

Agrupamento de médias via *bootstrap* de populações normais e não-normais¹

Patrícia de Siqueira Ramos²
Daniel Furtado Ferreira³

RESUMO

Este trabalho teve por objetivos propor uma alternativa *bootstrap* para um dos procedimentos de comparações múltiplas de Caliński e Corsten (1985) baseado em análise de agrupamento e avaliar as versões original e *bootstrap* desse procedimento por simulação Monte Carlo, considerando modelos probabilísticos normais e não-normais. Foi avaliado o desempenho em relação às taxas de erro tipo I por experimento e poder. Os dois testes são exatos sob H_0 completa e normalidade. A versão *bootstrap* em não-normalidade e H_0 completa controla o erro tipo I por experimento, enquanto o teste original é conservativo para valores pequenos de k e liberal para grandes valores de k . Em relação a H_0 parcial, os dois testes são liberais para menores diferenças entre os grupos e conservativos para maiores diferenças. O poder do teste *bootstrap* é maior do que o do teste original em H_0 parcial ou H_1 , seu desempenho é considerado superior e é robusto em relação à ausência de normalidade, sendo recomendada a sua utilização.

Palavras-chave: Comparações múltiplas, análise de agrupamento, erro tipo I por experimento.

ABSTRACT

Grouping means via bootstrap for normal and non-normal populations

The present work aimed to propose a bootstrap alternative to Caliński & Corsten (1985) multiple comparison procedure based on cluster analysis and evaluate the original and bootstrap tests by Monte Carlo simulation considering normal and non-normal probabilistic models. The performance was evaluated in relation to the type I error rates per experiment and in relation to the power. The two tests are exact under complete H_0 and normality. Under non-normality and complete the bootstrap test controlled the experiment wise error while the original test is conservative for small values and is liberal for large values. Under partial situation, the two tests are liberal for smaller differences between groups and are conservative for larger differences. The power of bootstrap test is higher than the original test under partial and, its performance is considered superior to the original test and is robust with respect to normality absence, being therefore recommended.

Key words: multiple comparison procedure, cluster analysis, type I experimental error.

Recebido para publicação em abril de 2007 e aprovado em janeiro de 2009

¹ Trabalho extraído da dissertação: Ramos PS (2007). Agrupamento de médias via *bootstrap* para populações normais e não-normais. Dissertação de Mestrado. Universidade Federal de Lavras, Lavras, 76p.

² Departamento de Ciências Exatas, Universidade Federal de Lavras. Lavras, MG. siqueirapaty@gmail.com

³ Departamento de Ciências Exatas, Universidade Federal de Lavras. Lavras, MG. danielff@ufla.br

INTRODUÇÃO

Na experimentação, o pesquisador se depara com o problema de comparar médias de níveis do fator em estudo. Admite-se a hipótese de nulidade $H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu_k$ como verdadeira de que não há diferenças entre os níveis de um fator qualquer, de efeitos fixos, em que k é o número de níveis do fator (Machado *et al.*, 2005). O teste F é utilizado para testar se existem diferenças reais entre níveis do fator, mas não indica quais as médias se diferem. Quando há rejeição da hipótese H_0 , a hipótese alternativa (H_1 ou H_A) de que há pelo menos uma diferença é considerada verdadeira. Se o fator for qualitativo e não-estruturado, procedimentos de comparações múltiplas (PCM) devem ser aplicados (Banzatto & Kronka, 1989).

Os PCM são procedimentos estatísticos que realizam todas as comparações par a par entre as médias e possuem vasta bibliografia a seu respeito (Hochberg & Tamhane, 1987; Machado *et al.*, 2005; Ramalho *et al.*, 2000; entre outros). Todos os PCM, permitem analisar diferenças entre médias após a conclusão do experimento para detectar possíveis grupos em um conjunto de níveis do fatores não-estruturados. Um problema é a ambiguidade que dificulta a interpretação dos resultados, ocorrendo quando dois níveis de um fator considerados como diferentes entre si não diferem de um terceiro (Machado *et al.*, 2005). Outra questão é a do controle do erro tipo I. Alguns testes, como t de Student & Duncan, possuem elevadas taxas de erro tipo I por experimento, sendo liberais, enquanto outros, como Scheffé & Tukey, possuem taxas de erro tipo I por experimento inferiores ao nível nominal de significância e são considerados conservativos. Em relação ao teste SNK, apesar de controlar o erro tipo I por experimento em H_0 completa, é liberal em H_0 parcial (Carmer & Swanson, 1973).

Ao testar H_0 , o pesquisador corre o risco de cometer erros de decisão, cujos principais são dos erros tipos I e II (Mood *et al.*, 1974). O tipo I é cometido ao se rejeitar a hipótese nula verdadeira, quando deveria ser aceita, e sua probabilidade é α . O erro tipo II é cometido ao não se rejeitar a hipótese nula falsa, e sua probabilidade é β , e o poder do teste é a probabilidade de se rejeitar H_0 , sendo H_0 falsa e igual a $1 - \beta$. Existem duas formas muito utilizadas para se medir as taxas de erro tipo I nos PCM. A primeira mede essa taxa por comparação (TPC), que é a probabilidade de se rejeitar uma hipótese verdadeira em todas as possíveis combinações de médias de níveis do fator, tomadas duas a duas. Na segunda, calcula-se a taxa de erro tipo I por experimento (TPE), que é a probabilidade de se realizar pelo menos uma inferência errada por experimento no sentido de rejeitar a hipótese nula verdadeira de igualdade entre médias de dois níveis quaisquer (Steel & Torrie, 1980).

Uma qualidade desejada para um procedimento é a robustez, e esse é considerado robusto se, ao violar uma das pressuposições básicas da análise de variância, ele mantiver, aproximadamente, o desempenho delineado originalmente na elaboração de sua teoria (Borges & Ferreira, 2003).

Alguns dos PCMs resultam em grupos sobrepostos; para contornar esse problema utilizam-se ideias de análise de agrupamento, combinadas com os testes de hipóteses (Hinkelmann & Kempthorne, 1987). Esses procedimentos separam as médias dos níveis do fator em grupos homogêneos, pela minimização da variação dentro e maximização da variação entre grupos. Um exemplo é o teste de Scott-Knott (1974), e dois outros PCMs, também baseados em técnicas de agrupamento, que foram apresentados por Caliński & Corsten (1985). O primeiro é fundamentado na distribuição da amplitude estudentizada (Pachares, 1959) e o segundo, na distribuição F . O problema referente à utilização desses testes é que apenas são válidos sob normalidade, porém nenhuma avaliação teórica foi realizada no trabalho original.

A aplicação dos PCMs depende da pressuposição de normalidade, que nem sempre é verificada, havendo assim fortes indicativos de que a maioria desses procedimentos não é robusta em relação à violação de normalidade. Em estudos de desempenho de testes estatísticos, devido à dificuldade de se obter analiticamente informações sobre as taxas de erro tipo I e poder, a simulação Monte Carlo é uma alternativa viável e que foi utilizada em vários estudos de avaliação de PCM, como o de Carmer & Swanson (1973), Perecin & Barbosa (1988), Silva *et al.* (1999), Borges & Ferreira (2003), entre outros. O método de Monte Carlo é uma técnica de simulação de dados através da geração de números pseudoaleatórios, de acordo com determinada distribuição de probabilidade. Esse método objetiva avaliar o comportamento de diferentes técnicas estatísticas que poderiam ser empregadas num problema específico (Dachs, 1988).

Métodos de reamostragem *bootstrap* podem ser utilizados em estudos sobre comparações múltiplas das médias dos níveis de um fator em situações de heterogeneidade de variâncias de modelos probabilísticos normais ou não-normais (Keselman *et al.*, 2002). Na ausência de qualquer conhecimento sobre a população, é realizada reamostragem com reposição de tamanho n da amostra original. Em cada reamostragem é obtida uma estimativa do parâmetro de interesse. Esse procedimento é executado repetidas vezes, obtendo-se assim inúmeras estimativas dos parâmetros que serão usadas para gerar a distribuição denominada distribuição *bootstrap* (Manly, 1998). A distribuição *bootstrap* de algum estimador de interesse é utilizada no lugar da “distribuição teórica” desse mesmo estimador para se realizar a inferência.

Assim, este trabalho tem por objetivos propor uma alternativa *bootstrap* para um dos procedimentos de comparações múltiplas de Caliński & Corsten (1985) e avaliar as versões original e *bootstrap* desse procedimento por meio de simulação Monte Carlo, considerando modelos probabilísticos normais e não-normais.

MATERIAL E MÉTODOS

Foram comparados por simulação Monte Carlo os desempenhos de dois métodos de agrupamento de médias na análise de variância. Para isso, foram realizadas simulações nas hipóteses nula (H_0) completa e parcial e alternativa (H_1), em que foram mensurados os erros tipo I e poder de acordo com a situação, permitindo a avaliação do desempenho dos dois testes. Os programas de simulação foram desenvolvidos no programa R (R Development Core Team, 2005).

O primeiro método, de Caliński & Corsten (1985), utiliza um teste simultâneo baseado em uma extensão da amplitude estudentizada. Sejam as amostras aleatórias $X_{11}, X_{12}, \dots, X_{1r}, X_{21}, \dots, X_{2r}, \dots, X_{ij}$, em que X_{ij} representa a observação do i ésimio nível do fator na j -ésima unidade amostral ou experimental ($i = 1, 2, \dots, k$), e $j = 1, 2, \dots, r$. O estimador da média de cada i -ésima população é dado por:

$$\bar{X}_i = \frac{\sum_{j=1}^r X_{ij}}{r} \quad (1)$$

em que $\bar{X}_{(1)}, \bar{X}_{(2)}, \dots, \bar{X}_{(k)}$ são as médias ordenadas dos níveis do fator. Foram calculadas as $m = k(k-1)/2$ distâncias entre médias definidas por: $d_{i' i} = \bar{X}_{(i')} - \bar{X}_{(i)}$, para $i' > i = 1, 2, \dots, k$. Assim, a matriz de distâncias $D = \{d_{ij}\}$ com dimensão $(k \times k)$ foi criada, sendo $d_{ij} = d_{ji}$. Na sequência, o método de agrupamento do vizinho mais distante foi aplicado (Johnson e Wichern, 1998). Em cada passo de fusão dos grupos a distância R_ℓ ($\ell = 1, 2, \dots, k-1$) entre eles é registrada. O próximo passo foi verificar a significância para o teste de hipótese da igualdade dos dois grupos que são candidatos a ser unidos. O valor- p é calculado por:

$$\text{valor-}p = \Pr \left(q_\alpha(k, \nu) > \frac{R_\ell}{\sqrt{\frac{QME}{r}}} \right) \quad \ell = 1, 2, \dots, k-1 \quad (2)$$

em que $q_\alpha(k, \nu)$ é o quantil superior da distribuição da amplitude estudentizada, k é o número de níveis do fator, ν são os graus de liberdade associados ao QME e QME é o quadrado médio do resíduo da análise de variância. O primeiro passo de fusão i para o qual o valor- p é menor do

que o nível de significância α foi considerado o ponto de corte do dendrograma, e os grupos de níveis do fator foram obtidos.

O segundo método é uma proposta do presente trabalho, na qual o procedimento de Caliński & Corsten (1985) é estendido para situações não-normais por meio de *bootstrap*. As rk observações amostrais foram combinadas em uma única amostra onde foram realizados r sorteios com reposição para compor a amostra de cada população. Esse procedimento foi repetido k vezes para formar um novo conjunto experimental *bootstrap*, composto de k níveis do fator com r observações. Novas médias foram computadas nessas amostras, e a amplitude estudentizada foi calculada por:

$$q_b = \frac{\bar{X}_{(k)}^b - \bar{X}_{(1)}^b}{\sqrt{\frac{QME_b}{r}}} \quad (3)$$

em que QME_b é o quadrado médio do resíduo da b -ésima amostra *bootstrap* e $\bar{X}_{(1)}^b$ e $\bar{X}_{(k)}^b$ são as médias amostrais ordenadas da primeira e k -ésima reamostragens de *bootstrap*, sendo o sobrescrito b utilizado para indicar a b -ésima amostra de *bootstrap*. Este processo foi repetido $B = 1000$ vezes, e o conjunto com todos os valores q_b obtidos ($b = 1, 2, \dots, B$) foi utilizado para computar o valor- p por:

$$\text{valor-}p = \frac{\sum_{b=1}^B I \left(\frac{R_\ell}{\sqrt{\frac{QME}{r}}} < q_b \right)}{B} \quad \ell = 1, 2, \dots, k-1 \quad (4)$$

em que $I(x)$ é a função indicadora de x representa a desigualdade na expressão (4). Quando essa desigualdade for satisfeita, a função indicadora retorna 1 e, caso contrário, retorna 0. O mesmo critério utilizado no teste de Caliński & Corsten (1985) foi aplicado para obtenção dos grupos de níveis homogêneos do fator.

Foram consideradas $N = 1000$ simulações de k populações não-estruturadas e qualitativas. Em cada simulação foram obtidas amostras de tamanho r de cada uma das k populações (níveis do fator). Consideraram-se os modelos probabilísticos normal, lognormal e exponencial para as populações, sendo as duas últimas distribuições comuns na experimentação e utilizadas para avaliar a robustez dos testes. A densidade da variável aleatória X_{ij} é representada por $f_{xi}(x_{ij})$.

Foram consideradas diversas configurações envolvendo o número k de níveis do fator, número de repetições r e número (δ) de erros-padrões de diferença entre médias consecutivas. Assim, foram utilizados valores de k iguais a 5, 10, 20 e 80 e de r iguais a 4, 10 e 20. Esses

valores foram combinados fatorialmente, totalizando 12 situações para cada distribuição tanto para o poder quanto para o erro tipo I. Em relação a H_1 , diferentes valores de δ foram considerados. Em todos os casos os testes foram aplicados considerando níveis nominais de significância α iguais a 0,01 e 0,05.

Em relação a H_0 completa, os parâmetros dos modelos $f_{xi}(x_{ij})$ $i = 1, 2, \dots, k$ foram considerados iguais para todos os valores de i . Assim, os níveis do fator apresentavam médias e variâncias comuns. Com relação a H_0 parcial foram considerados dois grupos, dentro dos quais o procedimento utilizado para H_0 completa é aplicado. As densidades entre os grupos tiveram valores diferentes para os parâmetros. Em H_1 , as densidades foram consideradas todas diferentes. No caso normal, elas se diferenciaram apenas pela média (μ), mantendo constante a variância σ^2 .

Os dois testes foram aplicados em todas as configurações simuladas, e nos N experimentos oriundos de cada uma delas foram estimados as taxas de erro tipo I por experimento e o poder. Foram avaliadas somente as taxas de erro tipo I por experimento, uma vez que todos os testes controlam as taxas de erro por comparação. As taxas de erro tipo I por experimento foram avaliadas em relação a H_0 completa e parcial, e o poder foi avaliado em H_0 parcial e H_1 . A situação de distribuição normal foi utilizada como ambiente favorável, pois o teste de Caliński & Corsten (1985) foi idealizado na pressuposição de normalidade.

A fim de se aplicar os testes, foi utilizado um exemplo real. O objetivo do estudo original foi testar diferentes tipos de combinações de culturas de bactérias fixadoras de Nitrogênio em plantas de trevo-vermelho. Havia cinco culturas de *Rhizobium trifolii*, que é própria do trevo-vermelho, mais um composto com cinco culturas de *Rhizobium meliloti*, típica da alfafa. O experimento foi conduzido em casa de vegetação, utilizando-se delineamento inteiramente casualizado, com seis níveis do fator (combinações das culturas) e cinco repetições, sendo cada uma constituída de um vaso. Na Tabela 1 encontram-se os teores de Nitrogênio em miligramas obtidos no estudo.

RESULTADOS EM DISCUSSÃO

Na Tabela 2 são apresentadas as taxas de erro tipo I por experimento (TPE) em normalidade, em função do número r de repetições, k níveis do fator e nível α de significância dos testes de Caliński & Corsten (C) e sua versão *bootstrap* (CB) para H_0 completa. As taxas observadas foram confrontadas com os intervalos de 99% de confiança para proporções (Leemis & Trivedi, 1996), considerando os níveis nominais de 0,01 e 0,05. Em todas as situações, ambos os testes apresentaram taxas não significativamente diferentes das dos níveis nominais correspondentes, não havendo tendência de os dois testes serem liberais ou conservativos, sendo exatos nessa situação. Na Figura 1 são apresentadas as TPEs para a situação de H_0 completa, na distribuição lognormal (0, 1) para $r = 20$ em função de k . Para fins de comparação, foram adicionados resultados em normalidade dos testes C e de Scott-Knott (SK) (Scott e Knott, 1974) obtidos por Silva *et al.* (1999). O teste C foi liberal para maior número de níveis do fator. O teste SK apresentou controle do erro tipo I para $k \leq 20$, mas foi conservativo para $k = 80$. Assim, o teste CB mostrou melhor controle do erro tipo I por experimento em normalidade ou distribuição lognormal (0, 1). A situação lognormal (0, 5) foi simulada, sendo o teste C conservativo para pequenos valores de k associados a pequenos valores de r . O aumento do número de níveis k do fator acentuou o aspecto liberal do teste C, com taxas de erro superiores a 0,80 e 0,90 para os níveis nominais de 0,01 e 0,05, respectivamente. Assim, esse teste não pode ser considerado robusto à violação de normalidade quando se considera a distribuição lognormal. O fato de o teste ser liberal nas situações de maior número de níveis agrava ainda mais o problema, pois essas representam as situações ideais para a aplicação de testes de agrupamento de médias. O teste CB também apresentou taxas de erro inferiores a 0,05 e 0,01 para α , em geral para valores de k pequenos ($k \leq 20$), associados a valores menores de r ($r \leq 10$). A característica conservativa desse teste desapareceu à medida que o número de níveis k do fator aumentou de 20 para 80. O teste CB não se tornou liberal com $k = 80$, o que o classifica como robusto.

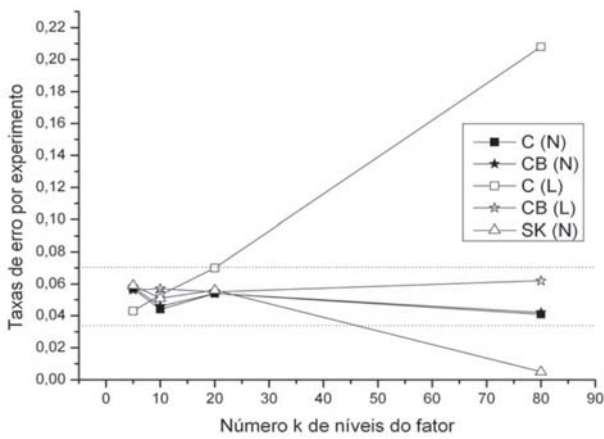
Tabela 1. Teor de Nitrogênio em mg de plantas de trevo vermelho inoculadas com combinações de culturas de *R. trifolii* e *R. meliloti*

	Repetições				
	1	2	3	4	5
3Dok1	19,4	32,6	27,0	32,1	33,0
3Dok5	17,7	24,8	27,9	25,2	24,3
3Dok4	17,0	19,4	9,1	11,9	15,8
3Dok7	20,7	21,0	20,5	18,8	18,6
3Dok13	14,3	14,4	11,8	11,6	14,2
Composto	17,3	19,4	19,1	16,9	20,8

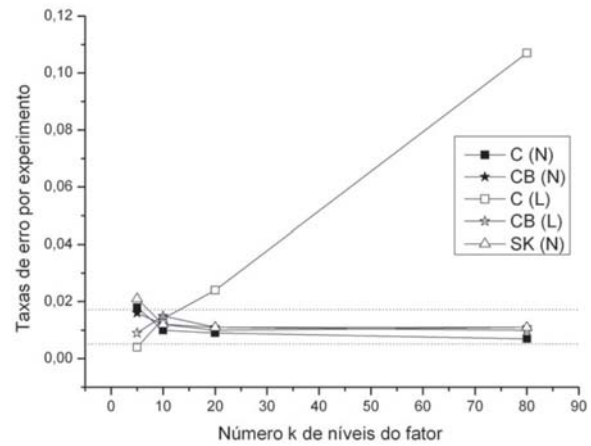
FONTE: Steel & Torrie, p. 141, 1980

Tabela 2. Taxas de erro por experimento (TPE) dos testes de Caliński & Corsten (C), sua versão *bootstrap* (CB) em função do número de repetições r , número de níveis k do fator e níveis nominais de significância $\alpha = 0,05$ e $\alpha = 0,01$, sob H_0 completa, considerando-se a distribuição normal (10, 1)

k	r	C		CB	
		0,05	0,01	0,05	0,01
5	4	0,049	0,009	0,053	0,010
5	10	0,047	0,009	0,046	0,012
5	20	0,057	0,018	0,058	0,016
10	4	0,055	0,009	0,057	0,011
10	10	0,042	0,007	0,042	0,007
10	20	0,044	0,010	0,046	0,012
20	4	0,047	0,006	0,050	0,008
20	10	0,048	0,010	0,054	0,013
20	20	0,054	0,009	0,054	0,010
80	4	0,044	0,010	0,051	0,015
80	10	0,048	0,008	0,050	0,009
80	20	0,041	0,007	0,042	0,011



(a) $\alpha = 0,05$



(b) $\alpha = 0,01$

Figura 1. Taxas de erro tipo I por experimento dos testes de Caliński & Corsten (C) e sua versão *bootstrap* (CB) e do teste de Scott-knott (SK), em função do número de níveis k do fator, considerando-se as distribuições normal (N) (10, 1) e lognormal (L) (0, 1) com H_0 completa, $r = 20$, para $\alpha = 0,05$ (a) e $\alpha = 0,01$ (b). As linhas pontilhadas representam os limites superior e inferior do IC exato para proporções com 99% de confiança.

Para a distribuição exponencial, com H_0 completa, cujos resultados são apresentados na Tabela 3, o teste C foi conservativo para $k = 5$ e $r = 4$ e liberal para $k = 80$, associado a um pequeno número de repetições ($r = 4$). O teste CB apresentou tamanho do teste sempre igual a α (0,01 ou 0,05).

Na Tabela 4 estão apresentadas as taxas de erro tipo I por experimento em função de k , r , α e δ , em H_0 parcial. Essas taxas foram mensuradas nas comparações entre níveis do fator de um mesmo grupo. Ambos os testes apresentaram o mesmo desempenho para $\alpha = 0,01$ e $\alpha = 0,05$. Pode-se observar, fixados α e r , que as TPEs aumentam com o incremento do número k de 5 para 80, sendo os testes liberais com δ pequenos ($\delta \leq 4$). Com $\delta \leq 8$, os testes tenderam a ser conservativos, exceto o teste CB, principalmente com maiores valores de k associados a menores

valores de r . Com H_0 parcial e normalidade, nenhum dos testes apresentou controle do erro tipo I por experimento, principalmente para $\delta \leq 4$. Nessa circunstância, os testes apresentaram TPE consideradas de pouco ($k = 5$) a muito liberais ($k = 80$). Para $\delta \geq 8$, os testes passaram para conservativos, exceto CB com $k = 80$ e $r \leq 10$.

Na Figura 2 estão apresentados os gráficos para poder dos testes C e CB em função de r , k e δ , considerando a distribuição normal de H_1 e $\alpha = 0,05$. Pode-se observar que há pequeno aumento no valor do poder à medida que o valor de k cresce, fixadas as outras variáveis do estudo. Como já é preconizado pela teoria, o poder cresce com o aumento de δ . Em quase todos os casos, praticamente não houve diferenças entre os valores de poder dos testes C e CB. O efeito de repetição só é evidente no caso de $k = 5$. Já era esperado que não houvesse efeito do número

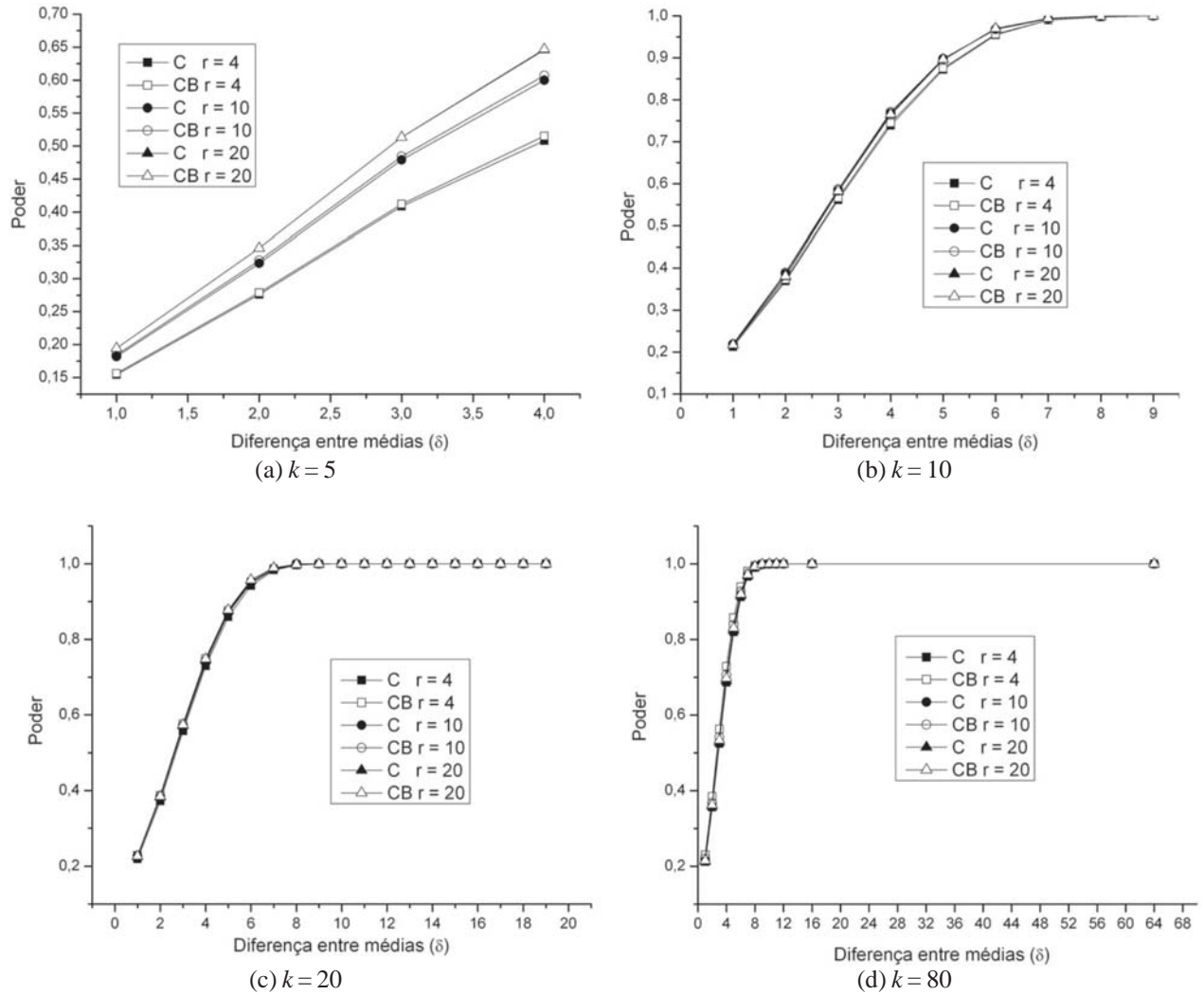


Figura 2. Poder dos testes de Caliński & Corsten (C) e sua versão *bootstrap* (CB) para diferentes números de repetições r , considerando-se a distribuição normal (10, 1) em função das diferenças entre médias δ para diferentes números de níveis k do fator, com H_1 e $\alpha = 0,05$.

Tabela 3. Taxas de erro por experimento (TPE) dos testes de Caliński e Corsten (C), sua versão *bootstrap* (CB) em função do número de repetições r , número de níveis k do fator e níveis nominais de significância $\alpha = 0,05$ e $\alpha = 0,01$, com H_0 completa, considerando-se a distribuição exponencial (0,1)

k	r	C		CB	
		0,05	0,01	0,05	0,01
5	4	0,033++	0,008	0,040	0,009
5	10	0,049	0,011	0,051	0,014
5	20	0,042	0,004	0,048	0,006
10	4	0,038	0,005	0,048	0,007
10	10	0,046	0,008	0,053	0,011
10	20	0,043	0,010	0,047	0,014
20	4	0,050	0,011	0,048	0,008
20	10	0,044	0,013	0,045	0,012
20	20	0,061	0,014	0,059	0,014
80	4	0,076**	0,029**	0,050	0,007
80	10	0,070	0,019	0,051	0,010
80	20	0,069	0,017	0,066	0,011

**Ultrapassou o LS do IC exato, com 99% de confiança para $\alpha = 0,01$ (0,021276) e $\alpha = 0,05$ (0,070504).

++Não atingiu o LI do IC exato, com 99% de confiança para $\alpha = 0,01$ (0,003727) e $\alpha = 0,05$ (0,033927).

Tabela 4. Taxas de erro tipo I dos testes Caliński & Corsten (C) e *bootstrap* (CB), em função do número de níveis k do fator, número de repetições r , diferenças d entre as médias para os níveis nominais de significância $\alpha = 0,05$ e $\alpha = 0,01$, com a distribuição normal e H_0 parcial

δ	r	k	0,05		0,01	
			C	CB	C	CB
2	4	5	0,164**	0,169**	0,061**	0,059**
		10	0,307**	0,307**	0,112**	0,123**
		20	0,407**	0,419**	0,170**	0,187**
		80	0,634**	0,666**	0,308**	0,352**
	10	5	0,204**	0,199**	0,075**	0,073**
		10	0,358**	0,364**	0,142**	0,153**
		20	0,474**	0,475**	0,302**	0,224**
		80	0,683**	0,685**	0,328**	0,341**
	20	5	0,175**	0,178**	0,077**	0,077**
		10	0,358**	0,366**	0,160**	0,166**
		20	0,478**	0,478**	0,234**	0,234**
4	4	5	0,156**	0,157**	0,095**	0,095**
		10	0,302**	0,308**	0,224**	0,233**
		20	0,547**	0,549**	0,468**	0,488**
		80	0,914**	0,914**	0,909**	0,912**
	10	5	0,158**	0,160**	0,117**	0,117**
		10	0,311**	0,311**	0,264**	0,267**
		20	0,521**	0,520**	0,493**	0,495**
		80	0,914**	0,914**	0,908**	0,910**
	20	5	0,168**	0,168**	0,129**	0,129**
		10	0,309**	0,310**	0,268**	0,268**
		20	0,500**	0,502**	0,476**	0,479**
8	4	5	0,016++	0,016++	0,003++	0,003++
		10	0,028++	0,032++	0,008++	0,011
		20	0,023++	0,057	0,006	0,016
		80	0,037	0,156**	0,017	0,067**
	10	5	0,027++	0,026++	0,004	0,005
		10	0,022++	0,025++	0,007	0,007
		20	0,037	0,040	0,007	0,011
		80	0,040	0,048	0,013	0,022**
	20	5	0,022++	0,021++	0,007	0,008
		10	0,030++	0,032++	0,011	0,011
		20	0,038	0,036++	0,009	0,011
32	4	5	0,032++	0,028++	0,002++	0,002++
		10	0,023++	0,033+	0,005	0,014
		20	0,028++	0,103**	0,009	0,059**
		80	0,024++	0,582**	0,006	0,497**
	10	5	0,033++	0,036	0,004	0,005
		10	0,024++	0,029++	0,003++	0,005
		20	0,026++	0,033++	0,005	0,005
		80	0,040	0,048	0,013	0,022**
	20	5	0,024++	0,027++	0,005	0,005
		10	0,027++	0,031++	0,007	0,006
		20	0,024++	0,030++	0,007	0,009

**Ultrapassou o LS do IC exato, com 99% de confiança para $\alpha = 0,01$ (0,021276) e $\alpha = 0,05$ (0,070504).

++Não atingiu o LI do IC exato, com 99% de confiança para $\alpha = 0,01$ (0,003727) e $\alpha = 0,05$ (0,033927).

de repetições, uma vez que as diferenças paramétricas entre médias dos níveis dos fatores foram fixadas em um valor constante (δ) em erros padrões. Em todos os casos, o valor 0,80 de poder foi atingido ou ultrapassado somente com $\delta > 4$. Os resultados para $\alpha = 0,01$ seguem o mesmo padrão de resposta e não são apresentados.

Na Figura 3 estão apresentados os gráficos para o poder dos testes C e CB em função do número de erros padrões (σ_x), repetições r e número k de níveis do fator de distribuição lognormal (0, 1), H_1 e $\alpha = 0,05$. Pode-se observar que os valores de poder dos testes aumentam com o aumento de δ , r e k , exceto quando k passa de 5 para 10. A taxa de aumento em função de δ para o valor de poder é menor para situações com valores de k maiores. O padrão de resposta dos testes quanto ao poder foi semelhante ao observado em normalidade (Figura 2). Sobre a lognormal, houve efeito do número de repetições para todos os valores de k , mas principalmente para $k \leq 20$.

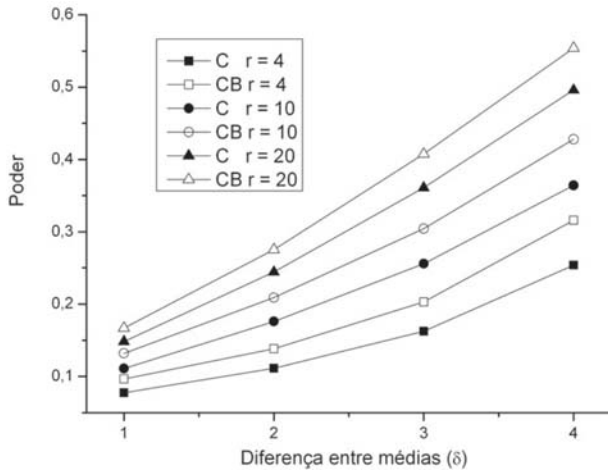
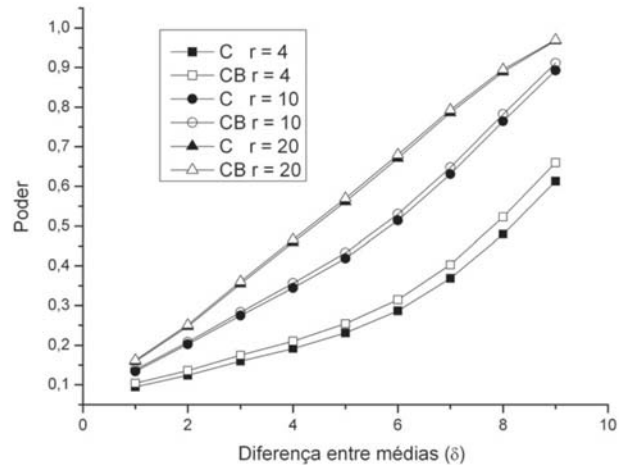
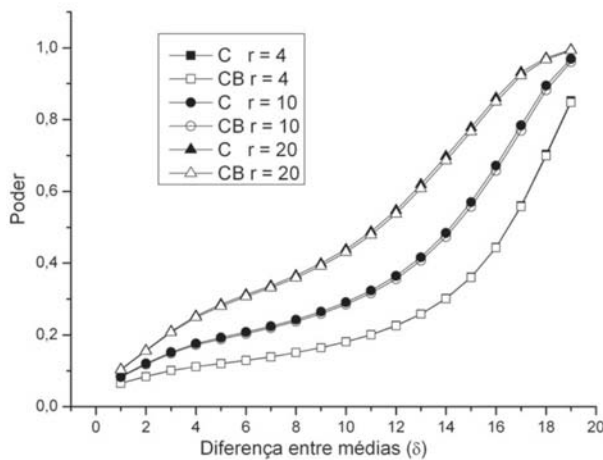
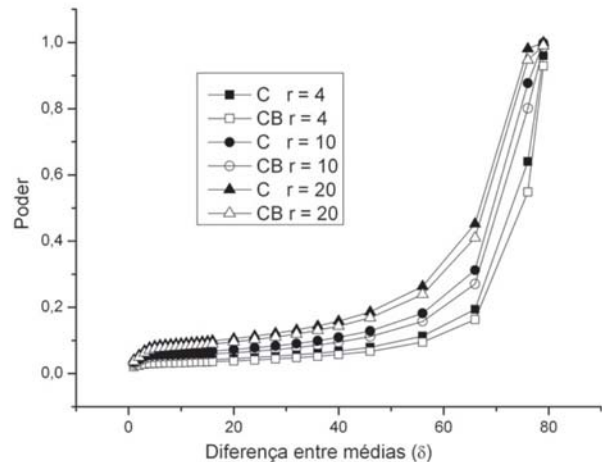
(a) $k = 5$ (b) $k = 10$ (c) $k = 20$ (d) $k = 80$

Figura 3. Poder dos testes de Caliński e Corsten (C) e sua versão *bootstrap* (CB) para diferentes números de repetições r , considerando-se a distribuição lognormal (0, 1) em função das diferenças entre médias δ para diferentes números de níveis k do fator, com H_1 e $\alpha = 0,05$.

Acredita-se que tal efeito seja consequência da heterogeneidade das variâncias na primeira situação. Os resultados de poder dos testes para a distribuição exponencial (0,1) foram similares aos obtidos com a lognormal, por isso não são apresentados.

Na Figura 4 estão apresentados os valores de poder dos testes C e CB em função de δ , r e k , sob normalidade e H_0 parcial para $\alpha = 0,05$ e $0,01$. Verifica-se aumento de poder à medida que δ aumenta, como é esperado. A taxa de aumento do poder é grande, e com $\delta \geq 4$ o poder se aproxima ou ultrapassa 0,80, principalmente para maiores valores de k e para $\alpha = 0,05$. Os valores de poder são maiores para $\alpha = 0,05$ em relação a $\alpha = 0,01$. Para $\delta \geq 8$, os valores de poder alcançam 1,00 tanto para $\alpha = 0,01$ quanto para $\alpha = 0,05$. Praticamente não há diferenças entre os dois testes estudados. Como as taxas de erro tipo I foram elevadas nesta situação de H_0 parcial, esperava-se que o poder fosse realmente maior do que sob H_1 , o que ocorreu.

Para os dados do experimento da Tabela 1, foram obtidas as médias ordenadas dos níveis do fator: $\bar{X}_{(1)} = 13,26(3D13)$, $\bar{X}_{(2)} = 14,64(3D4)$, $\bar{X}_{(3)} = 18,70(\text{composto})$, $\bar{X}_{(4)} = 19,92(3D7)$, e $\bar{X}_{(5)} = 23,98(3D5)$ e $\bar{X}_{(6)} = 28,82(3D1)$. A matriz de distâncias $D = \{d_{ij}\}$ com dimensão (6 x 6) foi criada:

$$D = \begin{pmatrix} 0,00 & 1,38 & 5,44 & 6,66 & 10,72 & 15,56 \\ 1,38 & 0,00 & 4,06 & 5,28 & 9,34 & 14,18 \\ 5,44 & 4,06 & 0,00 & 1,22 & 5,28 & 10,12 \\ 6,66 & 5,28 & 1,22 & 0,00 & 4,06 & 8,90 \\ 10,72 & 9,34 & 5,28 & 4,06 & 0,00 & 4,84 \\ 15,56 & 14,18 & 10,12 & 8,90 & 4,84 & 0,00 \end{pmatrix} \quad (5)$$

Na Tabela 5 são apresentados os valores obtidos ao se aplicarem os testes C e CB aos dados do exemplo. Esses valores se referem às amplitudes R_{ϵ} entre as médias e aos valores- p obtidos em cada passo de fusão pelos dois pro-

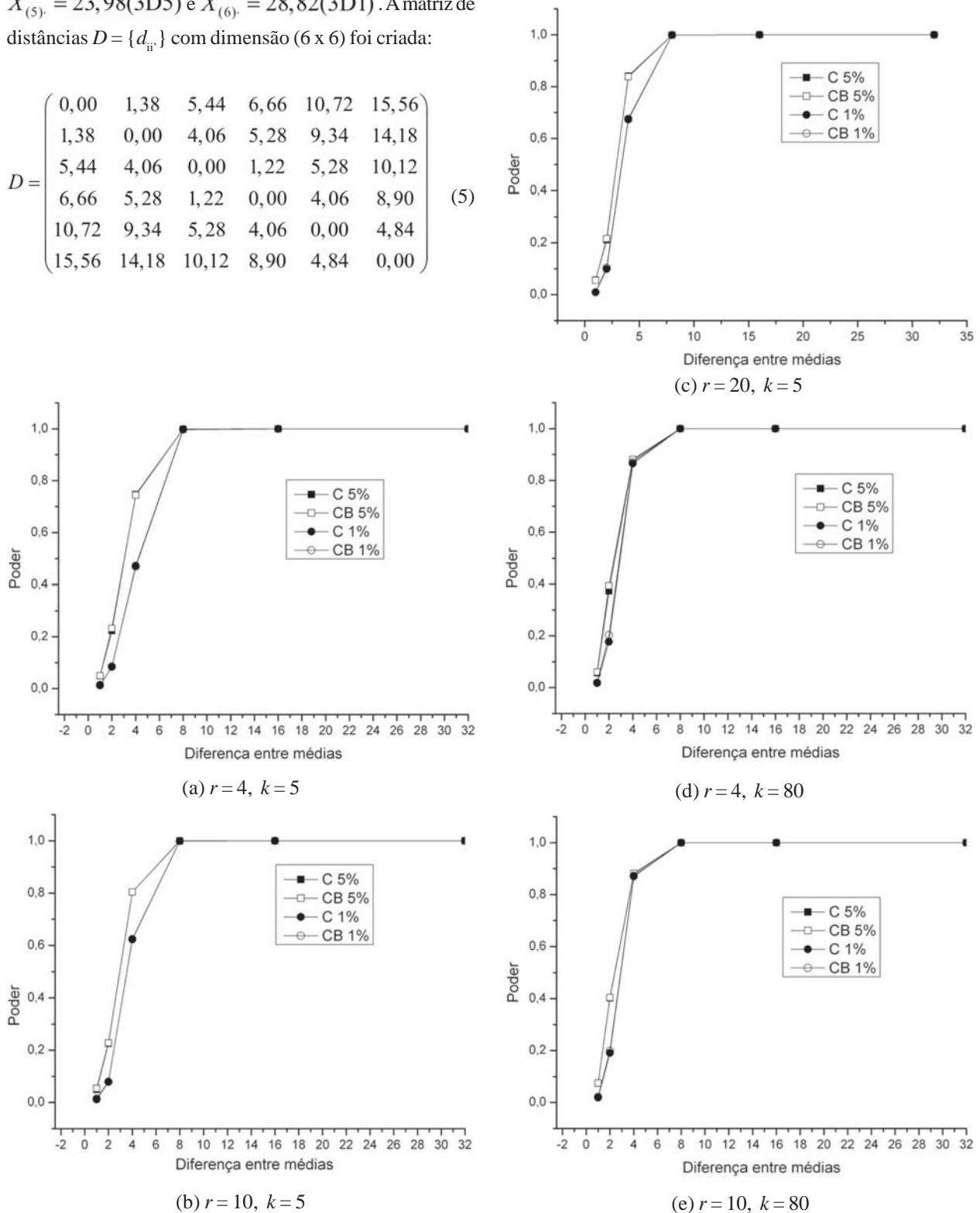
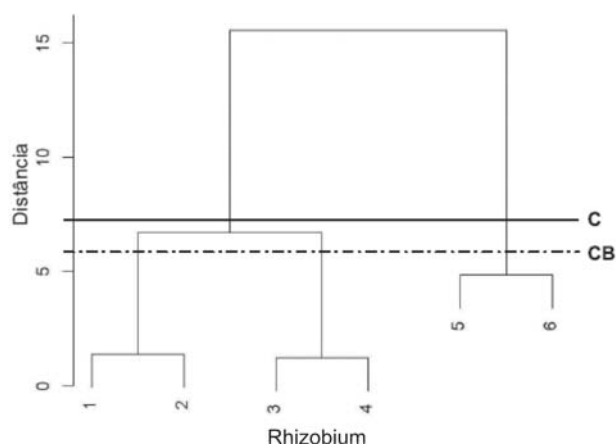


Figura 4. Poder dos testes de Caliński & Corsten (C) e sua versão *bootstrap* (CB) para $k = 5$ e $k = 80$ e níveis do fator, considerando-se a distribuição normal (N) (10, 1), com H_0 parcial, para $\alpha = 0,01$ e $0,05$.

Tabela 5. Valores das amplitudes R_i e dos valores- p obtidos em cada passo de fusão pelos procedimentos de Caliński & Corsten (C) e sua versão *bootstrap* (CB) para $B = 1000$ reamostragens

Passo	Médias unidas	R_i	Valor- p (C)	Valor- p (CB)
1	3 – 4	1,22	0,993	0,991
2	1 – 2	1,38	0,987	0,989
3	5 – 6	4,84	0,262	0,254
4	1 – 4	6,66	0,053	0,046
5	1 – 6	15,56	0,000	0,000

**Figura 5.** Dendrograma com os pontos de corte obtidos pelos testes C e CB para $\alpha = 0,05$.

cedimentos. Na Figura 5 encontra-se o dendrograma obtido com a utilização da função *agnes* do programa R (R Development Core Team, 2005) aplicada à matriz de distâncias obtida com as médias dos dados. Os pontos de corte considerados foram aqueles em que $\alpha = 0,05$ e resultaram em diferentes agrupamentos para os dois testes: para o teste C foram obtidos dois grupos de médias (1-4)(5-6) e para o teste CB, três grupos (1-2)(3-4)(5-6), com $B = 1000$ reamostragens.

CONCLUSÕES

O desempenho do teste *bootstrap* é, em geral, considerado superior ao do teste original e é robusto, sendo, portanto, recomendada a sua utilização rotineiramente. Os testes C e CB são exatos em H_0 completa e normalidade. O teste CB em não-normalidade e H_0 completa controlou o erro tipo I por experimento e é considerado robusto. O teste C sob não-normalidade e H_0 completa é conservativo para valores pequenos de k e liberal para grandes valores de k . Em relação a H_0 parcial, os testes C e CB são, em geral, liberais para $\delta \geq 4$ e conservativos para $\delta \geq 8$. O poder do teste CB supera de forma inexpressiva o poder do teste C sob H_0 parcial ou sob H_1 .

REFERÊNCIAS

Banzatto DA & Kronka SN (1989) Experimentação Agrícola. Jaboticabal, FUNEP. 247p.

Borges LC & Ferreira DF (2003) 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, 21:67-83.

Carmer SG & Swanson MR (1973) An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal American Statistical Association, 68:66-74.

Caliński T & Corsten LCA (1985) Clustering means in ANOVA by Simultaneous Testing. Biometrics, 41:39-48.

Dachs JNW (1988) Estatística computacional: uma introdução em turbo Pascal. Rio de Janeiro. 236p.

Hinkelmann K & Kempthorne O (1987) Design and analysis of experiments. v. 1. J. Wiley & Sons, New York. 495p.

Hochberg Y & Tamhane AC (1987) Multiple comparison procedures. J. Wiley & Sons, New York. 450p.

Johnson RA & Wichern DW (1998) Applied multivariate statistical analysis, 4ª ed. Englewood Cliffs, Prentice Hall. 816p.

Keselman HJ, Cribbie RA & Wilcox RR (2002) Pairwise Multiple Comparison Tests when Data are Nonnormal. Educational and Psychological Measurement, 62: 420-434.

Leemis, L & Trivedi KS (1996) A comparison of approximate interval estimators of the Bernoulli parameter. The American Statistician, 50:63-68.

Machado AA, Demétrio CGB, Ferreira DF & Silva JGC (2005) Estatística experimental: uma abordagem fundamentada no planejamento e no uso de recursos computacionais. In: 50ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, Londrina. Curso, Londrina: ISBN. 290p.

Manly BFJ (1998) Randomization, bootstrap and Monte Carlo methods in biology, 2ª ed. Chapman-Hall, London. 399p.

Mood AM, Graybill FA & Boes DC (1974) Introduction to the theory of statistics, 3ª ed. J. Wiley & Sons, New York. 564 p.

Pachares J (1959) Table of the upper 10 % points of the studentized range. Biometrika, 46:461-466.

Perecin D & Barbosa JC (1988) Uma avaliação de seis procedimentos para comparações múltiplas. Revista de Matemática e Estatística, 6:95-103.

R Development Core Team (2005) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, <http://www.R-project.org>.

Ramalho MAP, Ferreira DF & Oliveira AC (2000) Experimentação em genética e melhoramento de plantas. UFLA, Lavras. 303p.

Scott AJ & Knott M (1974) A cluster analysis method for grouping means in the analysis of variance. Biometrics, 30:507-512.

Silva EC da, Ferreira DF & Bearzoti E (1999) Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. Ciência Agrotécnica, 23:687-696.

Steel RGD & Torrie JH (1980) Principles and procedures of statistics. 2ª ed. McGraw-Hill Book, New York. 633p.