Um estudo sobre os níveis de significância por comparação e por experimento de procedimentos de comparações múltiplas de acordo com o resultado do teste-F na ANOVA

Josiane Rodrigues[1], Sônia Maria De Stefano Piedade[2], Idemauro Antonio Rodrigues de Lara[3]

Introdução

Em experimentos agrícolas, um problema muito comum é a comparação das médias de alguns tratamentos de interesse, para determinar quais deles diferem entre si. De acordo com Girardi et al. (2009), a maneira mais usual para tratar esse problema é a análise de variância (ANOVA).

Se os tratamentos do ensaio são de efeito fixo, o teste-F global da ANOVA testa a hipótese de igualdade entre as médias populacionais dos tratamentos. Caso o teste-F seja significativo, sendo os tratamentos de natureza qualitativa, então a aplicação de testes de comparação de médias é feita, com o intuito de investigar eventuais diferenças entre pares de médias específicos ou combinações lineares dessas médias.

Um dos dilemas envolvidos com os testes de médias é justamente a sua aplicação condicional a um resultado preliminar significativo do teste-F na ANOVA (Hsu, 1996). De acordo com Cardellino e Siewerdt (1992), essa é uma questão polêmica e responsável por dividir a opinião de muitos estudiosos, e que deve, portanto, ser melhor investigada.

No entanto, não é possível desvincular esse estudo dos possíveis erros que podem ser cometidos num teste de hipótese, afinal, ao analisar a hipótese para um contraste de médias, o teste, seja ele aplicado apenas mediante um resultado significativo do teste-F global ou não, possui probabilidades para os erros tipo I e II.

Segundo Girardi et al. (2009), há diferentes formas de avaliar a taxa de erro tipo I para um procedimento de comparação de médias, tais como o erro tipo I por comparação e o erro tipo I por experimento, por exemplo, o que acaba dificultando a avaliação do desempenho desses procedimentos de uma maneira geral.

Diversos trabalhos têm sido realizados para avaliar procedimentos de comparação de médias no que diz respeito ao controle das taxas de erro tipo I (Souza et al., 2012). Todavia, o estudo desses conceitos aliado a um resultado preliminar significativo do teste-F global ou não carece de estudos adicionais.

Neste trabalho, a partir de experimentos simulados no delineamento inteiramente casualizado pelo método de Monte Carlo, estudou-se e comparou-se os testes de Tukey, Duncan, DMS de Fisher, Student-Newman-Keuls (SNK) e Scheffé no que se refere ao controle das taxas de erro tipo I condicionais e incondicionais a um resultado significativo do teste-F global na ANOVA.

[1] Departamento de Tecnologia Agroindustrial e Socioeconomia Rural, Centro de Ciências Agrárias – CCA, Universidade Federal de São Carlos – UFSCar, CEP 13600-970, Araras, SP, Brasil. E-mail: *josirodrigues@ufscar.br*.
[2] Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ, Universidade de São Paulo – USP, CEP 13418-900, Piracicaba, SP, Brasil. E-mail: *soniamsp@usp.br*.
[3] Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ, Universidade de São Paulo – USP, CEP 13418-900, Piracicaba, SP, Brasil. E-mail: *idemauro@usp.br*.

Materiais e métodos

Foram simulados, via método de Monte Carlo, 128000 experimentos, 2000 para cada cenário, sendo estes num total de 64 casos, formados pela combinação dos seguintes fatores: 3, 5, 7 e 9 tratamentos, 3, 4, 10 e 20 repetições e coeficiente de variação de 1%, 5%, 10% e 20%. Os experimentos foram simulados de acordo com o modelo estatístico referente ao delineamento inteiramente casualizado, da seguinte forma:

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

em que y_{ij} representa o valor simulado da resposta obtida no *i*-ésimo tratamento em sua *j*-ésima repetição, μ é a média geral, arbitrada como sendo 100, τ_i é o efeito fixo do *i*-ésimo tratamento (o qual será considerado nulo), e ε_{ij} é o erro aleatório, gerado independentemente com distribuição normal de média zero e desvio padrão (σ) variando conforme o coeficiente de variação desejado.

Tomou-se o cuidado de, para todos os cenários simulados, realizar as análises com base na mesma semente aleatória, de modo a garantir que possíveis diferenças não ocorressem devido ao erro aleatório do processo de simulação, mas sim devido às diferenças entre os testes de comparações múltiplas avaliados. Ainda, o nível de significância nominal adotado em todos os casos foi o de 5%. Para simular os dados experimentais e realizar as análises estatísticas foi desenvolvido um algoritmo usando o *software* R (2013).

Dentro de cada um dos cenários simulados foram então estimadas as taxas de erro tipo I por comparação e por experimento para cada um dos testes de Tukey, Duncan, SNK, DMS de Fisher e Scheffé. Ainda, essas taxas de erro foram consideradas de duas formas distintas: i) quando o procedimento de comparação múltipla foi aplicado independentemente do resultado do teste-F global; ii) quando o mesmo foi aplicado apenas se o teste-F da ANOVA foi significativo.

Boardman e Moffitt (1971) definiram o nível de significância por comparação (α_c) como sendo a proporção do número de inferências errôneas (concluir que $\mu_i \neq \mu_{i'}$ quando $\mu_i = \mu_{i'}$), dividido pelo número total de comparações realizadas.

Já quando os testes de comparação de médias são aplicados apenas mediante um resultado preliminar significativo do teste-F global, o nível de significância por comparação, agora denominado condicional (α_1), pode ser estimado empiricamente tomando o número total de erros tipo I cometidos nas comparações das a médias duas a duas, dentre os experimentos que apresentaram resultado significativo para o teste-F, e o número total de inferências realizadas (Bernhardson, 1975).

Ainda, uma segunda maneira de estimar o nível de significância por comparação condicional (α_2) é tomar o número total de inferências errôneas dentro dos experimentos com resultado significativo para o teste-F global e dividi-lo pelo número total de inferências realizadas dentro desses ensaios (Bernhardson, 1975).

Boardman e Moffitt (1971) definiram ainda o nível de significância por experimento (α_e) como sendo a proporção do número de experimentos com no mínimo uma inferência errônea dividido pelo número total de experimentos.

O nível de significância por experimento para a combinação do teste-F da ANOVA e os testes de comparação de médias considerados, ao qual chamar-se-á nível de significância por experimento condicional (α_3), pode ser estimado tomando o número total de ensaios que apresentaram tanto um teste-F global significante e no mínimo uma comparação resultando em erro tipo I, dividido pelo número total de experimentos (Bernhardson, 1975).

O mesmo nível de significância pode ainda ser estimado baseado apenas nos experimentos que apresentaram um resultado significativo para o teste-F global. Este nível (α_4) pode ser então estimado dividindo o número total de ensaios que apresentaram tanto um teste-F significante e no mínimo uma comparação derivando em erro tipo I, pelo número total de ensaios dentre os 2000 com o teste-F global significativo (Bernhardson, 1975).

De forma a verificar se cada um dos níveis de significância estimados diferiu do nível de significância nominal estabelecido ($\alpha = 5\%$), foi utilizado o limite inferior de 0,0375 e o limite superior de 0,0625, obtidos com base no intervalo de confiança (IC) exato de 95% para uma proporção $\hat{p} = 0,05$. Deste modo, níveis de significância dentro desse intervalo não foram considerados como diferentes do nível nominal estabelecido.

Resultados e discussão

Com relação aos níveis de significância por comparação pode-se notar o seguinte comportamento dos resultados: as taxas incondicionais α_c foram sempre iguais as taxas condicionais α_1 para o procedimento de Scheffé, em todos os cenários simulados, enquanto houve uma tendência para as taxas α_c serem ligeiramente maiores que α_1 para cada um dos demais testes considerados. Por sua vez, as taxas condicionais α_2 foram sempre maiores que as taxas α_c nos cinco procedimentos de comparação de médias avaliados, em cada um dos cenários simulados. As diferenças entre essas duas últimas taxas de erro mostraram-se grandes quando o número de tratamentos do cenário foi pequeno, e passaram a diminuir à medida que o número de tratamentos dos ensaios aumentou.

No que diz respeito aos níveis de significância por experimento, bem como ocorreu para os níveis de significância por comparação, as taxas incondicionais α_e foram sempre iguais as taxas condicionais α_3 para o procedimento de Scheffé, enquanto que para cada um dos demais procedimentos estudados, as taxas α_e foram pouco maiores que as taxas α_3. Por sua vez, as taxas α_e mostraram-se sempre menores que as taxas condicionais α_4, e as diferenças entre elas, embora consideráveis, diminuíram à medida que o número de tratamentos dos cenários aumentou. Para o caso particular do procedimento de Scheffé, no entanto, as diferenças entre essas taxas para cenários com maiores números de tratamentos não foram consideravelmente grandes.

A fim de resumir o comportamento de cada um dos procedimentos no que diz respeito ao controle de cada um dos níveis de significância, as Figuras de 1 a 6 descrevem o que ocorre com cada um dos níveis para cada um dos testes à medida que o número de tratamentos dos ensaios aumenta. Apenas a variação no número de tratamentos foi considerada pois este foi o único fator que levou à alteração significativa dos resultados. Para tanto, fixou-se $r = 10$, $CV = 10\%$ e $\alpha = 5\%$. Todos os pontos nessas figuras foram obtidos por meio do estudo de Monte Carlo. As linhas entre os pontos foram desenhadas para representar o que ocorre com os níveis com o aumento do número de tratamentos apenas para fins de apresentação. Apenas o caso de $r = 10$ e $CV = 10\%$ foi reportado aqui a fim de economizar espaço. No entanto, os resultados para os demais cenários simulados foram muito similares em sua aparência.

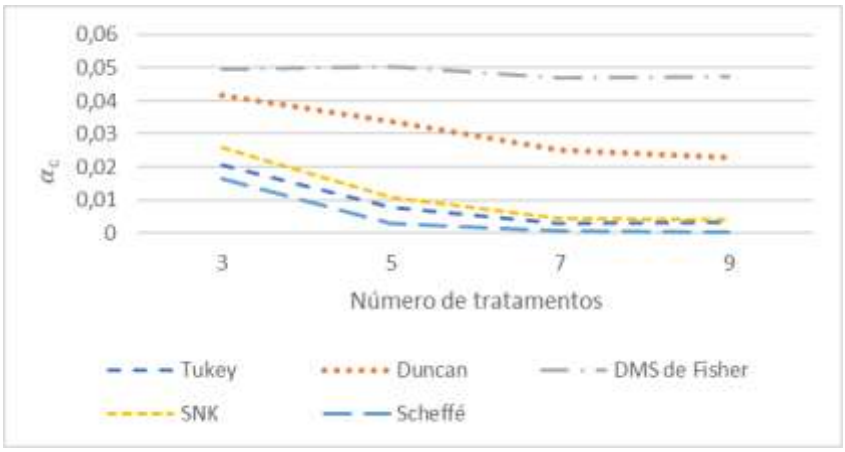

Figura 1 – Comportamento da taxa α_c para $r = 10$, $CV = 10\ \%$ e $\alpha = 5\ \%$, com respeito à variação do número de tratamentos.

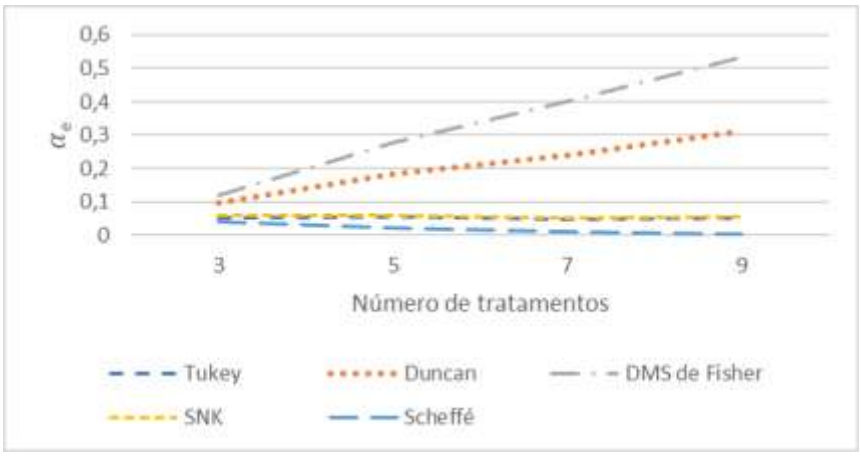

Figura 2 – Comportamento da taxa α_e para $r = 10$, $CV = 10\ \%$ e $\alpha = 5\ \%$, com respeito à variação do número de tratamentos.

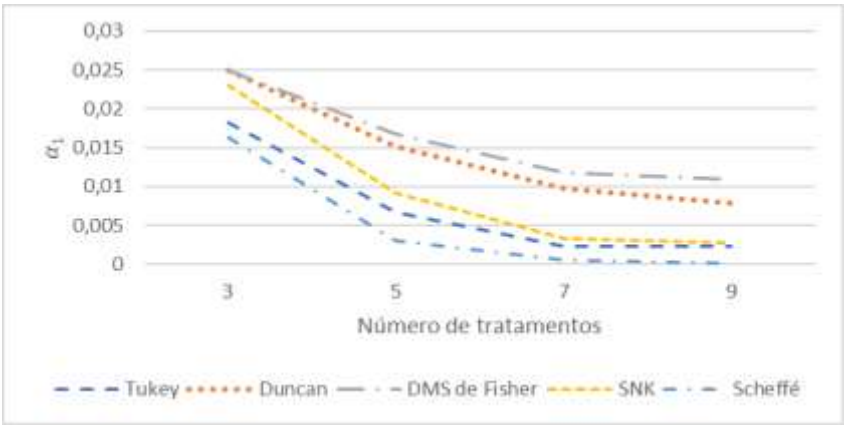

Figura 3 – Comportamento da taxa α_1 para $r = 10$, $CV = 10\ \%$ e $\alpha = 5\ \%$, com respeito à variação do número de tratamentos.

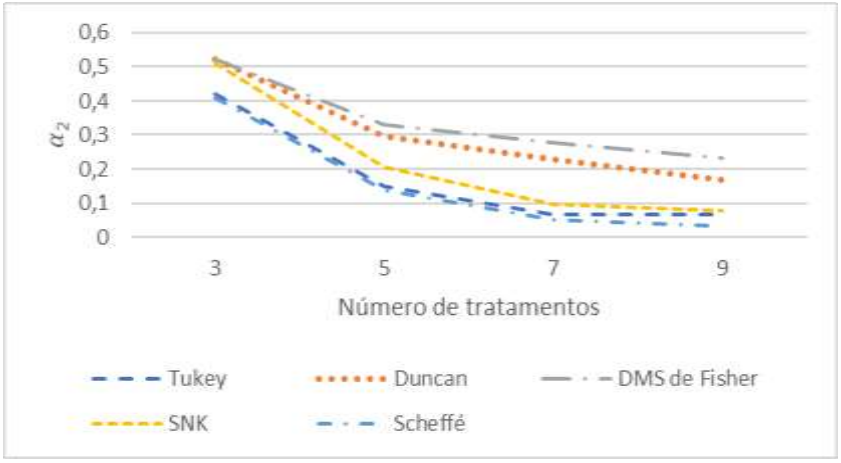

Figura 4 – Comportamento da taxa α_2 para $r = 10$, $CV = 10\ \%$ e $\alpha = 5\ \%$, com respeito à variação do número de tratamentos.

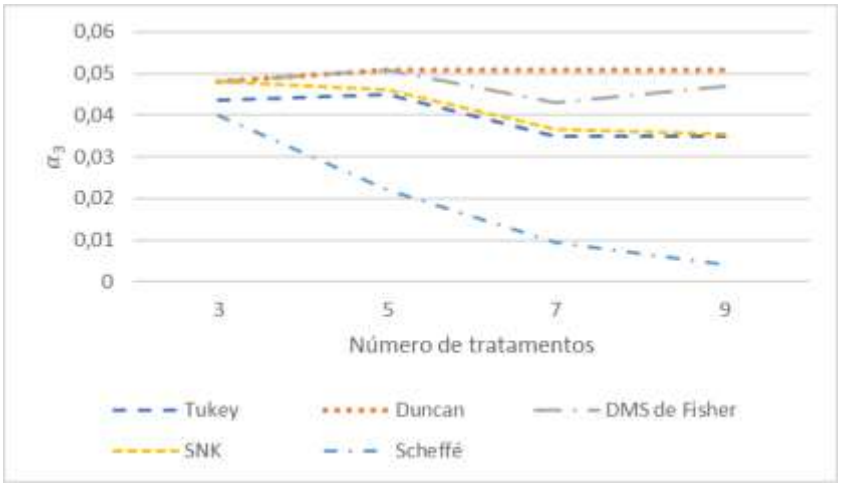

Figura 5 – Comportamento da taxa α_3 para $r = 10$, $CV = 10\ \%$ e $\alpha = 5\ \%$, com respeito à variação do número de tratamentos.

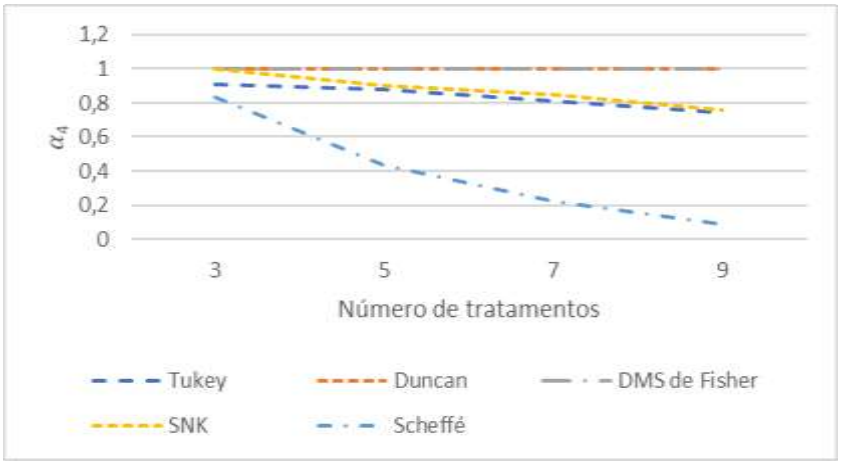

Figura 6 – Comportamento da taxa α_4 para $r = 10$, $CV = 10\ \%$ e $\alpha = 5\ \%$, com respeito à variação do número de tratamentos.

De um modo geral, foi possível verificar que o teste DMS de Fisher controla o nível de significância por comparação α_c, enquanto que os testes de Tukey e SNK controlam o nível de significância por experimento α_e. Os testes de Duncan e DMS de Fisher, por sua vez, controlam o nível de significância por experimento condicional α_3. Já o procedimento de Scheffé mostrou não controlar nenhuma das taxas de erro consideradas.

Além disso, verificou-se que a aplicação dos testes de médias apenas mediante um resultado preliminar significativo do teste-F ou não pode alterar os níveis de significância por comparação e por experimento e, portanto, isso deve ser considerado no momento da determinação das taxas de erro.

Referências Bibliográficas

BERNHARDSON, C.S. 1975. Type I error rates when multiple comparison procedures follow a significant *F* test of ANOVA. Biometrics 31: 229-232.

BOARDMAN, T.J.; MOFFITT, D.R. 1971. Graphical Monte Carlo Type I error rates for multiple comparison procedures. Biometrics 27: 738-744.

CARDELLINO, R.A.; SIEWERDT, F. 1992. Utilização correta e incorreta dos testes de comparação de médias. Revista da Sociedade Brasileira de Zootecnia 21: 985-995.

GIRARDI, L.H.; CARGNELUTTI Filho, A.; STORCK, L. 2009. Erro tipo I e poder de cinco testes de comparação múltipla de médias. Revista Brasileira de Biometria 27: 23-36.

HSU, J.C. 1996. Multiple Comparisons: theory and methods. Chapman & Hall, London, UK, ENG.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0, URL http://www.R-project.org/.