INFLUÊNCIA DO VALOR CRÍTICO NA DETECÇÃO DE VALORES DISCREPANTES EM ARQUEOMETRIA

Paulo Tadeu Meira e Silva de OLIVEIRA, aluno do programa de doutorado em estatística. Depto de Estatística, Instituto de Matemática e Estatística – USP, C. P.: 66281, CEP. 05315-970, São Paulo/SP, e-mail: poliveir@ime.usp.br, Tel. 0 XX (11) 3106.7760.

Casimiro S. A. Sepúlveda MUNITA, Pesquisador, Instituto de Pesquisas Energéticas e Nucleares – IPEN-CNEN/SP, C. P. 11049 CEP.: 05422-970, São Paulo/SP, e-mail: camunita@net.ipen.br Tel. 0 XX (11) 3816.9289

RESUMO: Este trabalho teve como objetivo detectar valores discrepantes (outliers) na determinação de 13 variáveis (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th e U) em 41 amostras de fragmentos cerâmicos usando a distância Mahalanobis $\left(D_i^2\right)$. Os D_i^2 calculados para cada amostra foram comparados com 3 critérios (teste F, teste do χ^2 e lambda Wilks) para verificar o critério mais conveniente.

Palavras-chave: distância Mahalanobis, resultados discrepantes, teste F, teste χ^2, lambda Wilks

1. INTRODUÇÃO

Desde a década do 50 tem sido grande a preocupação dos estatísticos para detectar e tratar resultados experimentais atípicos; isto é, resultados discrepantes (outliers), tanto em amostras univariadas, envolvendo apenas uma variável Bacon-Shone & Fung (1987) e multivariadas, envolvendo duas ou mais variáveis Jonhson & Wichern (1998).

Resultados discrepantes podem ser gerados por processo fora de controle, técnica analítica errada, contaminação durante a preparação da amostra, medida com alto erro, etc. Em geral, a identificação dos valores discrepantes é subjetiva, embora existam diferentes métodos estatísticos Barnett & Lewis (1994).

Na literatura poucos trabalhos tem sido publicados sobre a identificação de valores discrepantes em amostras que envolvem mais de uma variável. A maioria dos métodos propostos são gráficos, como dendrograma (árvore) de análise por agrupamento (cluster analysis) e é subjetivo. Não há um estudo comparativo da influência do valor crítico na detecção de valores discrepantes em resultados arqueométricos, embora esses resultados podem influir na interpretação. Neste trabalho aplica-se o método da distância Mahalanobis

para identificar valores discrepantes em resultados elementares de 41 amostras de fragmentos cerâmicos.

2. DESENVOLVIMENTO

O método apresentado neste trabalho foi aplicado nos resultados de treze variáveis (As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th e U) determinadas em um conjunto de 41 amostras de fragmentos cerâmicos coletados em um sítio arqueológico. Os detalhes sobre preparação das amostras e o método analítico foram publicados por Munita et al. (2000). Na Tabela 1 apresentam-se os resultados.

Inicialmente, os resultados foram normalizados mediante transformação por logaritmo base 10 para compensar as diferenças em magnitude de elementos em porcentagem dos que estão ao nível de traços Sayre (1975). Vários autores sugerem a distância Mahalanobis, D_i^2, como método para detecção de resultados discrepantes quando são determinadas várias variáveis Penny (1987). Para cada uma das n amostras no conjunto de p variáveis, a distância D_i^2 é calculada. Se $\bar{x}$ é a média do vetor e S é a matriz de covariância amostral, então

$$S = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \Big/ (n-1) \quad (1) \quad \text{e} \quad D_i = \sqrt{\{(x_i - \bar{x})^T S^{-1}(x_i - \bar{x})\}} \quad \text{para i =1....n} \quad (2)$$

onde $(x_i - \bar{x})$ é o vetor da diferença entre os valores da medida em um grupo e a média dos valores do outro grupo.

Jackson (1991) sugeriu a expressão $[p(n-1)/(n-p)]F_{p,n-p}$ (3) para determinar o valor crítico na determinação de valores discrepantes. Nesse caso deve-se partir da premissa que os resultados pertencem a uma distribuição normal.

Por outro lado, Krzanowski (1988) mostrou que $(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \sim \chi_p^2$ (4)

onde $\underset{\sim}{x}$ tem uma distribuição normal multivariada com média μ e matriz de dispersão Σ. Na prática deve-se estimar μ e Σ com o vetor da média amostral, $\underset{\sim}{\bar{x}}$, e matriz de covariância amostral, S. Substituindo Σ por S em (3), tem-se:

$$(x_i - \mu)^T S^{-1}(x_i - \mu) \sim \frac{p(n-1)}{n-p} F_{p,n-p} \quad (5)$$

Neste trabalho, a distância Mahalanobis ao quadrado foi calculada para cada uma das 41 amostras e os resultados estão apresentados nas três últimas colunas da Tabela 1.

Freqüentemente é sugerido em estatística que a distribuição F para calcular o valor crítico é mais adequada que a distribuição χ^2, especialmente quando o número de amostras é pequeno. Por exemplo, por meio da expressão (2) comparando os valores de D_i^2 calculados

Tabela 1. Resultados em partes por milhão, exceto quando indicado e valores da distância Mahalanobis.

Amostra	As	Ce	Cr	Eu	Fe	Hf	La	Na	Nd	Sc	Sm	Th	U	D_1^2	D_2^2	D_3^2
1	2,6	67,8	212,0	2,9	1,3	10,8	31,8	132,0	41,0	39,9	9,4	6,4	1,3	6,3	6,1	6,6
2	1,7	75,8	205,0	2,9	0,9	12,5	31,8	121,0	45,0	41,8	9,0	6,9	1,6	5,9	8,5	8,7
3	1,9	61,1	215,0	2,9	1,0	10,9	30,8	176,0	45,0	46,2	9,1	7,3	1,5	3,5	5,6	5,6
4	1,6	56,4	183,0	2,4	0,8	10,8	28,0	120,0	35,0	43,4	7,5	6,4	1,5	14,1	16,5	19,0
5	2,8	68,6	215,0	3,0	1,2	11,8	34,0	145,0	48,0	45,0	9,3	6,3	1,4	2,2	3,3	3,5
6	2,0	61,7	212,0	3,0	0,9	10,8	34,0	125,0	48,0	47,4	9,2	6,9	1,7	6,1	7,8	7,6
7	2,2	62,5	195,0	2,8	0,9	11,3	29,3	92,0	46,0	42,5	9,2	7,1	1,3	8,9	8,2	8,0
8	1,6	82,0	187,0	3,2	1,1	10,8	37,2	260,0	47,0	37,2	9,8	4,8	1,2	10,5	11,8	11,8
9	1,5	90,8	303,0	3,2	1,2	11,0	39,5	266,0	52,0	41,7	10,2	5,6	1,1	14,0	14,1	13,7
10	1,3	20,2	57,7	0,8	0,2	2,5	39,0	244,0	45,0	9,5	10,1	1,2	0,9	38,0		
11	2,8	64,6	216,0	2,8	1,2	10,9	31,2	271,0	40,0	44,5	9,7	7,4	1,5	7,3	14,1	15,4
12	2,4	85,2	214,0	3,3	1,6	10,8	37,6	155,0	53,0	43,9	10,8	5,2	1,2	9,3	9,6	9,5
13	1,9	135,0	150,0	4,6	0,7	12,3	54,5	160,0	67,0	50,9	14,1	5,7	1,2	19,4	21,2	20,6
14	1,8	101,5	230,0	3,4	1,4	11,7	45,5	144,0	51,0	45,0	11,4	7,7	1,3	10,8	11,2	14,0
15	0,5	104,5	214,0	3,5	1,4	11,8	46,6	144,0	59,0	48,1	12,2	6,5	1,5	6,7	7,5	7,5
16	1,4	95,2	245,0	3,5	1,2	12,1	44,0	187,0	57,0	43,0	11,3	5,8	1,4	6,3	5,9	6,2
17	2,0	67,5	205,0	2,8	1,2	10,6	36,6	93,0	42,0	43,5	10,1	7,0	1,9	9,7	14,8	14,5
18	1,2	63,4	183,0	2,9	1,0	10,5	33,9	130,0	44,0	40,7	9,6	6,7	1,7	6,0	6,4	6,3
19	3,0	65,3	212,0	2,9	1,3	10,5	33,5	138,0	50,0	42,6	9,7	6,8	1,6	3,1	3,4	3,4
20	2,7	67,8	236,0	3,0	1,1	11,0	33,8	139,0	55,0	41,2	10,0	6,3	1,4	4,1	4,0	3,9
21	1,5	83,5	82,0	2,5	0,4	6,6	29,2	976,0	45,0	34,2	8,6	5,2	1,6	24,6	27,2	
22	1,9	52,5	195,0	2,7	0,9	11,6	26,2	136,0	43,0	43,2	8,5	7,3	1,4	7,6	9,4	9,1
23	1,9	109,7	218,0	3,3	0,8	11,7	37,8	181,0	60,0	39,4	10,3	5,2	1,1	10,0	10,4	10,3
24	1,7	87,8	241,0	3,3	1,2	10,9	40,8	200,0	71,0	45,6	11,0	7,0	1,3	9,5	10,7	10,4
25	1,6	78,9	230,0	3,2	0,9	10,9	41,1	189,0	69,0	40,0	11,3	5,1	1,1	7,8	15,4	17,3
26	1,9	112,9	48,0	2,9	0,8	11,5	47,7	1583,0	68,0	35,1	10,7	10,0	2,0	26,7		
27	1,2	68,9	204,0	2,9	0,8	11,4	32,8	191,0	51,0	44,3	10,2	6,8	1,6	3,4	6,4	9,9
28	2,5	54,5	203,0	3,0	1,3	10,9	34,1	138,0	44,0	44,7	9,6	6,8	1,2	7,4	8,5	8,8
29	1,1	245,0	162,0	4,7	0,7	11,6	53,5	172,0	72,0	48,4	15,6	5,3	1,0	23,5	23,8	23,8
30	1,4	70,9	192,0	3,0	0,8	11,9	36,1	117,0	61,0	46,1	10,3	7,4	1,5	8,0	9,9	9,9
31	1,4	93,2	243,0	3,4	1,3	12,8	40,9	189,0	54,0	45,8	11,4	6,1	1,2	2,3	3,3	3,5
32	1,8	129,0	95,0	3,8	1,3	14,1	54,1	1339,0	62,0	40,5	12,3	7,0	1,1	12,0	20,6	21,2
33	1,6	110,0	260,0	3,8	1,3	12,3	48,3	159,0	59,0	44,1	13,2	5,8	0,9	7,6	11,2	12,2
34	1,7	95,2	204,0	3,4	1,4	12,5	43,5	192,0	48,0	50,1	11,1	6,8	1,2	9,5	9,1	9,2
35	3,1	104,0	99,0	3,9	1,6	11,1	48,9	583,0	65,0	37,8	12,6	6,2	0,8	15,6	16,0	15,7
36	3,0	137,9	94,0	3,7	1,3	11,0	51,7	535,0	54,0	37,3	12,7	5,6	0,8	12,1	14,7	16,1
37	3,0	134,0	70,0	4,4	1,3	14,4	56,8	360,0	67,0	45,1	9,2	7,7	1,1	30,0		
38	1,3	89,2	249,0	3,4	1,5	12,3	39,5	165,0	62,0	48,9	11,1	5,7	1,4	12,3	12,4	12,5
39	2,4	123,2	224,0	4,3	9,2	1,3	51,5	176,0	58,0	47,8	14,0	7,4	1,6	20,0	19,3	18,8
40	1,8	97,5	238,0	3,3	8,0	1,2	38,0	167,0	52,0	42,3	10,4	6,2	1,8	15,8	14,8	16,6
41	1,8	92,7	253,0	3,6	14,9	1,3	44,2	125,0	63,0	48,3	11,7	6,4	1,2	21,7	20,8	20,9

$\chi^2_{p;\alpha/n;\alpha=5\%}$ 33,9 ($n = 41$)

$\frac{p(n-1)}{n-p} F_{p,n-p;\alpha/n}$ 72,6 ($n = 41$)

$\frac{p(n-1)^2 F_{p,n-p-1;\alpha/n}}{n(n-p-1+pF_{p,n-p-1;\alpha/n})}$ 25,6 ($n = 41$) 24,9 ($n = 38$) 24,6 ($n = 37$)

(Tabela 1) com o valor crítico no nível de significância de 5%, tem-se $\chi^2_{p;\alpha/n}$ é de 33,9. Por esse valor, a amostra 10 seria considerada como valor discrepante, enquanto que ao ser comparado com o valor crítico obtido por (5), nenhuma das amostras seriam consideradas discrepantes porque no nível de significância de 5% o valor da expressão $p(n-1)/n-p F_{p,n-p;\alpha/n}$ é de 72,6. Assim, é possível concluir que neste caso o valor crítico obtido por (3) é mais adequado para a determinação de valores discrepantes que o valor crítico obtido por meio da equação (5).

Por outro lado, Wilks (1963) desenvolveu um método estatístico para determinar valores discrepantes em resultados multivariados, quando a distribuição é normal, para pesquisar valores discrepantes seqüencialmente, um a um. Para isso Wilks determinou a razão de dispersão R_i, $R_i = |A_i|/|A|$ e $A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$, onde $|A|$ = determinante(A), A_i é calculado para A com o valor i eliminado da amostra. O maior resultado discrepante é o que tem a menor razão de dispersão R_1, onde $R_1 = min(R_i)$, que é o resultado removido.

Wilks (1963) mostrou que $R_i \sim B_e\left(\frac{n-p-1}{2}, \frac{p}{2}\right)$ (6)

Os valores críticos para usar R_i como um teste para detecção de valores discrepantes são aproximados pelos limites de Bonferroni obtidos dos valores menores que $\frac{100\,\alpha}{n}\%$. Para os valores superiores que $\frac{100\,\alpha}{n}\%$ é por meio de $\left(1+\lfloor p/(n-p-1)\rfloor F_{p/n-p-1}\right)^{-1}$ (7) conforme foi demostrado por Caroni e Prescott (1992). Esses autores propuseram um teste seqüencial para valores discrepantes multivariados, comparando o valor crítico no nível de significância com o valor da distância Mahalanobis. Removendo o valor de D_i^2 quando é maior que o valor crítico. O mesmo procedimento é usado para os casos restantes.

O critério de Wilks para a detecção de um único valor discrepante está relacionado a D_i^2 pela equação $R_i = 1 - \left[n/(n-1)^2\right] D_i^2$ (8), conforme mostrado por Barnett e Lewis (1984). Os valores críticos de Wilks derivados de uma distribuição beta podem ser convertidos por uso da equação (6) para valores críticos maiores por meio da distância de Mahalanobis ao

quadrado $D_c^2 = [(n-1)^2/n](1-R_c)$ (9)

Substituindo a expressão (5) em (7) temos $\frac{p(n-1)^2 F_{p,n-p-1;\alpha/n}}{n(n-p-1+pF_{p,n-p-1;\alpha/n})}$ (10) que é o valor crítico de Wilks.

Aplicando a expressão (10) nos resultados da Tabela 1, tem-se que o valor crítico D_i^2, no nível de significância de 5% é de 25,6. Por este procedimento 3 amostras: 10; 26 e 37 são valores discrepantes. Isto confirma que a expressão (10) é mais adequada que a expressão (5) para determinar valores críticos na identificação de valores discrepantes.

Após retirar as 3 amostras (10; 26 e 37) por meio da expressão (10) encontrou-se que o valor crítico foi de 24,9. Em uma única amostra (21) a distância Mahalanobis é maior que o valor crítico; portanto, é um valor discrepante. Retirando essa amostra (21) e recalculando a distância Mahalanobis, resultados apresentados na coluna D_3^2 da Tabela 1, encontrou-se que o valor crítico para esse conjunto de resultados no nível de confiança de 95% de 24,6. Após eliminar essa amostra, em todos os casos a distância Mahalanobis é menor que o valor crítico.

3. CONCLUSÃO

O exemplo apresentado ilustra, claramente, a importância da metodologia para determinar o valor crítico quando se usa a distância Mahalanobis. Usando o valor do χ^2 só uma amostra tem valor discrepante. Entretanto, por meio da expressão (10) foram identificadas 4 amostras como valores discrepantes.

4. REFERÊNCIAS BIBLIOGRÁFICAS

BARNETT, V.; LEWIS, T.; Outliers in Statistical Data, Wiley & Sons, New York (1994).

BACON-SHONE, J.; FUNG, W.K. A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data. In: Applied Statistics, 35, Royal Statistical Society, UK, 1987. p.153-162.

JONHSON, R. A.; WICHERN, D. W. Applied Multivariate Analysis, New Jersey: Prentice Hall, 1998. p.486-497.

MARDIA, K. V.; KENT, J. T.; BIBBY J. M. Multivariate Analysis, London, Academic Press: Prentice 1998.

MUNITA, C. S.; PAIVA, ALVES M. A.; OLIVEIRA, P. M. S.; MOMOSE, E. F.Contribution of Neutron Activation Analisys to Archaelogical studys. J. Trace and Microprobe Techniques, 18(3), 381-387, 2000.

PENNY, KAY I., Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. In: Applied Statistics, 35, Royal Statistical Society, UK, 1987. p.153-162.

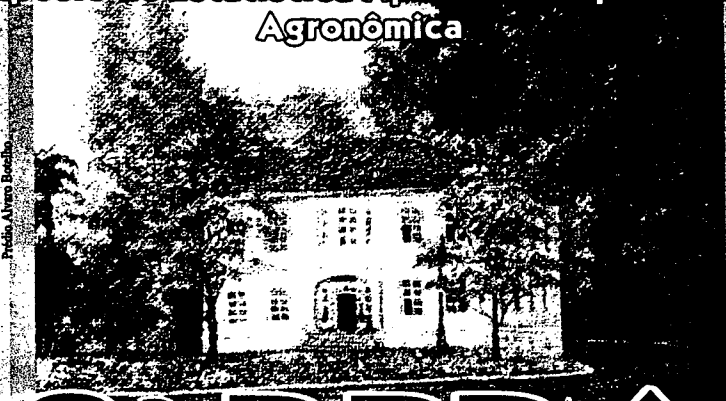
SAYRE, E. V. Brookhaven Procedures for Statistical Analyses of Multivariate Archaelmetric Data. Brookhaven National Laboratory Report BNL-21693, New York, 1975.

WILKS, S. S. (1963) Multivariate Statistical Outliers, Sankhya A, 25, 407-426.

Universidade Federal de Lavras
Departamento de Ciências Exatas

10º SEAGRO

10º Simpósio de Estatística Aplicada à Experimentação Agronômica



Pintura Óleo sobre tela: Sônia Maria de Stefano Piedade
Prédio Alvaro Botelho

48ª RBRAS

48ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria

07 a 11 julho de 2003

Programa e Resumos

Campus Universitário — Lavras-MG-Brasil

2003 TC

IPEN/CNEN-SP
BIBLIOTECA
"TEREZINE ARANTES FERRAZ"

Formulário de envio de trabalhos produzidos pelos pesquisadores do IPEN para inclusão na Produção Técnico Científica

AUTOR(ES) DO TRABALHO:
P.T.M. Oliveira, C.S. Munita

LOTAÇÃO: CRN RAMAL: 9289

TIPO DE REGISTRO:

art. / periód.:	☐	Publ. IPEN	☐	. resumo	☐
cap. de livro	☐	art. conf	☒	outros (folheto, relatório, etc...)	☐

TITULO DO TRABALHO:
Influência do valor crítico na detecção de valores discrepantes em arqueometria

APRESENTADO EM (informar os dados completos - no caso de artigos de conf.. informar o título da conferência. local. data. organizador. etc..)
48ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, Lavras, 7-11 julho 2003

PALAVRAS CHAVES PARA IDENTIFICAR O TRABALHO:
Distância Mahalanobis, teste F, teste x^2, lambda Wilks.

ASSINATURA:

DATA: 27/07/03