UNIVERSIDADE FEDERAL DE ALFENAS
UNIFAL-MG

BRUNA DE OLIVEIRA GONÇALVES

TESTE DE STUDENT-NEWMAN-KEULS *BOOTSTRAP*: PROPOSTA, AVALIAÇÃO E APLICAÇÃO EM DADOS DE PRODUTIVIDADE DA GRAVIOLA

ALFENAS - MG
2015

BRUNA DE OLIVEIRA GONÇALVES

TESTE DE STUDENT-NEWMAN-KEULS *BOOTSTRAP*: PROPOSTA, AVALIAÇÃO E APLICAÇÃO EM DADOS DE PRODUTIVIDADE DA GRAVIOLA

Dissertação apresentada ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, área de concentração em Estatística e Experimentação Agropecuária da Universidade Federal de Alfenas, MG, como parte dos requisitos para a obtenção do título de Mestre.
Linha de Pesquisa: Modelagem Estatística e Estatística Computacional
Orientador: Profa. Doutora Patrícia de Siqueira Ramos
Coorientador: Prof. Doutor Fabricio Goecking Avelar

ALFENAS - MG
2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
Biblioteca Central da Universidade Federal de Alfenas

Gonçalves, Bruna de Oliveira.
Teste de Student-Newman-Keuls bootstrap: proposta, avaliação e aplicação e dados de produtividade da graviola. / Bruna de Oliveira Gonçalves. -- Alfenas/MG, 2015.
77 f.

Orientadora: Patrícia de Siqueira Ramos.
Dissertação (mestrado em Estatística Aplicada e Biometrica) - Universidade Federal de Alfenas, 2015.
Bibliografia.

1. Comparações múltiplas (Estatística). 2. Monte Carlo, metodo de. I. Ramos, Patrícia de Siqueira. II. Título.

CDD-519.5

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Alfenas / UNIFAL-MG
Programa de Pós-graduação em Estatística Aplicada e Biometria
Rua Gabriel Monteiro da Silva, 700. Alfenas - MG CEP 37130-000
Fone: (35) 3299-1392 (Secretaria) (35) 3299-1121(Coordenação)
https://www.unifal-mg.edu.br/ppgeab/

1914 100 anos 2014 Universidade Federal de Alfenas

BRUNA DE OLIVEIRA GONÇALVES

"TESTE DE STUDENT-NEWMAN-KEULS *BOOTSTRAP*: PROPOSTA, AVALIAÇÃO E APLICAÇÃO EM DADOS DE PRODUTIVIDADE DA GRAVIOLEIRA".

A Banca Examinadora, abaixo assinada, aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Linha de Pesquisa: Modelagem Estatística e Estatística Computacional.

Aprovado em: 12 de fevereiro de 2015.

Prof.ª Dr.ª Patrícia de Siqueira Ramos Assinatura: Patrícia S Ramos
Instituição: UNIFAL-MG

Prof. Dr. Flávio Bittencourt Assinatura:
Instituição: UNIFAL-MG

Prof. Dr. Denismar Alves Nogueira Assinatura:
Instituição: UNIFAL-MG

Aos meus pais, Maria e Sebastião.

DEDICO

AGRADECIMENTOS

A Deus, por tudo.

Aos meus pais pelo amor, carinho e apoio em todos os momentos.

À minha irmã Daiane, pela amizade e conselhos.

Ao meu namorado Flávio, pelo seu amor, paciência e compreensão.

À Universidade Federal de Alfenas e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, juntamente com seus docentes, pela oportunidade concedida para a realização do mestrado.

À FAPEMIG, pela concessão da bolsa de estudos.

À minha orientadora, professora Patrícia de Siqueira Ramos, pela amizade, apoio, confiança e ensinamentos. Fica aqui o meu agradecimento e a minha admiração pelo seu exemplo de competência.

Ao meu coorientador, professor Fabricio Goecking Avelar, por seu apoio e amizade, além de sua ajuda e dedicação, fatores fundamentais para a conclusão deste trabalho.

Aos professores do Instituto de ciências exatas, pela atenção e contribuição para a minha formação.

Aos membros das bancas do exame de qualificação e da banca de defesa de mestrado, professores Denismar Alves Nogueira e Flávio Bittencourt, pela participação na banca e pelas sugestões e contribuições para o desenvolvimento deste trabalho.

Ao professor Luiz Alberto Beijo, pela participação nas bancas de qualificação e de defesa da dissertação, pelas orientações e pelo incentivo.

Aos meus colegas de turma, Mariana, Michelle, Lislaine e José Marcio, pela amizade e companheirismo.

A todos que, de alguma forma, contribuíram para a realização desse trabalho.

"Sem sonhos, a vida não tem brilho. Sem metas, os sonhos não têm alicerces. Sem prioridades, os sonhos não se tornam reais."

Augusto Cury.

RESUMO

Os Procedimentos de Comparações Múltiplas (PCM) podem ser utilizados para comparar médias de tratamentos. Há muitos testes de comparações múltiplas e, para escolher o melhor, devem ser levados em conta o controle do erro tipo I (testes exatos, conservadores ou liberais) e o poder desses testes. Para melhorar o seu desempenho, em relação ao erro tipo I e poder, métodos de reamostragem *bootstrap* têm sido utilizados em alguns estudos sobre PCM. O teste de Student-Newman-Keuls (SNK) possui boas qualidades estatísticas que poderiam ser melhoradas com o uso do *bootstrap*. Assim, os objetivos deste trabalho foram propor uma versão utilizando o *bootstrap* paramétrico do teste de comparações múltiplas SNK (SNK$_B$), avaliar o desempenho do teste SNK$_B$ e compará-lo com o teste SNK. O desempenho foi avaliado pelas taxas de erro tipo I por experimento e pelo poder por meio de um estudo de simulação Monte Carlo em condições de normalidade e não normalidade dos resíduos. Foram realizadas $N = 1000$ simulações de experimento com k tratamentos ($k = 5, 10, 20$ e 80) com r repetições ($r = 4, 10$ e 20). Diferentes hipóteses sobre as médias foram consideradas. Sob H_0 completa, as médias foram consideradas todas iguais, sob H_1, as médias foram todas diferentes, considerando a mesma variância, e, sob H_0 parcial, foram considerados dois grupos cujas médias eram diferentes entre si. Ambos os testes apresentaram valores de taxas de erro tipo I próximos do nível nominal de $0{,}05$ sob H_0 completa e normalidade. Sob H_0 completa e não normalidade, os testes SNK e SNK$_B$ controlaram as taxas de erro tipo I por experimento na maior parte dos casos simulados para $k = 5$ e $k = 10$, enquanto que, para $k = 20$ e $k = 80$, ambos os testes foram considerados liberais em alguns cenários. Sob H_0 parcial, o teste SNK$_B$ foi liberal em todos os casos simulados, enquanto que o teste SNK foi, em geral, conservador para $\delta \leq 2$ e liberal para os demais valores de δ. O poder do teste proposto em geral superou o poder do teste original nas situações de normalidade e não normalidade. Assim, em situações práticas, se as diferenças entre as médias dos tratamentos forem pequenas ($\delta \leq 2$), o teste SNK é mais indicado por controlar o erro tipo I e apresentar valores de poder satisfatórios. Nos demais casos, o teste SNK$_B$ é mais recomendado, apesar de ambos serem liberais para $\delta \geq 4$, se a situação for de H_0 parcial. Além disso, os testes SNK e SNK$_B$ foram aplicados em dados reais de um experimento delineado para avaliar os controles químico e mecânico de pragas da gravioleira com o objetivo de comparar os resultados obtidos pelos dois testes.

Palavras-chave: Comparações múltiplas. Reamostragem. Simulação Monte Carlo. Erro tipo I. Poder.

ABSTRACT

Multiple Comparisons Procedures (MCP) are used to compare treatment means. There are many tests with this purpose and to choose the best one, two features must be analysed: the control of type I error rate (exact, conservative or liberal tests) and the power. Bootstrap resampling methods have been used in some studies to improve the performance of MCP. The Student-Newman-Keuls (SNK) test shows good statistical qualities that can be improved with the use of bootstrap. Therefore, this study aimed to propose a SNK parametric bootstrap version (SNK_B) and compare it with the original SNK test. The performance was evaluated by experimentwise error rates and power using a Monte Carlo simulation study considering normal and non-normal situations. We considered $N = 1000$ simulations of k treatments ($k = 5, 10, 20$ e 80) with r repetitions ($r = 4, 10$ and 20). Under null hypothesis, the means were considered all equal, under H_1 the means were all different, but the variance was the same and, under partial H_0, we considered two groups with different means. Both tests showed type I error rates values close to the nominal level of 0.05 under H_0 and normality. Under H_0 and non-normality, both tests controlled the experimentwise error rates in most simulated cases for $k = 5$ and $k = 10$, whereas, for $k = 20$ and $k = 80$, the tests were considered liberal in some scenarios. Under H_0 partial, the SNK_B test was liberal in all simulated cases, while SNK test was generally conservative for $\delta \leq 2$ and liberal to other δ values. In general, the power of the proposed test surpassed the power of original test under normality and non-normality. Thus, in practice, if the differences between the treatment means are small ($\delta \leq 2$), the SNK test works better given that it controls the type I error and the power is satisfactory. In the other cases, the SNK_B test is recommended, although both are liberal for $\delta \geq 4$, if we are under partial H_0. Furthermore, the tests were applied to a real experiment designed to evaluate the chemical and mechanical controls of pests soursop in order to compare the results of both tests.

Key-words: Multiple comparisons. Resampling. Monte Carlo simulation. Type I error rate. Power.

LISTA DE TABELAS

Tabela 1 - Peso médio de frutos colhidos de graviola em kg ordenados de forma decrescente. Sítio Aldeia Verde, Maceió - AL, outubro / 1999 a fevereiro/2000. 48

Tabela 2 - Taxa de erro por experimento (TPE) dos testes SNK original e SNK *bootstrap* (SNK_B) sob H_0 completa, considerando-se a distribuição normal (10, 1) e nível de significância $0{,}05$, em função do número de tratamentos k e número de repetições r. 50

Tabela 3 - Taxas de erros por experimentos (TPE) sob H_0 completa para os testes *t* de Student, Tukey e Duncan, em função do número de tratamentos *k*, número de repetições *r*, coeficientes de variação (CV) e nível nominal de significância de 0,05. 51

Tabela 4 - Taxas de erro tipo I por experimento (TPE) para o teste de Scott-Knott, em função do número de repetições *r*, número de tratamentos *k*, coeficientes de variação (CV) e nível nominal de significância de 0,05..... 52

Tabela 5 - Taxas de erro por experimento (TPE) dos testes SNK original e SNK *bootstrap* (SNK_B) sob H_0 completa, considerando-se a distribuição lognormal (0, 1) e nível de significância $0{,}05$, em função do número de tratamentos *k* e número de repetições r. 53

Tabela 6 - Taxas de erro por experimento (TPE) dos testes SNK original (SNK) e SNK *bootstrap* (SNK_B) sob H_0 completa, considerando-se a distribuição exponencial (0, 1) e nível de significância $0{,}05$, em função do número de tratamentos *k* e número de repetições r. 54

Tabela 7 - Taxas de erro por experimento (TPE) dos testes de Calinski e Corsten (C) e sua versão *bootstrap* (CB), em função do número de repetições, número de tratamentos e nível nominal de significância de $0{,}05$ sob H_0 completa, considerando-se as distribuições normal (10, 1), exponencial (0, 1) e lognormal (0,1). 55

Tabela 8 - Taxas de erro tipo I dos testes SNK e SNK_B, em função do número de tratamentos *k*, número de repetições *r*, diferenças entre as médias δ, para o nível nominal de significância $\alpha = 0{,}05$, sob a distribuição normal e H_0 parcial. 57

Tabela 9 - Taxas de erro tipo I do teste Scott-Knott para o nível nominal de significância de $0{,}05$, em função do número de repetições r e do número de tratamentos k sob H_0 parcial e distribuição normal. 59
Tabela 10 - Taxas de erro tipo I dos testes CF, CFB, C e CB, em função do número de tratamentos k, número de repetições r, diferenças entre as médias δ, para o nível nominal de significância $\alpha = 0{,}05$, sob a distribuição normal e H_0 parcial. 60
Tabela 11 - Poder do teste de Scott-Knott, ao nível nominal de significância de $0{,}05$, em função do número de repetições r, número de tratamentos k e do erro padrão da média (δ). 64
Tabela 12 - Poder para diferentes testes de comparações múltiplas de médias em função do número de tratamentos k e do erro padrão da média de um tratamento $\sigma_{\bar{Y}}$, com 20 repetições e coeficientes de variação de 10%. 65
Tabela 13 - Valores-p dos testes SNK e $SNK_{B.}$ 72
Tabela 14 - Peso médio (kg) de frutos colhidos de graviola com 20 repetições e resultados obtidos pelos testes SNK e $SNK_{B.}$ 73

LISTA DE FIGURAS

Figura 1 - Ilusração do método do *bootstrap* não paramétrico. 39

Figura 2 - Ilustração do método do *bootstrap* paramétrico. 40

Figura 3 - Ilustração do processo de reamostragem *bootstrap* e obtenção de q_b.....44

Figura 4 - Funções de densidade de probabilidade das distribuições normal (10,1), lognormal (0, 1) e exponencial (0, 1). 46

Figura 5 - Poder dos testes de Student-Newman-Keuls (SNK) e sua versão *bootstrap* (SNK_B), em função das diferenças entre as médias δ, diferentes números de repetições *r* e diferentes números de tratamentos *k*, considerando-se a distribuição normal (10, 1) sob H_1 e $\alpha = 0{,}05$. 62

Figura 6 - Poder dos testes de Calinski e Corsten (C) e sua versão *bootstrap* (CB) em função da diferença entre as médias do númerp de repetições e do número de tratamentos, considerando a distribuição normal e H_1. 63

Figura 7 - Poder dos testes de Student-Newman-Keul (SNK) e sua versão *bootstrap* (SNK_B), em função das diferenças entre as médias δ e diferentes números de tratamentos *k*, considerando-se a distribuição lognormal (0, 1), sob H_1 e $\alpha = 0{,}05$. 66

Figura 8 - Poder dos testes de Student-Newman-Keul (SNK) e sua versão *bootstrap* (SNK_B), em função das diferenças entre as médias δ, diferentes números de repetições *r* e para diferentes números de tratamentos *k*, considerando-se a distribuição exponencial, sob H_1 e $\alpha = 0{,}05$. 67

Figura 9 - Poder dos testes de Student-Newman-Keul (SNK) e sua versão *bootstrap* (SNK_B), em função das diferenças entre as médias δ, diferentes números de repetições *r* e para diferentes números de tratamentos *k*, considerando-se a distribuição normal, sob H_0 parcial e $\alpha = 0{,}05$. 68

Figura 10 - Poder dos testes de C, CB, CF e CFB em função das diferenças entre as médias δ, diferentes números de repetições *r* e para diferentes números de tratamentos *k*, considerando-se a distribuição normal, sob H_0 parcial, $\alpha = 0{,}05$, $k = 5$ e $k = 10$. 69

Figura 11 - Poder dos testes C, CB, CF e CFB em função das diferenças entre médias δ, diferentes números de repetições *r* e para diferentes números de tratamentos *k*, considerando-se a distribuição normal, sob H_0 parcial, $\alpha = 0{,}05$, $k = 20$ e $k = 80$. 70

SUMÁRIO

1 INTRODUÇÃO 14
2 REFERENCIAL TEÓRICO 16
2.1 Distribuições de probabilidade 16
2.1.1 Distribuição normal 16
2.1.2 Distribuição lognormal 17
2.1.3 Distribuição exponencial 18
2.2 Testes de hipóteses 19
2.2.1 Teste de hipóteses exato sobre o parâmetro *p* de uma binomial 23
2.2.2 Valor-*p* 24
2.3 Análise de variância 25
2.4 Procedimentos de comparações múltiplas 27
2.4.1 Contrastes 27
2.4.2 Amplitude estudentizada 29
2.4.3 Tipos de testes de comparações múltiplas 30
2.4.3.1 Teste t de Student 30
2.4.3.2 Teste de Tukey 31
2.4.3.3 Teste de Duncan 31
2.4.3.4 Teste de Student-Newman-Keuls (SNK) 32
2.4.3.5 Testes baseados em análise de agrupamento: teste de Scott-Knott e dois testes propostos por Caliński e Corsten (1985) 34
2.5 Método de simulação Monte Carlo 36
2.6 *Bootstrap* 37
2.6.1 *Bootstrap* não paramétrico 38
2.6.2 *Bootstrap* paramétrico 39
2.6.3 *Bootstrap* semiparamétrico 40
3 METODOLOGIA 42
3.1 Teste de Student-Newman-Keuls (SNK) original 42

3.2 Teste SNK *bootstrap* 43
3.3 Simulações 45
3.4 Aplicação 47
4 RESULTADOS E DISCUSSÕES 49
4.1 Erro tipo I sob H_0 completa 49
4.1.1 Distribuição normal 49
4.1.2 Distribuições não normais 53
4.2 Erro tipo I sob H_0 parcial 56
4.3 Poder sob H_1 61
4.3.1 Distribuição normal 62
4.3.2 Distribuições não normais 65
4.4 Poder sob H_0 parcial 68
4.5 Aplicação 71
5 CONCLUSÕES 74
REFERÊNCIAS BIBLIOGRÁFICAS 75

1 INTRODUÇÃO

Comumente, ao se analisar um conjunto de dados, o pesquisador precisa decidir se, em média, os contrastes ou tratamentos aplicados produziram resultados iguais ou quais produzem um resultado superior. Para verificar se existe pelo menos uma diferença significativa entre as médias dos tratamentos utiliza-se a análise de variância, porém, esta não indica quais médias diferem entre si. Os procedimentos de comparações múltiplas (PCM) permitem identificar essas diferenças.

Há muitos PCM e eles diferem quanto ao controle de erro tipo I e poder. Para escolher o melhor teste, deve-se considerar as qualidades estatísticas dos procedimentos (taxa de erro tipo I e poder) de acordo com a taxa de erro tipo I os testes podem ser classificados como exatos, conservadores ou liberais.

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 para comparar os testes de comparações múltiplas.

O que se espera de um teste é que ele controle o erro tipo I na maior parte das situações, e apresente altas taxas de poder. Para melhorar o seu desempenho, em relação ao erro tipo I e poder, métodos de reamostragem *bootstrap* têm sido utilizados em alguns estudos sobre testes de comparações múltiplas de médias. Ramos e Ferreira (2009) utilizaram *bootstrap* para um dos procedimentos de comparações múltiplas de Caliński e Corsten (1985) e seu desempenho foi considerado superior ao do teste original.

A ideia básica de *bootstrap*, na ausência de qualquer conhecimento sobre a população, é realizar reamostragem com reposição de tamanho n da amostra original. A distribuição *bootstrap* de algum estimador de interesse é utilizada no lugar da distribuição teórica deste mesmo estimador, em função da dificuldade de desenvolvê-la ou do desconhecimento da distribuição da população de onde foi obtida a amostra aleatória.

O teste de Student-Newman-Keuls (SNK) é um teste de comparações múltiplas que controla as taxas de erro tipo I por experimento sob H_0 completa mas se torna liberal sob H_0 parcial. Além disso, seu poder é superior aos dos testes de Tukey, t protegido de Bonferroni e Scheffé (RAMOS; FERREIRA, 2009). Portanto, é um teste com boas qualidades mas que podem ser melhoradas com o uso do *bootstrap*.

Assim, os objetivos desse trabalho foram propor uma versão do teste de comparações múltiplas SNK usando *bootstrap* paramétrico e avaliar os testes SNK e SNK *bootstrap* por meio de simulação Monte Carlo. Além disso, as duas versões foram aplicadas em dados reais de um experimento para comparar os controles químico e mecânico de pragas da gravioleira com o objetivo de comparar os resultados obtidos pelos dois testes.

2 REFERENCIAL TEÓRICO

O referencial teórico está dividido nas seguintes seções: 2.1 Distribuições de probabilidade, 2.2 Teste de hipóteses, 2.3 Análise de variância, 2.4 Procedimentos de comparações múltiplas, 2.5 Método de simulação Monte Carlo, 2.6 *Bootstrap*.

2.1 Distribuições de probabilidade

Nesta seção serão apresentados conceitos, como variável aleatória e algumas distribuições de probabilidade, necessários para o entendimento da teoria deste trabalho.

As distribuições de probabilidade, contínuas ou discretas, ficam completamente definidas conhecendo-se os diversos valores em que a variável aleatória pode assumir, dentro do seu intervalo de definição, e as respectivas probabilidades (FERREIRA, 2009). Segundo Meyer (1983), uma variável aleatória pode ser definida como

Definição 2.1 *Sejam ξ um experimento e S um espaço amostral associado ao experimento. Uma função Y, que associe a cada elemento $s \in S$ à um número real, $Y(s)$, é denominada variável aleatória.*

A distribuição de probabilidades associa uma probabilidade a cada resultado numérico de um experimento. Nesta seção serão considerados alguns modelos de probabilidade: normal, lognormal e exponencial, que foram utilizados neste trabalho. A varável aleatória $Y(S)$ será denotada neste trabalho como Y.

2.1.1 Distribuição normal

A distribuição normal, entre as distribuições de variável aleatória contínua, é uma das mais importantes e amplamente utilizada em diversas áreas. Mood, Graybill e Boes (1974) definem a distribuição normal como

Definição 2.2 *Uma variável aleatória Y tem distribuição normal com parâmetros μ e σ^2, se sua função densidade de probabilidade é dada por:*

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, -\infty < y < \infty$$

em que os parâmetros μ e σ satisfazem $-\infty < \mu < \infty$ e $\sigma > 0$.

A função de distribuição de probabilidade acumulada da normal pode ser representada pela seguinte expressão

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du.$$

A normal é uma distribuição que apresenta forma de sino e é simétrica em relação à média. Sua importância se deve a vários fatores e, segundo Ferreira (2009), o principal deles é que essa distribuição é limitante de muitas outras distribuições de probabilidade contínuas ou até mesmo discretas. Segundo o mesmo autor, a justificativa é dada pelo teorema central do limite, que é um resultado fundamental em aplicações práticas e teóricas.

2.1.2 Distribuição lognormal

Mood, Graybill e Boes (1974) definem a distribuição lognormal como

Definição 2.3 *Uma variável aleatória Y tem distribuição lognormal se sua função densidade de probabilidade for dada por:*

$$f(y;\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}},$$

em que $y > 0$ e os parâmetros μ e σ satisfazem $-\infty < \mu < \infty$ e $\sigma > 0$.

Segundo Ferreira (2009), como na distribuição normal, a função de densidade de probabilidade acumulada da distribuição lognormal não possui forma explícita e seus valores devem ser obtidos por meio de métodos numéricos. A média e a variância da variável aleatória lognormal Y são dadas por

$$E(Y) = e^{\mu+\sigma^2/2} \qquad \text{e} \qquad V(Y) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}.$$

Existe uma relação entre as distribuições lognormal e normal: se Y é uma variável aleatória positiva e X tem uma distribuição normal definida por $X = ln\ Y$, então Y, definida por $Y = e^X$, tem uma distribuição lognormal. Se X segue uma distribuição normal padronizada, por exemplo, então $Y = e^X$ segue uma lognormal com parâmetros $\mu = 0$ e $\sigma = 1$.

2.1.3 Distribuição exponencial

Bolfarine e Sandoval (2001) definem uma distribuição exponencial da seguinte forma

Definição 2.4 *Uma variável aleatória Y tem distribuição exponencial com parâmetro λ se a sua função densidade de probabilidade é dada por:*

$$f(y;\lambda) = \lambda e^{-\lambda y}, y > 0,$$

em que $\lambda > 0$.

A função de distribuição de probabilidade acumulada, $F(y)$, possui uma forma simples, ao contrário do que acontece com as distribuições normal e lognormal.

$$F(y) = 1 - e^{-\lambda y}.$$

A média e a variância da variável aleatória exponencial Y são dadas por:

$$\mu = \frac{1}{\lambda} \qquad \text{e} \qquad \sigma^2 = \frac{1}{\lambda^2}.$$

Outras distribuições, como por exemplo, as distribuições t de Student, F de Snedecor, qui-quadrado e binomial serão mencionadas ao decorrer do trabalho. Porém, as definições

dessas distribuições, bem como suas propriedades, serão consideradas como pré-requisito para a leitura das próximas seções. Essas distribuições são importantes em Inferência Estatística, parte da Estatística que objetiva tirar conclusões sobre uma população com base em uma amostra, e um dos métodos que podem ser utilizados são os testes de hipóteses.

2.2 Testes de hipóteses

Em muitas situações, pode-se ter o interesse de tomar a decisão de rejeitar ou não rejeitar determinada afirmação com base em alguma evidência e isso pode ser feito por meio da formulação de hipóteses. Uma hipótese estatística, segundo Mood, Graybill e Boes (1974), pode ser definida como

Definição 2.5 *Hipótese estatística é qualquer afirmação acerca da distribuição de uma ou mais variáveis aleatórias.*

Segundo Mood, Graybill e Boes (1974), uma hipótese pode ser simples ou composta. Uma hipótese denotada por *H* é denominada hipótese simples se especifica completamente uma distribuição como, por exemplo, $H : \theta = \theta_0$. Caso contrário, a hipótese é denominada composta como, por exemplo, $H : \theta \leq \theta_0$.

Existem duas hipóteses, a hipótese nula e a hipótese alternativa. A hipótese nula é a informação, que se pressupõe verdadeira, a respeito de algum parâmetro relacionado a uma variável aleatória e é denotada por H_0. Se a hipótese nula (H_0) é considerada falsa, então, uma hipótese alternativa, denotada por H_1 ou H_A, é considerada como verdadeira. O teste de hipóteses é um procedimento de inferência estatística amplamente usado em várias áreas do conhecimento para decidir, com base em uma amostra da população, qual dessas hipóteses é verdadeira. Segundo Bolfarine e Sandoval (2001), um teste de hipóteses pode ser definido como

Definição 2.6 *Um teste de hipótese estatística, denotado por τ, é uma função de decisão $d : \chi \to \{a_0,a_1\}$, em que χ denota o espaço amostral associado à amostra $Y_1, Y_2, \ldots, Y_n$; a_0 corresponde à ação de considerar a hipótese H_0 como a verdadeira e a_1 corresponde à ação de considerar a hipótese H_0 como a falsa.*

A função de decisão d divide o espaço amostral χ em dois subconjuntos:

$$A_0 = \{(y_1,y_2,\ldots,y_n) \in \chi; d(y_1,y_2,\ldots,y_n) = a_0\}$$

e

$$A_1 = \{(y_1,y_2,\ldots,y_n) \in \chi; d(y_1,y_2,\ldots,y_n) = a_1\},$$

em que $A_0 \cup A_1 = \chi$ e $A_0 \cap A_1 = \emptyset$. A_0 é chamado de região de não rejeição, enquanto A_1 de região de rejeição.

É importante observar que, por exemplo, não rejeitar H_0 não quer dizer que H_0 seja verdadeira, mas sim que os dados amostrais não fornecem indícios suficientes para rejeitá-la. Portanto, nesse processo, ao se rejeitar ou não rejeitar H_0, a decisão poderá ser concordante ou discordante da situação real, podendo ser correta ou incorreta. Segundo Mood, Graybill e Boes (1974), é conveniente investigar as probabilidades de tomar essas decisões erradas. De acordo com Carmer e Swanson (1973), os possíveis erros são

- Erro Tipo I: erro cometido ao se rejeitar a hipótese nula verdadeira. A probabilidade de se cometer esse erro é dada por: $P[\text{Erro Tipo I}] = P[\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}] = \alpha$;
- Erro Tipo II: erro cometido ao não se rejeitar a hipótese nula falsa. Não é controlável diretamente pelo pesquisador. A probabilidade de se cometer esse erro é dada por: $P[\text{Erro Tipo II}] = P[\text{não rejeitar } H_0 \mid H_0 \text{ falsa}] = \beta$.

As probabilidades de se cometerem os erros tipo I e II crescem de forma inversa, sendo impossível controlá-las ao mesmo tempo em um único experimento. Assim, a baixa probabilidade de se incorrer no erro tipo I está associada à alta probabilidade de se cometer o erro tipo II e o único modo de causar a redução simultânea de ambos os erros é aumentar o tamanho amostral (FERREIRA, 2009).

A probabilidade fixada para o erro tipo I é denominada nível de significância do teste e seu valor deve ser estabelecido antes do experimento ser realizado. Quando se rejeita H_0 e ela de fato é falsa, considera-se que uma decisão correta foi tomada, isto ocorre com probabilidade $1 - \beta$, valor que recebe o nome de poder do teste. Segundo Bolfarine e Sandoval (2001), o poder de um teste pode ser definido como

Definição 2.7 *O poder do teste com região crítica A_1 para testar $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$*

é dado por

$$P_{H_1}\left[Y \in A_1\right] = P\left[Y \in A_1 | \theta_1\right], \tag{2.1}$$

em que $P_{H_1}\left[Y \in A_1\right] = 1 - \beta$ *e* β *é a probabilidade de se cometer o erro tipo II.*

Segundo Machado et al. (2005), poder é a capacidade do teste de detectar todas as reais diferenças entre os efeitos dos tratamentos, ou seja, é a probabilidade de se rejeitar H_0 dado que H_0 falsa.

As características probabilísticas de um teste podem ser descritas através de uma função que associa a cada valor de θ a probabilidade $\pi(\theta)$, em que $\pi(\theta)$ é a probabilidade de rejeitar H_0. A função $\pi(\theta)$ é chamada função poder e, de acordo com Mood, Graybill e Boes (1974), essa função poder pode ser definida como

Definição 2.8 (Função Poder) *A função poder de um teste de hipótese* τ*, denotada por* $\pi_\tau(\theta)$*, é definida como a probabilidade de* H_0 *ser rejeitada ao longo dos possíveis valores de* θ*.*

A função poder indica a qualidade do teste e, portanto, é utilizada para comparar dois testes. Uma função poder ideal seria tal que $\pi(\theta) = 0$ para H_0 verdadeira e $\pi(\theta) = 1$ para H_0 falsa. Em um problema prático, no entanto, raramente existirá um teste com estas características. Portanto, segundo Mood, Graybill e Boes (1974), deve-se verificar qual o teste mais poderoso.

Definição 2.9 (Teste mais poderoso) *Um teste* $\tau*$ *de* $H_0 : \theta = \theta_0$ *e* $H_1 : \theta = \theta_1$ *é dito ser um teste mais poderoso de nível de significância* α *se, e somente se:*
(i)$\pi_{\tau*}(\theta_0) = \alpha$ *(erro tipo I)*
(ii)$\pi_{\tau*}(\theta_1) \geq \pi_\tau(\theta_1)$ *(poder do teste),* $\forall\ \tau$ *com* $\pi_\tau(\theta_0) \leq \alpha$,

em que $\pi_{\tau*}(\theta_0)$ é a probabilidade do teste $\tau*$ rejeitar H_0 verdadeira e $\pi_{\tau*}(\theta_1)$ é a probabilidade do teste $\tau*$ rejeitar H_0 dado que H_0 é falsa.

Além de seus valores de poder, uma das maneiras mais utilizadas para comparar dois testes é por meio de suas taxas de erro tipo I. De acordo com Carmer e Swanson (1973), existem várias dificuldades para se medirem essas taxas nos experimentos. Segundo Steel e Torrie (1980), existem duas formas mais comuns no contexto de procedimentos de comparações múltiplas. A primeira é definida como taxa de erro tipo I por comparação, TPC (*comparisonwise* ou *per-comparison error rate*), que pode ser definida como

Definição 2.10 *A taxa de erro tipo I por comparação é a razão entre o número de erros e o número de comparações realizadas:*

$$TPC = \frac{\textit{número de comparações erradas}}{\textit{número total de comparações}}.$$

A TPC indica a probabilidade de se rejeitar uma hipótese verdadeira em todas as possíveis combinações de médias de tratamentos, tomadas duas a duas.

A segunda forma é definida como taxa do erro tipo I por experimento, TPE (*experimentwise error rate*)

Definição 2.11 *A taxa de erro tipo I por experimento é a probabilidade de se realizar pelo menos uma inferência errada por experimento:*

$$TPE = \frac{\textit{número de experimetos com, pelo menos, uma comparação errada}}{\textit{número total de experimentos}}.$$

É desejável que os PCM controlem a taxa de ocorrência do erro tipo I com a mesma eficiência, tanto para comparações, como para experimentos. Porém, segundo Vieira (2006), alguns testes de comparações múltiplas controlam a taxa de erro tipo I por comparações, enquanto outros controlam a taxa de erro tipo I por experimentos.

Os testes de comparações múltiplas de médias podem ser exatos, conservadores ou liberais. De acordo com Biase e Ferreira (2011), eles podem ser definidos como

Definição 2.12 *Um teste é considerado conservador quando a taxa de erro tipo I é menor que o nível de significância adotado.*

Definição 2.13 *Um teste é considerado liberal quando a taxa de erro tipo I é maior que o nível de significância adotado.*

Definição 2.14 *Um teste é considerado exato quando a taxa de erro tipo I é igual ao nível de significância adotado.*

Para decidir se um teste é conservador, liberal ou exato, deve-se realizar um teste de hipóteses para saber se a taxa de erro tipo I pode ser considerada menor, maior ou igual ao nível de significância adotado. Para essa comparação pode ser utilizado o teste de hipóteses exato sobre o parâmetro *p* de uma binomial.

2.2.1 Teste de hipóteses exato sobre o parâmetro *p* de uma binomial

Em um teste de hipóteses sobre proporções (p), deseja-se testar se a proporção populacional p é igual a um valor que se suspeita p_0. As hipóteses que são testadas a respeito do valor p são as seguintes:

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}, \tag{2.2}$$

ou

$$\begin{cases} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{cases}, \tag{2.3}$$

ou

$$\begin{cases} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{cases}. \tag{2.4}$$

O teste de hipóteses expresso por (2.2) é denominado bilateral e os testes expressos em (2.3) e (2.4) são denominados unilaterais.

Para realizar o teste exato, denominado teste binomial exato, é usada a relação exata entre a binomial e a distribuição F. O termo exato é usado no contexto de que o valor real da significância é, no máximo, igual ao valor nominal de significância (α), devido à natureza discreta da distribuição binomial (FERREIRA, 2009). De acordo com a hipótese, calcula-se, por meio da distribuição binomial, a probabilidade de ocorrência de valores mais extremos do que o observado (valor-p):

$$P(Y \leq y) = \sum_{i=0}^{y} \binom{N}{i} p^i (1-p)^{N-i}, \quad P(Y \geq y) = \sum_{i=y}^{N} \binom{N}{i} p^i (1-p)^{N-i},$$

dependendo da hipótese alternativa, em que N é o número de ensaios independentes de Bernoulli e p a probabilidade de sucesso de cada ensaio.

Segundo Ferreira (2009), algumas dificuldades podem ser apontadas para a obtenção

dessas probabilidades. A necessidade de computar binomiais com valores de N grandes é uma delas, sendo necessários recursos computacionais ou utilizar uma aproximação para a determinação do valor-p. Mesmo com essa aproximação, o teste ainda pode ser considerado exato.

Para exemplificar um teste binomial exato, pode-se considerar, por exemplo, um experimento que foi repetido 1000 vezes com o objetivo de verificar as propriedades de um teste. A taxa de erro tipo I por experimento foi igual a 6,5% e deseja-se saber com 5% de significância se o teste pode ser considerado exato. De acordo com a definição 2.14, um teste é exato se a taxa de erro tipo I for igual ao nível de significância, portanto as hipóteses seriam:

$$\begin{cases} H_0 : p = 0{,}05 \\ H_1 : p \neq 0{,}05 \end{cases}, \tag{2.5}$$

em que $\widehat{p} = 0{,}065$.

A decisão sobre um teste de hipóteses de não rejeitar ou de rejeitar a hipótese nula pode ser tomada com base no valor-p.

2.2.2 Valor-*p*

O valor-p é utilizado para se tirar conclusões a respeito de um teste de hipóteses. Formalmente, Bolfarine e Sandoval (2001) definem

Definição 2.15 *O valor-p é o menor nível de significância para o qual a hipótese nula H_0 seria rejeitada.*

Dessa forma, se o nível de significância (α) proposto para o teste for menor que o valor-*p*, não se rejeita a hipótese H_0. Por outro lado, se o nível de significância for maior que o valor-*p*, então se rejeita H_0.

De acordo com Bussab e Morettin (2004), o valor-*p* também pode ser interpretado como a probabilidade de que a estatística do teste, valor calculado a partir de uma amostra de dados usado para decidir se a hipótese nula será rejeitada ou não, tenha valor extremo em relação ao valor observado quando a hipótese H_0 é verdadeira. Nesse sentido, se o valor-*p* é menor que o

nível de significância proposto (α), então a estatística do teste está na região crítica e, portanto, rejeita-se H_0, caso contrário, não se rejeita H_0.

Segundo Casella e Berger (2010), uma vantagem de relatar o resultado do teste por intermédio de um valor-*p* é que o pesquisador pode escolher o nível de significância α que considerar mais apropriado. Além disso, segundo os mesmos autores, um valor-*p* relata os resultados de um teste em uma escala contínua, em vez de apenas decisões dicotômicas de rejeitar ou não H_0.

As conclusões de um teste de hipóteses podem ser obtidas pelo método tradicional, que compara a estatística do teste com os valores críticos, ou pelo método do valor-*p*, que compara o valor-*p* com o nível de significância. Estes métodos são equivalentes no sentido de que sempre resultam na mesma conclusão (TRIOLA, 2013).

2.3 Análise de variância

A análise da variância ou ANAVA é uma técnica estatística desenvolvida por R. A. Fisher em 1935, com o objetivo de estudar fatores ou tratamentos sobre os quais se suspeita que produzam mudanças sistemáticas em alguma variável de interesse. Montgomery (1991) define análise de variância como

Definição 2.16 *A Análise de Variância (ANAVA) é um procedimento utilizado para a comparação de vários grupos ou estratos de interesse, que permite investigar a existência de diferenças significativas entre os fatores estudados.*

De acordo com Freund (2006), a análise de variância baseia-se na decomposição da variação total da variável resposta em partes que podem ser atribuídas aos tratamentos (variação entre tratamentos) e ao erro experimental (variação dentro dos tratamentos). Essas variações podem ser medidas pelas somas de quadrados e podem ser organizadas em uma tabela, chamada de tabela de análise de variância.

Segundo Casella e Berger (2010), para a realização da análise de variância devem ser consideradas as seguintes pressuposições: homocedasticidade, normalidade e independência dos resíduos e aditividade do modelo. Segundo Vieira (2006), na prática, dificilmente essas pressuposições são todas satisfeitas e a ANAVA pode ser realizada quando existem pequenos desvios das pressuposições básicas, mas nunca quando nenhuma das pressuposições é válida.

Para a aplicação da análise de variância, é necessário calcular a estatística de teste F, apresentada por Fisher em 1924, que é utilizada para testar se existem diferenças reais entre os tratamentos. A estatística F pode ser definida como a razão entre a variância atribuída à fonte estudada e a variância atribuída ao resíduo do modelo, ou seja,

$$F = \frac{\text{Variância da fonte testada}}{\text{Variância do resíduo}}.$$

Essa estatística segue uma distribuição F com parâmetros ν_1 e ν_2 em que ν_1 são os graus de liberdade da fonte testada e ν_2 os graus de liberdade do resíduo.

Pode parecer estranho um procedimento que compara médias ser chamado de análise de variância. Segundo Roberts e Russo (2014), uma estimativa da variância do resíduo, chamada variação dentro dos tratamentos, não é afetada por efeito dos tratamentos, logo, se a variação entre os tratamentos for maior que a variação dentro dos tratamentos , então houve efeito dos tratamentos. Portanto, em uma análise de variância, efetivamente há a comparação de variâncias.

Por exemplo, no caso de k tratamentos (níveis do fator) com médias populacionais $\mu_1, \mu_2, \ldots, \mu_k$, as hipóteses estatísticas seriam

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: & \text{pelo menos uma das médias difere das demais.} \end{cases}$$

O teste F será utilizado para testar se existem diferenças reais entre tratamentos, ou seja, testar a rejeição ou não de H_0. Se H_0 não é rejeitada, não há mais questionamentos a serem feitos. Porém, quando H_0 é rejeitada, há pelo menos uma diferença entre médias dos tratamentos.

Portanto, a análise de variância fornece um método para determinar diferenças entre as k médias dos tratamentos. Contudo, não se sabe quais médias são significativamente diferentes (FREUND, 2006). O próximo passo é investigar onde se encontram essas diferenças entre os tratamentos, por meio dos procedimentos de comparações múltiplas (PCM).

2.4 Procedimentos de comparações múltiplas

Um grande número de PCM tem surgido durante as últimas décadas para comparar médias de tratamentos quando o teste F da análise da variância é significativo. A literatura é bastante ampla no que diz respeito a esses testes, o que auxilia sua aplicação por pesquisadores de diferentes áreas.

De acordo com Vieira (2006), os testes de comparações múltiplas de médias podem ser adequados para comparação das médias duas a duas (*pairwise comparisons*), para comparação das médias de grupos tratados com o controle (*comparisons with control*) e para comparações múltiplas (*multiple comparisons*).

Uma qualidade desejada para um procedimento de comparações múltiplas 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). De acordo com Ramos e Ferreira (2009), há fortes indicativos de que a maioria dos procedimentos não é robusta em relação à violação de normalidade.

A aplicação dos procedimentos de comparações múltiplas depende da natureza da estrutura dos tratamentos. Se os tratamentos forem qualitativos, a utilização de contrastes seguida de um teste específico é aconselhada (MACHADO et al., 2005).

2.4.1 Contrastes

Com o uso de contrastes é possível estabelecer comparações, entre tratamentos ou grupos de tratamentos, que sejam de interesse. Cochran e Cox (1992) definem contraste da seguinte forma

Definição 2.17 *Uma comparação entre k médias de tratamentos é denominada contraste quando puder ser expressa por uma função linear destas médias $c = a_1y_1 + a_2y_2 + \cdots + a_ky_k$, e se $\sum_{i=1}^{k} a_i = a_1 + a_2 + \cdots + a_k = 0$.*

De forma geral, não se conhecem as médias verdadeiras dos tratamentos e, por isso, não se conhece o valor do contraste dessas médias. Segundo Gomes (1989), é possível calcular as

estimativas dos contrastes conhecendo-se as estimativas das médias. O valor paramétrico do contraste é definido por

$$c = a_1\mu_1 + a_2\mu_2 + \cdots + a_k\mu_k \ , \ \sum_{i=1}^{k} a_i = 0,$$

cujo estimador é

$$\widehat{c} = a_1\bar{Y}_1 + a_2\bar{Y}_2 + \cdots + a_k\bar{Y}_k,$$

em que $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$ são as médias amostrais dos k tratamentos.

Se houver mais de duas médias, sempre existirá mais do que um contraste entre elas. Uma propriedade importante desses contrastes é a ortogonalidade. Cochran e Cox (1992) definem contrastes ortogonais como

Definição 2.18 *Dois contrastes* $\widehat{c}_1 = a_1\bar{Y}_1 + a_2\bar{Y}_2 + \cdots + a_k\bar{Y}_k$ *e* $\widehat{c}_2 = b_1\bar{Y}_1 + b_2\bar{Y}_2 + \cdots + b_k\bar{Y}_k$ *são ortogonais se*

$$\sum_{i=1}^{k} a_i b_i = 0.$$

Segundo Gomes (1989), na análise de variância os contrates ortogonais são de grande importância, pois indicam que a variação de um contraste é inteiramente independente da variação de outro qualquer que seja ortogonal a ele.

Para comparar cada um dos k tratamentos existem $N = \dfrac{k(k-1)}{2}$ comparações possíveis e apenas $k-1$ contrastes são ortogonais. Com o aumento do número k de tratamentos e o consequente aumento da amplitude da maior para a menor média, percebe-se que a probabilidade de se encontrarem diferenças significativas entre os N contrastes também aumenta.

De acordo com Machado et al. (2005), um teste de comparações múltiplas consiste em testar os contrastes envolvendo duas médias e todas as combinações entre elas. Para encontrar essa diferença pode ser utilizado o valor da diferença mínima significativa (DMS) ou valor crítico. A DMS varia para cada teste na sua teoria. Sua forma geral é

$$DMS = \gamma S_d,$$

em que $S_d = \sqrt{\frac{QMR}{r}}$ é o estimador do erro padrão da diferença de duas médias em um delineamento balanceado, QMR é o quadrado médio do resíduo da análise de variância associado a ν graus de liberdade, r é o número de repetições dos tratamentos e γ é o quantil superior da distribuição que depende do método, dos graus de liberdade do resíduo e do número de comparações simultâneas.

A diferença entre duas médias é considerada significativa se a estimativa do contraste $\widehat{c} = \left|\bar{Y}_i - \bar{Y}_j\right|$ excede o valor da DMS. Se a estimativa do contraste não exceder o valor da DMS, então a diferença é não significativa. Como o valor da DMS é diferente para cada teste, um mesmo grupo de médias pode apresentar resultados diferentes.

Segundo Banzatto e Kronka (1989), quatro dos mais comuns testes de comparações múltiplas são relacionados à distribuição da amplitude estudentizada.

2.4.2 Amplitude estudentizada

Alguns dos principais métodos de comparações múltiplas são funções da distribuição da amplitude estudentizada. Segundo Freund (2006), essa distribuição foi projetada para controlar a probabilidade global de se cometer pelo menos um erro tipo I quando são comparados os diferentes pares de médias. A amplitude estudentizada, representada pela variável aleatória Q, é definida por

$$Q = \frac{Y_{(k)} - Y_{(1)}}{S},$$

em que $Y_{(k)}$ e $Y_{(1)}$ são a maior e a menor observação, respectivamente, em uma amostra aleatória de tamanho k, S é o estimador do desvio padrão amostral, com ν graus de liberdade.

Segundo Machado et al. (2005) os quatro testes dependentes dessa distribuição são: t de Student (LSD), Tukey, Duncan e SNK, os quais necessitam da obtenção dos quantis superiores $q_\alpha(p,\nu)$, em que α é o nível de significância, p é o número de médias abrangidas e ν os graus de liberdade do resíduo. A escolha do teste que se deve adotar depende de suas qualidades estatísticas, sendo função do tipo de erro que é controlado e da forma como os erros

são controlados (SOUSA; LIMA JUNIOR; FERREIRA, 2012).

2.4.3 Tipos de testes de comparações múltiplas

A seguir serão apresentados alguns dos procedimentos de comparações múltiplas mais utilizados: LSD, Tukey, Duncan, SNK, Scott-Knott e dois testes apresentados por Caliński e Corsten (1985).

2.4.3.1 Teste t de Student

O teste t de Student (LSD - *Least Significant Difference*) é um teste baseado na amplitude estudentizada utilizado para comparar médias. Segundo Gomes (1989), o teste é utilizado quando os contrastes são ortogonais e o número máximo de comparações a serem feiras será o número de graus de liberdade para tratamentos. O valor crítico do teste t de Student é dado por

$$LSD = q_{\alpha}(2,\nu)S_d,$$

em que $q_{\alpha}(2,\nu)$ é o quantil superior da amplitude estudentizada, ν são os graus de liberdade do resíduo, $S_d = \sqrt{QMR/r}$ é o estimador do erro padrão da diferença de duas médias e QMR é o quadrado médio do resíduo associado a ν graus de liberdade.

O teste t de Student controla o erro tipo I no nível nominal α em um contraste pareado testado individualmente, porém, não controla esse tipo de erro para todos os testes pareados possíveis, nesse mesmo nível α. Esse teste controla o erro por comparação em um nível nominal máximo igual à α e, por esse motivo, é um método recomendado para realizar comparações planejadas a *priori* (MACHADO et al., 2005).

2.4.3.2 Teste de Tukey

O teste de Tukey, proposto por Tukey (1953), também é baseado na amplitude estudentizada e pode ser utilizado para comparar todo e qualquer contraste entre duas médias de tratamento. O valor crítico para o teste é

$$TSD = q_{\alpha}(k,\nu)S_d,$$

em que $q_{\alpha}(k,\nu)$ é o quantil da amplitude estudentizada, k é o número de tratamentos e ν são os graus de liberdade.

Se k for maior do que 2, o valor TSD será maior do que o valor LSD, o que torna o teste de Tukey mais conservador do que o teste LSD (CARMER; SWANSON, 1973).

Segundo Santos, Santos e Mesquita (2010), muito raramente, pode acontecer de se obterem um ou mais contrastes significativos pelo teste de Tukey, embora o teste F não tenha sido significativo na análise de variância. Fatos semelhantes ocorrem com o teste de Duncan, que não concorda inteiramente com o teste F.

2.4.3.3 Teste de Duncan

O teste de Duncan introduzido por Duncan (1955), da mesma forma que o teste SNK a ser visto, exige o cálculo de $k-1$ valores críticos. É um teste que apresenta amplitudes estudentizadas especiais, em que o nível de significância α varia de acordo com o número de médias abrangidas. O valor crítico de teste pode ser calculado por

$$MRT_p = q_{\alpha}(p,\nu)S_d,$$

em que $q_{\alpha}(p,\nu)$ é quantil, p é o número de médias abrangidas ($p = 2,\ldots,k$) e ν são os graus de liberdade do resíduo. Cada contraste testado envolve apenas duas médias, embora a amplitude do contraste possa abranger um número maior de médias.

Sua aplicação é mais trabalhosa do que o teste de Tukey, mas encontra mais diferenças entre os tratamentos, isto é, o teste de Duncan indica resultados significativos em casos em que o teste de Tukey não indicaria (SANTOS; SANTOS; MESQUITA, 2010).

Segundo Machado et al. (2005), os valores críticos MTR_p são maiores do que o valor crítico LSD e menores do que o valor crítico TSD. Portanto, o teste de Duncan é mais conservador do que o teste t de Student, mas mais liberal do que o teste de Tukey.

2.4.3.4 Teste de Student-Newman-Keuls (SNK)

Newman (1939) apresentou um teste que contornava os inconvenientes do teste t de Student para ensaios com mais de dois tratamentos e que foi modificado por Keuls (1952). Ajustava o valor de t dependendo da distância entre as médias então ordenadas. Enquanto os testes LSD e de Tukey requerem o cálculo de um único valor crítico, o teste SNK, como o teste de Duncan, exige o cálculo de $k-1$ valores críticos (DMS), dados por

$$SNK_p = q_\alpha(p,\nu)S_d,$$

em que $p = 2, 3, \ldots, k$ é o número de médias abrangidas pelo contraste, $q_\alpha(p,\nu)$ é o quantil da amplitude estudentizada e ν são os graus de liberdade do resíduo. O teste SNK segue a mesma estrutura de procedimento do teste de Duncan, a diferença está no cálculo dos valores críticos.

SNK_2 é igual ao valor crítico LSD e SNK_k é igual ao valor crítico TSD. Para valores intermediários de p, o valor SNK_p é um valor intermediário entre os valores críticos LSD e TSD (MACHADO et al., 2005).

O teste SNK é indicado para a comparação entre duas médias. É menos conservador que o teste de Tukey e pode evidenciar diferenças não discriminadas por ele (SANTOS; SANTOS; MESQUITA, 2010).

Os passos para o teste de SNK serão apresentados a seguir. Seja uma amostra aleatória em que se tenham k tratamentos qualitativos não estruturados e r repetições Y_{11}, Y_{12},..., $Y_{1r}, Y_{21}, \ldots, Y_{2r}, \ldots, Y_{kr}$, em que Y_{ij} é a observação advinda da j-ésima repetição $(j = 1, 2, \ldots,$

r) do i-ésimo tratamento ($i = 1, 2, \ldots, k$). O estimador da média de cada i-ésimo tratamento é dado por

$$\overline{Y}_i = \frac{\sum_{j=1}^{r} Y_{ij}}{r}.$$

As médias de tratamentos ordenadas são representadas por $\overline{Y}_{(1)}, \ldots, \overline{Y}_{(k)}$. Essas médias devem ser organizadas em ordem decrescente de magnitude. A seguir, devem ser obtidos os valores dos contrastes $|\widehat{c}|_p$, o número de médias p abrangidas por esse contraste e os valores das DMS das comparações, seguindo o seguinte roteiro:

1. $\bar{Y}_{(k)}$ é a primeira média base;
2. Calcular todas as SNK_p;
3. Calcular o valor do contraste $|\widehat{c}|_p$ entre a média base e a menor média:
   a. Se $|\widehat{c}|_p < SNK_p$, todas as médias abrangidas pelo contraste recebem a mesma letra e a primeira diferente recebe outra letra;
   b. Se $|\widehat{c}|_p \geq SNK_p$, repete-se o passo 2 tomando a média anterior na comparação com a média base até obter uma $SNK_p < |\widehat{c}|_p$ ou até terminarem as médias;
4. Muda-se a média base para a próxima e repete-se até que a média base seja a penúltima.

Na seção 3.1 será apresentado um roteiro para a aplicação do teste SNK utilizando o valor-p para decidir se o contraste é significativo, já que essa foi a forma utilizada para aplicar o teste neste trabalho.

Muitas vezes, torna-se difícil a interpretação dos resultados do teste SNK por não haver uma real separação das médias em grupos. Isso ocorre devido à ambiguidade nos resultados, assim como acontece com os testes LSD, Tukey e Duncan. Segundo Borges e Ferreira (2003), uma alternativa é a utilização de técnicas de análise de agrupamento, como por exemplo, os testes de Scott e Knott (1974) e os de Caliński e Corsten (1985). Esses procedimentos separam as médias dos níveis do fator em grupos homogêneos, pela minimização da variação dentro e maximização da variação entre grupos.

2.4.3.5 Testes baseados em análise de agrupamento: teste de Scott-Knott e dois testes propostos por Caliński e Corsten (1985)

Segundo Bhering et al. (2008), o teste de Scott-Knott começa por compartimentar os grupos para maximizar a soma dos quadrados entre grupos. A soma dos quadrados é definida como B_0, de acordo com a expressão:

$$B_0 = \frac{T_1}{K_1} + \frac{T_2}{K_2} - \frac{(T_1 + T_2)^2}{K_1 + K_2},$$

em que T_1 e T_2 são os totais dos dois grupos e K_1 e K_2 o número de médias dentro de cada grupo.

O procedimento apresenta λ como estatística

$$\lambda = \frac{\pi}{2(\pi - 2)} \times \frac{B_0}{\widehat{\sigma_0^2}},$$

em que π é uma constante de valor $3{,}141593\ldots$, B_0 é o valor máximo da soma de quadrados máxima entre 2 grupos de médias sobre todas as $(p-1)$ partições e $\widehat{\sigma_0^2}$ é a estimativa de máxima verossimilhança de σ^2.

Scott e Knott (1974) mostram que a distribuição da estatística λ pode ser aproximada por uma qui quadrado (χ^2) com $\nu = \dfrac{k}{(\pi - 2)}$ graus de liberdade. Se $\lambda < \chi^2_\alpha(\nu)$, conclui-se que as médias dos tratamentos pertencem ao mesmo grupo e, se, $\lambda \geq \chi^2_\alpha(\nu)$ indica que os dois grupos são estatisticamente diferentes e devem ser testados separadamente para novas divisões possíveis. O procedimento termina quando não há grupos remanescentes para serem divididos.

Caliński e Corsten (1985) propuseram dois procedimentos: o primeiro é uma extensão da amplitude estudentizada e o segundo é baseado na distribuição F. De acordo com Ramos e Ferreira (2009), o primeiro método é hierárquico, aglomerativo. É denominado hierárquico porque, quando uma média é incluída em um grupo homogêneo, ela não será retirada dele, e aglomerativo porque, a cada passo, dois grupos são unidos para formar um novo grupo. O algoritmo inicia-se com os k tratamentos ordenados em relação às suas médias, formando k

grupos e é então calculado a amplitude c_0

$$c_0 = |\bar{Y}_i - \bar{Y}_j|.$$

Compara-se c_0 com o valor crítico $R_\alpha = q_\alpha S_d$, em que $q_\alpha(k,\nu)$ é o quantil superior da distribuição estudentizada para k médias e ν graus de liberdade do resíduo. A homogeneidade dentro de cada grupo é testada comparando-se novamente a amplitude com R_α.

O segundo procedimento de Caliński e Corsten (1985) é não hierárquico e com regra de parada baseada na razão F estendida. O processo se inicia com os k tratamentos formando um grupo único. Neste procedimento, para encontrar a melhor partição em r grupos, realiza-se a minimização da soma de quadrados dentro dos grupos por meio de

$$r \sum_{j=1}^{n} \sum_{i \in P_j} (\overline{Y_i} \, \overline{\overline{Y}}_{P_j})^2, \tag{2.6}$$

em que n é o número de passos, P_j as partições, r é o número de repetições, $\overline{Y_i}$ é a média amostral do $i-$ésimo tratamento, pertencente ao j-ésimo subgrupo P, e $\overline{\overline{Y}}_{P_j}$ é a média das médias amostrais do j-ésimo subgrupo P. O agrupamento é aceito se o valor crítico, $(k-1)F_\alpha(k-1,\nu)QMR$, não ultrapassa o valor calculado em (2.6).

Como existem diversos PCM propostos na literatura e cada um apresenta características desejáveis e indesejáveis, não há um teste melhor (VIEIRA, 2006). Para escolher um teste de comparações múltiplas é necessário levar em conta suas propriedades. Segundo Carmer e Swanson (1973), é possível comparar os testes de comparações múltiplas por meio de simulação Monte Carlo. Vários trabalhos, como os de Carmer e Swanson (1973), Silva, Ferreira e Bearzoti (1999), Borges e Ferreira (2003), Ramos e Ferreira (2009), utilizaram simulação Monte Carlo na avaliação de procedimentos de comparações múltiplas.

2.5 Método de simulação Monte Carlo

De acordo com Pegden, Shannon e Sadowski (1995), uma simulação pode ser definida como

Definição 2.19 *A simulação é um processo de projetar um modelo computacional de um sistema real e conduzir experimentos com este modelo com o propósito de entender seu comportamento e/ou avaliar estratégias para sua operação.*

Portanto, simulação é um processo que tenta reproduzir o comportamento de um sistema real, em geral por meio de programas de computadores, e é usada para a solução de problemas. Se a solução alcançada for mais rápida, com eficiência igual ou superior, de menor custo e de fácil interpretação em relação a outro método qualquer, o uso da simulação é justificável.

Com o crescimento da complexidade dos problemas reais e a evolução dos sistemas computacionais, a simulação aparece como um instrumento cada vez mais utilizado nas mais variadas áreas de conhecimento (GARCIA; LUSTOSA; BARROS, 2010).

Existem dois tipos de modelos: o determinístico e o estocástico ou probabilístico. Os modelos determinísticos não contêm nenhuma variável aleatória e os modelos estocásticos possuem uma ou mais variáveis aleatórias como entrada, que levam à saídas aleatórias. De acordo com Pegden, Shannon e Sadowski (1995), para modelos estocásticos, o mecanismo para geração de variáveis aleatórias é chamado de método de Monte Carlo.

O método de Monte Carlo, idealizado pelos matemáticos Von Neumann e Ulam, surgiu aproximadamente em 1944, período da Segunda Guerra Mundial, em que foi ferramenta de pesquisa para o desenvolvimento da bomba atômica. O nome simulação de Monte Carlo é devido à famosa roleta de Monte Carlo, no Principado de Mônaco, uma vez que a roleta seria comparável a um mecanismo gerador de números aleatórios (CIRILLO, 2013).

Segundo Dachs (1988), a simulação de Monte Carlo pode ser definida como

Definição 2.20 *A simulação de Monte Carlo é uma técnica que consiste na simulação de dados por meio da geração de números pseudo-aleatórios, por meio de algum algoritmo em alguma linguagem de programação, de acordo com determinada distribuição de probabilidade.*

Segundo Cirillo (2013), o suporte estatístico para validação dos resultados obtidos via simulação Monte Carlo é verificado na lei dos grandes números, sendo possível interpretar que,

à medida que o número de simulações aumenta, a estimativa converge para o verdadeiro valor da variável.

Em estudos de desempenho de testes estatísticos, devido à complicação de se obter analiticamente informações sobre as taxas de erro tipo I e poder, a simulação de Monte Carlo é uma alternativa utilizada em estudos de avaliação dos testes de comparações múltiplas.

2.6 *Bootstrap*

O *bootstrap*, desenvolvido por Efron no final da década de 1970, é um método computacional usado principalmente na obtenção de estimativas de parâmetros (RAMOS; FERREIRA, 2009). Este método se baseia na construção de distribuições amostrais por reamostragem, e é muito utilizado para encontrar as estimativas intervalares dos parâmetros. O método *bootstrap* também pode ser utilizado, por exemplo, para calcular empiricamente o viés e variância de um estimador θ.

A técnica consiste em se retirar uma amostra de tamanho n da população e reamostrá-la com reposição, obtendo novas amostras de tamanho n da amostra original. Cada uma das amostras obtidas pelas reamostragens é uma amostra *bootstrap* e calcula-se uma estatística de interesse. Com a repetição desse procedimento obtêm-se as estimativas dos parâmetros que serão usadas para gerar a distribuição denominada distribuição *bootstrap*. A distribuição *bootstrap* do estimador de interesse é então utilizada no lugar da "distribuição teórica" deste mesmo estimador.

O método de *bootstrap* tem por base a ideia de que o pesquisador pode tratar sua amostra como se ela fosse a população que deu origem aos dados e usar amostragem com reposição da amostra original para gerar pseudo-amostras. Essencialmente, fundamenta-se na ideia de que, na ausência de conhecimento sobre a população, a distribuição de valores de uma amostra aleatória de tamanho n é a melhor orientação da distribuição da população (FERREIRA, 2009).

Segundo Carpenter e Bithell (2000), a metodologia na realização de reamostragens diferencia a técnica *bootstrap* em três abordagens: *bootstrap* não paramétrico, *bootstrap* paramétrico e *bootstrap* semiparamétrico.

2.6.1 *Bootstrap* não paramétrico

A abordagem não paramétrica utiliza apenas as unidades amostrais, sem nenhuma pressuposição ou especificação sobre o modelo de distribuição para os dados. Carpenter e Bithell (2000) definem *bootstrap* não paramétrico como

Definição 2.21 *O método de bootstrap não paramétrico consiste na seleção de n amostras independentes* y_1^*,y_2^*, $\ldots$,y_n^**, denominadas de amostras bootstrap, de tamanho igual ao da amostra original, com reposição da mesma.*

O algoritmo geral para um *bootstrap* não paramétrico é o seguinte

1. Amostrar n observações aleatoriamente com reposição de $y_1,\ y_2, \ldots,\ y_n$ para obter um conjunto de dados de inicialização, denotado Y^*.

2. Calcular a estatística de interesse, $\widehat{\theta}^* = \widehat{\theta}(Y^*)$.

3. Repetir os passos 1 e 2 um grande número de vezes, por exemplo $B = 1000$, para se obter a distribuição *bootstrap*.

A notação * denota um valor *bootstrap*, ou reamostrado, θ é um parâmetro ou vetor de parâmetros e $\widehat{\theta}$ é uma estimativa do parâmetro ou vetor de estimativas do vetor de parâmetros.

Uma ilustração do *bootstrap* não paramétrico encontra-se na Figura 1.

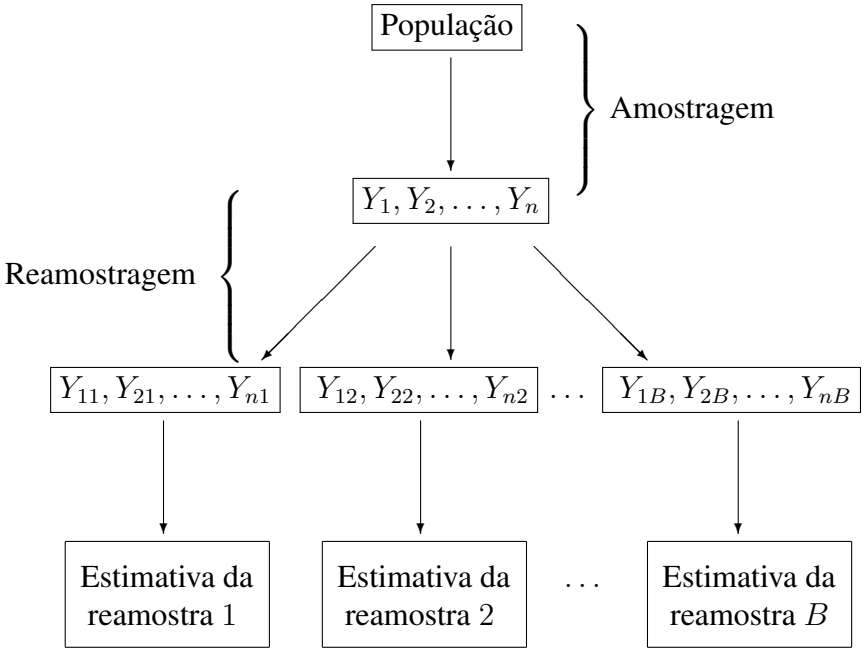


Figura 1 – Ilustração do método do *bootstrap* não paramétrico.
Fonte: Bastos (2013).

2.6.2 *Bootstrap* paramétrico

De acordo com Carpenter e Bithell (2000), no *bootstrap* paramétrico a distribuição da da amostra original é conhecida, de modo que os dados *bootstrap* são amostrados a partir dessa função.

Definição 2.22 *No método de bootstrap paramétrico, existe uma suposição sobre a distribuição que originou os dados amostrais e as B amostras boostrap são geradas utilizando esse modelo a partir dos parâmetros estimados com os dados da amostra original.*

O algoritmo para o *bootstrap* paramétrico é o seguinte

1. Seja $\widehat{\theta}$ o estimador de θ obtido dos dados. Amostram-se n observações $y_1,\ y_2, \ldots,\ y_n$, denotada por Y^*, a partir do modelo $f_y(\cdot; \widehat{\theta})$.

2. Calcular $\widehat{\theta}^* = \widehat{\theta}(Y^*)$.

3. Repita os passos 1 e 2 B vezes para obter uma estimativa da distribuição *bootstrap* paramétrica.

Uma ilustração do *bootstrap* paramétrico encontra-se representado na Figura 2.

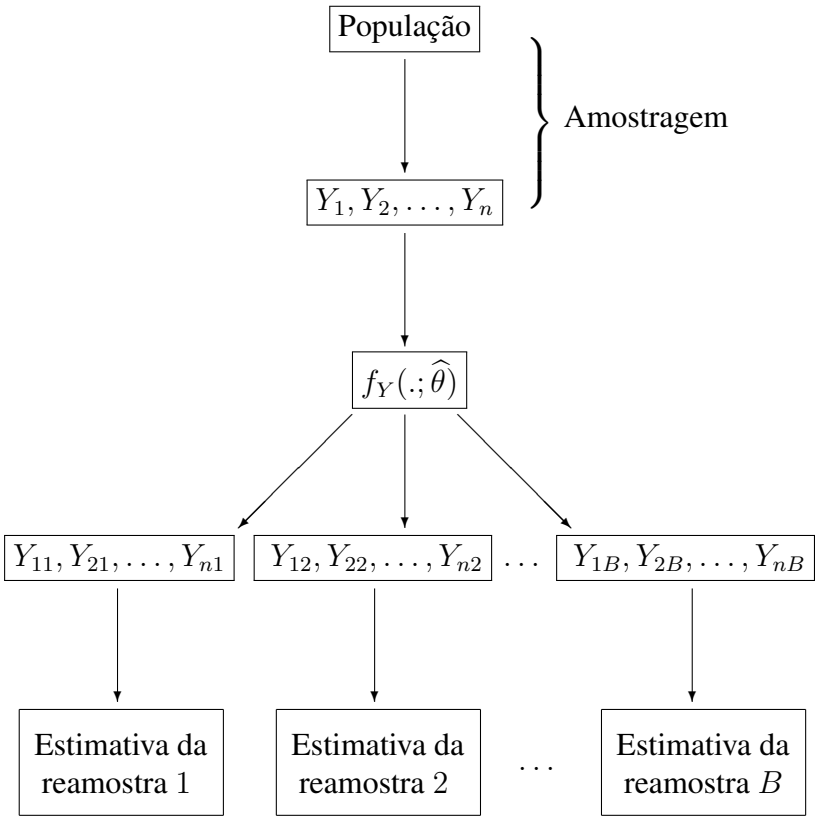


Figura 2 – Ilustração do método do *bootstrap* paramétrico.
Fonte: Bastos (2013).

2.6.3 *Bootstrap* semiparamétrico

Carpenter e Bithell (2000) definem o *bootstrap* semiparamétrico como

Definição 2.23 *O bootstrap semiparamétrico, apropriado para algumas formas de regressão, envolve reamostragem não paramétrica dos resíduos do modelo paramétrico.*

Segundo Cirillo (2013), a partir de um modelo como, por exemplo $Y = X\beta + \xi$, e considerando a amostra original composta por n resíduos, serão então extraídas as sub-amostras. Os passos são dados a seguir

1. Obtenha uma subamostra dos resíduos representada por $\xi_1^*, \xi_2^*, \ldots, \xi_n^*$.

2. Forme uma amostra com pseudo-observações, representadas por $y_1^*,\ y_2^*, \ldots,\ y_n^*$, utilizando o modelo $Y^* = X\widehat{\beta} + \xi^*$, em que X é o vetor de variáveis explicativas ou independentes.

3. Com a nova amostra obtida no passo 2, obtenha as estimativas de mínimos quadrados por $\widehat{\beta} = (X^t X)^{-1} X^t Y^*$.

4. Retorne ao passo 1 e repita o procedimento B vezes. Ao final, a distribuição empírica das estimativas dos parâmetros será gerada, podendo ser obtidas as estatísticas de interesse.

De acordo com Carpenter e Bithell (2000), esse tipo de reamostragem é apropriado somente quando é razoável supor que os resíduos ajustados são independentes e identicamente distribuídos.

3 METODOLOGIA

O desempenho de dois procedimentos de comparações múltiplas, SNK original e em sua versão *bootstrap*, foi avaliado por meio de simulação Monte Carlo. Foram delineadas simulações Monte Carlo sob a hipótese nula (H_0 completa), para mensurar as taxas de erro tipo I por experimento (TPE), e sob a hipótese alternativa (H_1), para mensurar o poder. Foram realizadas também simulações sob H_0 parcial, em que o erro tipo I foi medido dentro dos grupos de médias iguais e o poder foi medido entre os grupos de médias diferentes. Todas as rotinas necessárias para a implementação e avaliação dos testes foram realizadas utilizando o programa R em sua versão $3.0.2$ (R CORE TEAM, 2014).

Posteriormente, as taxas de erro tipo I e poder obtidas foram comparadas com os testes LSD, Tukey, Duncan, avaliados por Girardi, Cargnelutti Filho e Storck (2009) com o teste de Scott-Knott, avaliado por Borges e Ferreira (2003) e Silva, Ferreira e Bearzoti (1999), e os procedimentos de Caliński e Corsten baseados na amplitude estudentizada e na distribuição F, em suas versões original e *bootstrap*, avaliados respectivamente por Ramos e Ferreira (2009) e Ramos e Vieira (2014). Para tornar possível essa comparação, os cenários simulados neste trabalho foram semelhantes aos simulados pelos autores citados.

3.1 Teste de Student-Newman-Keuls (SNK) original

Para a aplicação do teste SNK original podem ser seguidos os passos descritos na seção $2.4.3.4$ utilizando os valores da DMS. O roteiro a seguir ilustra como a decisão pode ser tomada utilizando o valor-p para decidir se o contraste é significativo, considerando-se α o nível de significância adotado:

1. $\bar{Y}_{(k)}$ é a primeira média base;

2. Calcular o valor do contraste $|\widehat{c}|_p$ entre a média base e a menor média e o valor-p do contraste:

   a. Se valor-$p > \alpha$, todas as médias abrangidas pelo contraste recebem a mesma letra e a primeira diferente recebe outra letra;

   b. Se valor-$p \leq \alpha$, repete-se o passo 2 tomando a média anterior na comparação com a

média base até obter um valor-$p > \alpha$ ou até terminarem as médias;

3. Muda-se a média base para a próxima e repete-se até que a média base seja a penúltima.

O valor-p a ser calculado em cada contraste utiliza a distribuição da amplitude estudentizada ($q_\alpha(p,\nu)$) e é obtido por

$$\text{valor-}p = P\left[q_\alpha(p,\nu) > \frac{|\widehat{c}|_p}{\sqrt{QMR/r}}\right],$$

em que p é o número de médias abrangidas pelo contraste, r número de repetições, ν são os graus de liberdade associados ao QMR, sendo este o quadrado médio do resíduo da análise de variância.

3.2 Teste SNK *bootstrap*

O teste SNK *bootstrap* difere do teste original no cálculo do valor-p para decidir se as médias abrangidas pelo contraste em questão podem ser consideradas iguais ou diferentes, pois foi obtida uma distribuição da amplitude estudentizada q para cada reamostragem. Para implementar o *bootstrap* paramétrico, amostram-se rk observações de uma distribuição normal com média μ e variância σ^2, compondo novas amostras de cada tratamento. Novas médias são calculadas a partir das novas amostras obtidas e um valor q_b é obtido por meio de

$$q_b = \frac{\bar{Y}^b_{(k)} - \bar{Y}^b_{(1)}}{\sqrt{\frac{QMR_b}{r}}}, \tag{3.1}$$

em que QMR_b é o quadrado médio do resíduo da b-ésima amostra *bootstrap* e $\bar{Y}^b_{(1)}$ e $\bar{Y}^b_{(k)}$ são as médias amostrais ordenadas obtidas na b-ésima amostra. Este processo é repetido $B = 1.000$ vezes, como está representado na Figura 3.

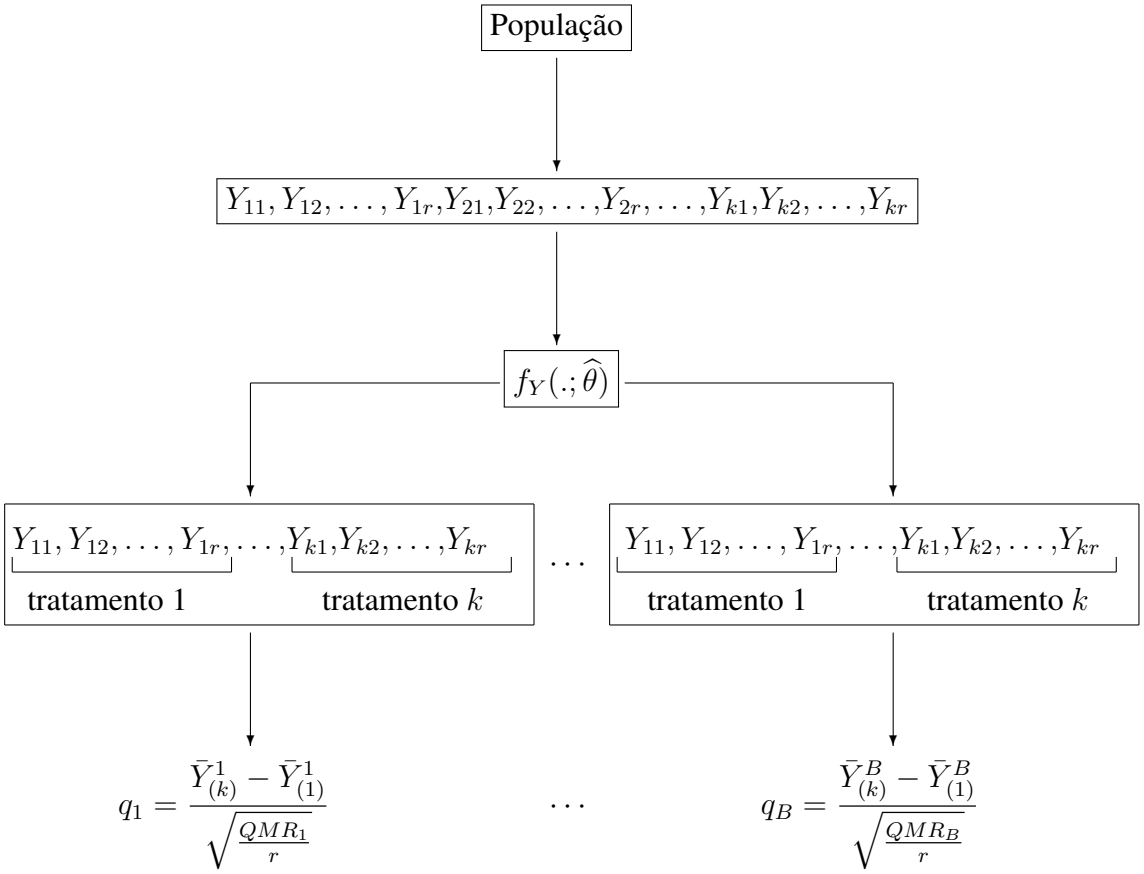


Figura 3 – Ilustração do processo de reamostragem *bootstrap* e obtenção de q_b.
Fonte: Da autora.

O conjunto com todos os valores q_b obtidos ($b = 1, 2, \ldots, B$) é utilizado para obter os valores-p por

$$\text{valor-}p = \frac{\sum_{b=1}^{B} I\left(q_b > \dfrac{|\widehat{c}|_p}{\sqrt{QMR/r}}\right)}{B}, \tag{3.2}$$

em que p é o número de médias abrangidas pelo contraste e $I(x)$ é a função indicadora de x, que satisfaz a seguinte regra:

$$\begin{cases} I(x) = 1 \text{ se } x \text{ for verdadeira} \\ I(x) = 0 \text{ se } x \text{ for falsa} \end{cases}.$$

O valor-p obtido em cada contraste entre médias foi comparado com o nível nominal de significância α e o mesmo critério do teste original foi aplicado para decidir se o contaste pode ser considerado significativo ou não.

3.3 Simulações

Foram consideradas $N = 1.000$ simulações de k tratamentos qualitativos não estruturados e número de repetições r, além de diferentes números de erros padrão δ de diferenças entre médias consecutivas, já que o poder foi avaliado sob H_1, em que as médias são todas diferentes, e sob H_0 parcial, em que algumas médias diferem entre si por δ erros padrão da média. O erro padrão da média é dado por

$$\sigma_{\bar{Y}} = \sqrt{\frac{\sigma^2}{r}}.$$

Os valores de k utilizados foram $k = 5$; 10; 20 e 80 tratamentos e valores de $r = 4$; 10 e 20 repetições. Essas combinações de valores tentam ilustrar o que acontece em situações reais, com poucos e muitos tratamentos e com números de repetições variados. Em todos os casos, o nível nominal de significância considerado foi de 5%.

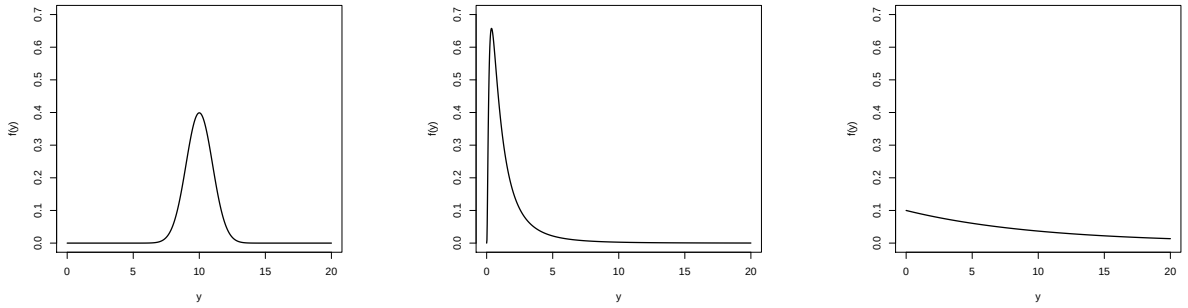
Para cada combinação de k e r utilizou-se o delineamento inteiramente casualizado (DIC), considerando o seguinte modelo:

$$y_{ij} = \mu + \tau_i + e_{ij},$$

em que y_{ij} é o valor observado para a variável resposta obtido para o i-ésimo tratamento em sua j-ésima repetição, μ é a constante associada a todas as observações, τ_i é o efeito do i-ésimo tratamento e e_{ij} erro aleatório associado a cada observação.

Foram considerados os modelos probabilísticos normal, lognormal e exponencial para descrever a variável aleatória do erro experimental e, assim, com as distribuições diferentes da normal, avaliar a robustez dos testes em situações adversas.

A Figura 4 ilustra as três distribuições de probabilidade utilizadas com os respectivos parâmetros. Sob H_0 completa, a média da normal foi considerada $\mu = 10$ sem perda de generalidade, bem como a variância $\sigma^2 = 1$. No modelo exponencial, o valor de λ foi $0{,}1$ e, no lognormal, a média foi igual a 0 e a variância igual a 1.



(a) Distribuição normal (b) Distribuição lognormal (c) Distribuição exponencial

Figura 4 – Funções de densidade de probabilidade das distribuições normal (10, 1), lognormal (0, 1) e exponencial (0,1).
Fonte: Da autora.

Além disso, ainda foram consideradas diferentes hipóteses sobre as médias. Sob H_0 completa, as médias dos tratamentos foram consideradas todas iguais. Dessa forma, para o modelo normal, o coeficiente de variação (CV) utilizado foi 10%. Uma ressalva deve ser feita em relação à escolha de apenas um valor de CV. De acordo com Silva, Ferreira e Bearzoti (1999) e Girardi, Cargnelutti Filho e Storck (2009), o CV não afetou as taxas de erro tipo I dos testes avaliados por esses autores, indicando que o esforço por parte de pesquisadores em diminuir o valor do CV não interfere no controle do erro tipo I dos testes de comparações múltiplas.

Para cada um dos valores obtidos das taxas de erro tipo I foram aplicados testes binomiais exatos considerando um nível nominal de significância de $0{,}01$ para as hipóteses $H_0 : \alpha = 0{,}05$ e $H_1 : \alpha \neq 0{,}05$. Se a hipótese nula for rejeitada (valor-$p \leq 0{,}01$) e as taxas de erro tipo I observadas forem consideradas significativamente inferiores ao nível nominal de significância considerado, o teste é considerado conservador. Se a hipótese nula for rejeitada (valor-$p \leq 0{,}01$) e as taxas de erro tipo I observadas forem consideradas significativamente superiores ao nível nominal de significância, o teste é considerado liberal. Porém, se as taxas de erro tipo I observadas não forem consideradas significativamente diferentes do nível nominal de significância considerado (valor-$p > 0{,}01$), o teste é considerado exato, ou seja, controla o

erro tipo I.

Sob a hipótese alternativa H_1, foram consideradas todas as k médias diferentes, mas a variância σ^2 foi constante. Entre duas médias de tratamentos consecutivas foi fixada a diferença de um erro padrão da média.

Sob a hipótese nula (H_0) parcial foram considerados dois grupos cujas médias são diferentes entre si por δ erros padrão, sendo $\delta = 2, 4, 8$ e 16. Dentro do grupo de médias iguais o mesmo procedimento para H_0 completa foi aplicado.

Os testes SNK original e *bootstrap* foram aplicados em todas as configurações simuladas e as taxas de erro tipo I por experimento e poder foram estimadas em todos os N experimentos gerados. Sob as hipóteses nulas completa e parcial, as taxas de erro tipo I são obtidas e, sob as hipóteses parcial e alternativa, o poder é estimado.

Após a implementação e avaliação do desempenho do teste SNK *bootstrap* será possível, num trabalho futuro, incluí-lo no pacote *ExpDes* (*Experimental Designs*) (FERREIRA; CAVALCANTI; NOGUEIRA, 2011), caso o desempenho desse teste seja considerado satisfatório, ou seja, se apresentar controle do erro tipo I e altos valores de poder.

3.4 Aplicação

Para avaliar o desempenho dos testes descritos neste trabalho foram utilizados dados reais de controles químico e mecânico de *Cerconota anonella* (Sepp.) (Lepidoptera: Oecophoridae) e de *Bephratelloides pomorum* (Fab.) (Hymenoptera: Eurytomidae), principais pragas da gravioleira, do trabalho realizado por Micheletti et al. (2001). O delineamento experimental utilizado foi o inteiramente casualizado, com 9 tratamentos e 20 repetições. Foi avaliada a variável peso e os tratamentos utilizados foram: 1- frutos sem proteção (testemunha); 2- saco de papel kraft (43cm de comprimento x 24cm de largura); 3- saco plástico comum fechado na extremidade inferior, apenas com alguns orifícios feitos para permitir o escoamento da água (49cm de comprimento x 28cm de largura); 4- saco plástico comum aberto na extremidade inferior (49cm de comprimento x 28cm de largura; 5- saco plástico perfurado (49cm de comprimento x 28cm de largura); 6- saco de papel impermeável nas duas faces (45cm de comprimento x 32cm de largura); 7- pulverização dos frutos semanal e localizada, com triflumuron 250g/kg, sendo a dosagem 100g/ha; 8- pulverizações com triflumuron 250g/kg, na

dosagem 100g/ha, sendo posteriormente os frutos ensacados com saco plástico perfurado; 9- pulverização dos frutos semanal e localizada, com imidacloprid 700g/kgg, sendo a dosagem 100g/ha. Os tratamentos foram aplicados em frutos de 4 a 6 cm de comprimento, repetidos 20 vezes, em plantas diferentes.

Os sacos foram presos aos ramos acima dos frutos, por meio de arame plastificado. No caso do tratamento com os inseticidas, considerou-se o intervalo de segurança de 10 dias que antecedem a colheita, para suspender a aplicação dos produtos em campo. Na ocasião da colheita, foram avaliadas o número de orifícios causados por C. anonella e por B. pomorum.

A característica peso dos frutos apresentou a média geral de 1,217 kg e coeficiente de variação (CV) igual a $41{,}21\%$. Na Tabela 1 são apresentados os pesos médios dos frutos de graviola, médias das 20 repetições de acordo com os respectivos tratamentos, ordenadas de forma decrescente.

Tabela 1 – Peso médio de frutos colhidos de graviola em kg ordenados de forma decrescente. Sítio Aldeia Verde, Maceió-AL, outubro/1999 a fevereiro/2000.

Tratamento	Peso médio (kg)
Saco de papel kraft	1,49
Saco de papel impermeável	1,40
Saco plástico aberto	1,37
Triflumuron + Saco plástico perfurado	1,37
Saco plástico fechado	1,36
Saco plástico perfurado	1,27
Testemunha	0,99
Imidacloprid	0,94
Triflumuron	0,76

Fonte: Micheletti et al. (2001).

Realizou-se a análise estatística dos dados, aplicando-se as duas versões do teste de SNK original e *bootstrap*, verificando as pressuposições dos testes, sendo adotado o nível nominal de $0{,}05$ de significância, com o objetivo de ilustrar a aplicação de ambos os testes.

4 RESULTADOS E DISCUSSÕES

Na sequência são apresentados os resultados em relação ao erro tipo I e ao poder, em cenários formados por combinações de número de tratamentos, número de repetições em situações de normalidade e não normalidade dos resíduos, considerando o nível nominal de significância de 0,05 e diferentes hipóteses sobre as médias, sob H_0 completa, sob H_1 e sob H_0 parcial.

4.1 Erro tipo I sob H_0 completa

Neste tópico são apresentados os resultados do erro tipo I por experimento em condições de normalidade e não normalidade dos resíduos.

4.1.1 Distribuição normal

Na Tabela 2 são apresentadas as taxas de erro tipo I por experimento (TPE) sob normalidade em função do número de tratamentos (k) e número de repetições (r) para H_0 completa dos testes SNK original (SNK) e sua versão *bootstrap* (SNK$_B$). Foram realizados testes binomiais exatos considerando um nível nominal de significância de $0{,}01$ para cada taxa de erro tipo I observada.

Tabela 2 – Taxas de erro por experimento (TPE) dos testes SNK original (SNK) e SNK *bootstrap* (SNK$_B$) sob H_0 completa, considerando-se a distribuição normal $(10, 1)$ e nível de significância $0{,}05$, em função do número de tratamentos k e número de repetições r.

k	r	SNK	SNK$_B$
5	4	0,045	0,043
	10	0,053	0,053
	20	0,052	0,051
10	4	0,052	0,051
	10	0,066	0,067
	20	0,054	0,055
20	4	0,047	0,050
	10	0,051	0,051
	20	0,057	0,056
80	4	0,047	0,053
	10	0,043	0,046
	20	0,050	0,048

Fonte: Da autora.

$^*CV = 10\%$

Sob H_0 completa e normalidade dos resíduos, as taxas de erro tipo I obtidas pelos testes SNK e SNK$_B$ não foram significativamente diferentes do nível nominal de significância adotado. Este resultado está de acordo com Machado et al. (2005) que afirma que o teste SNK controla as taxas de erro tipo I por experimento sob H_0 completa e normalidade. O mesmo resultado foi obtido por Ramos e Ferreira (2009) para os testes de Caliński e Corsten e a versão *bootstrap* do teste. Os resultados também se assemelham aos de Girardi, Cargnelutti Filho e Storck (2009) para o teste SNK, em que as taxas de erro tipo I por experimento não foram consideradas diferentes do nível nominal de significância, com exceção de alguns cenários que envolviam 3 e 5 tratamentos.

Na Tabela 3 estão apresentados os resultados obtidos por Girardi, Cargnelutti Filho e Storck (2009) para a comparação das taxas de erro tipo I dos testes t de Student, Tukey, Duncan e SNK.

Tabela 3 – Taxas de erros por experimento (TPE) sob H_0 completa para os testes t de Student, Tukey e Duncan, em função do número de tratamentos k, número de repetições r, coeficientes de variação (CV) e nível nominal de significância de $0{,}05$.

k	r	CV	t	Tukey	Duncan	SNK
5	4	1	0,266**	0,049	0,190**	0,052
		10	0,258**	0,051	0,189**	0,054
		20	0,267**	0,055	0,192**	0,060
5	10	1	0,269**	0,046	0,182**	0,050
		10	0,283**	0,045	0,189**	0,055
		20	0,282**	0,053	0,196**	0,059
5	20	1	0,289**	0,051	0,186**	0,054
		10	0,267**	0,049	0,178**	0,054
		20	0,276**	0,049	0,185**	0,052
10	4	1	0,573**	0,057	0,375**	0,057
		10	0,595**	0,052	0,368**	0,053
		20	0,574**	0,046	0,368**	0,047
10	10	1	0,607**	0,056	0,378**	0,056
		10	0,621**	0,046	0,379**	0,046
		20	0,607**	0,058	0,387**	0,059
10	20	1	0,612**	0,046	0,366**	0,046
		10	0,634**	0,043	0,367**	0,044
		20	0,613**	0,047	0,360**	0,047

Fonte: Girardi, Filho e Storck (2009).

* Taxas de erro tipo I que ficaram abaixo do limite inferior do IC para proporções usando a aproximação normal, com 99% de confiança.

** Taxas de erro tipo I por experimento que ficaram acima do limite superior do IC para proporções usando a aproximação normal, com 99% de confiança.

As taxas de erro tipo I dos testes t de Student e Duncan foram sempre superiores ao nível nominal de significância estabelecido de 5%, evidenciando que esses procedimentos não controlam a TPE. Já os testes de Tukey e SNK não apresentaram nenhum valor de TPE diferente do nível nominal de significância adotado, independentemente da variação do número de tratamentos, repetições e CV.

Um estudo em relação às taxas de erro tipo I do teste de Scott-Knott (1974) por meio do método de Monte Carlo foi realizado por Silva, Ferreira e Bearzoti (1999), como pode ser observado na Tabela 4, em que são apresentadas as taxas de erro tipo I por experimento e por comparação obtidas com nível nominal de significância de 5%.

Tabela 4 – Taxas de erro tipo I por experimento (TPE) para o teste de Scott-Knott, em função do número de repetições r, número de tratamentos k, coeficientes de variação (CV) e nível nominal de significância de $0{,}05$.

k	r	CV	TPE
5	4	1	0,0660*
		10	0,0505
		20	0,0615
		30	0,0540
	10	1	0,0590
		10	0,0665*
		20	0,0630
		30	0,0635
	20	1	0,0625
		10	0,0585
		20	0,0665*
		30	0,0640*
10	4	1	0,0420
		10	0,0425
		20	0,0435
		30	0,0490
	10	1	0,0525
		10	0,0615
		20	0,0570
		30	0,0545
	20	1	0,0505
		10	0,0510
		20	0,0565
		30	0,0515
80	4	1	0,0355**
		10	0,0355**
		20	0,0345**
		30	0,0370**
	10	1	0,0460
		10	0,0515
		20	0,0490
		30	0,0355**
	20	1	0,0470
		10	0,0500
		20	0,0495
		30	0,0550

Fonte: Silva, Ferreira e Bearzoti (1999).

* Ultrapassou o limite superior do IC exato, com 99% de confiança para o nível nominal de significância de 0,05 (0,06391386).

** Ultrapassou o limite inferior do IC exato, com 99% de confiança para o nível nominal de significância de 0,05 (0,03828164).

Com esses resultados, Silva, Ferreira e Bearzoti (1999) concluíram que o teste de Scott-

Knott controlou o erro tipo I. Ao se comparar tais resultados com os obtidos neste trabalho para os testes SNK e SNK$_B$ e os de Girardi, Cargnelutti Filho e Storck (2009), pode-se dizer que o teste de Scott-Knott controla a taxa de erro tipo I por experimento para a maior parte dos casos simulados, embora não tanto quanto os testes de SNK e SNK$_B$. Porém, também não atinge níveis elevados como os procedimentos de Duncan e t de Student.

O comportamento dos testes aqui analisados, SNK e SNK$_B$, foram semelhantes, logo, o teste SNK original seria mais indicado devido ao menor esforço computacional exigido. Porém, outras situações devem ser consideradas.

4.1.2 Distribuições não normais

Na Tabela 5 são apresentadas as TPE sob H_0 completa considerando-se a distribuição lognormal com parâmetros de posição e escala iguais a 0 e 1, respectivamente, na escala logarítmica. Foram considerados diferentes números de tratamentos k , números de repetições r e nível de significância de $0{,}05$.

Tabela 5 – Taxas de erro por experimento (TPE) dos testes SNK original (SNK) e SNK *bootstrap* (SNK$_B$) sob H_0 completa, considerando-se a distribuição lognormal $(0,\ 1)$ e nível de significância $0{,}05$, em função do número de tratamentos k e número de repetições r.

k	r	SNK	SNK$_B$
5	4	0,041	0,043
	10	0,035^{++}	0,032^{++}
	20	0,040	0,041
10	4	0,040	0,039
	10	0,041	0,042
	20	0,044	0,039
20	4	0,055	0,055
	10	0,068**	0,074**
	20	0,062	0,065**
80	4	0,282**	0,284**
	10	0,230**	0,232**
	20	0,176**	0,177**

Fonte: Da autora.
$^{++}$ Significativamente diferente (valor-$p < 0{,}01$) e considerado menor do que o nível de significância $\alpha = 0{,}05$.
** Significativamente diferente (valor-$p < 0{,}01$) e considerado maior do que o nível de significância $\alpha = 0{,}05$.

As taxas de erro por experimento do teste SNK original e SNK *bootstrap* com $k = 5$ e

$r = 10$ são menores do que o nível de significância adotado e, para $k = 20$, $r = 10$ e $k = 80$ e todos os valores de r são maiores do que o nível de significância adotado. Com 20 repetições e 20 tratamentos, o teste SNK$_B$ apresentou taxa de erro tipo I maior que o nível significância, sendo considerado liberal. Este resultado confirma o que Borges e Ferreira (2003) obtiveram considerando a distribuição lognormal, em que os testes de Tukey e SNK tenderam a apresentar taxas de erro por experimento bastante altas em situações de maior número de tratamentos.

Na Tabela 6 são apresentadas as TPE sob H_0 completa, considerando a distribuição exponencial com parâmetro $\lambda = 0{,}1$. Foram considerados diferentes números de tratamentos k e número de repetições r.

Tabela 6 – Taxas de erro por experimento (TPE) dos testes SNK original (SNK) e SNK *bootstrap* (SNK$_B$) sob H_0 completa, considerando-se a distribuição exponencial $(0{,}1)$ e nível de significância $0{,}05$, em função do número de tratamentos k e número de repetições r.

k	r	SNK	SNK$_B$
5	4	0,035^{++}	0,035^{++}
	10	0,041	0,040
	20	0,045	0,048
10	4	0,046	0,047
	10	0,040	0,038
	20	0,055	0,054
20	4	0,058	0,056
	10	0,045	0,047
	20	0,046	0,043
80	4	0,071**	0,071**
	10	0,052	0,054
	20	0,062	0,064**

Fonte: Da autora.
$^{++}$ Significativamente diferente (valor-$p < 0{,}01$) e considerado menor do que o nível de significância $\alpha = 0{,}05$.
** Significativamente diferente (valor-$p < 0{,}01$) e considerado maior do que o nível de significância $\alpha = 0{,}05$.

O teste SNK se mostrou conservador para 5 tratamentos e 4 repetições, apresentando taxa de erro tipo I por experimento significativamente inferior ao valor nominal de 0,05 e, para 80 tratamentos, o teste pode ser considerado liberal quando o número de repetições é igual a 4, apresentando taxa de erro tipo I por experimento significativamente maior que o valor nominal de $0{,}05$. Nas demais situações, as taxas de erro tipo I por experimento não foram significativamente diferentes dos valores nominais de significância.

O teste de SNK$_B$, como o SNK, apresentou uma certa robustez no controle da taxa de

erro tipo I por experimento sob H_0 completa, considerando-se a distribuição exponencial, em praticamente todos os casos analisados para 5, 10 e 20 tratamentos, sendo conservador apenas para $k = 5$ e $r = 4$ e liberal para $k = 80$, $r = 4$ e $r = 20$, com valores não muito distantes de $0{,}05$. De maneira geral, o desempenho do teste SNK *bootstrap* foi semelhante ao do SNK original nas situações de não normalidade aqui consideradas e sob H_0 completa.

Ramos e Ferreira (2009) também avaliaram o poder e as taxas de erro tipo I para comparar as versões original e *bootstrap*, proposta pelos autores, para um dos procedimentos de comparações múltiplas de Caliński e Corsten (1985) por meio de simulação Monte Carlo, considerando modelos probabilísticos normais e não normais. Para fins de comparação, na Tabela 7 são apresentadas essas taxas de erro tipo I por experimento, em função do número de repetições e tratamentos dos testes de Caliński e Corsten e sua versão *bootstrap* (CB) para H_0 completa.

Tabela 7 – Taxas de erro por experimento (TPE) dos testes de Caliński e Corsten (C) e sua versão *bootstrap* (CB), em função do número de repetições, número de tratamentos e nível nominal de significância de $0{,}05$ sob H_0 completa, considerando-se as distribuições normal $(10, 1)$, exponencial $(0{,}1)$ e lognormal $(0, 1)$.

r	k	normal		exponencial		lognormal	
		C	CB	C	CB	C	CB
4	5	0,049	0,053	$0{,}033^{++}$	0,040	$0{,}029^{++}$	0,045
	10	0,055	0,057	0,038	0,048	0,039	0,048
	20	0,047	0,050	0,050	0,048	0,067	0,054
	80	0,044	0,051	$0{,}076^{**}$	0,050	$0{,}307^{**}$	0,052
10	5	0,047	0,046	0,049	0,051	0,026	0,041
	10	0,042	0,042	0,046	0,053	0,036	0,041
	20	0,048	0,054	0,044	0,045	0,066	0,053
	80	0,048	0,050	0,070	0,051	$0{,}236^{**}$	0,058
20	5	0,057	0,058	0,042	0,058	0,043	0,053
	10	0,044	0,046	0,043	0,047	0,053	0,057
	20	0,054	0,054	0,061	0,059	0,070	0,055
	80	0,041	0,042	0,069	0,066	$0{,}208^{**}$	0,062

Fonte: Ramos e Ferreira (2009).
$^{++}$ Não atingiu o LI do IC exato, com 99% de confiança para $\alpha = 0{,}05$ $(0{,}033927)$.
** Não atingiu o LS do IC exato, com 99% de confiança para $\alpha = 0{,}05$ $(0{,}070504)$.

Em todas as situações considerando normalidade, ambos os testes apresentaram taxas não significativamente diferentes das do nível nominal correspondente sendo, portanto, exatos nessas situações. Para a distribuição exponencial, o teste C foi conservador para 5 tratamentos e 4 repetições e liberal para 80 tratamentos, associado a um pequeno número de repetições, já

o teste CB apresentou taxas não significativamente diferentes do nível nominal correspondente de 5%. Para a distribuição lognormal, o teste CB controlou a taxa de erro tipo I por experimento em todos os cenários e o teste C foi conservador para $k = 5$ e $r = 4$ e foi liberal para $k = 80$ e todas as repetições.

Com esses resultados, conclui-se que não houve grande diferença nas taxas de erro tipo I por experimento entre os testes SNK e SNK$_B$, enquanto que o teste CB mostrou melhor controle do erro tipo I por experimento que o teste C nos cenários analisados.

4.2 Erro tipo I sob H_0 parcial

Na Tabela 8 são apresentadas as taxas de erro tipo I por experimento em função do número de tratamentos k, números de repetições r e δ erros padrões de diferença entre as médias. As taxas de erro tipo I foram mensuradas nas comparações entre médias de um mesmo grupo, portanto, as diferenças de δ erros padrão são apenas consideradas para médias entre os grupos. Foram realizados testes binomiais exatos considerando um nível nominal de significância de $0{,}01$ para cada taxa de erro tipo I observada.

Tabela 8 – Taxas de erro tipo I dos testes SNK e SNK$_B$, em função do número de tratamentos k, número de repetições r, diferenças entre as médias δ, para o nível nominal de significância $\alpha = 0{,}05$, sob a distribuição normal e H_0 parcial.

(continua)

k	r	δ	SNK	SNK$_B$
5	4	2	0,067	0,351**
		4	0,091**	0,438**
		8	0,111**	0,433**
		16	0,116**	0,433**
	10	2	0,059	0,379**
		4	0,113**	0,479**
		8	0,088**	0,429**
		16	0,104**	0,432**
	20	2	0,047	0,373**
		4	0,105**	0,460**
		8	0,109**	0,456**
		16	0,101**	0,463**
10	4	2	0,039	0,749**
		4	0,082**	0,895**
		8	0,101**	0,878**
		16	0,101**	0,870**
	10	2	0,031++	0,779**
		4	0,118**	0,893**
		8	0,100**	0,899**
		16	0,100**	0,904**
	20	2	0,031++	0,809**
		4	0,106**	0,908**
		8	0,114**	0,910**
		16	0,108**	0,915**
20	4	2	0,043	0,969**
		4	0,093**	0,997**
		8	0,100**	0,995**
		16	0,103**	0,999**
	10	2	0,048	0,982**
		4	0,105**	1,000**
		8	0,100**	1,000**
		16	0,086**	0,997**
	20	2	0,049	0,974**
		4	0,095**	0,999**
		8	0,089**	0,997**
		16	0,095**	0,998**

Tabela 8 - Taxas de erro tipo I dos testes SNK e SNK$_B$, em função do número de tratamentos k, número de repetições r, diferenças entre as médias δ, para o nível nominal de significância $\alpha = 0{,}05$, sob a distribuição normal e H_0 parcial.

(continuação)

k	r	δ	SNK	SNK$_B$
80	4	2	0,001^{++}	1,000**
		4	0,026^{++}	1,000**
		8	0,101**	1,000**
		16	0,090**	1,000**
	10	2	0,001^{++}	1,000**
		4	0,086**	1,000**
		8	0,089**	1,000**
		16	0,117**	1,000**
	20	2	0,035	1,000**
		4	0,089**	1,000**
		8	0,090**	1,000**
		16	0,100**	1,000**

$^{++}$ Significativamente diferente (valor-$p < 0{,}01$) e considerado menor do que o nível de significância $\alpha = 0{,}05$.
** Significativamente diferente (valor-$p < 0{,}01$) e considerado maior do que o nível de significância $\alpha = 0{,}05$.

O teste SNK$_B$ apresentou altas taxas de erro tipo I por experimento, sendo considerado liberal em todas os cenários avaliados e essa taxa aumenta com o aumento do número de tratamentos k. Para o teste SNK original pode-se observar, fixando-se $\delta = 2$, $k = 5$ e $k = 20$, que as TPE não são significativamente diferentes do nível nominal de significância, para $k = 10$, $r = 4$ e $\delta = 2$ o teste também foi considerado exato, assim como para $k = 80$, $r = 20$ e $\delta = 2$, nos demais cenários em que $\delta = 2$ o teste SNK foi conservador. Para valores de δ diferentes de 2, as taxas de erro tipo I foram maiores que α, sendo então o teste SNK um teste liberal nesses cenários.

Silva, Ferreira e Bearzoti (1999) avaliaram a TPE do teste Scott-Knott (SK) sob H_0 parcial, conforme apresenta a Tabela 9 considerando o nível nominal de significância de $0{,}05$.

Tabela 9 – Taxas de erro tipo I do teste Scott-Knott para o nível nominal de significância de $0{,}05$, em função do número de repetições r e do número de tratamentos k sob H_0 parcial e distribuição normal.

r	k	TPE
4	5	0,0620
	10	0,0715*
	20	0,0790*
	10	0,0800*
	96	0,0820*
10	5	0,0665*
	10	0,0795*
	20	0,0815*
	40	0,0845*
	96	0,0830*
20	5	0,0725*
	10	0,0725*
	20	0,0735*
	40	0,0705*
	96	0,0730*

Fonte: Silva, Ferreira e Bearzoti (1999).
* Ultrapassou o limite superior do IC exato, com 99% de confiança para o nível nominal de significância de 0,05 (0,06391386).
** Ultrapassou o limite inferior do IC exato, com 99% de confiança para o nível nominal de significância de 0,05 (0,03828164).

O teste SK apresentou taxas de erro tipo I por experimento maiores que o nível nominal de significância, apesar de apresentar taxas menores que os teste SNK_B o teste SK também foi considerado liberal na maioria dos cenários avaliados.

Os testes C e CB, avaliados por Ramos e Ferreira (2009), foram considerados liberais para valores de δ iguais a 2 e 4 e para δ igual a 8 e 32, com $r = 4$ e $k = 80$, nos demais cenários ambos os testes foram conservadores. Ramos e Vieira (2014) avaliaram o teste de Caliński e Corsten baseado na distribuição F (CF) e sua versão *bootstrap* (CFB) sob H_0 parcial. Na Tabela 10 estão apresentadas as TPE em função da diferença entre as médias, do número de repetições e do número de tratamentos dos testes C, CB, CF e CFB.

Tabela 10 – Taxas de erro tipo I dos testes CF, CFB, C e CB, em função do número de tratamentos k, número de repetições r, diferenças entre as médias δ, para o nível nominal de significância $\alpha = 0{,}05$, sob a distribuição normal e H_0 parcial.

(continua)

k	r	δ	CF	CFB	C	CB
5	4	2	0,139**	0,216**	0,164**	0,169**
		4	0,126**	0,469**	0,156**	0,157**
		8	0,026++	0,504**	0,016++	0,016++
		16	0,021**	0,492**	0,030++	0,031++
	10	2	0,188**	0,279**	0,204**	0,199**
		4	0,147**	0,515**	0,158**	0,160**
		8	0,027++	0,543**	0,027++	0,026++
		16	0,026++	0,568**	0,025++	0,024++
	20	2	0,201**	0,270**	0,175**	0,178**
		4	0,147**	0,515**	0,168**	0,168**
		8	0,019++	0,547**	0,022++	0,021++
		16	0,023++	0,556**	0,022++	0,023++
10	4	2	0,0,277**	0,332**	0,307**	0,307**
		4	0,211**	0,677**	0,302**	0,308**
		8	0,036	0,650**	0,028++	0,032++
		16	0,040	0,707**	0,031++	0,040
	10	2	0,365++	0,438**	0,358**	0,364**
		4	0,239**	0,721**	0,311**	0,311**
		8	0,038**	0,734**	0,022++	0,025++
		16	0,030++	0,736**	0,029++	0,031++
	20	2	0,392++	0,474**	0,358**	0,366**
		4	0,257**	0,735**	0,309**	0,310**
		8	0,027**	0,751**	0,030++	0,032++
		16	0,034**	0,769**	0,024++	0,026++
20	4	2	0,553**	0,629**	0,407**	0,419**
		4	0,405**	0,894**	0,547**	0,549**
		8	0,049	0,898**	0,023++	0,057
		16	0,030++	0,912**	0,032++	0,070
	10	2	0,591**	0,678**	0,474**	0,475**
		4	0,368**	0,912**	0,521**	0,520**
		8	0,037	0,909**	0,037	0,040
		16	0,041	0,931**	0,034	0,045
	20	2	0,635**	0,707**	0,478**	0,478**
		4	0,377**	0,914**	0,500**	0,502**
		8	0,024++	0,924**	0,038	0,036
		16	0,038	0,914**	0,018++	0,021++

Tabela 10 - Taxas de erro tipo I dos testes CF, CFB, C e CB, em função do número de tratamentos k, número de repetições r, diferenças entre as médias δ, para o nível nominal de significância $\alpha = 0{,}05$, sob a distribuição normal e H_0 parcial.

(continuação)

k	r	δ	CF	CFB	C	CB
80	4	2	0,912**	0,986	0,634	0,666
		4	0,736**	1,000**	0,914**	0,914**
		8	0,050	1,000**	0,037	0,156**
		16	0,044	1,000**	0,029++	0,335**
	10	2	0,921**	0,984**	0,683**	0,685**
		4	0,708**	1,000**	0,914**	0,914**
		8	0,053	1,000**	0,040	0,048
		16	0,055	1,000**	0,028++	0,057
	20	2	0,920**	0,995**	-	-
		4	0,744**	1,000**	-	-
		8	0,048**	1,000**	-	-
		16	0,044**	1,000**	-	-

Fonte: Ramos e Vieira (2014).
++ Significativamente diferente (valor-$p < 0{,}01$) e considerado menor do que o nível de significância $\alpha = 0{,}05$.
** Significativamente diferente (valor-$p < 0{,}01$) e considerado maior do que o nível de significância $\alpha = 0{,}05$.

O teste CFB apresentou um comportamento semelhante ao teste SNK$_B$, apresentando altas taxas de erro tipo I por experimento e o CF foi liberal para menores valores de δ enquanto que o SNK foi liberal para maiores valores de δ. O teste CF é conservador em alguns cenários, assim como o teste SNK porém em cenários diferentes, e o teste CFB, como o SNK$_B$, é liberal em todos os cenários analisados. Já os testes C e CB são liberais para valores de δ pequenos ($\delta \leq 4$) e conservadores em alguns cenários em que $\delta \geq 8$. O mesmo efeito de aumento da taxa de erro tipo I por experimento, observado no teste SNK$_B$, foi observado para os testes C, CB, CF, CFB e SK. Com o aumento do número de tratamentos há um aumento das TPE.

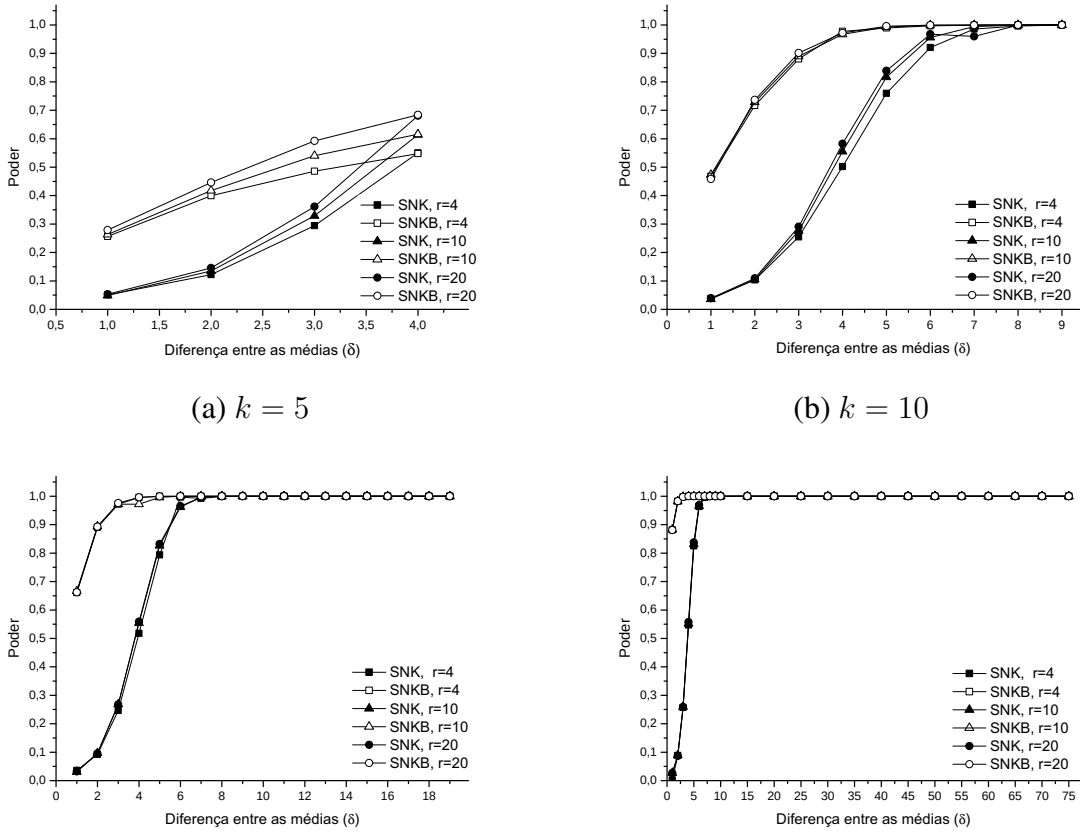
Com H_0 parcial e normalidade, nenhum dos testes apresentou controle do erro tipo I por experimento. O trabalho de Girardi, Cargnelutti Filho e Storck (2009), que avaliou os testes t de Student, Tukey, Duncan não considerou situações de nulidade parcial, logo não foi possível comparar as TPE sob H_0 parcial desses testes com as do teste proposto.

4.3 Poder sob H_1

Neste tópico são apresentados os resultados do poder para as distribuições normal, lognormal e exponencial.

4.3.1 Distribuição normal

Na Figura 5 são apresentados os gráficos de poder dos testes SNK e SNK$_B$, em função da diferença δ em erros padrão entre médias, para diferentes números de repetições (r) e tratamentos (k) , considerando a distribuição normal sob H_1 e nível de significância de 0,05.



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

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

Figura 5 – Poder dos testes de Student-Newman-Keuls (SNK) e sua versão *bootstrap* (SNK$_B$), em função das diferenças entre médias δ, diferentes números de repetições r e diferentes números de tratamentos k, considerando-se a distribuição normal (10,1) sob H_1 e $\alpha = 0{,}05$.

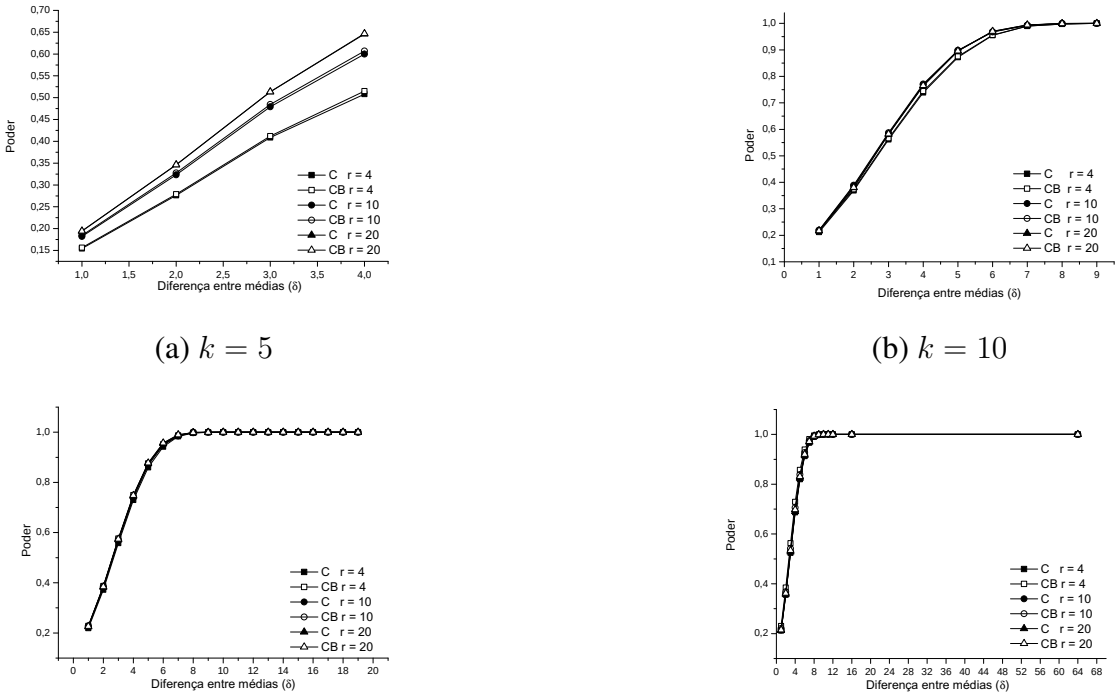
Fonte: Da autora.

Pode-se observar, comparando-se as Figuras 5a, 5b, 5c e 5d, que há um aumento no valor do poder à medida que o valor de k aumenta. Ao se considerar o efeito do número de repetições nos valores do poder dos testes, observou-se que o efeito de repetição só é evidente no caso de $k = 5$ e, para os demais casos, praticamente não houve diferenças entre esses valores. Como já preconizado, o poder de um teste aumenta com o aumento de δ. Essa mesma tendência, porém

em diferentes magnitudes, pode ser verificada em diversos trabalhos que avaliaram vários testes já apresentados, como Ramos e Ferreira (2009), Girardi, Cargnelutti Filho e Storck (2009), Conagin, Barbin e Demetrio (2008) e Silva, Ferreira e Bearzoti (1999). Quando as diferenças entre as médias são menores $(\delta \leq 4)$, o poder do teste SNK$_B$ é maior do que o do teste SNK que, por sua vez, apresenta valores de poder abaixo de 0,05 quando $\delta = 1$ para todos os valores de k.

Ao se comparar tais resultados com os obtidos por Ramos e Ferreira (2009), pode-se verificar o mesmo padrão de comportamento do poder do teste de Caliński e Corsten (C) e sua versão *bootstrap* (CB), ocorrendo um aumento do poder à medida que aumentam os valores de r, k e δ. Porém, em quase todos os casos, praticamente não houve diferenças entre os valores de poder dos testes C e CB, enquanto que, nos testes SNK e SNK$_B$, essa diferença é evidente quando os valores de δ são pequenos.

Na Figura 6 estão apresentados os gráficos para poder dos testes C e CB em função do número de repetições, do número de tratamentos e da diferença entre médias, considerando a distribuição normal e sob H_1, apresentados por Ramos e Ferreira (2009).



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

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

Figura 6 – Poder dos testes de Caliński e Corsten (C) e sua versão *bootstrap* (CB) em função da diferença entre médias do número de repetições e do número de tratamentos, considerando a distribuição normal e H_1.
Fonte: Ramos e Ferreira (2009).

Silva, Ferreira e Bearzoti (1999) avaliaram os valores do poder do teste de Scoott-Knott (1974) em função do número de repetições, do número de tratamentos e erros padrão da média, esses valores estão apresentados na Tabela 11.

Tabela 11 – Poder do teste de Scott-Knott, ao nível nominal de significância de $0{,}05$, em função do número de repetições r, número de tratamentos k e do erro padrão da média (δ).

r	k	Diferença real entre médias (δ)				
		2	4	6	8	10
	5	0,3945	0,8142	0,9578	-	-
	10	0,4434	0,8236	0,9767	0,9993	0,9999
4	20	0,4639	0,8346	0,9824	0,9998	0,1000
	40	0,4720	0,8361	0,9827	0,9997	0,1000
	96	0,4845	0,8429	0,9835	0,9998	0,1000
	5	0,4054	0,8402	0,9840	-	-
	10	0,4503	0,8379	0,9829	0,9996	0,1000
10	20	0,4651	0,8461	0,9844	0,9998	0,1000
	40	0,4740	0,8492	0,9840	0,9997	0,1000
	96	0,4856	0,8442	0,9838	0,9998	0,1000
	5	0,4124	0,8467	0,9878	-	-
	10	0,4498	0,8360	0,9848	0,9998	0,1000
20	20	0,4645	0,8386	0,9842	0,9998	0,1000
	40	0,4743	0,8382	0,9842	0,9998	0,1000
	96	0,4862	0,8446	0,9841	0,9997	0,1000

Fonte: Silva, Ferreira e Bearzoti (1999).

Os autores observaram que o poder do teste tendeu a aumentar com o aumento do número de tratamentos de 5 para 10 e do número de repetições de 4 para 10, e que essa tendência foi mais clara quando a diferença real entre médias foi de 2 erros padrão.

Comparando o teste de Scott-Knott avaliado por Silva, Ferreira e Bearzoti (1999) com os testes analisados no presente trabalho pode-se verificar que o poder do teste SK é maior do que do teste SNK e menor que o do teste SNK$_B$ nos diferenças entre as médias analisadas.

Girardi, Cargnelutti Filho e Storck (2009) avaliaram o poder dos testes t de Student, Tukey, Duncan e SNK sob H_1. Os valores do poder desses testes são apresentados na Tabela 12. Pode-se observar que os testes t e Duncan apresentaram maior poder em relação ao Tukey, já o teste SNK apresentou valores intermediários.

Tabela 12 – Poder para diferentes testes de comparações múltiplas de médias em função do número de tratamentos k e do erro padrão da média de um tratamento $\sigma_{\bar{Y}}$, com 20 repetições e coeficiente de variação de 10%.

k	Diferença real entre médias	t	Tukey	Duncan	SNK
4	1	0,3134	0,1356	0,2784	0,1860
	2	0,6499	0,4747	0,6887	0,6029
	4	0,9210	0,8027	0,9209	0,9208
	8	0,9999	0,9996	0,9999	0,9999
10	1	0,5914	0,3518	0,5585	0,4544
	2	0,8210	0,6694	0,8149	0,7932
	4	0,9610	0,8687	0,9610	0,9609
	8	0,9999	0,9988	0,9999	0,9999
50	1	0,9100	0,8028	0,9013	-0,8732
	2	0,9636	0,9082	0,9623	0,9576
	4	0,9923	0,9628	0,9923	0,9923
	8	0,9999	0,9975	0,9999	0,9999

Fonte: Girardi, Cargnelutti Filho e Storck (2009).

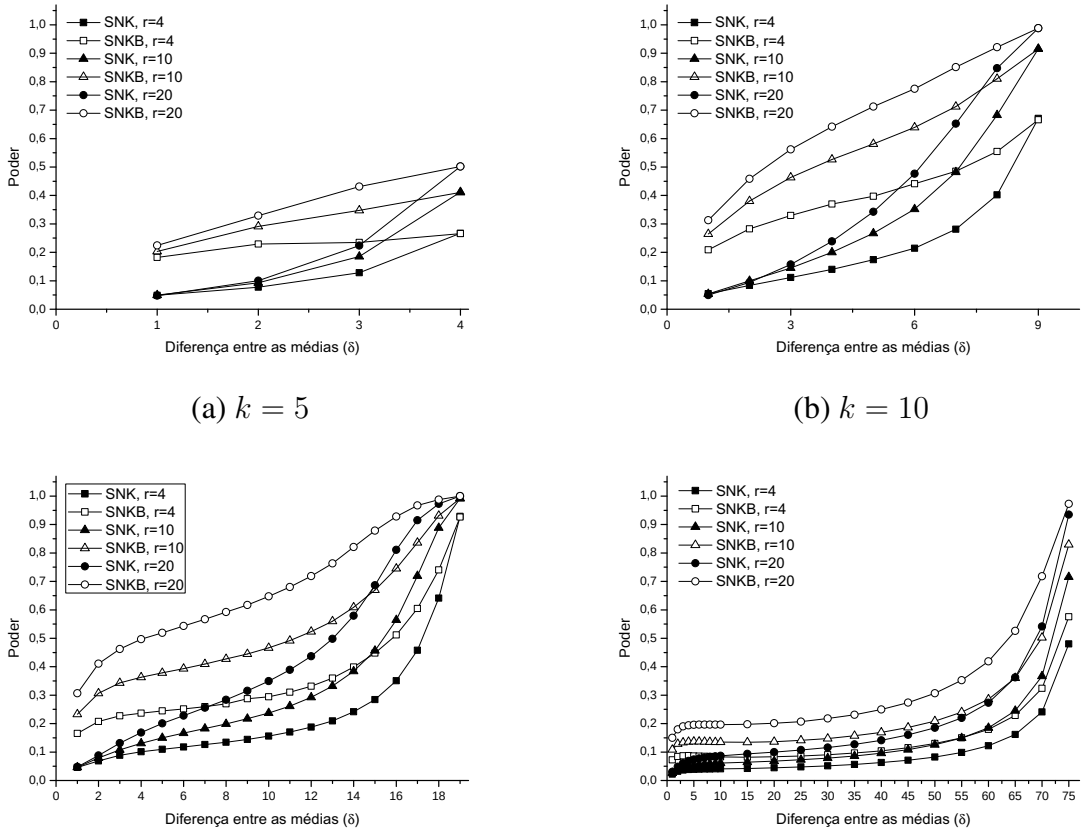
Observa-se que, em todos os testes analisados, quanto maior a magnitude da diferença entre médias consecutivas, a porcentagem de decisões corretas aumenta rapidamente. O teste SNK$_B$ apresentou maior poder que os demais testes apresentados.

4.3.2 Distribuições não normais

Na Figura 7 estão apresentados os gráficos de poder dos testes SNK e SNK$_B$, em função da diferença δ em erros padrão entre médias, com diferentes números de repetições (r) e tratamentos (k) , considerando a distribuição lognormal, com parâmetros de posição e escala iguais a 0 e 1, sob H_1 e nível de significância de 0,05.

Ambos os testes apresentaram comportamento do poder diferente do apresentado sob distribuição normal. O número de repetições (r) afetou o poder dos testes indicando que, quanto maior o r, maior o poder. Além disso, para cada valor de r, o teste SNK$_B$ apresentou maiores valores de poder do que o SNK. Ramos e Ferreira (2009) também observaram que houve efeito do número de repetições para todos os valores de k. Segundo os autores, isso se deve à heterogeneidade de variâncias. Como as médias e a variância não são independentes nas distribuições não normais, ao se determinar os parâmetros da distribuição para gerar outro tratamento, essa são alteradas, portanto, quanto maior o número de repetições, menor a heterogeneidade de

variâncias e maior o poder.



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

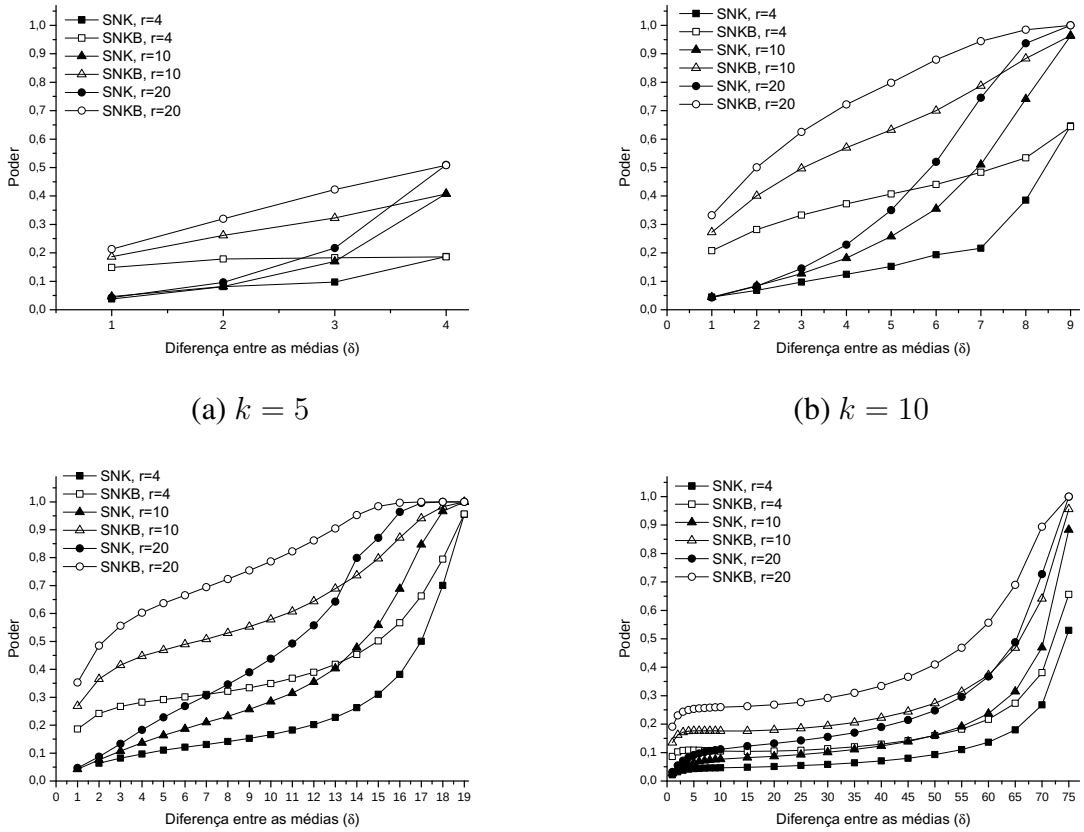
(c) $k = 20$ (d) $k = 80$

Figura 7 – Poder dos testes de Student-Newman-Keuls (SNK) e sua versão *bootstrap* (SNK$_B$), em função das diferenças entre médias δ e diferentes números de tratamentos k, considerando-se a distribuição lognormal (0,1), sob H_1 e $\alpha = 0{,}05$.

Fonte: Da autora.

O poder do teste SNK$_B$, como na distribuição normal, é maior do que o poder do teste SNK e, à medida que a diferença entre as médias aumenta, o poder de ambos os testes tendem a ser iguais. Como era de se esperar, à medida que a diferença entre médias consecutivas aumenta, a porcentagem de decisões corretas cresce rapidamente para valores de k maiores que 5. Para k = 5, o poder aumenta de forma mais lenta. Para valores de k iguais a 10 e 20, os valores do poder são mais altos mesmo para valores de δ pequenos.

Na Figura 8 estão apresentados os gráficos de poder dos testes SNK e SNK$_B$, em função da diferença em erros padrão entre as médias (δ), com diferentes números de repetições (r) e tratamentos (k), considerando a distribuição exponencial, com parâmetro $\lambda = 0{,}1$, sob H_1 e nível de significância de 0,05.



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

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

Figura 8 – Poder dos testes de Student-Newman-Keuls (SNK) e sua versão *bootstrap* (SNK$_B$), em função das diferenças entre médias δ, diferentes números de repetições r e para diferentes números de tratamentos k, considerando-se a distribuição exponencial, sob H_1 e $\alpha = 0{,}05$.

Fonte: Da autora.

Observa-se na Figura 8 que os resultados obtidos foram semelhantes aos resultados obtidos na distribuição lognormal, porém, de uma forma geral, os valores de poder na exponencial foram um pouco mais altos. O aumento do poder dos testes também foi bastante influenciado pelo aumento no número de repetições e, para o teste SNK original, isso acontece principalmente quando a diferença entre médias passa a ser maior do que 2.

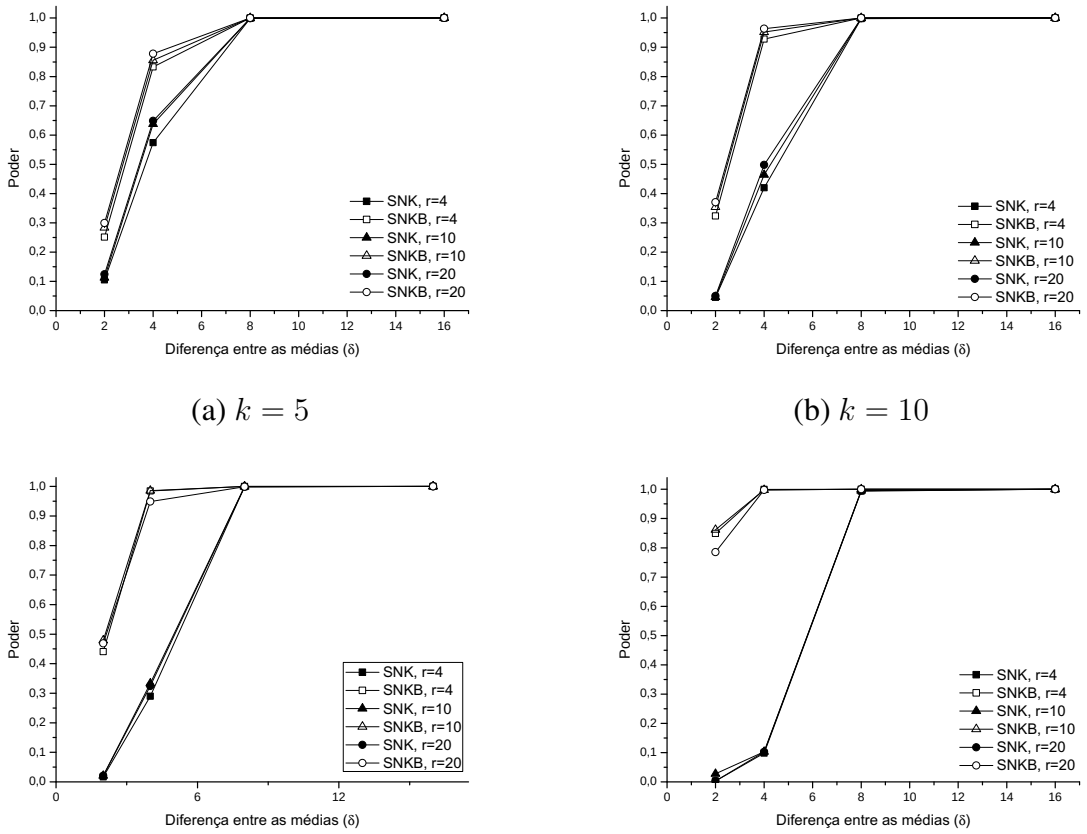
Como era de se esperar, à medida que a diferença entre médias aumenta, o poder dos testes também aumenta. O poder do teste SNK$_B$ é superior ao poder do teste SNK original em todas as situações analisadas, porém, a taxa de aumento do poder no teste SNK é maior do que no teste SNK$_B$.

O comportamento sob H_1 foi o mesmo para as duas distribuições não normais porém, o poder de ambos os testes apresentaram valores menores para a distribuição lognormal para todos os valores de δ analisados. Para as duas distribuições verifica-se a mesma tendência observada

na distribuição normal considerando a diferença entre as médias e número de tratamentos.

4.4 Poder sob H_0 parcial

Na Figura 9 estão apresentados os valores de poder dos testes SNK e SNK$_B$ em função da diferença de δ erros padrão, número de repetições r e número de tratamentos k, sob normalidade e H_0 parcial e nível de significância de 0,05.



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

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

Figura 9 – Poder dos testes de Student-Newman-Keuls (SNK) e sua versão *bootstrap* (SNK$_B$), em função das diferenças entre médias δ, diferentes números de repetições r e para diferentes números de tratamentos k, considerando-se a distribuição normal, sob H_0 parcial e $\alpha = 0{,}05$.

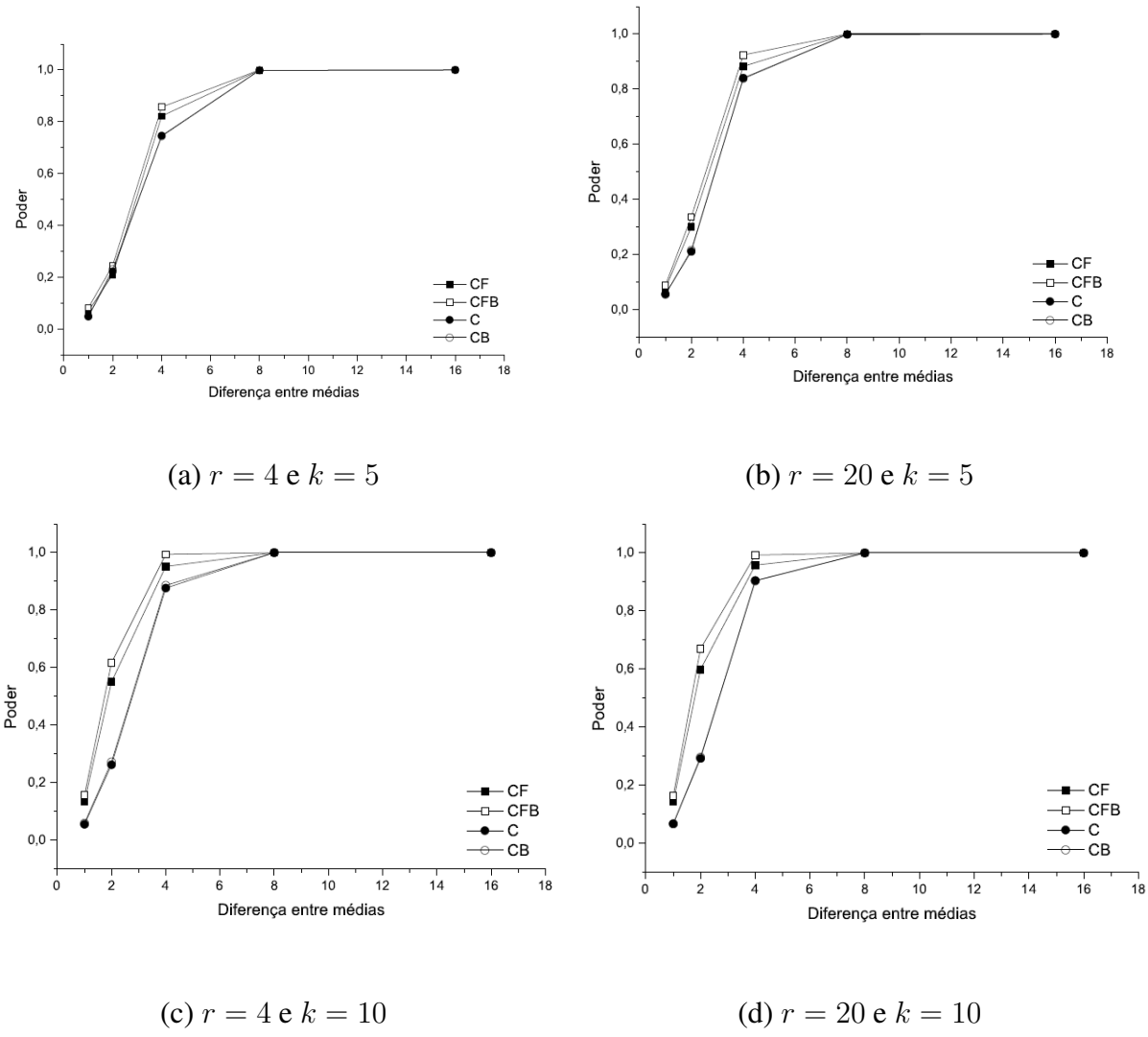
Fonte: Da autora.

Para valores de δ menores do que 4, os valores de poder do teste SNK foram menores do que os do SNK$_B$. Quando $\delta \geq 8$, os valores de poder se igualam. Ou seja, na situação de H_0 parcial, se as diferenças entre os grupos forem pequenas, o teste SNK$_B$ discrimina melhor

as diferenças do que o teste SNK.

Observa-se um aumento de poder à medida que a diferença δ entre as médias se torna maior. O mesmo comportamento foi observado por Ramos e Ferreira (2009) e Ramos e Vieira (2014).

Nas Figuras 10 e 11 estão apresentados os valores do poder dos testes C, CB, CF e CFB avaliados por Ramos e Vieira (2014) em situações similares às consideradas neste trabalho sob H_0 parcial em função da diferença de erros padrão δ entre as médias, número de repetições r e do número de tratamentos k.



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

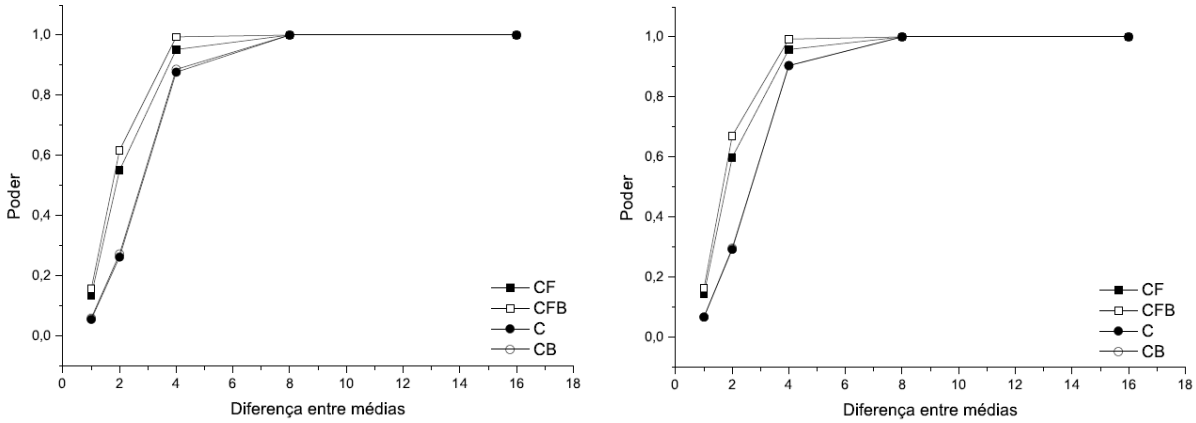
(b) $r = 20$ e $k = 5$

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

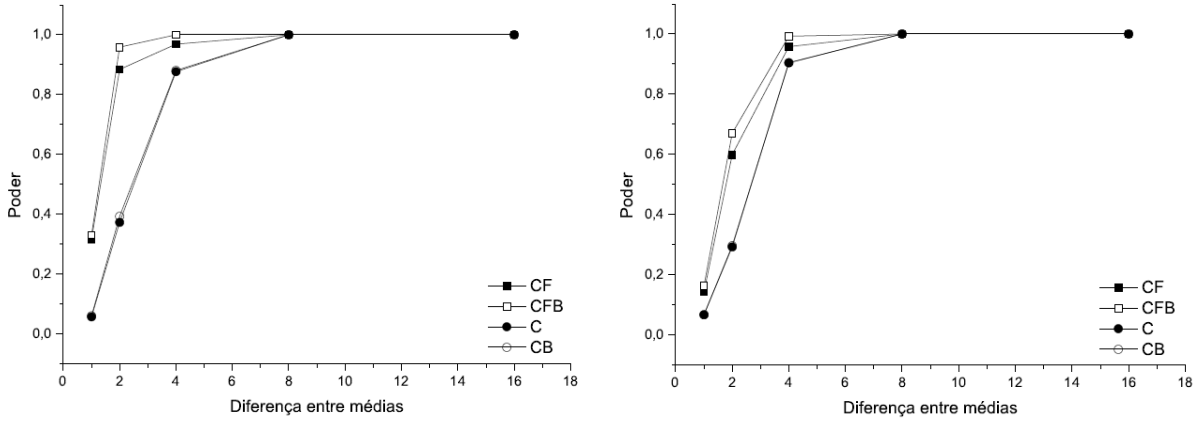
(d) $r = 20$ e $k = 10$

Figura 10 – Poder dos testes de C, CB, CF e CFB em função das diferenças entre médias δ, diferentes números de repetições r e para diferentes números de tratamentos k, considerando-se a distribuição normal, sob H_0 parcial, $\alpha = 0{,}05$, $k = 5$ e $k = 10$.

Fonte: Ramos e Vieira (2014).



(a) $r = 4$ e $k = 20$ (b) $r = 10$ e $k = 20$



(c) $r = 4$ e $k = 80$ (d) $r = 10$ e $k = 80$

Figura 11 – Poder dos testes de C, CB, CF e CFB em função das diferenças entre médias δ, diferentes números de repetições r e para diferentes números de tratamentos k, considerando-se a distribuição normal, sob H_0 parcial, $\alpha = 0{,}05$, $k = 20$ e $k = 80$.

Fonte: Ramos e Vieira (2014).

Como as taxas de erro tipo I foram mais elevadas para o teste SNK$_B$ do que para o teste SNK sob H_0 parcial, esperava-se que o poder fosse realmente maior, o que ocorreu. Para os testes C, CB, CF e CFB os valores do poder foram similares quando considerados os mesmos cenários e, assim como os teste SNK e SNK$_B$, o número de repetições não teve influência sobre o valor do poder.

O poder dos testes C, CB, CF e CFB, assim como os testes SNK e SNK$_B$, aumentou com o aumento de δ, principalmente quando $\delta \leq 4$. O teste SNK$_B$ foi o que apresentou maiores valores de poder, principalmente para valores de δ pequenos.

4.5 Aplicação

Para os dados do exemplo descrito na seção 3.4 foram obtidas as médias dos tratamentos, em que foi avaliada a variável peso de frutos de graviola em kg. As médias estão organizadas em ordem decrescente de magnitude:

$$\bar{Y}_{(1)} = 1{,}49$$
$$\bar{Y}_{(2)} = 1{,}40$$
$$\bar{Y}_{(3)} = 1{,}37$$
$$\bar{Y}_{(4)} = 1{,}37$$
$$\bar{Y}_{(5)} = 1{,}36$$
$$\bar{Y}_{(6)} = 1{,}27$$
$$\bar{Y}_{(7)} = 0{,}99$$
$$\bar{Y}_{(8)} = 0{,}94$$
$$\bar{Y}_{(9)} = 0{,}76$$

Para testar a hipótese de igualdade das médias dos tratamentos é necessário antes verificar se há homogeneidade de variâncias, testar se os resíduos têm distribuição normal e verificar se os resíduos são independentes.

Na Tabela 13 são apresentados os valores-p obtidos ao se aplicarem os testes SNK e SNK$_B$, apresentados nas seções 3.1 e 3.2, respectivamente, ao conjunto de dados. Para cada contraste entre as médias de tratamentos ordenados foi então calculado um valor-p.

Tabela 13 – Valores-p dos testes SNK e SNK$_B$.

Médias ordenadas	Valor-p (SNK)	Valor-p (SNK$_B$)
1 - 9	0,001	0,000
1 - 8	0,025	0,003
1 - 7	0,047	0,003
1 - 6	0,773	0,408
1 - 5	0,936	0,715
1 - 4	0,889	0,592
1 - 3	0,751	0,389
1 - 2	0,589	0,279
2 - 9	0,004	0,000
2 - 8	0,089	0,004
2 - 7	0,141	0,005
2 - 6	0,936	0,555
2 - 5	0,995	0,951
2 - 4	0,982	0,863
2 - 3	0,857	0,542
3 - 9	0,006	0,000
3 - 8	0,107	0,001
3 - 7	0,155	0,003
3 - 6	0,932	0,455
3 - 5	0,998	0,974
3 - 4	0,004	1,000
4 - 9	0,078	0,000
4 - 8	0,106	0,000
4 - 7	0,820	0,000
4 - 6	0,952	0,179
4 - 5	1,000	0,796
5 - 9	0,003	0,000
5 - 8	0,060	0,000
5 - 7	0,070	0,000
5 - 6	0,589	0,058
6 - 9	0,013	0,000
6 - 8	0,119	0,000
6 - 7	0,094	0,000
7 - 9	0,353	0,050
7 - 8	0,764	0,349
8 - 9	1,000	0,047

Fonte: Da autora.

Na Tabela 14 estão apresentados os resultados obtidos pelos testes SNK e SNK$_B$, prosseguindo com os passos descritos na seção 3.1 e 3.2.

Tabela 14 – Peso médio (kg) de frutos colhidos de graviola com 20 repetições e resultados obtidos pelos testes SNK e SNK$_B$.

Tratamento	Peso médio (kg)	SNK	SNK$_B$
1 - Saco de papel kraft	1,49	a	a
2 - Saco de papel impermeável	1,40	ab	a
3 - Saco plástico aberto	1,37	ab	a
4 - Triflumuron + Saco plástico perfurado	1,37	ab	a
5 - Saco plástico fechado	1,36	ab	a
6 - Saco plástico perfurado	1,27	ab	a
7 - Testemunha	0,99	bc	b
8 - Imidacloprid	1,94	bc	b
9 -Triflumuron	0,76	c	c

Fonte: Da autora.
*Médias seguidas de mesma letra na coluna não diferem entre si ao nível de 5% de significância de acordo com os testes SNK e SNK$_B$.

Pelo teste SNK$_B$, o tratamento Triflumuron diferiu estatisticamente dos demais e, pelo teste SNK, o mesmo tratamento não diferiu dos testes Imidacloprid e testemunha. Os tratamentos saco plástico perfurado, saco plástico fechado, Triflumuron + saco plástico perfurado, saco plástico aberto, saco de papel impermeável e de papel kraft não diferiram entre si pelos dois testes. De acordo com o teste SNK, a testemunha diferiu apenas dos tratamentos saco de papel kraft e Triflumuron e, pelo teste SNKB, a testemunha diferiu de todos os tratamentos, exceto o Imidacloprid. Observa-se pelos resultados obtidos que os sacos plásticos (comuns ou perfurados) foram as formas mais viáveis de controle de C. anonella e de B. pomorum.

Nessa aplicação, pode-se observar que o teste SNK$_B$ identificou mais diferenças significativas do que o teste SNK, o que pode ser notado pelo maior número de valores-p significativos da Tabela 13, e isso se deve provavelmente a sua característica mais liberal.

5 CONCLUSÕES

O teste de comparações múltiplas *bootstrap* foi proposto com sucesso. Seu desempenho foi considerado superior ao do teste SNK original sob H_0 completa e H_1 sob normalidade.

Os testes SNK e SNK$_B$ são exatos sob H_0 completa e normalidade. Sob H_0 completa e não normalidade, os testes SNK e SNK$_B$ controlam as taxas de erro tipo I por experimento e são considerados exatos na maior parte dos casos simulados para $k = 5$ e $k = 10$, enquanto que, para $k = 20$ e $k = 80$, ambos os testes em alguns cenários são considerados liberais.

Sob H_0 parcial, o teste SNK$_B$ foi liberal em todos os casos simulados, enquanto que o teste SNK foi, em geral, conservador para $\delta \leq 2$ e liberal para os demais valores de δ. O poder do teste SNK$_B$ é, na maioria das situações, superior ao do teste SNK sob H_1 e sob H_0 parcial.

Os dois testes apresentaram certa robustez.

Assim, em situações práticas, se as diferenças entre as médias dos tratamentos forem pequenas ($\delta \leq 2$), o teste SNK é mais indicado por controlar o erro tipo I e apresentar valores de poder satisfatórios. Nas demais situações, o teste SNK$_B$ é mais recomendado, apesar de ambos serem liberais para $\delta \geq 4$, se a situação for de H_0 parcial.

REFERÊNCIAS BIBLIOGRÁFICAS

BANZATTO, D. A.; KRONKA, S. N. **Experimentação Agrícola**. 11. ed. Jaboticabal: FUNEP, 1989. 247p.

BASTOS, R. L. **Proposição de testes bootstrap para o índice de qualidade sensorial**. Dissertação (Pós-Graduação em Estatística e experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2013.

BHERING, L. L. et al. Alternative methodology for scott-knott test. **Crop Breeding and Applied technology**, Brazilian Society of Plant Breeding, v. 8, n. 1, p. 9, 2008.

BIASE, N. G.; FERREIRA, D. F. Testes de igualdade e de comparações múltiplas para várias proporções binomiais independentes. **Revista Brasileira de Biometria**, v. 29, n. 4, p. 549–570, 2011.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001. 123p.

BORGES, L. C.; FERREIRA, D. F. Poder e taxas de erro tipo I dos testes scott-knott, tukey e student-newmankeuls sob distribuições normal e não normais dos resíduos. **Revista de Matemática e Estatística**, v. 21, p. 67–83, 2003.

BUSSAB, W.; MORETTIN, P. A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2004.

CALIŃSKI, T.; CORSTEN, L. C. A. Clustering means in anova by simultaneous testing. **Biometrics**, v. 41, n. 1, p. 39–48, 1985.

CARMER, S.; SWANSON, N. R. An evaluation of ten pairwise multiple comparison procedures by monte carlo methods. **Journal of the American Statistical Association**, v. 68, n. 341, p. 66–74, 1973.

CARPENTER, J.; BITHELL, J. *Bootstrap* confidence intervals: When, which, what? a practical guide for medical statistician. **Statistics in Medicine**, v. 19, p. 1141–1164, 2000.

CASELLA, G.; BERGER, R. L. **Inferência estatística**. 2. ed. São Paulo: Cengage Learning, 2010. 588p.

CIRILLO, M. A. Reamostragem e simulação: uma introdução. In: ENCONTRO MINEIRO DE ESTATÍSTICA.; II SEMANA DA MATEMÁTICA, 12., 2013. **Anais...** Uberlândia: UFU, 2013. p. 44.

COCHRAN, W. G.; COX, G. M. **Experimental Designs**. 2. ed. New York: John Wiley and Sons, 1992. 611p.

CONAGIN, A.; BARBIN, D.; DEMETRIO, C. G. B. Modifications for the tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. **Sci. Agric.**, v. 65, n. 4, p. 428–432, 2008.

DACHS, J. N. W. **Estatística computacional: uma introdução em turbo pascal**. Rio de Janeiro: LTC, 1988.

DUNCAN, D. B. Multiple range end multiple ftests. **Biometrics**, 1955.

FERREIRA, D. F. **Estatística Básica.** 2. ed. Lavras: Editora UFLA, 2009. 664 p.

FERREIRA, E. B.; CAVALCANTI, P. P.; NOGUEIRA, D. A. Experimental designs: um pacote R para análise de experimentos. **Revista da Estatística**, v. 1, p. 1–9, 2011.

FREUND, J. E. **Estatística Aplicada: economia, administração e contabilidade**. 11. ed. Porto Alegre: Bookman, 2006. 536p.

GARCIA, S.; LUSTOSA, P. R. B.; BARROS, N. S. Aplicabilidade do método de simulação de monte carlo na previsão dos custos de produção de companhias industriais: O caso da companhia vale do rio doce. **Revista de Contabilidade e Organizações**, v. 4, n. 10, p. 156–¨173, 2010.

GIRARDI, L. H.; CARGNELUTTI FILHO, A.; STORCK, F. Erro tipo I e poder de cinco testes de comparações múltipla de médias. **Revista Brasileira de Biometria**, v. 27, n. 1, p. 23–36, 2009.

GOMES, F. P. **Curso de Estatística Experimental**. 12. ed. São Paulo: Livraria Nobel S.A., 1989. 467p.

KEULS, M. The use of the studentized range in connection with an analysis of variance. **Euphytica**, v. 1, p. 112–122, 1952.

MACHADO, A. A. et al. Estatística experimental: uma abordagem fundamentada no planejamento e no uso de recursos computacionais. In: IN: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 50.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 11., 2005. **Anais...** Londrina, 2005. p. 290.

MEYER, P. L. **Probabilidade: aplicações à estatística**. 2. ed. Rio de Janeiro: Livros técnicos e Científicos, 1983.

MICHELETTI, S. M. F. B. et al. Controle de cerconota anonella (sepp.) (lep.: Oecophoridae) e de bephratelloides pomorum (fab.) (hym.: Eurytomidae) em frutos de graviola (annona muricata l.). **Revista Brasileira de Fruticultura**, v. 23, n. 3, 2001.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 3. ed. New York: John Wiley and Sons, 1991. 649p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3. ed. New York: McGraw Hill, 1974. 577 p.

NEWMAN, D. The distribution of range in samples from a normal population, expressed in terms of independent estimate of a standard deviation. **Biometrika**, v. 31, p. 20–30, 1939.

PEGDEN, C. D.; SHANNON, R. E.; SADOWSKI, R. P. **Introduction to simulation using SIMAN**. 2. ed. New York: McGraw-Hill, 1995.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

RAMOS, P. S.; FERREIRA, D. F. Agrupamento de médias via bootstrap de populações normais e não-normais. **Revista Ceres**, v. 56, n. 1, p. 140–149, 2009.

RAMOS, P. S.; VIEIRA, M. T. Bootstrap multiple comparison procedure based on the f distribution. **Revista Brasileira de Biometria**, v. 31, n. 4, p. 529–546, 2014.

ROBERTS, M.; RUSSO, R. **A Student's Guide to Analysis of Variance**. Taylor & Francis, 2014. ISBN 9781317725053. Disponível em: <https://books.google.com.br/books?id=s1u4AwAAQBAJ>.

SANTOS, É. C. d.; SANTOS, E. C. d.; MESQUITA, M. F. S. Fundamento dos testes estatísticos e sua aplicabilidade em ensaios experimentais com animais. **Revista Agrogeoambiental**, v. 2, n. 3, 2010.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, v. 30, n. 3, 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.

SOUSA, V. A.; LIMA JUNIOR, M. A.; FERREIRA, L. R. C. Avaliação de testes estatísticos de comparações múltiplas de médias. **Rev. Ceres**, v. 59, n. 3, p. 350–354, 2012.

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

TRIOLA, M. F. **Introdução à Estatística**. 11. ed. Rio de Janeiro: LTC, 2013.

TUKEY, J. W. **The problem of multiple comparisons**. Princeton, NJ: Princeton University, 1953.

VIEIRA, S. **Análise de variância: ANOVA**. São Paulo: Atlas, 2006. 204 p.