BOOTSTRAP MULTIPLE COMPARISON PROCEDURE BASED ON THE F DISTRIBUTION

Patrícia de Siqueira RAMOS[1]
Mariani Tabarim VIEIRA[2]

▪ *ABSTRACT: This work is aimed to propose a bootstrap version of the multiple comparison test based on the F distribution presented by Caliński and Corsten in* 1985 *and compare it to the original version by Monte Carlo simulation. The two procedures test the homogeneity of treatment means within each of two or more subsets using minimization of the sum of squares among partitions as criterion. Their performances are evaluated computing the Type I error rates and power, and the simulation algorithms were implemented in* `R` *software. Under* H_0 *and normality, both tests control the Type I error rates. Under* H_0 *and non-normality, the bootstrap test (CFB) controls the Type I error rates and therefore is robust while the original test (CF) is conservative under lognormal distribution for* $k = 10$ *and* $r = 10$*. Under partial* H_0*, the CF test is liberal for small differences,* δ*, between means and conservative for larger differences, while the CFB test is always liberal and with higher Type I error rates with* δ *increasing. Under* H_1*, the CFB test is more powerful due to its liberal trait for normal and non-normal distributions. The CF test is recommended in practical situations since it controls Type I error rates in most situations and shows higher power values.*

▪ *KEYWORDS: Comparison of means; resampling; cluster analysis; sum of squares; Monte Carlo simulation; Type I error rate; power.*

1 Introduction

In experimental research investigators usually need to compare more than two different treatment means, which is mostly done using some Multiple Comparison

[1]Universidade Federal de Alfenas – UNIFAL, Instituto de Ciências Exatas, CEP 37130-000, Alfenas, MG, Brasil. E-mail: *patricia.ramos@unifal-mg.edu.br*
[2]Universidade Federal de Alfenas – UNIFAL, Faculdade de Nutrição, CEP 37130-000, Alfenas, MG, Brasil. E-mail: *marianivieira@hotmail.com*

Procedure (MCP), applied after an F test (Machado et al., 2005). The MCPs are statistical methods used to perform all pairwise comparisons between means. There are many procedures of this type, but researchers have difficulties in working with them because most of them provide ambiguous results, which affects determining the best treatments (Silva et al., 1999). Care must be taken when conducting hypothesis testing or interval estimation to choose correct inference procedures. The main problem is the multiplicity effect which can lead to many tests with incorrect significant results (Hinkelmann and Kempthorne, 1987).

When testing the null hypothesis that all treatment means are equal,

$$H_0 : \mu_1 = \mu_2 = ... = \mu_i = ... = \mu_k, \tag{1}$$

where k is the number of treatments, there is a risk of committing decision errors, both Type I and Type II errors. The Type I error occurs when the true null hypothesis, H_0, is rejected and the Type II occurs when the false H_0 is not rejected (Mood et al., 1974). When the test rejects H_0 we investigate where differences between means are using MCPs. In experimental research, it is very important to use tests with greater ability to control Type I error rate and show high power values, that is, detect real differences between the treatment means. The researcher controls Type I error by determining the nominal significance level α (Machado et al., 2005).

According to Steel and Torrie (1980) there are two main ways to measure Type I error rates of MCPs. The first one measures the Type I error rate by comparison, called comparisonwise error rate, which is the probability of rejecting a true hypothesis in all possible combinations of treatment means, taken two by two. The second and more common way calculates Type I error rates by experiment. The experimentwise error rate is the probability of performing at least one wrong inference by experiment.

Resampling methods like bootstrap have been used in multiple comparison studies of treatment means in many situations such as heteroscedasticity and non-normal data (Keselman et al., 2002; Krishnamoorthy and Mathew, 2007; Ramos end Ferreira, 2009). Efron (1979) introduced the bootstrap method that uses statistical inference in its traditional form without so many mathematical assumptions, and it is a versatile and computationally intensive method. When there is no knowledge about the population, new samples of size n are taken from the original sample using sampling with replacement and, in each resampling, an estimate of the parameter is obtained. This procedure is performed many times obtaining several estimates of the parameters that will be used to generate the bootstrap distribution (Manly, 1978).

Caliński and Corsten (1985) and Scott and Knott (1974) proposed some examples of tests that are based on cluster analysis. This type of methodology avoids ambiguity in the results by making good discrimination of treatment means. Silva et al. (1999) evaluated the Scott-Knott procedure by simulation and recommend it due to its high power values and robustness. The Caliński and Corsten (1985) procedure, based on the studentized range, was evaluated by Ramos and Ferreira

(2009), and a bootstrap version was proposed and evaluated. Both tests control the Type I error in various situations and are powerful.

The use of analytical methods to evaluate the performance of statistical tests by Type I error rates and power is quite complicated. An alternative is the Monte Carlo simulation which avoids the analytical difficulties and helps obtaining results more easily. This is a helpful procedure to evaluate the performance of statistical tests because the values obtained are very close to those found analytically (Silva et al., 1999). Several studies evaluating MCPs analyzed Type I error rates and power with Monte Carlo simulations, such as: Carmer and Swanson (1973), Silva et al. (1999), Perecin and Barbosa (1988), Borges and Ferreira (2003), Ramos end Ferreira (2009), among others.

All MCPs test the same null hypothesis but each of them has a different approach to control the Type I error. Therefore, sometimes, the same data can give different results depending on the method chosen. Although many authors have evaluated the performance of MCPs, it is very difficult to give general recommendations about the procedure that should be chosen (Hinkelmann and Kempthorne, 1987).

According to these simulation studies there are tests, such as Duncan and Student's t-test, that have high experimentwise error rates, sometimes much larger than the nominal significance level α, and they are considered very liberal. Other tests, such as Tukey's and Scheffe's tests present experimentwise error rates generally smaller than the nominal significance level and are considered very conservative. Some tests, such as SNK's test, control the experimentwise error under null H_0 but it is very liberal under partial H_0 (Carmer and Swanson, 1973).

This paper aims to propose a bootstrap version of the multiple comparison test based on the F distribution presented by Caliński and Corsten in 1985 and compare it to the original version, relative to the experimentwise Type I error rate and power by Monte Carlo simulation.

2 Methodology

The performance of two multiple comparison procedures were compared using Monte Carlo simulation. The first one, proposed by Caliński and Corsten (1985), is an extension of the F distribution and the second one is a bootstrap version of the first. Monte Carlo simulations were made to evaluate the performance of the tests under null hypothesis and under partial null hypothesis by measuring Type I error rates and power. All algorithms needed to implement and evaluate the tests were performed using `R` software, version 2.14.0 (R Development Core Team, 2011).

2.1 Test based on the F distribution

Let $X_{11}, X_{12}, \ldots, X_{1r}, X_{21}, \ldots, X_{2r}, \ldots, X_{kr}$ be a random sample in which k are qualitative and unstructured treatments and r are replications where X_{ij} is the observation corresponding to the i-th treatment in the j-th sample or experimental

unit ($i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, r$). The mean estimator of each treatment ($\bar{X}_{i\cdot}$) is given by

$$\bar{X}_{i\cdot} = \frac{\sum_{j=1}^{r} X_{ij}}{r}. \tag{2}$$

The F test in the analysis of variance is used to verify if there are real differences between treatments. When the null hypothesis (1) of equality between all treatment means is rejected, the alternative hypothesis is considered. Thus, there is at least one difference between the treatment effects. The next step is to investigate which treatments are different.

The first procedure tests the homogeneity of means within every two or more subgroups. The first group comprises all means and, at each step, means are subsequently split and the number of partitions p grows, $p = 1, 2, 3, \ldots, k$.

The stopping criterion is based on the sum of squares SS_p of the considered subgroup. At each p, the partition chosen is that one with the smallest sum of squares within-group, which is equivalent to finding a partition in p groups for which the sum of squares between groups is the greatest. For this purpose we used `kmeans` function of `R` software and the method chosen was Hartigan-Wong (Hartigan and Wong, 1979). As at each stage different groups may be formed is therefore called a non-hierarchical method.

The k-means is a method for clustering that divides M points in k groups to minimize the sum of squares within-group. It is practically impossible to find, among all existing partitions, the one which returns the lowest SS_p and, thus, the methods look for a local optimum (Hartigan and Wong, 1979). Under this limitation, the method does not return the partition with the smallest SS_p. The R function `kmeans` allows the choice between four algorithms and we chose Hartigan-Wong because it is recommended by the `kmeans` function authors on the `R` software site. The value chosen for the parameter that sets the inicial values `nstart` was 10.

This test uses the F distribution to calculate the p-value by

$$p\text{-value} = Pr\left(F(k-1, \nu) > \frac{SS_p}{MSE(k-1)}\right) \quad p = 1, 2, 3, \ldots, k, \tag{3}$$

where MSE is the mean square error from the analysis of variance, k is the number of treatments and ν are the degrees of freedom associated with the MSE. One different p-value will be obtained for each partition p.

The p-value is compared to the nominal level of significance α. The first p partition where the p-value is greater than the significance level α is the cutoff and the group obtained will be considered.

2.2 The bootstrap test

The difference between the bootstrap version and the test presented above is the p-value calculation used to verify the significance of the partition. To perform the bootstrap resampling, the rk observations are combined into one sample where r resamplings with replacement form new samples of each treatment. From the samples, new means are calculated and a f_b value is obtained by

$$f_b = \frac{SS_p^b}{MSE^b(k-1)} \tag{4}$$

for each resampling, where SS_p^b is the sum of squares within group of the b-th bootstrap sample, MSE^b is the mean square of the b-th bootstrap sample in the p-th partition and $b = 1, \ldots, B$ (see below) refers to resamplings.

This process is repeated $1,000$ times ($B = 1,000$) and the set with all f_b values obtained is used to calculate the p-value by

$$p - \text{value} = \frac{\sum_{b=1}^{B} I\left(\frac{SS_p}{MSE(k-1)} < f_b\right)}{B} \quad p = 1, 2, 3, \ldots \tag{5}$$

where $I(x)$ is the indicator function of x. A $p-$value is computed for each partition p.

The p-value is compared with the nominal significance level α for each p partition and the same criteria of the original test is applied to obtain clusters with equal or homogeneous treatments.

2.3 Simulations

The number of simulations was set to $N = 1,000$. Several values of unstructured qualitative k treatments, r replications and δ standard errors of mean were used. The δ represents the differences between consecutive means, given that power was assessed under a partial null hypothesis, wherein some means differed by δ standard errors of mean and under the alternative hypothesis, in which mean values were all different. The standard error of mean is given by

$$\sigma_{\bar{X}} = \sqrt{\frac{MSE}{r}}.$$

As said before, Caliński and Corsten (1985) proposed two multiple comparison procedures, Ramos and Ferreira (2009) proposed a bootstrap version of the one based on the studentized range. These last authors evaluated the original and the modified version. To evaluate the performance of the tests, they used numbers of treatments k equal to 5, 10, 20 and 80, and r replications equal to $r = 4$, 10 and 20. In this paper, we use the same values of k and r to compare the procedures by them and by us. The nominal significance level adopted by Ramos and Ferreira (2009)

was $\alpha = 0.01$ and 0.05. However, as the results for the two values were similar, we adopted $\alpha = 0.05$ as it is the most common value. We used the normal distribution model, in which the test was originally conceived, but also the lognormal and the exponential models to evaluate the robustness of the two tests in adverse situations.

Moreover, we considered different hypotheses about means. Under the null hypothesis (H_0) and normal and lognormal models, the treatment means were simulated from a distribution having the same mean, $\mu = 10$, without loss of generality, and variance $\sigma^2 = 1$. To the exponential model, we adopted $\lambda = 0.1$, which is equal to mean $\mu = 10$ and variance $\sigma^2 = 100$.

Under the partial null hypothesis, we considered two groups whose means of the same group were equal and the means of the two groups were different from each other by δ standard errors of mean, and $\delta = 1, 2, 4, 8$ and 16. The first group was simulated from the same model as null H_0 and the second group had all the same treatment means but different from the first group. Under the alternative hypothesis (H_1), the nearest means were all different from each other by $\delta = 1$ standard error of mean, but the variance was considered constant and equal to 1.

We used exact binomial tests to decide if the observed Type I error rates were equal to $\alpha = 0.05$. The nominal significance level was 0.01 to test $H_0 : \alpha = 0.05$ against $H_1 : \alpha \neq 0.05$ and the hypothesis testing was applied to each of the observed Type I error rates. If the null hypothesis is rejected and the observed Type I error rates are significantly (p-value < 0.01) lower than the nominal significance level, the test will be considered conservative. If the null hypothesis is rejected and the observed Type I error rates are significantly (p-value < 0.01) higher than the nominal significance level, the test will be considered liberal. However, if the observed Type I error rates are not considered (p-value > 0.01) different from the nominal significance level, the test will be considered exact, which means that it controls the Type I error.

The two tests were applied to all simulated settings and the Type I error rates and power were estimated in all N generated experiments. Under the complete and partial null hypotheses the Type I error rates were obtained and, under the partial H_0 and H_1, the power was estimated.

3 Results

The results of the Type I error rates and power values of the tests under complete and partial H_0 and under H_1 are presented in the following. The nominal significance level was 0.05 and different numbers of replications, treatments and differences between treatments were considered.

3.1 Type I error rates under null H_0

Table 1 shows the experimentwise error rates as a function of r and k of the Caliński and Corsten test based on the F distribution (CF) and its bootstrap version (CFB) for null H_0 and normal distribution. We also show the results of Caliński

Table 1 - Experimentwise Type I error rates of tests CF, CFB, SK, C and CB as a function of number of replications r and number of treatments k, for $\alpha = 0.05$, under null H_0

r	k	CF	CFB	SK	C	CB
4	5	0.050	0.050	0.051	0.049	0.053
4	10	0.056	0.055	0.043	0.055	0.057
4	20	0.059	0.060	0.038	0.047	0.050
4	80	0.058	0.057	0.036	0.044	0.051
10	5	0.059	0.061	0.067	0.047	0.046
10	10	0.048	0.045	0.062	0.042	0.042
10	20	0.053	0.051	0.052	0.048	0.054
10	80	0.051	0.048	0.052	0.048	0.050
20	5	0.058	0.056	0.059	0.057	0.058
20	10	0.045	0.044	0.051	0.044	0.046
20	20	0.064	0.062	0.056	0.054	0.054
20	80	0.041	0.041	0.050	0.041	0.042

and Corsten test based on the studentized range (C) and its bootstrap version (CB) proposed by Ramos and Ferreira (2009). C and CB are the main competitors of the tests presented in this work because they control Type I error in most cases and they present high power. We show the results obtained by Silva et al. (1999) for the Scott-Knott test because it is a commonly used test, it shows unambiguous results and it has good qualities relative to Type I error and power.

The Caliński and Corsten test based on the F distribution showed no tendency to be liberal or conservative and the same occurred with its bootstrap version and with all tests compared. Thus, all tests here presented may be considered exact under the null H_0 and normal distribution.

Table 2 shows the experimentwise error rates as a function of r and k of the CF and CFB tests for null H_0 to the non-normal distributions used. The results for normal model were rewritten to be compared.

The CFB test was exact under the null H_0 and non-normal distributions and it is a robust test. The CF test was conservative for the lognormal model and $k = 10$ and $r = 20$. Ramos and Ferreira (2009) also used lognormal and exponential models with the same parameters of this work to evaluate the robustness of C and CB tests. According to their findings to $\alpha = 0.05$, for the exponential model, the C test was conservative for $k = 5$ and $r = 4$ and liberal for $k = 80$ and $r = 4$. To the lognormal distribution, the C test was liberal for $k = 80$ and $r = 20$. Therefore, CF is more robust than the C test.

3.2 Type I error under partial H_0

Type I error rates as a function of r, k and δ values equal to 1, 2 and 4 simulated under partial H_0 are shown in Table 3. For a small difference between

Table 2 - Experimentwise Type I error rates of tests CF and CFB as a function of number of replications r and number of treatments k, for $\alpha = 0.05$, under null H_0 for different probability models

		Normal		Lognormal		Exponential	
r	k	CF	CFB	CF	CFB	CF	CFB
4	5	0.050	0.050	0.037	0.051	0.043	0.047
4	10	0.056	0.055	0.035	0.038	0.039	0.043
4	20	0.059	0.060	0.054	0.059	0.052	0.050
4	80	0.058	0.057	0.043	0.041	0.061	0.059
10	5	0.059	0.061	0.040	0.046	0.049	0.064
10	10	0.048	0.045	0.042	0.052	0.045	0.047
10	20	0.053	0.051	0.049	0.052	0.034	0.038
10	80	0.051	0.048	0.042	0.043	0.056	0.054
20	5	0.058	0.056	0.042	0.052	0.053	0.057
20	10	0.045	0.044	0.031^{++}	0.039	0.042	0.047
20	20	0.064	0.062	0.046	0.049	0.060	0.062
20	80	0.041	0.041	0.050	0.050	0.052	0.052

$^{++}$ Significantly different (valor-$p < 0.01$) and smaller than the nominal significance level $\alpha = 0.05$.
** Significantly different (valor-$p < 0.01$) and larger than the nominal significance level $\alpha = 0.05$.

groups ($\delta = 1$), the CF test controlled Type I error for small values of k ($k = 5$) while the CFB was liberal in these cases. Ramos and Ferreira (2009) observed that the C and CB tests were liberal for all k and r values in the case of $\delta = 1$.

For $\delta = 2$ and $\delta = 4$ the CF and CFB tests showed similar performances with regard to the Type I error rates and similar to the values of C and CB. In general, with the value of r fixed, the Type I error rates rise with the increase of k, the tests become liberal for these differences between groups and the CFB test shows higher values of Type I error rates.

Borges and Ferreira (2003) evaluated the performance of the Scott-Knott test (SK) under partial H_0 in contexts similar to the ones simulated here. For $\delta = 4$, $r = 10$ and $\alpha = 0.05$, these authors obtained type I error rates equal to 0.10; 0.42; 0.61, 0.98 and 1.00 considering $k = 5, 10, 20, 40$ and 96 respectively. The CF test presents better performance than SK's because it has lower rates than SK's under similar conditions, except for $k = 5$. The same happened with C and CB.

Type I error rates as a function of r, k and δ values equal to 8 and 16 simulated under partial H_0 are shown in Table 4. The CF test shows behavior similar to C and CB for a greater difference between groups of means ($\delta = 8$ and $\delta = 16$). It is conservative for smaller values of k while the CFB maintains its liberal characteristic, with Type I error rates much higher with the increase of k, reaching values equal to 1 when $k = 80$. When we compare the CF and C tests, the former controls error rates for larger values of k more often than the latter.

Although the CFB test has presented similar behavior to the CF under null H_0 in relation to type I error rates, this similarity does not remain under partial

Table 3 - Experimentwise Type I error rates of tests CF, CFB, C and CB as a function of differences between means δ (1, 2 and 4), number of replications r and number of treatments k, for $\alpha = 0.05$, under partial H_0

δ	r	k	CF	CFB	C	CB
		5	0.058	0.077**	0.075**	0.074**
	4	10	0.078**	0.092**	0.134**	0.138**
		20	0.150**	0.165**	0.132**	0.138**
		80	0.338**	0.354**	0.135**	0.146**
		5	0.082	0.102**	0.077**	0.083**
1	10	10	0.107**	0.112**	0.118**	0.122**
		20	0.160**	0.177**	0.151**	0.156**
		80	0.361**	0.375**	0.177**	0.178**
		5	0.065	0.085**	0.091**	0.088**
	20	10	0.124**	0.130**	0.103**	0.104**
		20	0.180**	0.192**	0.137**	0.133**
		80	0.384**	0.403**	-	-
		5	0.139**	0.216**	0.164**	0.169**
	4	10	0.277**	0.332**	0.307**	0.307**
		20	0.553**	0.629**	0.407**	0.419**
		80	0.912**	0.986**	0.634**	0.666**
		5	0.188**	0.279**	0.204**	0.199**
2	10	10	0.365**	0.438**	0.358**	0.364**
		20	0.591**	0.678**	0.474**	0.475**
		80	0.921**	0.984**	0.683**	0.685**
		5	0.201**	0.270**	0.175**	0.178**
	20	10	0.392**	0.474**	0.358**	0.366**
		20	0.635**	0.707**	0.478**	0.478**
		80	0.920**	0.995**	-	-
		5	0.126**	0.469**	0.156**	0.157**
	4	10	0.211**	0.677**	0.302**	0.308**
		20	0.405**	0.894**	0.547**	0.549**
		80	0.736**	1.000**	0.914**	0.914**
		5	0.147**	0.515**	0.158**	0.160**
4	10	10	0.239**	0.721**	0.311**	0.311**
		20	0.368 **	0.912**	0.521**	0.520**
		80	0.708**	1.000**	0.914**	0.914**
		5	0.147**	0.552**	0.168**	0.168**
	20	10	0.257**	0.735**	0.309**	0.310**
		20	0.377**	0.914**	0.500**	0.502**
		80	0.744**	1.000**	-	-

$^{++}$ Significantly different (valor-$p < 0.01$) and smaller than the nominal significance level $\alpha = 0.05$.
** Significantly different (valor-$p < 0.01$) and larger than the nominal significance level $\alpha = 0.05$.

Table 4 - Experimentwise Type I error rates of tests CF, CFB, C and CB as a function of differences between means δ (8 and 16), number of replications r and number of treatments k, for $\alpha = 0.05$, under partial H_0

δ	r	k	CF	CFB	C	CB
8	4	5	0.026^{++}	0.504^{**}	0.016^{++}	0.016^{++}
		10	0.036	0.650^{**}	0.028^{++}	0.032^{++}
		20	0.049	0.898^{**}	0.023^{++}	0.057
		80	0.05	1.000^{**}	0.037	0.156^{**}
	10	5	0.027^{++}	0.543^{**}	0.027^{++}	0.026^{++}
		10	0.038	0.734^{**}	0.022^{++}	0.025^{++}
		20	0.037	0.909^{**}	0.037	0.040
		80	0.053	1.000^{**}	0.040	0.048
	20	5	0.019^{++}	0.547^{**}	0.022^{++}	0.021^{++}
		10	0.027^{++}	0.751^{**}	0.030^{++}	0.032^{++}
		20	0.024^{++}	0.924^{**}	0.038	0.036
		80	0.048	1.000^{**}	-	-
16	4	5	0.021^{++}	0.492^{**}	0.030^{++}	0.031^{++}
		10	0.040	0.707^{**}	0.031^{++}	0.040
		20	0.030^{++}	0.912^{**}	0.032^{++}	0.070
		80	0.044	1.000^{**}	0.029^{++}	0.335^{**}
	10	5	0.026^{++}	0.568^{**}	0.025^{++}	0.024^{++}
		10	0.030^{++}	0.736^{**}	0.029^{++}	0.031^{++}
		20	0.041	0.931^{**}	0.034	0.045
		80	0.055	1.000^{**}	0.028^{++}	0.057
	20	5	0.023^{++}	0.556^{**}	0.022^{++}	0.023^{++}
		10	0.034	0.769^{**}	0.024^{++}	0.026^{++}
		20	0.038	0.914^{**}	0.018^{++}	0.021^{++}
		80	0.044	1.000^{**}	-	-

$^{++}$ Significantly different (valor-$p < 0.01$) and smaller than the nominal significance level $\alpha = 0.05$.
** Significantly different (valor-$p < 0.01$) and larger than the nominal significance level $\alpha = 0.05$.

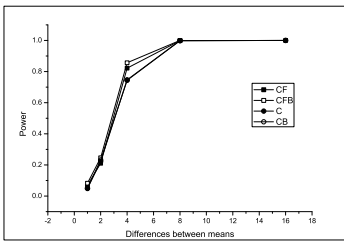
H_0. The CF test shows a liberal behavior with increasing Type I error rates until $\delta = 4$. For differences between groups $\delta \geq 8$ its behavior varies between being conservative to exact and its behavior is more conservative for $k = 5$, that is, for fewer treatments. The CFB test is even more liberal for $\delta \geq 8$ and the Type I error rates increase with larger values of k. The performance of the C test is more conservative than the CF's while the CB test presents more instability in the error rates ranging from conservative to exact or liberal.

Under partial H_0, the data are simulated in a way that there are two groups with differences of δ standard error of means between them. When these differences are large ($\delta \geq 8$), half of the treatments are simulated from a normal distribution with one value of mean and the half are simulated with a larger mean value.

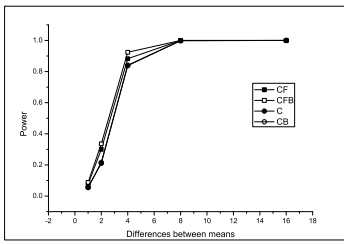
To obtain the bootstrap distribution of the CFB method, the simulated data are resampled B times, generating treatment groups with mean values very different from those simulated initially. This will produce very different sums of squares matrices, often, with lower values than the matrix obtained from the original data. The resulting p-values will be lower than those obtained by the original CF, and consequently the chosen partition (where p-value $> \alpha$) will not be the one which divides the treatments into two groups but rather more than two groups. This effect grows with increasing number of k treatments, which can be seen in Table 4. Thus, under partial H_0, when there are groups of treatments with different values, the bootstrap test tends to inflate the Type I error rates.

3.3 Power under partial H_0

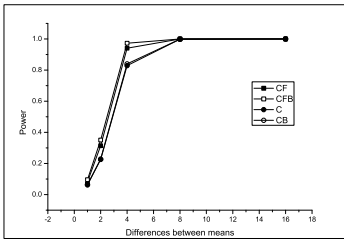
The power values of the CF, CFB, CB and C tests are shown in Figures 1 and 2, as a function of the δ standard error of mean, number of replications r and k treatments under partial H_0 and $\alpha = 0.05$.



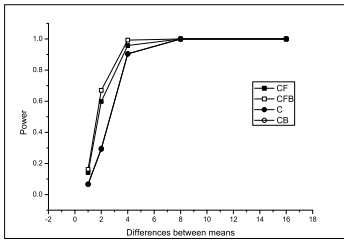
(a) $r = 4, k = 5$



(b) $r = 20, k = 5$



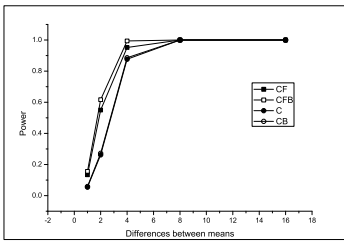
(c) $r = 4, k = 10$



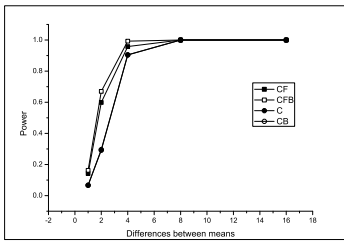
(d) $r = 20, k = 10$

Figure 1 - Power values of the CF, CFB, C and CB tests, under partial H_0 for $\alpha = 0.05$, $k = 5$ and $k = 10$.

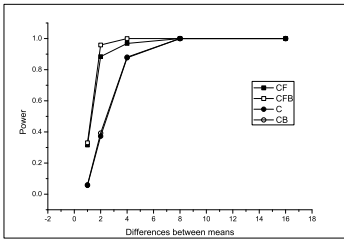
As expected, power values grew with increasing δ. The greater the difference ($\delta \geq 4$) the higher the power (greater than 0.80) and this is more evident with larger values of k and r. There are few differences between tests C and CF and their corresponding bootstrap version. The CF test is more powerful than the C test, despite similar Type I error rates, which is an advantage of the CF test. The



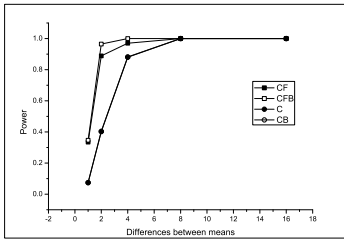
(a) $r = 4, k = 20$



(b) $r = 10, k = 20$



(c) $r = 4, k = 80$



(d) $r = 10, k = 80$

Figure 2 - Power values of the CF, CFB, C and CB tests, under partial H_0 for $\alpha = 0.05$, $k = 20$ and $k = 80$.

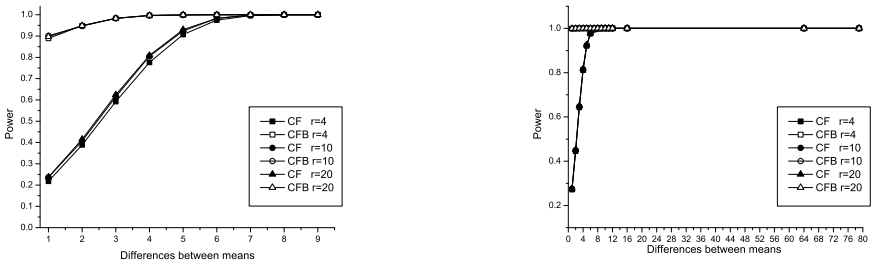
power values of the CFB test are the highest, which was expected, due to its liberal characteristics in relation to Type I error rates.

Tasaki et al. (1987) evaluated six multiple comparison procedures based on cluster analysis, including both methods of Caliński and Corsten and the Scott-Knott test. The authors performed 100 simulations with $\alpha = 0.05$, $k = 6$, $r = 5$ and different normal distribution mean values: $\mu = 1, \ldots, 6$ and variance 1. Under partial H_0 they considered four sets, where there were two groups in the two sets and three groups in the last two sets. With two groups, the CF test performed similar to or better than the C and SK tests, detecting correct partitions more often. With three groups, the CF, SK and C tests did not show good results. It is important to be cautious in interpreting these results because the number of simulations was small.

3.4 Power under H_1

In Figure 3 the power values of the CFB and CF tests are shown as a function of δ, for $k = 10$ and $k = 80$ and different values of r under H_1, for $\alpha = 0.05$ and normal distribution. In the CF test, for all values of k, power values around 0.80 were achieved for $\delta = 4$. For smaller values of δ, there were differences in the power values, with an increasing trend in such values with the rise of k. The CFB test shows very high power values, around 0.90, even with small differences between

the means, and this fact is more evident with $k = 80$. Thus, the bootstrap test showed the same trend of partial H_0, showing higher Type I error rates, especially for larger values of k. In general, the increase of δ causes an increase in power as already mentioned in the case of partial H_0. The differences between treatment means were fixed at a constant value in standard errors (δ), therefore, power is not affected by the number of replications r. This was as expected.



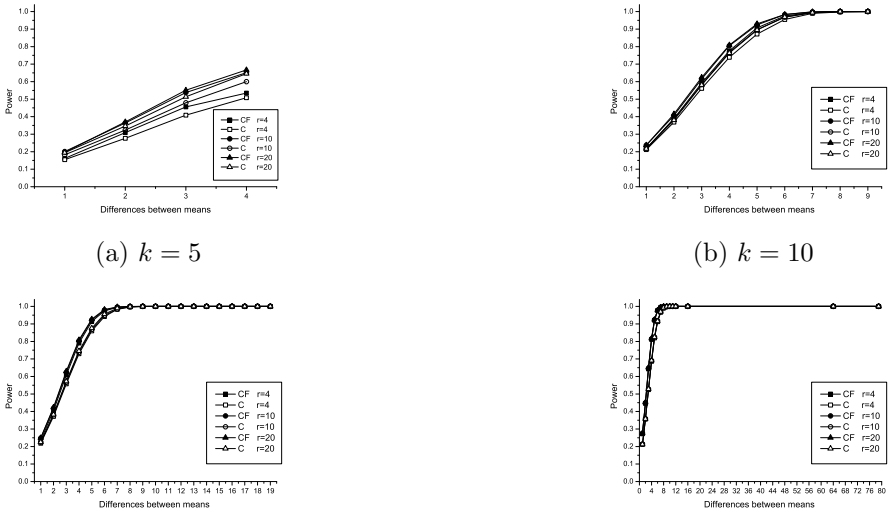
(a) $k = 10$ (b) $k = 80$

Figure 3 - Power values of the CF and CFB tests under H_1, for $\alpha = 0.05$, $k = 10$, $k = 80$ and different values of r considering the normal distribution.

In Figure 4 we show the power values for the comparison of the tests in their original formulations (CF and C). These tests show very similar behavior, where the increasing of the number of treatments k is the most influential factor to increase the power. The values are more variant at $k = 5$ (Figure 4a), in which case there is a slight effect of the number of replications r. Nonetheless, this stabilizes when k increases.

We show the power values to compare the bootstrap tests (CFB and CB) in Figure 5. The CFB test showed values of power higher than CB's in all situations for $\delta \leq 7$, and as the value of δ decreased, this difference was greater. For $\delta > 7$ both tests showed values of power equal to 1. Increasing k also caused an increase in the power values of the two tests.

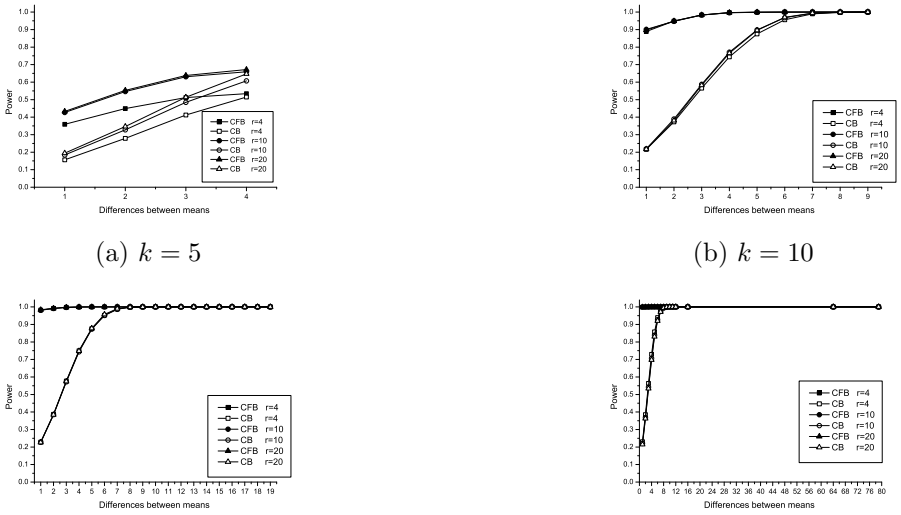
In Figures 6 and 7 we show the power values of the CF and CFB tests as a function of δ, number of treatments k and replications r under H_1, for $\alpha = 0.05$ and non-normal distributions. For both distributions, lognormal and exponential, the power values increase as δ, r and k raise. This pattern is similar to the normal distribution (Figure 3), except by the fact that the number of replications affects the power for the non-normal distributions. Between the non-normal distributions there are not expressive differences but the power values under exponential distribution are a little higher than under lognormal distribution.



(a) $k = 5$ (b) $k = 10$

(c) $k = 20$ (d) $k = 80$

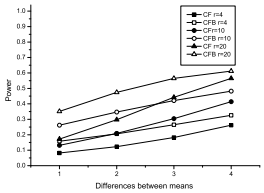
Figure 4 - Power values of CF and C tests, under H_1, for $\alpha = 0.05$, for different values of k and r considering the normal distribution.



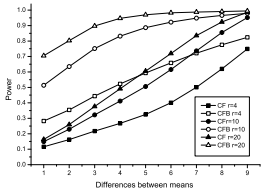
(a) $k = 5$ (b) $k = 10$

(c) $k = 20$ (d) $k = 80$

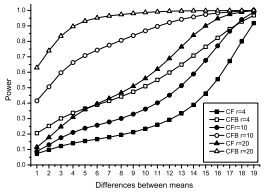
Figure 5 - Power values of CFB and CB tests, under H_1, for $\alpha = 0.05$, for different values of k and r considering the normal distribution.



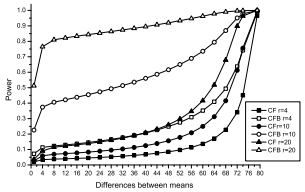
(a) $k = 5$



(b) $k = 10$

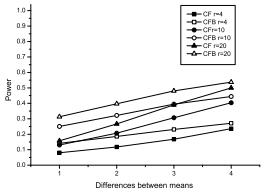


(c) $k = 20$

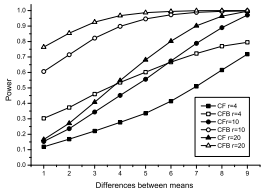


(d) $k = 80$

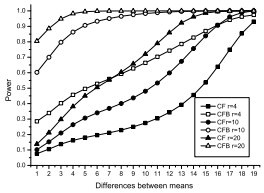
Figure 6 - Power values of CF and CFB tests, under H_1, for $\alpha = 0.05$, for different values of k and r considering the lognormal distribution.



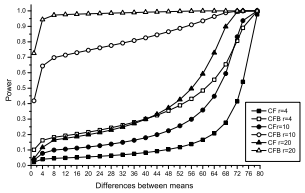
(a) $k = 5$



(b) $k = 10$



(c) $k = 20$



(d) $k = 80$

Figure 7 - Power values of CF and CFB tests, under H_1, for $\alpha = 0.05$, for different values of k and r considering the exponential distribution.

Conclusions

Under null H_0, the CFB and CF tests control the Type I error rates under normality. Under H_0 and non-normality, the CFB test controls the Type I error rates and therefore is robust while the CF test is conservative under lognormal distribution for $k = 10$ and $r = 10$. Under partial H_0, the original CF test is liberal for small differences between the means and conservative for larger differences, while the bootstrap test CFB is always liberal and Type I error rates raise with the increase of δ. The power is always higher in case of the CFB test than in the case of the CF test and the two tests reach power values equal to 1 at $\delta \geq 8$. Under H_1, the two tests show high values of power and the CFB test is always more powerful due to its liberal trait for normal and non-normal distributions. The CF test outperforms the CFB and it is recommended in practical situations since it controls Type I error rates in most situations and shows higher power values.

Acknowledgments

The authors thank Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for the financial support.

RAMOS, P. S.; VIEIRA, M. T. Procedimento de comparações múltiplas *bootstrap* baseado na distribuição *F*. *Rev. Bras. Biom.,* São Paulo, v.31, n.4, p.529-546, 2013.

- RESUMO: O objetivo do presente trabalho é propor um teste *bootstrap* baseado na distribuição F proposto por Caliński and Corsten em 1985 e compará-lo com sua versão original por meio de simulação Monte Carlo. Os procedimentos testam a homogeneidade das médias de tratamentos dentro de cada dois ou mais subgrupos utilizando a minimização da soma de quadrados da partição como critério. O desempenho é mensurado em relação às taxas de erro tipo I por experimento e ao poder e os programas para simulação foram implementados em `R`. Sob H_0 e normalidade, os dois testes controlam as taxas de erro tipo I. Sob H_0 e não-normalidade, o teste *bootstrap* CFB controla as taxas de erro tipo I e é robusto, enquanto o teste original CF é conservador sob a distribuição lognormal para $k = 10$ e $r = 10$. Sob H_0 parcial, o teste CF é liberal para pequenas diferenças δ entre as médias e conservador para maiores diferenças, enquanto o CFB se mostra liberal sempre e com taxas de erro maiores com o aumento de δ. Sob H_1, o poder do CF é maior devido a sua característica liberal. O teste original apresenta melhor desempenho e é mais recomendado em situações práticas porque controla as taxas de erro tipo I na maior parte dos casos e apresenta altos valores de poder.

- PALAVRAS-CHAVE: Comparação de médias; reamostragem; análise de agrupamento; soma de quadrados; simulação Monte Carlo; taxas de erro tipo I; poder.

References

BORGES, L. C.; FERREIRA, D. F. Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normais e não normais dos resíduos. *Revista de Matemática e Estatística*, v.21, p.67-83, 2003.

CALIŃSKI, T.; CORSTEN, L. C. A. Clustering means in ANOVA by Simultaneous Testing. *Biometrics*, v.41, p.39-48, 1985.

CARMER, S. G. ; SWANSON, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal American Statistical Association*, v.68, p.66-74, 1973.

EFRON, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.*, v.7, p.1-26, 1979.

HARTIGAN,J. A.; WONG M. A. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v.28, p.100-108, 1979.

HINKELMANN, K.; KEMPTHORNE, O. *Design and analysis of experiments*. v.1. New York: J. Wiley & Sons, 1987, 495p.

KESELMAN, H. J. ; CRIBBIE, R. A.; WILCOX, R. R. Pairwise Multiple Comparison Tests when Data are Non-normal. *Educational and Psychological Measurement*, v.62, p.420-434, 2002.

KRISHNAMOORTHY, K. F. L. ; MATHEW T. A Parametric Bootstrap Approach for ANOVA with Unequal Variances: Fixed and Random Models. *Computational Statistics & Data Analysis*, v.51, p.5731-5742, 2007.

MACHADO, A. A.; DEMÉTRIO, C. G. B.; FERREIRA D. F.; SILVA J. G. C. Estatística experimental: uma abordagem fundamentada no planejamento e no uso de recursos computacionais. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA SA SOCIEDADE INTERNACIONAL de BIOMETRIA, 50.,2005, Londrina. *Anais...* Londrina: Editora ISBN, 2005. 290p.

MANLY, B. F. J. *Randomization, bootstrap and Monte Carlo methods in biology*. 2.ed. London: Chapman-Hall, 1998. 399p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction to the theory of statistics*. 3.ed. New York: J. Wiley & Sons, 1974. 564p.

PERECIN, D. ; BARBOSA, J.C. Uma avaliação de seis procedimentos para comparações múltiplas. *Revista de Matemática e Estatística*, v.6, p.95-103, 1988.

R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing*. Disponível em: http://www.r-project.org. 2011, acesso em: 20 jan. 2011.

RAMOS, P. S.; FERREIRA, D. F. Agrupamento de médias via bootstrap de populações normais e não-normais. *Revista Ceres*, v.56, p.140-149, 2009.

SCOTT, A. J.; KNOTT, M. A Cluster analysis method for grouping means in the analysis of variance. *Biometrics*, v.30, p.507-512, 1974.

SILVA, E. C.; 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 Agrotécnica*, v.23, p.687-696, 1999.

STEEL, R. G. D.; TORRIE, J. H. *Principles and procedures of statistics*. 2.ed. New York: McGraw-Hill Book., 1980. 633 p.

TASAKI, A. Cluster analysis method for grouping means in the analysis of variance. *Biometrics*, v.30, p.507-512, 1974.

Received in 12.11.2013.

Approved after revised in 10.03.2014.