ERRO TIPO I E PODER DE CINCO TESTES DE COMPARAÇÃO MÚLTIPLA DE MÉDIAS

Luís Henrique GIRARDI[1]
Alberto CARGNELUTTI FILHO[2]
Lindolfo STORCK[2]

- RESUMO: O objetivo deste trabalho foi avaliar os testes t, Tukey, Bonferroni, Duncan e Student-Newman-Keuls (SNK) em relação ao erro tipo I e ao poder, em cenários formados por combinações de número de tratamentos, número de repetições e coeficientes de variação, em condição de normalidade dos resíduos. Foram simulados, via Monte Carlo, oitenta e vinte e cinco cenários, respectivamente, para avaliar o erro tipo I e o poder dos testes. O teste t controla a taxa de erro tipo I por comparação (TPC), e os testes de Tukey, Bonferroni e SNK controlam a taxa de erro tipo I por experimento (TPE). Já o teste de Duncan não controla a TPC e a TPE. Há decréscimo da TPC dos testes Tukey, Bonferroni, Duncan e SNK e acréscimo da TPE nos testes t e Duncan com o aumento do número de tratamentos. Há acréscimo do poder dos cinco testes com o incremento do tamanho da diferença entre médias adjacentes e com o aumento do número de tratamentos. Os testes t e Duncan apresentam maior poder em relação ao Tukey e Bonferroni, nessa ordem, e o SNK apresenta situação intermediária.
- PALAVRAS-CHAVE: Simulação; Monte Carlo; taxa de erro por comparação; taxa de erro por experimento.

1 Introdução

Um problema comum em diversas áreas da ciência é comparar o efeito de muitos tratamentos para determinar quais destes produzem resultados médios diferentes entre si, caso exista esta diferença. O caminho mais usual para tratar este problema é a análise de variância (ANOVA). O teste F da ANOVA aplicado aos resultados de um experimento testa a hipótese de igualdade de médias de todos os tratamentos (hipótese H_0). Ao rejeitar essa hipótese a um nível α de significância e sendo os tratamentos com mais de dois níveis e de natureza qualitativa, análises complementares como testes de comparações múltiplas de médias (TCMM) são adequadas para identificar quais tratamentos diferem.

Existem diversos TCMM (teste t ou LSD (*Least Significant Difference*), Tukey, Duncan, t de Bonferroni ou Bonferroni, Dunnett, Student-Newman-Keuls, Scheffé, Scott-Knott, Waller-Duncan, REGWK, Gabriel), cada um com as suas particularidades, e diferem fundamentalmente na filosofia de controle do erro tipo I (Ramalho, Ferreira e Oliveira, 2000). O teste de Scott-Knott, por exemplo, é indicado quando se deseja evitar ambigüidades nos resultados (Ferreira; Muniz e Aquino,

[1] Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul - UFRGS, CEP 91509-900, Porto Alegre, RS, Brasil. E-mail: *henrique.girardi@gmail.com*
[2] Departamento de Fitotecnia, Centro de Ciências Rurais - CCR, Universidade Federal de Santa Maria - UFSM, CEP 97105-900, Santa Maria, RS, Brasil. E-mail: *cargnelutti@pq.cnpq.br* / *lindolfo@pq.cnpq.br*

1999). Já o teste de Dunnett é indicado para comparar os tratamentos em relação a um tratamento testemunha (controle) (BANZATO e KRONKA, 2006).

Entre as decisões de um teste de hipóteses há erros e acertos. A probabilidade de rejeitar uma hipótese nula (H_0) dado que ela é verdadeira (deveria ser aceita), ou seja, afirmar estatisticamente que existe diferença significativa quando, de fato, esta não existe é o erro tipo I, representado por α e, normalmente, fixado pelo pesquisador. Outra decisão incorreta é o erro tipo II (representado por β) definido pela probabilidade de aceitar uma hipótese nula dado que ela é falsa (deveria ser rejeitada), isto é, afirmar estatisticamente que não existe diferença significativa quando, de fato, esta existe. Por outro lado, uma decisão correta é tomada ao afirmar que existe diferença significativa entre pelo menos duas médias comparadas quando esta realmente existe. A probabilidade de tomar essa decisão é o poder do teste $(1 - \beta)$.

É importante que o TCMM apresente um erro tipo I igual ao nominal, fixado previamente pelo pesquisador, e que este erro mantenha-se inalterado nas variações de cenário (combinações de número de tratamentos, número de repetições e coeficientes de variação) e, além disso, apresente um poder estatístico elevado (mais próximo possível de 1 ou 100%).

Há diferentes formas de medir a taxa de erro tipo I, dificultando a avaliação do mérito relativo dos TCMM. Experimentos simulados sem efeito de tratamento, ou seja, as k médias populacionais não diferem (situação de nulidade completa - $\mu_1 = \mu_2 = \ldots = \mu_{k-1} = \mu_k$) são submetidos aos TCMM, e assim a taxa de erro por comparação (comparisonwise) é estimada por meio da razão entre o número de inferências erradas (quantidade de vezes que o TCMM rejeitou H_0) e o número total de inferências. Já a taxa de erro por experimento (experimentwise) é estimada por meio da razão entre o número de experimentos com no mínimo uma inferência errada e o número total de experimentos.

Outro aspecto que dificulta a comparação de resultados dos TCMM é a possibilidade de diferentes configurações das k médias populacionais, citada por Einot e Gabriel (1975) e Rafter, Abell e Braselton (2002). Nesse sentido, uma configuração de amplitude mínima é o equivalente a separar o efeito de todos os tratamentos envolvidos na análise em dois grupos: a primeira metade das médias iguais entre si, mas diferente da segunda metade das médias, que também são iguais entre si, ou seja, $\mu_1 = \mu_2 = \ldots = \mu_{k/2} < \mu_{(k/2)+1} = \ldots = \mu_{k-1} = \mu_k$. Outra possibilidade é a configuração de amplitude máxima em que a primeira média é a menor, a última média é a maior e as médias intermediárias são superiores a menor, inferiores a maior e não diferem entre si, conforme: $\mu_1 < \mu_2 = \mu_3 = \ldots = \mu_{k-2} = \mu_{k-1} < \mu_k$. Já a configuração igualmente espaçada atribui para cada média adjacente um incremento fixo (número de erro padrão da média de um tratamento) em relação à média anterior, como $\mu_1 < \mu_2 < \mu_3 < \ldots < \mu_{k-2} < \mu_{k-1} < \mu_k$. Variações dessas configurações foram realizadas para avaliação do erro tipo I e do poder do teste em trabalhos de Carmer e Swanson (1973), Kemp (1973), Thomas (1974), Einot e Gabriel (1975), Perecin e Barbosa (1988), Conagim, Igue e Nagai (1999), Silva, Ferreira e Bearzoti (1999) e Borges e Ferreira (2003), dificultando a comparação dos resultados.

A configuração igualmente espaçada define que todos os contrastes de médias, duas a duas, apresentam uma determinada diferença que é testada por meio do TCMM em avaliação. Sendo assim, é adequada para estimar o poder do teste, uma vez que toda e qualquer diferença acusada como significativa pelo TCMM deve-se diretamente as diferenças de fato existentes, e não se confunde com a taxa de erro tipo I por comparação (fato que pode ocorrer nas configurações de amplitude mínima e de amplitude máxima). No entanto, mesmo em trabalhos que utilizam a configuração

igualmente espaçada, a forma como o poder do teste é calculado, dificulta a comparação dos resultados. Assim, como exemplo, em um experimento com cinco tratamentos, dez contrastes de médias podem ser testados (μ1 vs μ2, μ1 vs μ3, μ1 vs μ4, μ1 vs μ5, μ2 vs μ3, μ2 vs μ4, μ2 vs μ5, μ3 vs μ4, μ3 vs μ5 e μ4 vs μ5). A razão entre o número de diferenças significativas por um TCMM, em avaliação, e total de inferências (dez) é uma primeira possibilidade de estimativa poder do teste. Uma segunda possibilidade é estimar o poder do teste por meio da razão entre o número de contrastes significativos entre apenas os contrastes com médias adjacentes (μ1 vs μ2, μ2 vs μ3, μ3 vs μ4, μ4 vs μ5). Assim, espera-se maior e menor poder do teste, respectivamente, em relação à primeira e a segunda possibilidade.

Diversos TCMM sob diferentes cenários (variações de número de tratamentos e/ou de número de repetições e/ou ainda de coeficientes de variação) têm sido avaliados quanto ao erro tipo I e ao poder (Carmer e Swanson, 1973; Kemp, 1973; Thomas, 1974; Einot e Gabriel, 1975; Perecin e Barbosa, 1988; Conagin, Igue e Nagai, 1999; Silva, Ferreira e Bearzoti, 1999; Borges e Ferreira, 2003; Conagin e Barbin, 2006; Conagin, Barbin e Demétrio, 2008). No entanto, cenários mais extremos e com configuração igualmente espaçada para o estudo do poder do teste, obtido pela razão entre o número de contrastes significativas e total de contrastes, carecem de estudos adicionais. Assim, o objetivo deste trabalho foi avaliar os testes t ou LSD (Least Significant Difference), Tukey, t de Bonferroni ou Bonferroni, Duncan e Student-Newman-Keuls (SNK), muito utilizados em publicações e disponíveis em diversos softwares estatísticos, em relação ao erro tipo I e ao poder, em cenários formados por combinações de número de tratamentos, número de repetições e de coeficientes de variação.

2 Material e métodos

Foi utilizada simulação Monte Carlo para avaliar as taxas de erro tipo I por comparação (TPC) e por experimento (TPE) e o poder dos testes t ou LSD (*Least Significant Difference*), Tukey, t de Bonferroni ou Bonferroni, Duncan e Student-Newman-Keuls (SNK). Foi adotado o modelo matemático de um delineamento inteiramente casualizado: $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, com i = 1, 2, ... p e j = 1, 2, ...r, no qual, Y_{ij} é o valor simulado na j-ésima repetição do i-ésimo tratamento, μ é a média geral (constante) arbitrada como 100, τ_i é o efeito do tratamento i, estipulado de tal forma que $\sum_{i=1}^{p} \tau_i = 0$, e ε_{ij} é o resíduo ou erro aleatório na j-ésima repetição do i-ésimo tratamento, simulado com distribuição normal com média zero e desvio padrão (σ) variando conforme o coeficiente de variação desejado.

Para simular os dados experimentais, realizar as análises de variância e os testes de comparações múltiplas de médias (TCMM), calcular as taxas de erro tipo I (TPC e TPE) e o poder foi desenvolvido um algoritmo usando o software *Oracle Crystal Ball* 7.3.1 em conjunto com o aplicativo *Microsoft Excel* 2007.

Foram considerados dois grupos de estudo (I e II) e para cada cenário foram simulados 3000 experimentos. Tomou-se o cuidado de, para qualquer particular cenário, estimar as taxas de erro tipo I e o poder dos TCMM com os mesmos resultados simulados (com base na mesma semente aleatória), para assegurar que diferenças nessas estimativas não se devam ao erro aleatório do processo de simulação via método de Monte Carlo e sim as diferenças entre os TCMM. O nível nominal de significância adotado em todos os cenários foi de 5%.

No grupo I foram simulados 240000 experimentos (80 cenários x 3000 experimentos por cenário) sem efeito de tratamento, ou seja, em situação de nulidade completa ($\tau_1 = \tau_2 = ... = \tau_{p-1} = \tau_p$). Os 80 cenários foram formados pela combinação entre o número de tratamentos (p = 3, 5, 10, 50 e 100), o número de repetições (r = 3, 4, 10 e 20) e os coeficientes de variação (CV) iguais a 1%, 5%, 10% e 20%.

Nesse grupo foram estimadas as taxas de erro tipo I por comparação (TPC) e por experimento (TPE) dos testes t ou LSD (*Least Significant Difference*), Tukey, t de Bonferroni ou Bonferroni, Duncan e Student-Newman-Keuls (SNK) em cada cenário, de acordo com o seguinte cenário demonstrativo (p = 5; r = 4 e CV = 5%). Nesse cenário foram simulados 3000 experimentos com cinco tratamentos, e, em cada experimento dez estimativas de contrastes de médias, duas a duas, podem ser comparadas (μ_1 *vs* μ_2, μ_1 *vs* μ_3, μ_1 *vs* μ_4, μ_1 *vs* μ_5, μ_2 *vs* μ_3, μ_2 *vs* μ_4, μ_2 *vs* μ_5, μ_3 *vs* μ_4, μ_3 *vs* μ_5 e μ_4 *vs* μ_5) com o valor da diferença mínima significativa de cada um dos cinco TCMM considerados. A TPC foi estimada por meio da razão entre o número total de inferências erradas nos 3000 experimentos (quantidade de vezes que o TCMM rejeitou H_0) e o número total de inferências (nesse caso, 3000 experimentos x 10 contrastes por experimento = 30000 contrastes), multiplicada por 100. Já a TPE foi estimada por meio da razão entre o número de experimentos com no mínimo uma inferência errada entre as 10 testadas e o número total de experimentos (3000), multiplicada por 100.

Para verificar se a TPC e a TPE diferiu do nível nominal de significância estabelecido (α = 5%), utilizou-se o limite inferior (3,975%) e o limite superior (6,025%), com base no intervalo de confiança (IC) exato de 99% para uma proporção $\hat{p}$ = 0,05 expresso por: IC = $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, na qual n = 3000 experimentos e $Z_{\alpha/2}$ = 2,575829304 obtido da normal padrão com média zero e desvio padrão um. Assim, taxas de erro dentro deste intervalo não foram assumidas como diferentes do valor nominal.

No grupo II foram simulados 75000 experimentos (25 cenários x 3000 experimentos por cenário) com diferentes efeitos de tratamentos definidos de acordo com a configuração igualmente espaçada que atribui para cada média adjacente um incremento fixo (número de erro padrão da média de um tratamento) em relação à média anterior ($\tau_1 < \tau_2 < ... < \tau_{p-1} < \tau_p$). Buscou-se desta forma criar um ambiente adequado para o estudo do poder discriminativo dos cinco TCMM estudados, pois a situação $\tau_1 < \tau_2 < ... < \tau_{p-1} < \tau_p$ foi simulada para ser verdadeira, respeitando-se a restrição de que $\sum_{i=1}^{p} \tau_i = 0$. Os 25 cenários foram formados pela combinação entre o número de tratamentos (p = 3, 5, 10, 50 e 100) e o número de erro padrão da média de um tratamento definido por $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{r}}$ ($\sigma_{\bar{x}}$ = 0,5, 1, 2, 4 e 8) mantendo-se fixo o r = 20 e CV = 10%. A fixação do CV justifica-se em função dos efeitos de tratamentos serem vinculados ao erro padrão da média de um tratamento, o que não altera o poder em diferentes precisões experimentais (Silva, Ferreira e Bearzoti, 1999).

Nesse grupo foi estimado o poder dos testes t ou LSD (*Least Significant Difference*), Tukey, t de Bonferroni ou Bonferroni, Duncan e Student-Newman-Keuls (SNK) em cada cenário, de acordo com o seguinte cenário demonstrativo (p = 5; $\sigma_{\bar{x}}$ = 2; r = 20 e CV = 10%). Nesse cenário foram simulados 3000 experimentos com cinco tratamentos, e, em cada experimento dez estimativas de contrastes de médias, duas a duas, podem ser comparadas (μ_1 *vs* μ_2, μ_1 *vs* μ_3, μ_1 *vs* μ_4, μ_1 *vs* μ_5, μ_2 *vs* μ_3, μ_2 *vs*

μ_4, μ_2 *vs* μ_5, μ_3 *vs* μ_4, μ_3 *vs* μ_5 e μ_4 *vs* μ_5) com o valor da diferença mínima significativa de cada um dos cinco TCMM considerados. Assim, nesse caso, o erro padrão da média de um tratamento é $\sigma_{\bar{x}} = \frac{10}{\sqrt{20}} = 2{,}236$ e os efeitos de tratamentos com base em $2\,\sigma_{\bar{x}}$ são $\tau_1 = 91{,}056$, $\tau_2 = 95{,}528$, $\tau_3 = 100{,}000$, $\tau_4 = 104{,}472$ e $\tau_5 = 108{,}944$. O poder do teste foi estimado por meio da razão entre o número total de inferências corretas nos 3000 experimentos (quantidade de vezes que o TCMM rejeitou H_0) e o número total de inferências (nesse caso, 3000 experimentos x 10 contrastes por experimento = 30000 contrastes), multiplicada por 100.

3 Resultados e discussão

Nos 80 cenários formados pela combinação entre o número de tratamentos (p = 3, 5, 10, 50 e 100), o número de repetições (r = 3, 4, 10 e 20) e os coeficientes de variação (CV) iguais a 1%, 5%, 10% e 20%, as taxas de erro tipo I por comparação (TPC) e por experimento (TPE), de maneira geral, não apresentaram alterações de elevada magnitude com variabilidade do número de repetições e do CV (Tabelas 1 a 4). Menor variabilidade entre as TPC e TPE é observada com o acréscimo do número de tratamentos. Com elevado número de tratamentos o efeito das repetições sobre a estimativa da TPC e TPE pode ser considerado desprezível, pois mesmo com poucas repetições os graus de liberdade do erro experimental são altos. Esses resultados, de certa forma, estão de acordo e permitem comparações com Perecin e Barbosa (1988) que apresentaram as taxas de erro tipo I conjuntamente para CV = 10% e CV = 20% e com base em experimentos com 4 repetições e nível nominal de significância de 5%.

Em todos os cenários, nos cinco testes de comparações múltiplas de médias (TCMM) a TPC foi inferior a TPE, o que era esperado, uma vez que a igualdade entre essas taxas de erro tipo I, só seria obtida se em todos os experimentos com no mínimo um contraste significativo a totalidade dos contrastes fosse significativa (Tabelas 1 a 4). A igualdade da TCP e TPE e de acordo com o nível nominal de significância estabelecido seria o ideal em um TCMM, pois conforme Perecin e Barbosa (1988), um TCMM que controla a TPC torna-se muito "frouxo" ao ser aplicado a todo o experimento, enquanto que um teste que controla a TPE torna-se "conservador" ao ser olhado por comparação.

A estimativa da taxa de erro tipo I por comparação (TPC), do teste t, manteve-se entre o limite inferior (3,975%) e o superior (6,025%) do intervalo de confiança de 99% para uma proporção em todos os cenários (Tabelas 1 a 4). Então, deve-se interpretar que as TPC não diferiram do nível nominal de significância estabelecido (α = 5%), podendo-se inferir que o teste t controla essa taxa de erro tipo I (Tabelas 1 a 4). Esses resultados corroboram com Perecin e Barbosa (1988) e Ramalho, Ferreira e Oliveira (2000). Por outro lado, as taxas de erro tipo I por experimento (TPE) foram sempre superiores ao nível nominal de significância estabelecido (α = 5%) evidenciando que esse procedimento não controla a TPE, conforme destacado em Ramalho, Ferreira e Oliveira (2000). Houve aumento da TPE com o acréscimo do número de tratamentos chegando a 100% nos cenários com 100 tratamentos. Com 100 tratamentos sempre houve ao menos um contraste, entre pares de médias, significativo em cada experimento, independentemente do número de repetições e da precisão experimental. Resultados semelhantes aos de Perecin e Barbosa (1988) que encontraram 25,0%, 58,5%, 89,5%, 99,6% e 100%, respectivamente para 5, 10, 20, 40 e 100 tratamentos.

O teste de Tukey não apresentou nenhuma estimativa de TPE fora do intervalo de confiança de 99% [3,975%; 6,025%], independentemente da variação do número de tratamentos, repetições e CV, indicando que não difere do nível nominal de significância estabelecido (α = 5%) (Tabelas 1 a 4). Então, infere-se que esse procedimento controla a TPE, conforme destacado em Ramalho, Ferreira e Oliveira (2000) e de acordo com resultados de Perecin e Barbosa (1988) e Borges e Ferreira (2003). Já a TPC, em todos os cenários, foi sempre inferior a 3,975%, o que revela que o mesmo difere do nível nominal de significância estabelecido (α = 5%) e que não controla essa taxa de erro tipo I. Houve diminuição gradativa da TPC com o acréscimo do número de tratamentos conforme constatado em Perecin e Barbosa (1988), que encontraram 0,9%, 0,2%, 0,1%, 0,0% e 0,0%, respectivamente para 5, 10, 20, 40 e 100 tratamentos. Esses resultados concordam parcialmente, com Ramalho, Ferreira e Oliveira (2000) que afirmam que o teste de Tukey controla adequadamente a TPC e TPE, preservando o nível nominal de significância (α). No entanto, esta última taxa de erro tipo I é inferior ao nível nominal estabelecido.

Entre os 80 cenários, em apenas 15 (18,75%) a estimativa da TPE do teste de Bonferroni não diferiu do nível nominal de 5%, pois a mesma esteve entre 3,975% e 6,025%. Nos demais 65 (81,25%) a TPE foi inferior ao nível nominal (Tabelas 1 a 4). De certa forma, este é um resultado interessante, pois implica em afirmar que o teste de Bonferroni é "melhor" do que se propõe, uma vez que fornece taxas mais baixas de erro erro tipo I. Por outro lado, analisando as propriedades analíticas de um teste, espera-se que ele atenda a taxa de erro nominal fixada. Em relação a TPC, nos 80 cenários a estimativa foi inferior ao nível nominal de 5% e houve um decréscimo com o aumento do número de tratamentos. Ainda em relação a TPE, entre as 14 estimativas que não diferem de 5%, 11 estão entre os 20 cenários com três tratamentos o que demonstra que o número de tratamentos exerce uma certa influência na TPE.

Quanto ao teste de Duncan, de maneira geral, há falta de controle de ambas as taxas de erro tipo I, com excesão da TPC com até três tratamentos. Ocorreu aumento da TPE e diminuição da TPC com o aumento do número de tratamentos, independentemente do número de repetições e do CV. Em 100% dos cenários a TPE foi superior e em 80% a TPC foi inferior ao nível nominal de significância estabelecido (α = 5%) (Tabelas 1 a 4), de acordo com intervalo de confiança de 99% [3,975%; 6,025%]. Esses resultados se assemelham com os de Perecin e Barbosa (1988) que encontraram TPC de 3,9%, 2,6%, 1,9%, 1,4% e 1% e TPE de 18,4%, 39,4%, 62,0%, 88,8% e 100%, respectivamente para 5, 10, 20, 40 e 100 tratamentos.

O teste de Student-Newman-Keuls (SNK) controlou a TPE, com exceção de alguns cenários que envolvem 3 e 5 tratamentos, cujas estimativas excederam o limite superior de confiança de 99% (6,025%), e apresentou estimativas de TPC inferiores ao nível nominal de significância estabelecido (α = 5%) em todos os cenários (Tabelas 1 a 4). Os resultados se assemelham com os de Perecin e Barbosa (1988) que encontraram TPC de 1,1%, 0,2%, 0,1%, 0,0% e 0,0% e TPE de 5,5%, 4,8%, 4,5%, 6,4% e 4,4%, respectivamente para 5, 10, 20, 40 e 100 tratamentos.

De maneira geral, o teste t controla a taxa de erro tipo I por comparação, enquanto os testes de Tukey, Bonferroni e SNK controlam a taxa de erro tipo I por experimento. Esses últimos mantém a TPC em níveis inferiores ao nominal estabelecido com diminuição gradativa conforme o aumento do número de tratamentos. Já o teste de Duncan não controla a TPC e TPE.

O poder dos cinco TCMM, para um mesmo número de tratamentos, aumentou com o acréscimo do tamanho da diferença entre médias adjacentes atribuídas de acordo o número de erro padrão da média de um tratamento $\left(\sigma_{\bar{x}}\right)$, e para um mesmo

$\sigma_{\bar{x}}$ houve aumento do poder com o acréscimo do número de tratamentos (Tabela 5). Acréscimo do poder decorrente do aumento da diferença entre médias adjacentes é esperado e explicado, uma vez que quanto maior é a diferença entre duas médias, maior a probabilidade de ser considerada significativa por um TCMM. Essa mesma tendência, porém em diferentes magnitudes, foi verificada em Perecin e Barbosa (1988) e Silva, Ferreira e Bearzoti (1999) em relação aos testes t, Tukey, Duncan e SNK. Já o acréscimo do poder com o aumento do número de tratamentos, nesse trabalho, é esperado, por considerar na sua estimativa todos os contrastes de médias, duas a duas, do experimento e não somente os contrastes com médias adjacentes. Portanto, devido à diferença entre a metodologia utilizada neste trabalho e a descrita em outros, os poderes dos testes t, Tukey, Duncan e SNK, não podem diretamente serem comparados aos resultados de Perecin e Barbosa (1988), Silva, Ferreira e Bearzoti (1999), que encontraram inalteração do poder com o acréscimo do número de tratamentos em relação aos testes t, Duncan e SNK e decréscimo em relação ao teste Tukey. Não cabe aqui julgar a adequabilidade da forma de estimativa do poder do teste.

Os teste t e Duncan apresentaram maior poder em relação ao Tukey e Bonferroni, nessa ordem, e o SNK apresentou situação intermediária (Tabela 5). Resultado esperado e que pode ser explicado pela diferença mínima significativa (DMS) para um mesmo experimento, com mais de dois tratamentos, apresentar a seguinte ordem crescente em relação aos testes: teste t, Duncan (média das DMS), SNK (média das DMS), Tukey e Bonferroni.

Conclusões

Nos cenários formados pela combinação entre o número de tratamentos (3, 5, 10, 50 e 100), o número de repetições (3, 4, 10 e 20) e os coeficientes de variação iguais a 1%, 5%, 10% e 20%, o teste t controla a taxa de erro tipo I por comparação (TPC), e os testes de Tukey, Bonferroni e SNK controlam a taxa de erro tipo I por experimento (TPE). Já o teste de Duncan não controla a TPC e a TPE. Há decréscimo da TPC dos testes Tukey, Bonferroni, Duncan e SNK e acréscimo da TPE nos testes t e Duncan com o aumento do número de tratamentos, independentemente do número de repetições e da precisão experimental.

Há acréscimo do poder dos testes t, Tukey, Bonferroni, Duncan e SNK com o incremento do tamanho da diferença entre médias adjacentes e com o aumento do número de tratamentos, independentemente do número de repetições e da precisão experimental. Os testes t e Duncan apresentam maior poder em relação ao Tukey e Bonferroni, nessa ordem, e o SNK apresenta situação intermediária.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão de bolsa de produtividade em pesquisa a Alberto Cargnelutti Filho e Lindolfo Storck.

GIRARDI, L. H.; CARGNELUTTI FILHO, A.; STORCK, L. Type I error and power of five multiple comparison procedures for means. *Rev. Bras. Biom.*, São Paulo, v.27, n.1, p.23-36, 2009.

- ***ABSTRACT:*** *The objective of this work was to evaluate the t, Tukey's, Bonferroni's, Duncan's and Student-Newman-Keuls (SNK) tests for type I error and power, in scenarios formed by combinations of number of treatments, number of repetitions and coefficients of variation on condition of residuals normality. Were simulated, by Monte Carlo, eighty and twenty-five scenarios, respectively, for the type I error and to the power. The t-test controls the type I comparisonwise error rate (TPC) and tests Tukey, Bonferroni and SNK control the type I experimentwise error rate (TPE). The Duncan's tests does not control neither the TPC or TPE. There are decreasing TPC of Tukey's, Bonferroni's, Duncan's and SNK tests and the increase in TPE t and Duncan's tests with the increasing number of treatments. There is increased power of the five tests with the increase in the size of the difference between means adjacent to the increase in the number of treatment. The t and Duncan's tests have more power in relation to Tukey's and Bonferroni's, in that order, and SNK showed intermediate situation.*

- ***KEYWORDS:*** *Simulation, Monte Carlo, comparisonwise error rate, experimentwise error rate.*

Referências

BANZATO, D. A.; KRONKA, S. N. *Experimentação agrícola*. 4.ed. Jaboticabal: FUNEP, 2006. 250p.

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

CARMER, S. G.; SWANSON, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *J. Am. Stat. Assoc.*, New York, v.68, p.66-74, 1973.

CONAGIN, A.; BARBIN, D. Teste modificado de Bonferroni e Sidak. *Sci. Agric.,* Piracicaba, v.63, p.70-76, 2006.

CONAGIN, A.; BARBIN, D.; DEMÉTRIO, C. G. B. Modifications for the Tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. *Sci. Agric.,* Piracicaba, v.65, p.428-432, 2008.

CONAGIN, A. et al. *Poder discriminativo de diferentes testes de comparação de médias*. Campinas: Instituto Agronomico, 1999. 16p. (Boletim científico, 44).

EINOT, I.; GABRIEL, K. R. A study of the powers of several methods of multiple comparisons. *J. AM. Stat. Assoc.*, New York, v.70, p.574-583, 1975.

FERREIRA, D. F.; MUNIZ, J. A.; AQUINO, L. H. Comparações múltiplas em experimentos com grande número de tratamentos – utilização do teste de Scott Knott. *Ciênc. Agrotec.*, Lavras, v.23, p.745-752, 1999.

KEMP. K.E. Multiple comparisons: comparisonwise versus experimentwise Type I error rates and their relationship to power. *J. Dairy Sci.*, Champaign, v.58, p.1374-1378, 1973.

PERECIN, D.; BARBOSA, J. C. Uma avaliação de seis procedimentos para comparações múltiplas. *Rev. Mat. Estat.*, São Paulo, v.6, p.95-103, 1988.

RAFTER, J. A. et al. Multiple comparison methods for means. *SIAM Review*, Philadelphia, v.44, p.259-278, 2002.

RAMALHO, M. A. P. et al. *Experimentação em genética e melhoramento de plantas*. Lavras: UFLA, 2000. 326p.

SILVA, E. C. et al. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. *Ciênc. Agrotec.*, Lavras, v.23, p.687-696, 1999.

THOMAS, D. A. H. Error Rate in Multiple Comparisons among Means – Results of a Simulation Exercise. *Appl. Stat.,* Washington, v.23, p.284-294, 1974.

Recebido em 24.10.2008.

Aprovado após revisão 22.03.2009.

Tabela 1 - Taxas de erro tipo I por comparação (TPC) e por experimento (TPE), em percentual, para diferentes testes de comparação múltipla de médias ao nível nominal de significância de 5%, em função do número de tratamentos (TRAT) e de repetições (REP) e coeficiente de variação = 1%

TRAT	REP	Teste t		Tukey		Bonferroni		Duncan		SNK	
		TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE
3	3	5,144	11,467**	2,200*	4,867	1,489*	3,533*	4,622	9,900**	2,944*	5,633
3	4	5,389	12,233**	2,478*	5,967	1,944*	4,633	4,978	11,067**	3,300*	6,700**
3	10	5,111	11,900**	2,300*	5,633	1,922*	4,767	4,378	9,800**	3,022*	6,100**
3	20	5,011	12,233**	1,967*	4,967	1,711*	4,367	4,133	9,667**	2,589*	5,667
5	3	5,240	25,667**	0,867*	5,600	0,520*	3,467*	3,967*	19,067**	1,320*	6,267**
5	4	5,110	26,600**	0,790*	4,967	0,543*	3,767*	3,623*	19,000**	1,120*	5,233
5	10	4,967	26,933**	0,603*	4,667	0,433*	3,467*	3,347*	18,200**	0,890*	5,033
5	20	5,120	28,900**	0,640*	5,133	0,447*	3,667*	3,267*	18,667**	0,933*	5,467
10	3	4,959	54,933**	0,219*	5,667	0,106*	2,933*	2,754*	36,700**	0,287*	5,733
10	4	5,103	57,333**	0,223*	5,767	0,139*	3,833*	2,680*	37,467**	0,290*	5,767
10	10	5,061	60,667**	0,182*	5,567	0,120*	3,900*	2,418*	37,833**	0,232*	5,567
10	20	4,900	61,200**	0,157*	4,600	0,109*	3,267*	2,196*	36,633**	0,207*	4,633
50	3	4,978	99,800**	0,008*	4,467	0,004*	2,533*	1,341*	91,767**	0,009*	4,867
50	4	5,099	99,767**	0,008*	5,167	0,004*	3,133*	1,291*	92,733**	0,008*	5,167
50	10	5,006	99,933**	0,007*	5,267	0,004*	3,200*	1,131*	92,367**	0,007*	5,267
50	20	4,968	99,933**	0,006*	4,400	0,003*	2,367*	1,061*	92,500**	0,006*	4,400
100	3	5,028	100,000**	0,002*	5,733	0,001*	3,467*	0,939*	99,633**	0,002*	5,667
100	4	5,017	100,000**	0,002*	4,967	0,010*	2,967*	1,017*	99,167**	0,002*	4,967
100	10	4,999	100,000**	0,002*	4,767	0,001*	3,267*	0,905*	99,600**	0,002*	4,767
100	20	5,084	100,000**	0,002*	5,733	0,001*	4,133	0,905*	99,500**	0,002*	5,733

* Taxas de erro tipo I (por comparação ou por experimento) que ficaram abaixo do limite inferior do intervalo de confiança com 99% de probabilidade (3,975%) para a proporção empírica desta taxa.

** Taxas de erro tipo I por experimento que ficaram acima do limite superior do intervalo de confiança com 99% de probabilidade (6,025%) para a proporção empírica desta taxa.

Tabela 2 - Taxas de erro tipo I por comparação (TPC) e por experimento (TPE), em percentual, para diferentes testes de comparação múltipla de médias ao nível nominal de significância de 5%, em função do número de tratamentos (TRAT) e de repetições (REP) e coeficiente de variação = 5%

TRAT	REP	Teste t		Tukey		Bonferroni		Duncan		SNK	
		TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE
3	3	5,089	11,400**	2,244*	5,100	1,800*	4,033	4,656	10,200**	2,989*	5,967
3	4	4,756	11,200**	1,811*	4,400	1,556*	3,733*	4,211	9,633**	2,489*	5,100
3	10	4,789	11,533**	1,900*	4,600	1,556*	4,067	3,944*	9,100**	2,478*	5,100
3	20	5,000	11,933**	2,044*	5,333	1,633*	4,200	4,256	9,767**	2,800*	5,933
5	3	5,317	25,433**	0,850*	5,567	0,490*	3,333*	4,130	19,867**	1,327*	6,200**
5	4	4,800	24,833**	0,833*	5,533	0,547*	3,767*	3,500*	18,033**	1,217*	5,900
5	10	5,260	28,567**	0,733*	5,300	0,493*	3,700*	3,580*	19,467**	1,057*	5,667
5	20	4,897	27,067**	0,650*	4,900	0,470*	3,633*	3,230*	18,167**	0,993*	5,133
10	3	4,887	55,000**	0,215*	4,833	0,113*	2,700*	2,664*	35,933**	0,280*	4,867
10	4	4,944	57,233**	0,176*	4,833	0,101*	3,033*	2,524*	36,400**	0,236*	5,000
10	10	4,864	61,000**	0,130*	4,233	0,087*	2,900*	2,199*	36,433**	0,154*	4,267
10	20	5,052	62,400**	0,136*	4,900	0,096*	3,633*	2,231*	38,233**	0,177*	5,033
50	3	5,027	99,900**	0,008*	5,267	0,004*	3,133*	1,369*	92,000**	0,008*	5,267
50	4	4,986	99,867**	0,006*	4,833	0,004*	2,633*	1,230*	92,633**	0,007*	4,833
50	10	4,971	99,933**	0,008*	5,400	0,005*	3,433*	1,111*	91,233**	0,008*	5,400
50	20	5,010	100,000**	0,007*	5,400	0,004*	3,233*	1,088*	92,467**	0,007*	5,400
100	3	5,080	100,000**	0,002*	5,400	0,001*	3,167*	1,104*	99,533**	0,002*	5,400
100	4	4,987	100,000**	0,002*	5,267	0,001*	3,167*	0,996*	99,167**	0,002*	5,267
100	10	5,025	100,000**	0,002*	5,000	0,001*	3,400*	0,917*	99,500**	0,002*	5,067
100	20	4,999	100,000**	0,002*	4,967	0,001*	3,200*	0,891*	99,233**	0,002*	4,967

* Taxas de erro tipo I (por comparação ou por experimento) que ficaram abaixo do limite inferior do intervalo de confiança com 99% de probabilidade (3,975%) para a proporção empírica desta taxa.

** Taxas de erro tipo I por experimento que ficaram acima do limite superior do intervalo de confiança com 99% de probabilidade (6,025%) para a proporção empírica desta taxa.

Tabela 3 - Taxas de erro tipo I por comparação (TPC) e por experimento (TPE), em percentual, para diferentes testes de comparação múltipla de médias ao nível nominal de significância de 5%, em função do número de tratamentos (TRAT) e de repetições (REP) e coeficiente de variação = 10%

TRAT	REP	Teste t		Tukey		Bonferroni		Duncan		SNK	
		TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE
3	3	4,867	10,800**	2,056*	5,000	1,478*	3,500*	4,478	9,667**	2,933*	5,867
3	4	4,856	11,067**	2,044*	4,867	1,678*	4,100	4,311	9,433**	2,789*	5,600
3	10	5,311	12,467**	1,900*	4,667	1,567*	4,000	4,522	10,167**	2,711*	5,600
3	20	5,122	12,433**	1,956*	5,100	1,767*	4,600	4,344	10,300**	2,678*	6,033**
5	3	4,900	24,100**	0,717*	4,233	0,433*	2,567*	3,700*	17,967**	1,133*	4,667
5	4	4,977	25,833**	0,813*	5,100	0,603*	4,000	3,657*	18,933**	1,157*	5,467
5	10	4,930	28,333**	0,613*	4,567	0,447*	3,467*	3,277*	19,300**	0,903*	5,000
5	20	4,683	26,733**	0,610*	4,933	0,453*	3,733*	3,043*	17,833**	0,873*	5,400
10	3	4,766	54,900**	0,179*	4,733	0,097*	2,500*	2,588*	36,200**	0,249*	4,867
10	4	5,109	59,500**	0,196*	5,200	0,124*	3,500*	2,601*	36,833**	0,266*	5,367
10	10	5,010	62,133**	0,157*	4,633	0,107*	3,200*	2,284*	37,933**	0,191*	4,667
10	20	4,867	63,367**	0,138*	4,333	0,093*	3,033*	2,173*	36,733**	0,179*	4,400
50	3	4,969	99,867**	0,008*	4,767	0,004*	2,833*	1,328*	91,967**	0,009*	4,767
50	4	4,915	99,833**	0,006*	4,467	0,003*	2,467*	1,211*	92,333**	0,006*	4,467
50	10	5,041	99,933**	0,008*	5,000	0,005*	3,200*	1,139*	91,867**	0,008*	5,000
50	20	4,960	100,000**	0,007*	5,300	0,004*	3,267*	1,075*	91,867**	0,008*	5,300
100	3	5,060	100,000**	0,002*	4,567	0,001*	3,067*	1,086*	99,833**	0,002*	4,567
100	4	5,011	100,000**	0,002*	5,200	0,001*	2,600*	1,008*	99,433**	0,002*	5,200
100	10	5,023	100,000**	0,002*	5,400	0,001*	3,567*	0,910*	99,367**	0,002*	5,400
100	20	5,068	100,000**	0,002*	5,067	0,001*	3,600*	0,905*	99,767**	0,002*	5,067

* Taxas de erro tipo I (por comparação ou por experimento) que ficaram abaixo do limite inferior do intervalo de confiança com 99% de probabilidade (3,975%) para a proporção empírica desta taxa.

** Taxas de erro tipo I por experimento que ficaram acima do limite superior do intervalo de confiança com 99% de probabilidade (6,025%) para a proporção empírica desta taxa.

Tabela 4 - Taxas de erro tipo I por comparação (TPC) e por experimento (TPE), em percentual, para diferentes testes de comparação múltipla de médias ao nível nominal de significância de 5%, em função do número de tratamentos (TRAT) e de repetições (REP) e coeficiente de variação = 20%

TRAT	REP	Teste t		Tukey		Bonferroni		Duncan		SNK	
		TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE	TPC	TPE
3	3	4,789	10,767**	2,167*	4,967	1,589*	3,633*	4,444	9,767**	2,856*	5,633
3	4	4,467	10,433**	1,944*	4,600	1,456*	3,533*	3,933*	8,900**	2,522*	5,167
3	10	5,000	11,967**	1,878*	5,000	1,611*	4,300	4,244	9,800**	2,678*	5,667
3	20	5,111	12,433**	1,811*	4,833	1,600*	4,200	4,267	9,967**	2,578*	5,667
5	3	5,200	25,233**	0,830*	5,367	0,467*	3,233*	4,023	19,367**	1,247*	5,867
5	4	5,380	26,767**	0,857*	5,533	0,567*	3,733*	3,917*	19,267**	1,247*	6,033**
5	10	5,100	28,267**	0,717*	5,333	0,543*	4,167	3,547*	19,633**	1,027*	5,933
5	20	5,020	27,667**	0,610*	4,900	0,500*	4,067	3,357*	18,500**	0,910*	5,267
10	3	5,037	57,000**	0,199*	4,967	0,113*	3,200*	2,830*	38,967**	0,258*	5,167
10	4	5,018	57,400**	0,182*	4,633	0,111*	2,800*	2,536*	36,767**	0,240*	4,733
10	10	5,041	60,767**	0,196*	5,833	0,130*	3,933*	2,408*	36,867**	0,261*	5,967
10	20	4,804	61,333**	0,141*	4,700	0,100*	3,300*	2,106*	36,000**	0,178*	4,700
50	3	5,077	99,867**	0,009*	5,833	0,005*	3,433*	1,371*	91,933**	0,009*	5,833
50	4	5,024	99,967**	0,008*	5,533	0,004*	3,100*	1,256*	91,733**	0,008*	5,533
50	10	4,969	99,933**	0,006*	4,633	0,004*	3,233*	1,125*	91,600**	0,007*	4,633
50	20	5,084	99,933**	0,007*	5,267	0,004*	3,467*	1,118*	93,167**	0,007*	5,267
100	3	5,005	100,000**	0,002*	5,433	0,001*	3,267*	1,078*	99,333**	0,002*	5,433
100	4	5,000	100,000**	0,002*	5,367	0,001*	3,233*	1,005*	99,400**	0,002*	5,367
100	10	4,962	100,000**	0,001*	4,500	0,001*	2,700*	0,896*	99,267**	0,001*	4,500
100	20	5,002	100,000**	0,002*	4,267	0,001*	2,667*	0,873*	99,200**	0,002*	4,267

* Taxas de erro tipo I (por comparação ou por experimento) que ficaram abaixo do limite inferior do intervalo de confiança com 99% de probabilidade (3,975%) para a proporção empírica desta taxa.

** Taxas de erro tipo I por experimento que ficaram acima do limite superior do intervalo de confiança com 99% de probabilidade (6,025%) para a proporção empírica desta taxa.

Tabela 5 - Poder, em percentual, para diferentes testes de comparação múltipla de médias ao nível nominal de significância de 5%, em função do número de tratamentos (TRAT) e do erro padrão da média de um tratamento $(\sigma_{\bar{x}})$, com 20 repetições e coeficiente de variação de 10%

TRAT	Diferença real entre médias	Teste t	Tukey	Bonferroni	Duncan	SNK
3	0,5 $\sigma_{\bar{x}}$	7,444	3,356	2,922	6,622	4,500
3	1 $\sigma_{\bar{x}}$	15,356	8,433	7,678	14,000	10,689
3	2 $\sigma_{\bar{x}}$	45,278	32,644	30,989	44,256	40,389
3	4 $\sigma_{\bar{x}}$	86,033	76,567	75,333	86,033	86,011
3	8 $\sigma_{\bar{x}}$	100,000	100,000	99,989	100,000	100,000
5	0,5 $\sigma_{\bar{x}}$	11,960	2,863	2,377	9,083	4,073
5	1 $\sigma_{\bar{x}}$	31,343	13,560	11,930	27,840	18,597
5	2 $\sigma_{\bar{x}}$	64,997	47,470	45,463	63,870	60,287
5	4 $\sigma_{\bar{x}}$	92,093	80,270	78,753	92,090	92,083
5	8 $\sigma_{\bar{x}}$	99,993	99,957	99,933	99,993	99,993
10	0,5 $\sigma_{\bar{x}}$	29,385	7,162	6,116	23,010	9,591
10	1 $\sigma_{\bar{x}}$	59,136	35,176	33,344	55,846	45,439
10	2 $\sigma_{\bar{x}}$	82,097	66,941	65,613	81,487	79,325
10	4 $\sigma_{\bar{x}}$	96,096	86,871	86,072	96,095	96,090
10	8 $\sigma_{\bar{x}}$	99,998	99,876	99,841	99,998	99,998
50	0,5 $\sigma_{\bar{x}}$	80,840	61,339	60,317	77,863	69,290
50	1 $\sigma_{\bar{x}}$	91,000	80,282	79,690	90,126	87,323
50	2 $\sigma_{\bar{x}}$	96,361	90,818	90,505	96,228	95,758
50	4 $\sigma_{\bar{x}}$	99,228	96,283	96,149	99,228	99,227
50	8 $\sigma_{\bar{x}}$	99,999	99,802	99,749	99,999	99,999
100	0,5 $\sigma_{\bar{x}}$	90,101	77,916	77,347	88,505	83,874
100	1 $\sigma_{\bar{x}}$	95,456	89,090	88,777	95,013	93,559
100	2 $\sigma_{\bar{x}}$	98,187	94,960	94,803	98,118	97,882
100	4 $\sigma_{\bar{x}}$	99,610	97,953	97,885	99,610	99,609
100	8 $\sigma_{\bar{x}}$	99,999	99,805	99,761	99,999	99,999