Avaliação de testes estatísticos de comparações múltiplas de médias

Clayton Albuquerque de Sousa[1], Mario Andrade Lira Junior[2], Rinaldo Luiz Caraciolo Ferreira [3]*

RESUMO

O objetivo deste trabalho foi verificar a eficiência e uniformidade das respostas de diferentes testes de comparações múltiplas entre médias, bem como a taxa de ocorrência do erro tipo I. Quatro diferentes testes de comparações múltiplas de médias foram aplicados a 200 experimentos e a 10 variáveis, criadas com valores aleatórios, para avaliação do número de níveis; concordância com o teste F; taxa de ocorrências do erro tipo I por variável e por experimento. Os testes diferiram em seus resultados para todas as variáveis analisadas. As maiores divergências de resultados foram encontradas quando se utilizou o teste LSD. Quanto ao nível de significância adotado, maior concordância foi encontrada por meio do teste de Tukey.

Palavras-chave: Duncan, Dunnet, erro tipo I, LSD, Tukey.

ABSTRACT

Evaluation of multiple mean comparison tests

The objective was to verify the efficiency and uniformity of responses from different tests to compare means and the rate of occurrence of Type I error. Four different tests for multiple comparison of means were applied to 200 experiments and 10 variables created with random values, to assess the number of levels; agreement with the F test and rate of occurrence of type I error per variable and per experiment. The tests differed in their results for all variables. The greatest divergences in results were found when using the LSD test. For the level of significance, the greatest agreement was found by the Tukey test.

Key words: type I error, Tukey, Duncan, Dunnet, LSD.

Recebido para publicação em 30/12/2011 e aprovado em 01/06/2012

[1] Engenheiro-Agrônomo, Mestre. Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manoel de Medeiros, s/n, Dois Irmãos, 52171-900, Recife, Pernambuco, Brasil. clayton.adesousa@gmail.com

[2] Engenheiro-Agrônomo, Doutor. Bosista do CNPq. Departamento de Agronomia, Universidade Federal Rural de Pernambuco, Rua Dom Manoel de Medeiros, s/n, Dois Irmãos, 52171-900, Recife, Pernambuco, Brasil. mario.lira@depa.ufrpe.br

[3] Engenheiro Florestal, Doutor. Bosista do CNPq. Departamento de Ciência Florestal, Universidade Federal Rural de Pernambuco, Rua Dom Manoel de Medeiros, s/n, Dois Irmãos, 52171-900, Recife, Pernambuco, Brasil. rinaldo@dcfl.ufrpe.br

INTRODUÇÃO

Na pesquisa científica de diversas áreas do conhecimento, as análises estatísticas são utilizadas como ferramentas para avaliação de suas hipóteses. A comparação múltipla entre médias de tratamentos experimentais, precedida de análise de variância, é, sem dúvida, uma das mais comuns dessas análises (Araújo, 2003; Cargnelutti Filho *et al*., 2003; Couto *et al*., 2009).

Os testes de comparações múltiplas entre médias de tratamentos são de grande interesse na pesquisa aplicada (Conagin *et al*., 2008), quando o objetivo do trabalho é comparar tratamentos qualitativos. A utilização de um teste é realizada quando a análise de variância detecta existência de efeito significativo dos tratamentos a um determinado nível de significância, de modo que se decide rejeitar a hipótese de nulidade (pelo menos um contraste ortogonal entre tratamentos diferentes de zero).

Dentre os testes mais comuns para comprovação dos resultados experimentais, estão o Teste F, utilizado para verificar se existe diferença significativa entre contrastes ortogonais dos tratamentos, e os testes de Tukey, Duncan, Dunnet e o teste LSD, normalmente utilizados para detalhar esta informação, permitindo mostrar, especificadamente, quais tratamentos diferem, ou não, estatisticamente.

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 estes erros são controlados. Vale ressaltar que se deve procurar um teste que seja robusto frente a algumas violações dos pressupostos básicos para sua aplicação, tais como normalidade, homocedasticidade e independência de erros (Machado *et al*. 2005).

Uma característica desejável é que os testes apresentem controle da taxa de erro tipo I, que afirmem haver diferença entre os tratamentos, quando, na realidade, as diferenças são causadas pelo acaso. A probabilidade máxima de se rejeitar erroneamente uma hipótese nula (H_0) é determinada pelo nível de significância do teste e fundamenta o erro tipo I, sendo comumente utilizado o nível de 5% de probabilidade.

O comportamento de alguns testes estatísticos e seu desempenho em termos de taxa de erro tipo I tem sido avaliado, por exemplo, por Hsu (1996), Borges & Ferreira (2003), Conagin & Barbin (2006a,b), Conagin *et al*. (2008), dentre outros, mas ainda existem muitas perguntas a serem respondidas neste campo da investigação (Bretz *et al*., 2008).

Há muitas dificuldades em se comparar o erro tipo I nos procedimentos de comparações múltiplas (Carmer & Swanson, 1971). A razão entre o número de erros tipo I e o número de comparações é definida como taxa de erro por comparação, chamada de "comparisonwise", e a razão entre o número de experimentos com um ou mais erros tipo I e o número total de experimentos é definida como taxa de erro por experimento, chamada de "experimentwise" (Steel & Torrie, 1980). Espera-se que os testes de comparações múltiplas de médias controlem a taxa de ocorrência do erro tipo I com a mesma eficiência, tanto para comparações, como para experimentos. Alguns testes de comparações múltiplas controlam a taxa de erro tipo I para comparações, enquanto outros controlam a taxa de erro tipo I para experimentos (Vieira, 2006).

Resultados diferentes de taxas de ocorrência do erro tipo I também têm sido registrados quando se avaliam diferentes tamanhos de experimentos, de acordo com o número de tratamentos, repetições e coeficientes de variação (Conagin, 1998, 1999; Conagin *et al*., 1999 e Conagin & Gomes, 2004).

O objetivo deste trabalho foi verificar a eficiência e uniformidade das respostas de diferentes testes de comparações múltiplas entre médias, bem como a taxa de ocorrência do erro tipo I.

MATERIAL E MÉTODOS

Foram gerados dados para 200 experimentos conduzidos em delineamento inteiramente casualizado, composto de 32 tratamentos, cada experimento, com quatro repetições e dez variáveis, totalizando 2000 aplicações dos testes. Os tratamentos hipotéticos foram construídos com base em um experimento com um tratamento controle, para fim de designação de testemunha para o teste de Dunnett. O número de tratamentos e variáveis foi baseado em experimentos de algumas áreas das ciências agrárias que, normalmente, adotam estas quantidades, como, por exemplo, microbiologia do solo ou melhoramento vegetal.

Os dados foram gerados, utilizando-se a função de criação de dados aleatórios do Microsoft Excel 2007® para que os tratamentos, *a priori*, não apresentassem diferença significativa entre si. Mediante análise de variância, os testes utilizados para avaliação dos experimentos foram: o teste F, o teste de Tukey (procedimento means), o teste de Duncan, o teste de Dunnet e o teste LSD, todos aplicados utilizando-se o procedimento GLM do SAS (SAS Institute, 1999).

Foi determinada a quantidade total de resultados significativos e não significativos, em cada teste, e submetidos à análise pelo teste do qui-quadrado a 5% de probabilidade, para testar a hipótese de que não havia diferença entre a quantidade de resultados significativos obtidos e a quantidade de resultados significativos que era esperada em cada teste. Esta análise permitiu verificar a adequação das respostas obtidas dos testes, quanto ao nível de significância usado.

Além disso, foram avaliadas as seguintes variáveis: 1-número de níveis dos testes, determinado pelo número de

grupos formados por tratamentos que não diferiram entre si em cada variável ; 2- coeficiente de concordância dos demais testes com o teste F (CCF), determinado pela comparação em cada caso do resultado do teste F com o resultado de cada teste de comparação de médias, atribuindo-se valor 0 para o caso de apresentarem resultados concordantes e valor 1, no caso de resultados opostos (para essas duas variáveis não se realizou a análise para o teste F); 3- taxa de ocorrências do erro tipo I por variável (TIV), determinada pela atribuição, em cada caso, de valor 0 para as ocorrências de resultados não significativos e de valor 1, no caso de resultados significativos; e 4- taxa de ocorrências do erro tipo I por experimento (TIE), determinada pela atribuição, em cada caso, de valor 0, para os experimentos com resultados não significativos, e valor 1, no caso de resultados significativos. Como os dados foram criados por procedimento aleatório, para a determinação da TIV foi utilizado o conjunto total dos 2000 casos (200 experimentos com 10 variáveis cada) e, para determinação da TIE, o conjunto dos 200 experimentos.

Todas as análises estatísticas foram realizadas por meio do *software* SAS Learning Edition 4.1® (SAS Institute, 2006) a 5% de probabilidade.

RESULTADOS E DISCUSSÃO

Considerando-se todos os testes, a análise demonstrou comportamento divergente com relação à apresentação de diferença significativa para o conjunto dos 2000 casos. O teste F e o teste de Tukey apresentaram elevado número de avaliações com resultados não significativos, mais de 1900 cada teste, o que demonstra um comportamento coerente com o conjunto de dados construído, que teve como hipótese não haver diferença entre tratamentos (Tabela 1). Considerando-se que todos os tratamentos eram formados por números aleatórios, e adotando-se o nível de significância usual de 5%, podería esperar um máximo de 100 ocorrências de significância. Os dois testes adequaram-se a esta probabilidade máxima.

O teste de Dunnet não obteve resultados significativamente diferentes do esperado, de acordo com o teste do qui-quadrado, apesar de, em valores absolutos, não se adequar ao nível máximo de 100 ocorrências de resultados significativos.

Já os testes LSD e Duncan demonstraram um número de resultados significativos superior ao número de não significativos, com, respectivamente, 1942 e 1542, de um total de 2000 avaliações sendo significativas (97,1 e 77,1% do total, respectivamente). Esse resultado mostra que estes testes induzem a maior ocorrência do erro tipo I, do que o nível de significância pré-definido, com relação aos outros testes avaliados.

Pelo teste do qui-quadrado, os testes de Duncan e LSD demonstraram diferenças entre o número de observações que teriam resultado não significativo (dados gerados aleatoriamente) e o que realmente aconteceu com as comparações das médias (resultados observados).

A análise de variância das variáveis número de níveis dos testes de comparação múltipla de médias, coeficiente de concordância dos testes de médias com o teste F, taxas de ocorrência do erro tipo I por variável e por experimento demonstrou resultados altamente significativos sugerindo que, entre os testes de comparações múltiplas analisados, existe diferença na capacidade de verificar semelhanças entre tratamentos e, consequentemente, diferentes capacidades de controle do erro tipo I (Tabela 2).

O teste de médias que mais distinguiu os tratamentos, separando-os em mais níveis, foi o LSD, demonstrando ser um teste menos rigoroso, possuindo maior incidência de erros tipo I, ou seja, encontrar diferenças entre tratamentos quando esta diferença não existe (Tabela 3). O teste de Tukey e o teste de Dunnet confirmaram os resultados encontrados no teste de qui-quadrado, não apresentando taxa de erro do tipo I superior à esperada.

O teste LSD foi o que mais divergiu do teste F. Numa escala binária, em que o coeficiente 0 significa que o resultado do teste de comparação de médias concorda com o teste F, ou seja, os dois testes acusaram resultados significativos ou os dois testes resultaram não significativos, e o coeficiente 1, que os resultados do teste de comparação de médias e o teste F são opostos, o LSD obteve um coeficiente médio de 0,972, demonstrando que a maioria absoluta dos resultados foi divergente do teste F (Ta-

Tabela 1. Número de casos com resultados significativos e não significativos demonstrados pelos testes de significância, e teste do qui-quadrado em um conjunto de 2000 análises de dados aleatórios

Testes de significância	Número de casos		*Probabilidade > qui-quadrado**
	Significativo	Não significativo	
F	90	1910	0,4818
Tukey	98	1902	0,4424
Dunnet	184	1816	0,1855
LSD	1942	58	<0,0001
Duncan	1542	458	<0,0001

*Para 100 casos esperados como significativos, a 5% de probabilidade.

bela 3). Vale ressaltar o desempenho também do teste de Duncan, que obteve coeficiente de 0,771, obtendo também alto índice de discordância com os resultados do teste F. Os outros testes obtiveram coeficientes menores que 0,1, quando comparados com o teste F, demonstrando pouca divergência.

O teste F é considerado um teste robusto, pouco sensível a problemas de não adequação dos dados aos pré-requisitos da análise de variância (Moore, 2000). Como complemento do teste F, deseja-se que um teste de comparações múltiplas de médias tenha a mesma consistência nos resultados, concordando com os demonstrados pelo teste F.

A maior taxa de ocorrência do erro tipo I por variável foi encontrada no teste LSD, que obteve um coeficiente de 0,972, ou seja, em quase 100% dos casos analisados pelo teste LSD houve rejeição da hipótese nula, que afirmava não haver diferença entre os tratamentos, quando esta deveria ser aceita. Já nos testes F e Tukey, a taxa de ocorrência do erro tipo I foi de 0,045 e 0,049, respectivamente (Tabela 4). O resultado indica controle da taxa de erro do tipo I por esses testes, assim os resultados apresentados por esses dois testes são os que podem ser considerados mais confiáveis, com taxas de ocorrência do erro tipo I dentro da expectativa. Os resultados confirmam que o teste de Tukey pode ser considerado o teste de comparação múltipla de médias com o maior controle da taxa de ocorrência dos erros tipo I sob distribuições normais (Borges & Ferreira, 2003).

Já para a comparação por experimentos, os testes de Duncan e o LSD obtiveram resultado 1,0 para a taxa de ocorrência do erro tipo I, o que indica ausência de controle deste tipo de erro por parte destes testes (Tabela 4). Nos 200 experimentos avaliados, encontraram-se diferenças significativas que foram atribuídas aos tratamentos em uma ou em mais de uma das variáveis analisadas, porém, pela própria definição do erro tipo I, estas diferenças podem não ser devidas aos tratamentos, mas sim devidas à variação do acaso.

A escolha do teste de comparações de médias para controlar o nível de significância para variáveis ou para experimentos torna-se mais importante à medida que aumenta o número de tratamentos. Em experimentos com apenas dois tratamentos, o nível de significância para cada variável e para experimentos é o mesmo, já que cada experimento só permite uma única comparação de médias. No caso de experimentos com grande número de tratamentos, os valores dos dois níveis de significância podem ser muito diferentes (Vieira, 2006).

A menor taxa de ocorrência de erro tipo I foi verificada no teste F, sendo este, consequentemente, o teste que apresentou o melhor controle do erro tipo I, seguido pelo teste de Tukey.

Segundo Vieira (2006), os testes de Tukey e Dunnet são conservadores, pois possuem menor probabilidade de rejeição da hipótese nula, mantendo o nível de significância para experimentos em 5%, enquanto o nível de significância para as comparações de médias for menor que 5% em experimentos com mais de dois tratamentos.

Tabela 3. Número de níveis e coeficiente de concordância com o teste F, demonstrados pelos testes de comparações múltiplas de médias em um conjunto de 2000 análises

Testes de significância	Número de níveis	Coeficiente de concordância com teste F
Tukey	1 c	0,049 a
Dunnet	1 c	0,092 b
LSD	4 a	0,972 d
Duncan	2 b	0,771 c

Médias seguidas de mesma letra, na coluna, não diferem entre si a 5% de probabilidade pelo teste de Tukey.

Tabela 4. Taxas de ocorrência do erro do tipo I por variável (TIV) e taxas de ocorrência do erro do tipo I por experimentos (TIE) no teste F e nos testes de comparações múltiplas de médias em um conjunto de 2000 análises

Testes de significância	TIV	TIE
F	0,045 d	0,29 a
Tukey	0,049 d	0,43 b
Dunnet	0,092 c	0,71 c
LSD	0,972 a	1 d
Duncan	0,771 b	1 d

Médias seguidas de mesma letra, na coluna, não diferem entre si a 5% de probabilidade pelo teste de Tukey.

Tabela 2. Resumo da análise de variância para as variáveis: número de níveis dos testes de comparação múltipla de médias, coeficiente de concordância dos testes de médias com o teste F, taxas de ocorrência do erro tipo I por variável e por experimento em um conjunto de 2000 análises de dados aleatórios

Variáveis	*Pr > F*	*CV (%)*	*R^2*
Número de níveis	<0,0001	76	0,39
Coeficiente de concordância com teste F	<0,0001	92	0,61
Taxa de erro tipo I por variáveis	<0,0001	82	0,74
Taxa de erro tipo I por experimento	<0,0001	61	0,30

CV-Coeficiente de Variação; R^2-coeficiente de determinação.

Ao se comparar tais resultados com os obtidos por Bernhardson (1975), pode-se confirmar o caráter conservador do teste de Tukey. Porém, os procedimentos de Duncan e LSD atingiram níveis bem mais elevados que os encontrados pelo autor, que observou taxas de erro tipo I de mais de 40%, tanto para comparações como para experimentos.

Os resultados encontrados para TIV e TIE estão de acordo com todas as situações estudadas por Santos *et al*. (2001), em que a taxa de erro tipo I por variável foi sempre inferior à taxa de erro por experimento.

CONCLUSÕES

O teste de Tukey apresentou maior poder para controle do erro tipo I. A maior ocorrência do erro tipo I foi observada nos testes de Duncan e LSD.

REFERÊNCIAS

Araújo AP (2003) Analysis of variance of primary data on plant growth analysis. Pesquisa Agropecuária Brasileira, 38:1-10.

Bernhardson CS (1975) Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. Biometrics, 31:337-340.

Borges LC & Ferreira DF (2003) Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normais e não normais dos resíduos. Revista de Matemática e Estatística, 21:67-83.

Bretz F, Hothorn T & Westfall P (2008) Multiple comparison procedures in linear models. In: Brito P (ed.) Compstat 2008: Proceedings in Computational Statistics. Berlin, Physica-Verlag HD. p.423-431.

Cargnelutti Filho A, Storck L & Dal'col Lucio A (2003) Ajustes de quadrado médio do erro em ensaios de competição de cultivares de milho pelo método de Papadakis. Pesquisa Agropecuária Brasileira, 38:467-473.

Carmer SG & Swanson MR (1971) Detection of differences between means: a Monte Carlo study of five pairwise multiple comparison procedures. Agronomy Journal Madison, 63:940- 945.

Conagin A (1998) Discriminative power of the modified Bonferroni´s test. Revista de Agricultura, 73:31-46.

Conagin A, Igue T & Nagai V (1999) Poder discriminativo de diferentes testes de médias. Campinas, Instituto Agronômico. 16p. (Boletim Científico, 44).

Conagin A (1999) Discriminative power of the modified Bonferroni's test under general and partial null hypothesis. Revista de Agricultura, 74:117-126.

Conagin A & Gomes FP (2004) Escolha adequada dos testes estatísticos para comparações múltiplas. Revista de Agricultura, 79:288-295.

Conagin A & Barbin D (2006a) Bonferroni's and Sidak's modified tests. Scientia Agricola, 63:70-76.

Conagin A & Barbin D (2006b) Poder e eficiência dos diferentes testes estatísticos para comparações múltiplas. Revista de Agricultura, 81:118-137.

Conagin A, Barbin D & Demétrio CGB (2008) Modifications for the Tukey test procedure and evaluation of the power and efficiency of multiple comparison procederes. Scientia Agricola, 65:428-432.

Couto MRM, Lucio AD, Lopes SJ & Carpes RH (2009) Transformações de dados em experimentos com abobrinha italiana em ambiente protegido. Ciência Rural, 39:1701-1707.

Hsu JC (1996) Multiple comparisons. London, Chapman and Hall. 277p.

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

Moore DS (2000) A estatística básica e sua prática. Rio de Janeiro, LTC-Livros Técnicos e Científicos Editora. 482p.

Santos C, Ferreira DF & Bueno Filho JSS (2001) Novas alternativas de testes de agrupamento avaliadas por meio de simulação Monte Carlo. Ciência e Agrotecnologia, 25:1382-1392.

Sas Institute (1999) The SAS System for Windows. Cary, SAS Institute Inc.

Sas Institute (2006) Statistical Analysis System. SAS Learning Edition 4.1®, SAS Institute Inc.

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

Vieira S (2006) Análise de variância: ANOVA. São Paulo, Atlas. 204p.