Rev. Centro de Ciências Rurais, 12(2-3):163-173, 1982.
Rec. p/publ. em 09/82.

COMPARAÇÃO MÚLTIPLA DE MÉDIAS NA ANÁLISE DA VARIÂNCIA PELOS TESTES TUKEY, DUNCAN E DMS.

Multiple Comparison Among Means in a Analysis of Variance by Tukey, Duncan and LSD Test.

Ione Aydee Bernardes Pignataro*

RESUMO

Técnicas de simulação, em computador, foram usadas para estudar as porcentagens dos erros tipo I e II com três testes para comparação múltipla de médias.

Os testes empregados foram: Tukey, Duncan e DMS com e sem a exigência de um teste F anterior. A simulação foi feita para experimentos com quatro repetições de 6, 10, 15 e 20 tratamentos.

Os resultados mostraram que o teste Tukey apresentou a maior proteção contra o erro tipo I e muito pouca sensibilidade para detectar diferenças entre médias. O DMS sem um teste F preliminar apresentou a menor proteção contro o erro tipo I e a mais alta sensibilidade para detectar diferenças entre médias. O teste Duncan sem teste F preliminar apresentou maior proteção contra o erro tipo I e maior sensibilidade que o DMS com um teste F preliminar.

SUMMARY

Computer simulation techniques were used to study the type I and II error rates for three procedures in pairwise multiple comparison of treatment means.

The procedures studied were: Tukey, Duncan and LSD tests with and without a preliminary F test. Simulation was for experiments with 4 replicates of 6, 10, 15 and 20 treatments.

Results indicated that procedure gave the highest protection against type I error and very poor sensitivity in detecting differences among means. The LSD procedure without a preliminary F test gave the poorest protection against type I error and highest sensitivity for detecting differences among means. The Duncan procedure without preliminary F test gave higher protection against type I error and more sensitivity detecting difference than LSD with a preliminary F test.

INTRODUÇÃO

O uso dos testes para comparação de médias só é apropriado quando os tratamentos não são quantitativos e não é possível, antes do exame dos dados, a partição dos graus de liberdade para formar contrastes ortogonais, como é relembra-

* Professor Adjunto do Departamento de Fitotecnia, Centro de Ciências Rurais da Universidade Federal de Santa Maria. 97.100 - Santa Maria, RS.

do por CHEW (7).

Para essas situações onde os tratamentos não são quantitativos, e não é possível usar os contrastes ortogonais, existe um número grande de testes, sendo os mais conhecidos os testes Tukey, Duncan e DMS que se encontram em muitos textos de métodos estatísticos como FEDERER (11), STEEL & TORRIE (18) e GOMES (14).

Apesar do mais novo destes testes ter sido publicado a quase 30 anos, não há ainda acordo sobre qual o mais apropriado (DUNNETT, 9).

A escolha do teste depende em última análise da importância atribuída a cada tipo de erro e só o experimentador pode avaliar as consequências desses erros (WALDO, 20).

Na comparação entre duas médias há três decisões possíveis: a média Y_1 é igual a média Y_2, a média Y_1 é maior que a média Y_2 ou a média Y_1 é menor que a média Y_2. O acerto ou erro desse decisão depende das médias verdadeiras, que em geral não são conhecidas.

Podem ocorrer três tipos de decisão incorreta (HARTEN, 15). As médias são realmente iguais e a decisão tomada é de que são diferentes. Neste caso comete-se o erro tipo I. O erro tipo II é cometido quando as médias são diferentes e afirma-se que são iguais. Comete-se o erro tipo III quando se toma uma decisão inversa a real, a média Y_1 é maior que a Y_2 e toma-se a decisão de que Y_2 é maior que Y_1.

De acordo com CARMER & SWANSON (5) muitos experimentadores na hora de escolher o teste a ser empregado ficam inseguros, pois desconhecem as propriedades dos mesmos, o que não admira pois não há concordância nem entre os estatísticos.

Os testes Duncan e DMS com algumas restrições são recomendados por CARMER & SWANSON (5, 6), KEMP (17) e BALAAM (1), enquanto BOARDMAN & MOFFIT (3) consideram estes testes muito liberais. GILL (3) aconselha que o DMS e o Duncan não devem mais ser usados e recomenda o teste Tukey. EINOT & GABRIEL (10) também recomendam o teste Tukey. CHEW (7) considera que a última palavra na escolha dos testes para comparação múltipla não foi dada e recomenda para reduzir as objeções ao teste Duncan a exigência de um F significativo antes, enquanto que BERNHARDSON (2) considera que as proporções de erro ficam modificadas conforme se exige ou não um F significativo para o emprego dos mesmos. DUNNETT (9) considera que em alguns casos o teste F significativo pode ser aconselhado mas não deve ser usado como rotina, pois este procedimento modifica as propriedades dos testes e seu efeito ainda não é bem conhecido.

Na discussão dos testes de comparação múltipla FEDERER (12), STEEL (19) e DUNNETT (9) deram maior ênfase ao erro tipo I, enquanto CHEW (7) considera que em experimentos agrícolas o erro tipo II é no mínimo tão importante quanto o erro tipo I.

O erro tipo II é mais difícil de ser estudado porque depende do tamanho relativo da diferença entre as médias e do número de médias em comparação. No presente trabalho se estudará o comportamento dos testes mais comuns Tukey, Dun-

can e DMS para comparação múltipla de médias com e sem a exigência de F significativo determinando a porcentagem de erro tipo I e II cometidos.

MATERIAL E MÉTODOS

Os testes estudados foram os testes Tukey e DMS como estão descritos por STEEL & TORRIE (18) e teste Duncan descrito por DUNCAN (8).

Os experimentos foram feitos por simulação, para o que foi escrito um programa para o computador IBM 360 que gerou os dados e analisou cada experimento.

Os valores foram gerados usando-se o método de MARSAGLIA & BRAY (16)e transformados em variações independentes com distribuição normal usando o método de BOX & MULLER (4).

Foram feitos 500 experimentos, cada experimento com 4 repetições, para cada um dos 12 grupos de médias verdadeiras mostradas na Tabela 1, onde o número de tratamentos variou de 6 a 20.

Os dados simulados para cada experimento foram gerados a partir do modelo linear

$$Y_{ij} = M + T_i + E_{ij}$$

onde Y_{ij} é a observação simulada para a repetição j do tratamento i; M é a média geral do experimento que foi considerada igual a 100. T_i representa o efeito do tratamento i e é a diferença entre a média geral e a média verdadeira do tratamento, uma vez que a soma dos T_i foi feita igual a zero. E_{ij} é o desvio aleatório amostrado independentemente para cada observação, com distribuição normal, média zero e desvio padrão igual a 10, o que dá um coeficiente de variação verdadeiro de 10% para os experimentos.

Os testes foram aplicados com $\alpha = 0,05$ e $\alpha = 0,01$, com e sem a exigência de um F para tratamentos com $P < 0,05$. A decisão obtida para cada comparação entre médias foi classificada em correta, erro tipo I, II ou III.

Foi determinado, para cada tipo de erro, a porcentagem de erro cometido tendo como base o número possível de erros do tipo considerado.

Para o erro tipo II foi determinado a porcentagem de erro para as seguintes diferenças reais entre médias verdadeiras: 0,5; 1,0; 1,5; 2,0; 2,5; 3,0 e 3,5 desvios padrão.

RESULTADOS OBTIDOS E DISCUSSÃO

Os resultados que seguem foram obtidos usando 4 repetições, o número de repetições mais comum nos experimentos. A mudança do número de repetições modificará os resultados.

A porcentagem de erro tipo I apresentou influência do teste empregado, da exigência ou não de um F significativo antes do teste e do grupo de experimento (Tabela 2).

TABELA 1. Conjunto de médias verdadeiras para tratamentos usados no estudo de simulação.

NÚMERO DO GRUPO	NÚMERO DE MÉDIAS	MÉDIA DOS TRATAMENTOS									
1	6	100	100	100	100	100	100				
2	6	90	100	100	100	100	110				
3	6	85	90	95	100	110	120				
4	10	100	100	100	100	100	100	100	100	100	100
5	10	90	100	100	100	100	100	100	100	100	110
6	10	85	90	90	95	100	100	100	110	110	120
7	15	100	100	100	100	100	100	100	100	100	100
		100	100	100	100	100					
8	15	90	100	100	100	100	100	100	100	100	100
		100	100	100	100	110					
9	15	85	85	90	90	95	97	98	100	100	100
		105	105	110	120	120					
10	20	100	100	100	100	100	100	100	100	100	100
		100	100	100	100	100	100	100	100	100	100
11	20	90	100	100	100	100	100	100	100	100	100
		100	100	100	100	100	100	100	100	100	100
12	20	85	85	90	90	90	90	95	95	100	100
		100	100	100	100	110	110	110	110	120	120

TABELA 2. Porcentagem de erro tipo I usando os testes de Tukey, Duncan e DMS para comparação múltipla de médias.

Nº TRATAMENTOS	GRUPO	$\alpha = 0,05$						$\alpha = 0,01$						Experimentos F com P < 0,05 (Porcentagem)
		Tukey		Duncan		DMS		Tukey		Duncan		DMS		
		Sem F*	Com F**	Sem F	Com F	Sem F	Com F	Sem F	Com F	Sem F	Com F	Sem F	Com F	
6	1	0,15	0,11	0,97	0,17	2,45	0,39	0,027	0,027	0,24	0,19	0,37	0,16	1,4
	2	0,13	0,13	0,70	0,63	2,40	2,24	0,000	0,000	0,37	0,37	0,33	0,33	41,0
10	4	0,018	0,0044	1,04	0,13	2,56	0,16	0,000	0,000	0,07	0,027	0,29	0,058	0,6
	5	0,043	0,042	1,09	0,85	2,57	1,74	0,0071	0,0071	0,07	0,06	0,28	0,25	33,3
	6	0,08	0,08	2,00	2,00	3,20	3,20	0,000	0,000	0,20	0,20	0,52	0,52	100,0
15	7	0,0095	0,038	0,17	0,017	2,50	0,34	0,000	0,000	0,11	0,015	0,32	0,14	0,4
	8	0,010	0,010	0,24	0,15	2,58	1,39	0,000	0,000	0,15	0,11	0,36	0,26	23,2
	9	0,000	0,000	0,46	0,46	2,54	2,54	0,000	0,000	0,17	0,17	0,25	0,25	100,0
20	10	0,0063	0,0052	0,71	0,065	2,52	0,12	0,0042	0,0042	0,047	0,012	0,36	0,041	0,8
	11	0,0065	0,0052	0,73	0,31	2,49	0,74	0,000	0,000	0,057	0,037	0,33	0,16	12,4
	12	0,000	0,000	0,987	0,987	2,42	2,42	0,000	0,000	0,060	0,060	0,253	0,25	100,0

* Sem exigência de F com P < 0,05.

**Com exigência de F com P < 0,05.

A exigência do F significativo antes do teste reduziu o erro tipo I nos grupos de experimento em que as médias eram iguais (grupos 1, 4, 7 e 10) e em menor proporção quando apenas duas médias eram diferentes das demais (grupos 2, 5, 8 e 11). Os grupos onde havia diversas médias verdadeiras diferentes apresentaram 100% dos experimentos com F significativo nos grupos 6, 9 e 12 e 99,4% no grupo 3.

Usando-se α = 0,01 todos os testes apresentaram pequena porcentagem de erro tipo I, o que apóia a sugestão de CHEW (7) de se usar níveis de significância mais rígidos para reduzir o erro tipo I.

Com α = 0,05 o teste Tukey foi o que apresentou menor porcentagem de erro tipo I, exigindo-se ou não o F significativo, mostrando não haver necessidade da exigência de F significativo para o uso do teste Tukey.

A baixa porcentagem de erro tipo I com o teste Tukey explica a escolha deste teste por autores que se preocupam especialmente com o erro tipo I (EINOT & GABRIEL, 10; GILL, 13).

O teste Duncan apresentou resultados intermediários entre o Tukey e o DMS.

A exigência de F significativo antes do teste reduziu a porcentagem do erro principalmente nos grupos onde todas as médias verdadeiras eram iguais.

Estes resultados apóiam só em parte a sugestão de CHEW (7) de se usar a exigência de F significativo antes do teste Duncan para reduzir o erro tipo I, uma vez que isto só ocorre nos grupos de experimentos em que as médias verdadeiras são iguais ou semelhantes.

O DMS foi que apresentou maior porcentagem de erro tipo I, 3,2% quando não se exigiu o F significativo. A redução deste erro pela exigência do F significativo foi observado nos grupos de experimentos com todas as médias verdadeiras iguais.

A porcentagem de experimentos com um ou mais erros tipo I (Tabela 3) foi pequena no teste Tukey e elevada nos testes Duncan e DMS. Nos experimentos com todas as médias iguais a exigência do F significativo antes do teste reduziu esta porcentagem a valores menores que 1,5%, já nos outros grupos de experimentos esta redução foi menor. A porcentagem de erro tipo II para as diversas diferenças verdadeiras entre médias reais foram obtidas com os grupos de experimentos 3, 6, 9 e 12 onde havia uma grande variação entre médias verdadeiras (Tabela 4).

Nestes grupos não houve a influência da exigência ou não de F significativo uma vez que só o grupo 3 não tinha 100% dos experimentos com F significativo.

As diferenças reais entre médias verdadeiras apontadas pelos testes, com 5% ou menos de erro, foram as seguintes (em desvio padrão): Tukey 3,5%; Duncan 3,0 e DMS 2,5 independente do número de tratamentos usados. Para diferenças iguais ou menores que 2,5 desvios padrão para o teste Tukey, 2,0 para Duncan e 1,5 para o DMS, a porcentagem de erro foi de 30% ou mais, quando se usou α = 0,05. Usando α = 0,01 a porcentagem de erro tipo II foi bem maior.

As porcentagens de erro tipo II mostradas nos grupos 2, 5, 8 e 11 onde as médias verdadeiras eram iguais, com exceção da primeira e última, se encontram na Tabela 5. Nestes grupos havia apenas dois tamanhos de diferença real entre mé-

TABELA 3. Porcentagem de experimentos com um ou mais erro tipo I.

Nº TRATAMENTOS	GRUPO	$\alpha = 0,05$						$\alpha = 0,01$					
		Tukey		Duncan		DMS		Tukey		Duncan		DMS	
		Sem F*	Com F**	Sem F	Com F	Sem F	Com F	Sem F	Com F	Sem F	Com F	Sem F	Com F
6	1	1,8	1,2	13,4	1,4	18,6	1,4	0,4	0,4	1,6	1,2	4,6	1,4
	2	0,8	0,8	3,2	2,8	10,6	9,6	0,00	0,00	2,2	2,2	1,6	1,6
10	4	0,8	0,2	20,6	0,6	41,8	0,6	0,00	0,00	2,0	0,6	8,2	0,6
	5	1,0	1,0	17,6	12,4	32,4	19,2	0,2	0,2	1,6	1,4	5,6	5,0
	6	0,4	0,4	8,8	8,8	13,8	13,8	0,00	0,00	1,0	1,0	2,6	2,6
15	7	0,6	0,4	6,2	0,4	58,6	0,4	0,00	0,00	4,6	0,4	15,4	0,4
	8	0,6	0,6	13,0	7,4	53,6	20,2	0,00	0,00	9,8	6,6	12,2	8,4
	9	0,00	0,00	3,2	3,2	15,4	15,4	0,00	0,00	1,2	1,2	1,8	1,8
20	10	0,4	0,2	37,0	0,8	73,4	0,8	0,2	0,2	5,2	0,4	24,0	0,8
	11	0,8	0,6	35,8	9,2	67,0	11,2	0,00	0,00	5,0	2,8	21,2	7,0
	12	0,00	0,00	21,6	21,6	43,0	43,0	0,00	0,00	1,6	1,6	6,6	6,6

* Sem exigência de F com $P < 0,05$

** Com exigência de F com $P < 0,05$

TABELA 4. Porcentagem de erro tipo II encontrado nos grupos 3, 6, 9 e 12 de experimentos, considerando-se as diversas amplitudes de diferença real entre médias verdadeiras, usando os testes Tukey, Duncan e DMS.

Nº TRATAMENTOS	GRUPO	α = 0,05							α = 0,01						
		Diferença verdadeira entre tratamentos expressa em desvios padroes*													
		0,5	1,0	1,5	2,0	2,5	3,0	3,5	0,5	1,0	1,5	2,0	2,5	3,0	3,5
							TUKEY								
6	3	99,0	93,6	80,2	53,8	26,6	9,8	2,6	99,0	93,6	80,2	53,8	26,6	9,8	2,6
10	6	99,2	97,4	88,4	66,4	36,0	13,7	3,8	99,9	99,6	96,5	86,4	63,3	32,4	12,6
15	9	99,9	98,7	91,8	71,1	40,3	17,3	4,0	99,99	99,7	97,7	88,1	64,8	36,1	12,8
20	12	99,9	99,4	94,0	76,6	46,3	20,1	4,8	99,98	99,9	98,3	90,4	69,2	38,7	13,9
							DUNCAN								
6	3	96,2	85,0	60,5	35,6	14,5	5,4	1,0	97,2	90,2	72,5	39,2	18,6	6,8	2,8
10	6	94,6	78,1	50,1	20,5	6,7	0,9	0,2	98,8	94,2	79,5	52,4	23,6	6,7	1,2
15	9	98,3	91,4	53,5	41,2	16,2	4,6	1,0	99,1	94,6	77,9	49,5	21,9	7,4	1,8
20	12	95,9	81,2	53,1	24,4	6,2	0,7	0,1	99,4	95,5	81,6	53,0	23,8	7,2	1,2
							DMS								
6	3	91,2	71,7	39,3	15,4	4,1	8,8	0,0	98,0	90,0	70,6	41,5	16,4	4,6	1,4
10	6	92,0	71,9	40,3	14,4	3,4	0,7	0,2	98,1	90,8	69,9	39,0	14,5	2,4	0,2
15	9	91,6	71,4	38,6	13,3	2,4	0,5	0,0	98,3	90,3	67,3	35,8	11,9	2,7	0,5
20	12	91,0	70,5	37,3	12,6	2,6	0,2	0,0	98,1	89,5	66,2	34,1	11,1	2,1	0,2

*(Diferença entre duas médias verdadeiras)/(verdadeiro desvio padrão = 10).

TABELA 5. Porcentagem erro tipo II encontrado nos grupos de experimentos 2, 5, 8 e 11, usando os testes Tukey, Duncan e DMS, com e sem exigência de F significativo.

DIFERENÇA REAL ENTRE TRATAMENTOS	Nº TRATAMENTO	GRUPO	$\alpha = 0,05$						$\alpha = 0,01$					
			Tukey		Duncan		DMS		Tukey		Duncan		DMS	
			Sem F*	Com F**	Sem F	Com F	Sem F	Com F	Sem F	Com F	Sem F	Com F	Sem F	Com F
2	6	2	57,4	63,8	30,2	62,0	17,0	54,2	79,8	79,8	57,2	62,8	46,2	61,2
	10	5	64,4	73,0	21,2	66,8	11,2	66,8	83,6	83,8	55,0	69,2	35,6	67,6
	15	8	70,0	81,6	46,0	77,6	14,8	87,3	85,6	87,2	52,6	77,8	37,0	77,0
	20	11	77,8	90,0	27,6	87,8	13,2	87,6	90,4	93,0	56,6	88,2	37,4	87,8
1	6	2	94,9	95,0	90,2	91,5	73,3	80,4	98,8	98,8	89,8	92,6	92,1	92,5
	10	5	97,0	97,1	78,9	85,8	70,1	82,7	98,3	94,4	94,4	94,9	89,7	91,5
	15	8	98,3	98,5	91,6	94,9	69,8	93,9	99,6	99,6	93,9	95,8	89,0	93,3
	20	11	99,1	99,2	83,7	94,4	70,7	92,7	99,7	99,7	96,5	97,8	90,2	94,9

* Sem exigência de F com $P < 0,05$.

** Com exigência de F com $P < 0,05$

dias verdadeiras, um e dois desvios padrão. A exigência de F significativo para aplicação do teste aumentou bastante a porcentagem do erro tipo II, pois na maioria dos experimentos destes grupos não se encontrou F significativo. A porcentagem de erro tipo II usando o DMS com exigência de F significativo foi maior que a encontrada para o teste Duncan sem esta exigência.

A exigência de um F significativo antes da aplicação dos testes Duncan e DMS reduz o erro tipo I nos grupos com médias semelhantes, mas ao mesmo tempo aumenta a porcentagem do erro tipo II, tornando difícil a diferenciação de médias que tenham uma diferença real de até dois desvios padrão em grupos de experimentos com médias semelhantes.

A porcentagem de erro tipo III foi pequena, variou de zero a 0,13% o que concorda com os valores obtidos por CARMER & SWANSON (5, 6). Devido a pequena ocorrência deste erro não foi estudada a porcentagem levando em conta o tamanho da diferença real entre duas médias verdadeiras, pois esta porcentagem é tanto maior quanto menor for esta diferença.

CONCLUSÕES

Os resultados obtidos permitem as seguintes conclusões, para os grupos de tratamentos estudados:

1. A exigência do F significativo antes da aplicação dos testes diminui o erro tipo I e aumenta o erro tipo II e a amplitude entre médias que o teste é capaz de detectar.

2. O teste Tukey dá maior proteção contra o erro tipo I e maior porcentagem do erro tipo II.

3. Não há necessidade da exigência do F significativo antes da aplicação do teste Tukey.

4. O DMS é o teste que apresentou menor porcentagem de erro tipo II e maior porcentagem de erro tipo I. Os resultados são bastante modificados pela exigência de F significativo antes do teste.

5. O teste Duncan quando usado sem exigência de F significativo antes da aplicação apresenta melhores resultados que o DMS com a exigência de F significativo.

LITERATURA CITADA

1. BALAAM, L. N. Multiple comparisons-a sampling experiment. *Aust. J. Stat.*, 5:62-84, 1963.

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

3. BOARDMAN, T. J. & MOFFIT, D. R. Graphical Monte Carlo type I error rates for multiple comparison procedures. *Biometrics*, 27:738-744, 1971.

4. BOX, G. E. P. & MULLER, M. E. A note on the generation of randon normal deviates. *Ann. Math. Stat.*, 29:610-611, 1958.

5. CARMER, S. G. & SWANSON, M. R. Detetection of diferences between means. A Monte Carlo study of five pairwise multiple comparison procedures. *Agron. J.*, 71:940-945, 1971.

6. CARMER, S. G. & SWANSON, M. R. An evaluation of ten pairwise comparison procedures by Monte Carlos methods. *J. Amer. Stat. Assoc.*, 68:66-74, 1973.

7. CHEW, V. Comparing treatment means. A compendium. *Hort Science*, 11:348-356, 1976.

8. DUNCAN, D. B. Multiple range and multiple F test. *Biometrics*, 11:1-42. 1955.

9. DUNNETT, C. W. Query 272. Multiple comparison tests. *Biometrics*, 26:139-141, 1970.

10. EINOT, I. & GABRIEL, K. R. A study of the power os several methods of multiple comparisons. *J. Amer. Stat. Assoc.*, 70:574-583, 1975.

11. FEDERER, W. T. *Experimental design theory and aplication*. New York, Mac Millan, 1955. 544 p.

12. FEDERER, W. T. Experimental error rates. *Proc. Amer. Soc. Hort. Sci.*, 56: 973-979, 1961.

13. GILL, J. L. Current status of multiple comparison means in designed experiments. *J. Diary Sci.*, 56:973-977, 1973.

14. GOMES, F. P. *Curso de Estatística Experimental*, 6ª ed., São Paulo, Livraria Nobel, 1976. 430 p.

15. HARTER, H. L. Error rates and sample size for range test in multiple comparisons. *Biometrics*, 13:511-536, 1957.

16. MARSAGLIA, G. & BRAY, T. A. One-line random number generators and their use in combinations. *Comm. Assoc. Comp. Mach.*, 11:757-759, . 1968.

17. KEMP, K. E. Multiple comparisons: comparison wise versus experiment type I error rates and their relationship to power. *J. Dairy Sci.*, 58:1373-1378, 1975.

18. STEEL, R. G. D. & TORRIE, J. H. *Principles and procedures of statistics*. New York, MacGraw Hill, 1960. 481 p.

19. STEEL, R. G. D. Query 163. Error rates in multiple comparisons. *Biometrics*, 17:326-328, 1961.

20. WALDO, D. R. An evaluation of multiple comparison procedures. *J. Animal Sci.*, 42:539-544, 1976.