ESCOLHA ADEQUADA DOS TESTES ESTATÍSTICOS PARA COMPARAÇÕES MÚLTIPLAS

Armando Conagin [1]
F. Pimentel – Gomes [2]

RESUMO

Para avaliar o poder discriminativo dos testes estatísticos mais comuns, foram gerados, pelo método de Monte Carlo, seiscentos experimentos com 4 e 8 repetições e 8 tratamentos em blocos ao acaso; foram adotados o coeficiente de variação de 10% e o nível α de 5%. Pelo exame da Tabela I, verifica-se uma concordância de natureza geral para todos os testes efetuados; o poder discriminativo dos vários testes será tanto maior quanto maior for a diferença introduzida no processo. Entretanto, a eficiência dos testes na avaliação do poder discriminativo não é a mesma, para uma mesma diferença. A partir da Tabela I é possível verificar que os testes *t* de Student unilateral, *t* de Student bilateral e Waller exibem o poder discriminativo maior que os demais testes utilizados. Logo a seguir aparece o teste de Duncan, seguido dos de Dunnettu, Bonferroni Modificado, Bonferroni II, SNK, REGWF, REGWQ, Tukey, Sidak e Bonferroni I, a serem definidos. Comparações do tipo 7 – 6 e 8 – 6 (caso em que a hipótese H_o é verdadeira) que correspondem, na tabela, às colunas 0% permitem uma análise do comportamento de todos os testes em termos da concordância entre o poder discriminativo observado e o valor nominal do erro adotado para todos os testes, que foi $\alpha = 0,05$. Os resultados indicam que o teste de *t*, unilateral, *t*, bilateral e Waller exibem boa concordância entre o poder discriminativo que, no caso, é do "erro por comparação" e o valor nominal adotado, cinco por cento. O teste de Duncan exibiu valores um pouco menores que cinco por cento; os testes de Dunnettu, Bonferroni Modificado e SNK mostraram valores acima de um por cento; os testes REGWF, REGWQ, Tukey, Sidak e Bonferroni I exibem valores ao redor de um por cento ou menores, indicando a sua natureza conservadora (erro por experimento).

Palavras-Chaves: Comparações múltiplas, testes estatísticos, método de Monte Carlo, erro experimental.

ABSTRACT

SUITABLE SELECTION OF STATISTICAL TESTS FOR MULTIPLE COMPARISONS

In order to evaluate the discriminative power of most known statistical tests were generated by Monte Carlo Method, six hundred experiments in which different number of replications (r=4 and r=8), different values of treatments and different degrees of freedom of error were introduced, maintaining the coefficient of variation in 10 percent and the nominal error of the process in 5 percent ($\alpha = 0,05$).

From the results, one general and expected conclusion arises: that all tests exhibit certain uniformity of behaviour because their discriminative power are always

[1] Engenheiro Agrônomo, Pesquisador Científico VI, (aposentado) Instituto Agronômico, Secretaria de Agricultura e Abastecimento, Campinas, SP.
[2] Engenheiro Agrônomo, Professor Catedrático (aposentado), ESALQ/USP, Piracicaba, SP.

higher and crescent when the differences between the two means compared are of crescent magnitude in the model adopted (randomized blocks); the 8 treatments are so chosen that treatments 1, 2, 3, 4, and 5 are compared with the number six (control) with parametric differences of 30%, 20%, 15%, 10% and 5% and the treatments 6, 7 and 8 are identicals; so, the expected values of 7-6 and 8-6 equal zero percent. From the result (see table I), is possible to see that Student's t unilateral, Student's t bilateral and Waller tests exhibit the highest discriminative power (number of significative difference in percentage) for all differences; Duncan's test follows; in order of values comes next, Dunnettu, Modified Bonferroni II, SNK, REGWF, REGWQ, Tukey, Sidak and Bonferroni I tests. Comparisons of the type 7-6 and 8-6, cases in which H_o is true, the values included in Table I permit the evaluation of the tax of error "comparisonwise" for all the tests (nominal error, $\alpha = 0,05$). The results indicate that *t* unilateral, *t* bilateral and Waller exhibit the discriminative power well around the five percent nominal error. Duncan's test shows a value a little bellow the five percent; the Dunnettu's, Modified Bonferroni's, SNK's, show values above the one percent value; the REGWF, REGWQ, Tukey, Sidak and Bonferroni I exhibit values around the one percent or less, indicating the conservative nature of these tests (experimentwise error).

Key-words: Multiple comparisons, statistical tests, Monte Carlo Method, experimentwise error.

INTRODUÇÃO

Na experimentação agronômica, principalmente na de campo, devido às variações ecológicas (clima, solo), o coeficiente de variação dos experimentos é normalmente alto, bem acima de 10% (Igue, 1974). Já nas análises químicas, nos experimentos de estufa, com vasos, o coeficiente de variação se situa, normalmente, bem abaixo de 10%, ficando mais fácil detectar, estatisticamente, as diferenças entre os tratamentos pesquisados.

Todo teste estatístico é feito levando-se em consideração dois tipos de erro: erro tipo I ou de 1ª espécie que fixa a área de rejeição da hipótese de nulidade (verdadeira) em $\alpha = 0,05$ ou $\alpha = 0,01$; o segundo tipo de erro é o erro II, ou de segunda espécie, que é o que se comete quando se aceita uma hipótese falsa. Este erro é conhecido como erro β.

Vamos supor que a hipótese de nulidade, H_o seja falsa, e a H_a, hipótese alternativa seja verdadeira, e que o resultado do teste esteja situado fora da área de rejeição da hipótese H_o. Então, H_o não vai ser rejeitada e vai-se aceitar H_o como verdadeira. Estar-se-á cometendo um erro do tipo II, (aceitar uma hipótese, no caso a H_o, que é falsa).

A técnica de escolha da área de rejeição da hipótese de nulidade, de valor α, consiste em escolher, dentre as áreas de valor α uma área de rejeição de H_o, de tal forma, que, para α fixado, o valor β seja mínimo, (se H_a for verdadeira).

Modernamente existem dois tipos de erro de tipo I. O erro I, por comparação (*comparison wise*) e o erro I por experimento (*experimentwise*). Se em um experimento em que se comparam 6 tratamentos dois a dois, das 15 comparações possíveis se cometem erradamente cinco rejeições da hipótese de nulidade dentre as comparações efetuadas, o erro por comparação será de 5/15.

Se forem efetuados 3 experimentos com 15 comparações possíveis e em um deles se cometem um ou mais erros de rejeição da hipótese de nulidade, e nenhum erro nos outros dois experimentos, o erro por experimento será de 1/3 (um experimento com

comparações erradas, dividido pelo número total de experimentos feitos). O erro tipo MEER é do tipo *erro por experimento* em condições da hipótese de nulidade generalizada e da hipótese de nulidade parcial.

A avaliação do poder discriminativo dos vários testes estatísticos considerados e mais utilizados é de grande importância, pois possibilita a escolha, pelo pesquisador, do mais adequado para o julgamento das hipóteses existentes na pesquisa. Com esse objetivo foi efetuada a pesquisa atual.

MATERIAL E MÉTODO

A avaliação da eficiência dos vários testes foi bastante simplificada pelo uso do Método de Monte Carlo seguido da análise estatística proporcionada pelos modernos computadores.

Na literatura, os primeiros pesquisadores que trabalharam nesse campo foram: Gabriel (1965), Carmer & Swanson (1971 e 1975), Boardman & Moffitt (1971), Bernardson (1975), Chew (1977), Hotchberg & Tamhane (1974) e outros mais.

No Brasil, Cordeline & Siewerd (1992), Perecin & Barbosa (1988), Conagin, Igue & Nagai (1999), também pesquisaram a eficiência de alguns dos testes existentes.

Certos pesquisadores procuraram avaliar, também, o desempenho dos vários testes com relação aos tipos de erro I por comparação e por experimento.

Além da comparação de médias pelos vários testes, Boardman & Moffitt (1971) e Bernardson (1975) incluíram tratamentos com médias idênticas, o que possibilitou a avaliação também dos erros tipo I por comparação e por experimento, utilizando a taxa de erro nominal fixada em $\alpha = 0,05$. Como resultado das pesquisas, constataram que o teste *t* de Student do tipo *erro por comparação* manteve a freqüência de rejeição da hipótese de nulidade ao redor de cinco por cento para número de tratamentos variando de um a dez.

O teste de Duncan teve porcentagem variável exibindo valores gradualmente menores que cinco por cento, ficando, para dez tratamentos, com o valor próximo de 0,025. Os testes SNK e Tukey exibiram uma freqüência decrescente, chegando perto de um por cento, para dez tratamentos.

Na avaliação do tipo de erro cometido, por experimento, a situação se inverteu. Os testes *t* de Student e Duncan foram apresentando erros maiores e crescentes a partir de 0,05, para atingir, respectivamente , valores próximos de 50 por cento para o teste *t* e 25 por cento para o teste de Duncan, enquanto as freqüências de SNK e Tukey mantiveram-se sempre próximas do valor nominal $\alpha = 0,05$.

Perecin & Barbosa (1988) também estudaram o problema, incluindo, além dos 4 testes citados, os testes SNK modificado e o teste de Waller. A magnitude das diferenças das médias contíguas diferiam por $2\sigma(\bar{x})$, 4 $\sigma(\bar{x})$, 6 $\sigma(\bar{x})$ e 8 $\sigma(\bar{x})$; utilizaram 4 repetições e número de tratamentos, 5, 10, 20, 40 e 100. Concluíram que o aumento do número de tratamentos tem influência no aumento do poder dos testes pesquisados, e que a eficiência na detecção de diferenças foi, na ordem decrescente: Waller, *t*, Duncan, SNK modificado, SNK e Tukey. Concluíram que, para diferenças crescentes entre médias, o poder discriminativo de todos os testes aumentava.

Verificaram que, para diferenças de $2\sigma(\bar{x})$ e 4 $\sigma(\bar{x})$, a eficiência do teste de Tukey se revelou bem inferior ao dos demais testes; foi ainda verificado, que nas condições pesquisadas, o teste Tukey foi fortemente afetado pelo número de tratamentos, e que quanto maior foi o número de tratamentos avaliados, menor foi o poder do teste. Em geral a ordem de classificação foi: Waller, t, Duncan, SNK

modificado, SNK e Tukey, este último, tendo apresentado algumas características desfavoráveis.

O presente trabalho visou avaliar a eficiência de um número bem maior de testes estatísticos usando o método de Monte Carlo; foram incluídos 4 e 8 repetições e 8 tratamentos no modelo. Os tratamentos apresentavam valor médio de 5200, 4800, 4600, 4400, 4200, 4000, 4000 e 4000 para os tratamentos 1, 2, 3, 4, 5, 6, 7 e 8, respectivamente.

O coeficiente de variação usado foi de 10%, tendo sido realizados 400 experimentos com 4 repetições e 200 com 8 repetições, possibilitando avaliar diferença de médias de 30%(1 – 6), 20%(2 – 6), 15%(3 – 6), 10%(4-6), 5%(5 – 6), 0%(7 – 6) e 0%(8 – 6). Os testes utilizados foram: Waller, *t* unilateral, *t* bilateral, Duncan, SNK, REGWF, REGWQ, Tukey, Sidak, Bonferroni I, Bonferroni II, Dunnettu e Bonferroni Modificado.

O teste Waller mencionado é o de Waller – Duncan; o teste *t* é o de *t* de Student. O teste SNK é devido a Student - Newman - Keuls; os testes REGWF e REGWQ foram desenvolvidos por Ryan – Einot – Gabriel – Welsh.

Os livros de texto em Estatística mais comumente utilizados, Pimentel - Gomes (2000), Steel & Torrie (1981), trazem com detalhes instruções de uso para teste t de Student, e para os testes de Duncan, Tukey, SNK, Scheffe, Bonferroni e Dunnett.

O aplicativo SAS (versão 6.04, 1990) possibilita o cálculo desses testes e ainda os testes de Sidak, REGWF, REGWQ, Gabriel, GT_2, Waller e outros.

O valor crítico F_c a 0,05 para 8 tratamentos e 4 repetições é 2,49. Dos 400 experimentos com quatro repetições, 22 apresentaram testes de F da análise da variância (valor F_o), não significativos. Para 8 repetições e 8 tratamentos, $F_c = 2,25$; em 200 experimentos estudados, todos os valores de F obtidos na Análise da Variância, foram significativos, isto é, todos os F_o foram maiores que 2,25.

No processo de simulação, adotou-se um coeficiente de variação de 10%, onde $CV=(s/\bar{x})100$.

Esse coeficiente avalia a variabilidade média do experimento. Outra medida da variação é o *índice de variação* (Pimentel – Gomes, 2000) definido como:

$IV = [s(\bar{x})/\bar{x}]100$. Este índice avalia a variabilidade média dos tratamentos já que leva em consideração, o desvio padrão e o número de repetições do experimento.

Um exemplo, das diferenças entre os dois coeficientes, é possível ser verificado abaixo, se levarmos em consideração as médias dos tratamentos, a média geral dos dois experimentos 3 e 4 e a análise conjunta dos dois experimentos, os desvios padrão e o erro padrão (erro padrão da média) dos experimentos respectivos:

Exp 3	**Exp 4**	**Exp (3 e 4)**
r = 4 repetições	r = 4 repetições	r = 8 repetições
$\bar{x}_1 = 5304,5$	$\bar{x}_1 = 5139,0$	$\bar{x}_1 = 5221,8$
$\bar{x}_2 = 4535,5$	$\bar{x}_2 = 4700,5$	$\bar{x}_2 = 4618,0$
$\bar{x}_3 = 4616,8$	$\bar{x}_3 = 4367,8$	$\bar{x}_3 = 4492,3$
$\bar{x}_4 = 4093,5$	$\bar{x}_4 = 4123,3$	$\bar{x}_4 = 4108,4$
$\bar{x}_5 = 4490,0$	$\bar{x}_5 = 3791,3$	$\bar{x}_5 = 4140,6$
$\bar{x}_{6(c)} = 3638,3$ (controle)	$\bar{x}_{6(c)} = 4111,3$ (controle)	$\bar{x}_{6(c)} = 3874,8$ (controle)
$\bar{x}_7 = 4134,0$	$\bar{x}_7 = 4122,3$	$\bar{x}_7 = 4128,1$

$\bar{x}_8 = 3901,0$	$\bar{x}_8 = 4503,0$	$\bar{x}_8 = 4202,1$
$\bar{x} = 4339,2$	$\bar{x} = 4357,3$	$\bar{x} = 4348,3$
s = 365,74	s = 479,24	s = 426,28
$s\bar{x} = 182,87$	$s\bar{x} = 239,62$	$s\bar{x} = 150,74$
CV = 8,43%	CV = 11,00%	CV = 9,80%
IV = 4,21%	IV = 5,50%	IV = 3,47%

Os três coeficientes de variação são estimativas do coeficiente de variação paramétrico que é CV= 10,0%; já o Índice de Variação do experimento conjunto (r = 8) é bem menor que os outros dois Índices de Variação (em que r = 4), o que indica uma maior precisão das estimativas das médias dos tratamentos do experimento com 8 repetições.

A diferença entre os testes Bonferroni I e II é a seguinte: o número de diferenças testadas para o I, é $k=C_8^2=28$ e para o II, k=7 (7 diferenças em relação ao controle).

Os testes REGWF e REGWQ foram desenvolvidos para comparações múltiplas em função do valor F obtido e do valor Q (amplitude existente entre os tratamentos).

O teste de Sidak é fundamentalmente parecido com o de Bonferroni, utilizando uma desigualdade desenvolvida por Sidak (detalhes, no SAS. 1990).

O teste de Bonferroni Modificado foi calculado separadamente, conforme Conagin (1999).

Fazendo-se um estudo comparativo entre os vários testes a partir da Tabela I, verifica-se que os testes *t* unilateral, Waller, *t* bilateral e Duncan apresentam os maiores poderes discriminativos e valores bastante próximos para as várias magnitudes de diferenças; em seguida situam-se os testes Dunnettu, Bonferroni Modificado e Bonferroni II; seguem pela ordem dos valores, SNK, REGWF, REGWQ, Tukey, Sidak e Bonferroni I.

Verifica-se que, para todos os testes, o poder discriminativo de cada um deles é tanto maior quanto maior for a magnitude da diferença pesquisada; o poder dos testes é maior para r = 8 que para r = 4, em todos os casos e para todos os testes.

Nas colunas 0% aparece a porcentagem de rejeição da hipótese de nulidade (verdadeira) para as duas diferenças 7 – 6 e 8 – 6.

Verifica-se que os testes *t* unilateral, Waller e *t* bilateral, mantiveram a estimativa observada de rejeição da hipótese de nulidade próxima dos cinco por cento; Duncan exibe um valor um pouco menor; os testes SNK, Dunnettu e Bonferroni Modificado apresentam valores intermediários entre cinco e um por cento, menores que os de Duncan; os testes REGWF, REGWQ, Tukey, Sidak e Bonferroni I apresentam valores ainda menores, ao redor ou inferior a um por cento; esses resultados concordam com os apresentados por Boardman & Moffitt (1971) e Bernardson (1975), os quais usaram em suas pesquisas, somente os testes *t*, Duncan, SNK, Tukey e Scheffé.

Em função dos resultados, cabe uma recomendação aos pesquisadores da área agronômica: na hora do planejamento de seus experimentos, no caso de experimento único de campo, sempre que possível utilizar oito, em vez de quatro repetições (como é freqüente), mesmo que seja preciso reduzir à metade o tamanho do canteiro, pois o aumento do poder discriminativo do teste é altamente compensador, (Ver Tabela I).

CONCLUSÕES

Nas condições pesquisadas, as seguintes conclusões são possíveis:

1. O aumento do número de repetições aumenta o poder discriminativo do teste, seja este qual for.
2. No teste de diferença de médias, quanto maior for à magnitude da diferença, maior será o poder discriminativo do teste utilizado.
3. A escolha adequada dos testes em função da natureza das pesquisas permite que se consiga um aumento do poder discriminativo (caso do Bonferroni II em relação ao Bonferroni I, teste Dunnettu em relação ao Dunnett quando se testam diferenças positivas em relação ao controle; teste de *t* unilateral em relação ao bilateral quando se conhece o sentido da diferença).
4. Os testes *t* unilateral, Waller e *t* bilateral são altamente eficientes quando se está disposto a cometer erros do tipo I por comparação; são bem mais discriminativos que SNK, Tukey e Sidak. O teste de Duncan apresenta eficiência próxima à de Waller e *t* unilateral e bilateral.
5. Dos testes do tipo erro por experimento, SNK, REGWF e REGWQ se situam em posição intermediária.
6. Dos testes do tipo MEER, na fase de *screening* o teste de Tukey é superior ao de Sidak e do Bonferroni I.
7. Para comparações escolhidas, predeterminadas ou em relação a um controle estabelecido, os testes Dunnettu, Bonferroni Modificado e Bonferroni II são bem superiores aos de Tukey, Sidak e Bonferroni I.
8. Dentre os testes mais comumente utilizados pelos pesquisadores, (os testes *t*, Duncan e Tukey), na fase de *screening*, em que nada se sabe *a priori* sobre a natureza das respostas, o teste de Tukey, apesar de mais exigente, é o que deve ser preferido, por ser do tipo MEER, isto é, capaz de proteger o pesquisador contra uma rejeição da hipótese nula no caso em que ela é, na realidade, verdadeira.
9. Para comparações predeterminadas (caso das 7 diferenças que abrangem todos os tratamentos), o teste Bonferroni II apresenta poder discriminativo maior que o teste de Tukey.

REFERÊNCIAS BIBLIOGRÁFICAS

BERNARDSON, C.S., 1975. Type I Error Rates when Multiple Comparison Procedures Follows a Significant F Test of Anova. **Biometrics**, **31**:337-340.

BOARDMAN, T.J.; MOFFITT, D.R., 1971. Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedures. **Biometrics**, **27**:738-744.

CORDELINO, R.A.; SIEWERT, F., 1992. Utilização Correta e Incorreta dos Testes de Comparação das Médias. **Rev. Soc. Bras. Zoot.**, **21**:985-995.

CARMER, S.G.; SWANSON, M.R., 1971. Detection of Differences between Means: a Monte Carlo Study of Five Pairwise, Multiple Comparison Procedure. **Agronomy Journal**, **63**:940-945.

CARMER, S.G.; SWANSON, M.R., 1973. Evaluation of Ten Pairwise, Multiple Comparison Procedures by Monte Carlo Methods. **Journal of Am. Statist. Assoc.**, **68**:66-74.

CHEW, V., 1977. Comparisons Among Treatment Means in na Analysis of Variance. **USDA Bulletin, H-6**:64.

CONAGIN, A., 1998. Discriminative Power of 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. (Boletim Científico, 44).

GABRIEL, K.R., 1964. A Procedure for Testing the Homogeneity of All Ses of Means in Analysis of Variance. **Biometrics, 20**:459-477.

HOCHBERG, Y.; TAMHANE, A.C., 1987. **Multiple Comparisons Procedures**. New York: J. Wiley, 450p.

IGUE, T., 1974. Tabelas de Probabilidades. Campinas: Instituto Agronômico, 12p. (Circular, 41).

PERECIN, D.; BARBOSA, J.C., 1988. Uma Avaliação de Seis Procedimentos para Comparações Múltiplas. **Revista de Matemática e Estatística, 6**:95-103.

PIMENTEL-GOMES, F., 2000. **Curso de Estatística Experimental**. ESALQ/USP, Piracicaba.

STATISTICAL ANALYSIS SYSTEM INSTITUTE – SAS/STAT, 1990. **User's Guide**. 4.ed., Cary, USA, 890p.

STEER, R.G.D.; TORRIE, J.H., 1981. **Principles and Procedures of Statistics**. McGraw-Hill, 663p.

WALLER, R.A.; DUNCAN, D.B., 1972. A Bayes Rule for the Symmetric Comparison Problem. **American Statistical Association, 67**:253-255.

Tabela I - Poder discriminativo dos testes pesquisados para experimentos com 4 e 8 repetições, diferenças de 30% (1-6), 20% (1-6), 15% (3-6), 10% (4-6), 5% (5-6) e 0% (7-6 e 8-6). O coeficiente de variação utilizado foi de 10%.

Repetições	r = 4						r = 8					
Testes \ Dif %	**30%**	**20%**	**15%**	**10%**	**5%**	**0%**	**30%**	**20%**	**15%**	**10%**	**5%**	**0%**
Waller	93,3	68,0	44,0	23,8	9,3	4,4	100,0	96,5	79,5	44,5	15,5	5,8
t unilateral	98,0	82,5	62,0	38,3	17,5	5,3	100,0	97,5	84,5	54,5	21,0	5,5
t bilateral	96,0	71,5	48,3	26,3	9,0	4,6	100,0	95,5	79,0	44,0	12,0	4,8
Duncan	94,5	65,0	39,8	20,0	5,5	4,3	100,0	94,5	73,5	37,5	11,0	3,8
SNK	79,8	36,0	18,3	5,8	1,8	1,6	99,5	79,5	48,5	22,0	3,5	1,5
REGWF	79,0	34,3	17,5	5,3	0,8	0,8	99,5	79,0	47,0	21,5	2,0	1,3
REGWQ	76,5	30,8	16,3	3,8	1,3	0,6	99,5	75,5	44,0	18,0	0,5	1,0
Tukey	72,8	28,3	15,3	2,3	0,5	0,3	99,5	70,0	36,5	14,0	1,5	0,8
Sidak	67,5	21,8	10,5	1,3	0,3	0,1	99,5	67,5	35,5	12,0	1,0	0,3
Bonferroni I	67,0	21,3	10,0	1,3	0,3	0,1	99,5	67,5	35,5	11,0	1,0	0,2
Bonferroni II	87,3	48,3	29,5	11,8	2,5	1,5	100,0	85,0	48,5	26,5	5,5	0,8
Dunnettu	90,8	56,3	33,5	13,3	3,8	2,3	100,0	91,5	63,5	31,5	8,0	1,3
Bonferroni Modif.	87,8	53,5	32,5	13,0	3,8	2,4	100,0	91,5	64,0	33,5	8,0	1,3