power of the CF and CR tests is greater than the power of the MGM test. Considering, a scenario similar to this experiment, with 5 treatments and 4 replications under complete H_1*, the power of the CF and CR tests to detect the difference between means of* $2\sigma_{barY}$ *or more, is greater than approximately 30%. For the MGM and MGR tests, the power of the tests is in the order of 25% and 21%, respectively. Still, the tests proposed by Calinski & Corsten (1985) could not detect separation of treatment 1 form treatments 2 to 4. This can be explained because for the CF and CR tests to have reached the formation of this group, they had a type I error rate per experiment ranging from 1% up to 53.1%.*

Thus, the fact that the groups formed by the MGM test were no longer homogeneous because they separated more groups, but rather because they formed groups with similar characteristics.

These results presented in the two examples do not mean that this will always happen for the MGM and MGR tests. However, one can observe the good characteristics that these tests, whereas the classical tests such as Scott-Knott could not detect such differences. Thus, the idea is not to show that these proposed tests are better than those already present in the literature, but to be an alternative for the use of procedures that present good characteristics, with the control of type I error per experiment, under complete H_1, high power, and without ambiguity in its results, that is, a new alternative for the formation of groups of means.

Based on Carmer & Swanson (1973), one should not develop an MCP giving total emphasis to type I error, because in this view one can see the fragility of Tukey's test. Nor even less want the formation of smaller groups in a more homogeneous way, since it ends up generating some inconsistencies as shown in the Example 2. However, we cannot make choices like those made by Carmer & Swanson (1973 when choosing Fisher's protected T-test or the t-Bayesian test, because it performs well in some evaluations, however, very high rates of type I error. One should be very cautious in the usage choices of an MCP because the search for good multiple comparison procedures continues since this still represents a gap in science.

5. Conclusions

The proposed MGM and MGR tests performed better than the Skott-Knott's test, in most of the evaluations performed, except for the type I error per experiment under partial H_0, which even the Skott-Knott test does not control. Among all the tests evaluated in this study, the MGM test presented the best performance in almost all evaluation configurations, adding the advantage of not presenting ambiguity in its results.

However, we noticed some limitations of the proposed tests (MGM and MGR tests). As the number of treatments increases in the simulated scenarios, under partial H_0, the empirical type I error rate increases. While the only test presented that controls the type I error rate at the nominal level of significance is the Tukey's test, but with decreasing power the more the number of treatments in the simulated scenarios is increased. This means that the decision to choose the best test to be used must be taken with caution and verification of the advantages and disadvantages and in which experimental scenario the test is being applied.

We are presenting two more possible tests to be used in multiple comparison procedures, allowing the researcher more test options for decision making.

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. Batista, B. D. O. & Ferreira, D. F. Alternative to Tukey test. **44,** 1–10 (2020).
2. Batista, B. D. O. & Ferreira, D. F. SMR: An R Package for Computing the Externally Studentized Normal Midrange Distribution. *The R Journal* **6,** 123–136 (2014).
3. Batista, B. D. O. & Ferreira, D. F. *SMR: Externally Studentized Midrange Distribution* R package version 2.0.1 (Vienna, Austria, 2014). http://cran.r-project.org/web/packages/SMR/index.html.
4. Batista, B. D. O. & Ferreira, D. F. The Externally Studentized Normal Midrange Distribution. *Ciência e Agrotecnologia* **41,** 378–389 (2017).
5. Bernhardson, C. S. 375: Type I Error Rates When Multiple Comparison Procedures Follow a Significant F Test of ANOVA. English. *Biometrics* **31,** 229–232 (1975).
6. Bhering, L. L., Cruz, C. D. a., Vasconcelos, E. S. d., Ferreira, A. & Resende Jr, M. F. R. d. Alternative methodology for Scott Knott test. *Crop Breeding and Applied Biotechnology* **8,** 9–16 (2008).
7. Boardman, T. J. & Moffitt, D. R. Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedures. *Biometrics* **27,** 738–744 (1971).
8. Borges, L. C. & Ferreira, D. F. Poder e taxas de erro tipo I dos tetes Scott-Knott, Tukey e Student-Newman-Keuls sob Distribuições normal e não normais dos resíduos. *Revista Matemática e Estatística* **21.** (Portuguese), 67–83 (2003).
9. Calinski, T. & Corsten, L. C. A. Clustering Means in Anova by Simultaneous Testing. *Biometrics* **41,** 39–48 (1985).
10. 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).
11. Conrado, T. V., Ferreira, D. F., Scapim, C. A. & Maluf, W. R. Adjusting the Scott-Knott cluster analyses for unbalanced designs. *Crop Breeding and Applied Biotechnology* **17,** 1–9 (2017).
12. Cui, X., Dickhaus, T., Ding, Y. & Hsu, J. *Handbook of Multiple Comparisons* 418 (CRC Press, Boca Raton, 2021).
13. David, H. A., Hartley, H. O. & Pearson, E. S. The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation. English. *Biometrika* **41,** 482–493 (1954).
14. David, H. A. & Nagaraja, H. N. *Order Statistics* 458 (John Wiley & Sons, Canada, 2003).
15. Duncan, D. B. Multiple range and multiple F tests. *Biometrics* **11,** 1–42 (1955).
16. Einot, I. & Gabriel, K. R. A Study of the Powers of Several Methods of Multiple Comparisons. *Journal of the American Statistical Association* **70,** 574–583 (1975).

17. Figueiredo, U. J., Nunes, J. A. R., Parrella, R. A. C., Souza, E. D., Silva, A. R., Emygdio, B. M., Machado, J. R. A. & Tardin, F. D. Adaptability and stability of genotypes of sweet sorghum by GGEBiplot and Toler methods. *Genetics and Molecular Research* **14,** 11211–11221 (2015).

18. Graybill, F. *An introduction to linear statistical models* **1,** 463 (McGraw-Hill, New York, 1961).

19. Gumbel, E. J. *Statistics of Extremes* 375 (Columbia University Press, New York, 1958).

20. Hartley, H. O. The Range in Random Samples. English. *Biometrika* **32,** 334–348 (1942).

21. Keuls, M. The use of the "studentized range" in connection with an analysis of variance. *Euphytica* **1,** 112–122 (1952).

22. Leemis, L. M. & Trivedi, K. S. A Comparison of Approximate Interval Estimators for the Bernoulli Parameter. *The American Statistician* **50,** 63–68 (1996).

23. Newman, D. The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. *Biometrika* **31,** 20–30 (1939).

24. 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).

25. 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).

26. Pearson, E. S. & Hartley, H. O. Tables of the Probability Integral of the Studentized Range. English. *Biometrika* **33,** 89–99 (1943).

27. Pearson, E. S. & Hartley, H. O. The Probability Integral of the Range in Samples of n Observations From a Normal Population. *Biometrika* **32,** 301–310 (1942).

28. Pearson, E. S. A Further Note on the Distribution of Range in Samples Taken from a Normal Population. *Biometrika* **18,** 173–194 (1926).

29. Pearson, E. S. The Percentage Limits for the Distribution of Range in Samples from a Normal Population. *Biometrika* **24,** 404–417 (1932).

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

31. R CORE TEAM. *R. A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022). https://www.R-project.org/.

32. Ramos, P. S. & Ferreira, D. F. Agrupamento de médias via bootstrap de populações normais e não-normais. **56.** (Portuguese), 140–149 (2009).

33. Ramos, P. S. & Vieira, M. T. Bootstrap multiple comparison procedure based on the F distribution. **31,** 529–546 (2014).

34. Rider, P. R. The midrange of a sample as an estimator of the population midrange. *Journal of the American Statistical Association* **52,** 537–542 (1957).

35. Sauder, D. C. & DeMars, C. E. An Updated Recommendation for Multiple Comparisons. *Advances in Methods and Practices in Psychological Science* **2,** 26–44 (2019).

36. Scott, A. J. & Knott, M. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. English. *Biometrics* **30,** 507–512 (1974).

37. Searle, S. R. *Linear models for unbalanced data* 536 (Wiley, New York, 1987).

38. Shimokawa, T. & Goto, M. Hierarchical cluster analyis for multi-sample comparisons based on the power-normal distribution. *Behaviometrika* **38,** 125–138 (2011).

39. 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.** (Portuguese), 687–696 (1999).

40. Student. Errors in routine analysis. *Biometrika* **19,** 151–164 (1927).

41. Tukey, J. W. The problem of multiple comparisons. *Mimeographed monograph.* Unpublished memorandum in private circulation (1953).

42. Waller, R. A. & Duncan, D. B. A Bayes Rule for the Symmetric Multiple Comparisons Problems. *Journal of the American Statistical Association* **64,** 1484–1503 (Dec. 1969).