TESTING MULTIVARIATE NORMALITY OF ELEMENTAL CONCENTRATION DATA OBTAINED BY INAA: A CASE STUDY OF THE LAGO GRANDE ARCHAEOLOGICAL SITE

Roberto Hazenfratz, Casimiro S. Munita, Eduardo G. Neves

Instituto de Pesquisas Energéticas e Nucleares, IPEN – CNEN/SP
Av. Prof. Lineu Prestes, 2242 – Cidade Universitária – CEP 05508-000
São Paulo – SP – Brasil
robertohm@usp.br

ABSTRACT

The ceramic provenience studies are important in Archaeology due to its potential to infer the materials source used in the ceramic manufacture, which is an indicative of the socio-cultural interactions among ancient societies. A study of provenience was carried out using the elemental concentrations (Na, Lu, U, Yb, La, Th, Cr, Cs, Sc, Fe, Eu, Ce, Hf, Tb) of 50 ceramic fragment samples obtained by Instrumental Neutron Activation Analysis (INAA), from the Lago Grande archaeological site, in the central Amazon, Brazil. The preliminary multivariate statistical analysis of the ceramics elemental concentration suggested the existence of two distinct groups of different elemental composition. In archaeometric studies, the elemental concentrations are usually logarithmically transformed, based on the assumption that the trace elements have a natural lognormal distribution in nature. The normality is often desirable in multivariate analysis. In this paper, the lognormal distribution of concentrations was tested for multivariate normality, in order to validate the application of some multivariate statistical techniques.

1. Introduction and Justification

Many multivariate techniques rely on the assumption of normality of the data, or at least it is considered to be desirable (Baxter, 1997). In archaeometric studies based on the artifacts chemical composition, the elemental concentrations are frequently logarithmically transformed in the hope that this will make the data more normally distributed, based on the assumption that the trace elements have a lognormal distribution in nature.

A lognormal distribution is a continuous distribution obtained when the logarithms of observation follow a normal distribution. This distribution is followed by many sets of geological data, as trace elements. They are characterized by the fact that most of the observations are small, but a few are very large compared with the mean value. The equation that represents the lognormal frequency distribution is (Koch and Link, 2002):

$$f(w) = \frac{1}{w\beta\sqrt{2\pi}} \exp\left[-\frac{1}{2\beta^2} (\ln w - \alpha)^2 \right] \quad (1)$$

with α and β^2 representing the mean and variance, respectively, of the natural logarithm of the observation w.

2. Data Acquisition and Treatment

Fifty ceramic fragments from the Lago Grande archaeological site were analyzed by INAA in order to characterize their elemental composition. The experimental procedure employed to obtain the elemental concentrations by INAA can be found in Munita *et al* (2003). Three multivariate techniques were employed to analyze and interpret the results: Cluster, Principal Components and Discriminant Analysis. The latter two techniques are implemented by many softwares in a way that rely on the assumption of multivariate normality.

After some considerations about the relative standard deviation obtained and conditions of the experimental determinations of some radioisotopes, fourteen elements were selected for multivariate analysis: Na, Lu, U, Yb, La, Th, Cr, Cs, Sc, Fe, Eu, Ce, Hf, Tb. After this data pre-treatment, an outlier removal procedure was carried out, based on the Wilk´s multivariate outlier test (Oliveira, 2003).

Based on the assumption that trace elements follow a lognormal distribution in nature (Beier & Mommsem, 1994; Sayre, 1975; Koch and Link, 2002), the elemental concentration results were normalized to base-10 logarithm to improve the normality condition and to compensate for differences in the magnitude of elemental concentrations which are percentage from the ones in trace level.

The cluster analysis results were considered to assess data partition and to apply the multivariate normality tests in each potential group (Hazenfratz et al, 2009).

3. Multivariate Tests Selection

Two multivariate tests were selected to assess multivariate normality: the Shapiro-Wilk test, implemented in *R*, a software for statistical modeling, and a version of Shenton-Bowman multivariate test proposed by Doornik and Hansen (1994), whose calculation routine was implemented in *Scilab*. There are works in the literature that study the power and size of both tests, showing their good properties and limitations (Doornik & Hansen, 1994; Santos & Ferreira, 2003; Cantelmo e Ferreira, 2007).

3.1. Shapiro-Wilk test

The function implemented in *R* is a multivariate generalization of Shapiro-Wilk univariate test proposed by Domanski. Cantelmo & Ferreira (2007) showed by Monte Carlo simulations that the multivariate normality test implemented in R is extremely liberal and its application is unadvisable in routine calculations. (Cantelmo and Furtado, 2007).

3.1. Shenton-Bowman test

For comparison, it was implemented a modified version of the omnibus test for univariate and multivariate normality based on Shenton & Bowman (1977), proposed by Doornik and Hansen (2008). Their proposed multivariate statistic is:

$$E_p = Z_1^{'} Z_1 + Z_2^{'} Z_2 \underset{app}{\sim} \chi^2(2p) \quad (2)$$

where Z_1 and Z_2 are respectively the vectors of skewness and kurtosis coefficients transformed to standard normals.

The advantages of this statistic are its easy implementation and the fact that it only requires tables of the χ^2 distribution. The condition to apply this test is the number of samples $n \geq 7$. The implementation of this multivariate test was performed in *Scilab*.

4. Results and Discussion

It was performed an univariate normality test in order to study its influence in the multivariate normality tests. The test selected was the Anderson-Darling test implemented in the *R* software. The results are presented in the next table.

Table 1: Univariate normality test

Number	Variable	Statistic	p	Status
1	Na	1.35	1.46E-03	Rejected
2	Lu	2.14	1.61E-05	Rejected
3	U	0.37	4.07E-01	Accepted
4	Yb	0.45	2.57E-01	Accepted
5	La	0.32	5.24E-01	Accepted
6	Th	1.05	8.51E-03	Rejected
7	Cr	0.60	1.12E-01	Accepted
8	Cs	2.08	2.27E-05	Rejected
9	Sc	0.65	8.52E-02	Accepted
10	Fe	0.57	1.29E-01	Accepted
11	Eu	0.70	6.13E-02	Accepted
12	Ce	0.27	6.71E-01	Accepted
13	Hf	0.36	4.27E-01	Accepted
14	Tb	0.91	1.92E-02	Rejected

The multivariate tests were performed firstly with all the elements included, and then with successive variable extraction in order to observe how the departure from univariate normality influences the multivariate test.

Additionally, a previous cluster analysis was considered in order to establish potential groups in the data set, divided according their elemental composition. As the cluster analysis does not rely on the assumption of data normality, it was used to divide data in groups in order to perform the multivariate test per group. For more details about the cluster analysis and archaeological significance of the groups, refer to Hazenfratz *et al* (2009).

The results of Shenton-Bowman and Shapiro-Wilk multivariate tests applied according the strategies previously discussed are presented in sequence. The level of significance adopted was $\alpha = 5$ %.

By the observation of table 2, it is possible to reach some conclusions regarding the application of the multivariate tests in the whole data set. The test 1 shows that the observations considered as an unique group do not follow a multivariate normal.Tests 1, 2, 5 and 6 shows that extracting the elements which present the highest departures from univariate normality (see table 1) improves both statistics, but all the results are still far from normality.

Concerning scale, from the tests 2 and 3, it can be seen that the assumption that data normalization improves the condition of normality is plausible. The two statistics worsen in test 3, when the multivariate tests are performed in the original data. From comparison between tests 2 and 4, it is observed that there is no difference between the bases used in the logarithmic normalization. The results for base-10 and base-e are the same.

The tests 7, 8, 9, 10 and 11 concerns the data cluster classification in 2 groups. Test 7 presented numerical problems in both Shapiro-Wilk and Shenton-Bowman tests. The comparison of Shapiro-Wild tests 8, 9 and 10 establishes an unexpected scenario. The removal of elements according the strategy adopted yielded a worse result for group 1, when comparing with the utilization of 9 elements. On the other hand, the Shenton-Bowman test presented the expected results: the sequential removal yielded a better statistic, with the consequence that the tests 9 and 10 statistics are situated inside the critical region for 5% of significance. For these tests in the group 1, the null hypothesis cannot be rejected, and multivariate normality can be considered for the elements U, Yb, La, Cr, Fe, Ce and Hf. From test 11, it is not possible to consider multivariate normality for the group 2 (for $\alpha = 5$ %), considering the elements U, Yb, La, Cr, Ce, Hf after the removal of Na, Lu, Th, Cs, Tb, Sc, Eu, Fe.

The tests 12-17 are related to the data cluster classification in 3 groups. From comparison among tests 12, 13, 14 and 15 related to group 1, it is observed that the Shapiro-Wilk statistic improves with sequential elemental removal. On the other hand, the Shenton-Bowman statistic improves along tests 12, 13 and 14, but worsen in test 15 in comparison with the others. Furthermore, the statistic moves to the left side of the χ^2 critical region in tests 14 and 15, when compared with 12 and 13. At this moment,

Table 2: Multivariate Normality Tests

Test	Scale	Group	n	Variables	Removed Elements	Elements for Analysis	Shapiro-Wilk		Shenton-Bowman (α = 0.05)		
							W	p	Ep	Ep, crit (i)	Ep, crit (s)
1	$\log_{10}$	All	43	14	none	Na, Lu, U, Yb, La, Th, Cr, Cs, Sc, Fe, Eu, Ce, Hf, Tb	0.7438	2.69E-07	144.18	16.93	41.34
2	$\log_{10}$	All	43	9	Na, Lu, Th, Cs, Tb	U, Yb, La, Cr, Sc, Fe, Eu, Ce, Hf	0.7757	1.12E-06	102.48	9.39	28.87
3	original	All	43	9	Na, Lu, Th, Cs, Tb	U, Yb, La, Cr, Sc, Fe, Eu, Ce, Hf	0.6754	1.78E-08	142.61	9.39	28.87
4	ln	All	43	9	Na, Lu, Th, Cs, Tb	U, Yb, La, Cr, Sc, Fe, Eu, Ce, Hf	0.7757	1.12E-06	102.48	9.39	28.86
5	$\log_{10}$	All	43	7	Na, Lu, Th, Cs, Tb, Sc, Eu	U, Yb, La, Cr, Fe, Ce, Hf	0.8162	8.18E-06	63.39	6.57	23.68
6	$\log_{10}$	All	43	6	Na, Lu, Th, Cs, Tb, Sc, Eu, Fe	U, Yb, La, Cr, Ce, Hf	0.8510	5.55E-05	61.13	5.23	21.03
7	$\log_{10}$	1/2	9	14	none	Na, Lu, U, Yb, La, Th, Cr, Cs, Sc, Fe, Eu, Ce, Hf, Tb	-	-	-	-	-
8	$\log_{10}$	1/2	9	9	Na, Lu, Th, Cs, Tb	U, Yb, La, Cr, Sc, Fe, Eu, Ce, Hf	0.6234	1.80E-04	41.92	9.39	28.87
9	$\log_{10}$	1/2	9	7	Na, Lu, Th, Cs, Tb, Sc, Eu	U, Yb, La, Cr, Fe, Ce, Hf	0.4178	6.89E-07	21.20	6.57	23.68
10	$\log_{10}$	1/2	9	6	Na, Lu, Th, Cs, Tb, Sc, Eu, Fe	U, Yb, La, Cr, Ce, Hf	0.4929	5.28E-06	14.06	5.23	21.03
11	$\log_{10}$	2/2	34	6	Na, Lu, Th, Cs, Tb, Sc, Eu, Fe	U, Yb, La, Cr, Ce, Hf	0.8519	3.09E-04	49.88	5.23	21.03
12	$\log_{10}$	1/3	15	14	none	Na, Lu, U, Yb, La, Th, Cr, Cs, Sc, Fe, Eu, Ce, Hf, Tb	0.2841	9.83E-08	48.83	16.93	41.34
13	$\log_{10}$	1/3	15	9	Na, Lu, Th, Cs, Tb	U, Yb, La, Cr, Sc, Fe, Eu, Ce, Hf	0.4138	7.65E-07	32.35	9.39	28.87
14	$\log_{10}$	1/3	15	7	Na, Lu, Th, Cs, Tb, Sc, Eu	U, Yb, La, Cr, Fe, Ce, Hf	0.7880	2.59E-03	5.15	6.57	23.68
15	$\log_{10}$	1/3	15	6	Na, Lu, Th, Cs, Tb, Sc, Eu, Fe	U, Yb, La, Cr, Ce, Hf	0.8379	1.17E-02	1.74	5.23	21.03
16	$\log_{10}$	2/3	19	6	Na, Lu, Th, Cs, Tb, Sc, Eu, Fe	U, Yb, La, Cr, Ce, Hf	0.8323	3.50E-03	30.72	5.23	21.03
17	$\log_{10}$	3/3	9	6	Na, Lu, Th, Cs, Tb, Sc, Eu, Fe	U, Yb, La, Cr, Ce, Hf	0.4929	5.28E-06	14.06	5.23	21.03

it is not possible to draw a plausible conclusion for this behavior. Nevertheless, it is expected to be a consequence of the Shenton-Bowman algorithm, as proposed by Doornik and Hansen (2008), rather than something related to the elements distribution. The test 16 for group 2 presents both statistics outside the critical region for the 5 % of significance adopted. In contrast, test 17 presents a poor result in Shapiro-Wilk test, but a good Shenton-Bowman statistic, inside the critical region for group 3.

5. Conclusion and Future Perspectives

It is not possible to assume a multivariate normality condition for the data in compositional archaeometric studies without a previous statistical treatment in order to assess the partitioning patterns of data. It is true even when the objective is to assess multivariate normality to justify the application of some multivariate techniques, as principal components and discriminant analysis. The best results from the multivariate normality tests were achieved dividing the data in groups, according to the cluster analysis, a multivariate technique that does not rely on the normality assumption.

Even with data clustering prior to the tests, the null hypothesis of multivariate normality can be rejected inside some of them. Nevertheless, in some cases the Shapiro-Wilk and Shenton-Bowman tests do not present similar results. The reasons for those differences must be investigated in order to check if any of the tests is not adequate in some cases.

It is observed that for the data set employed at this work, there is a correlation between univariate normality and multivariate normality of the studied elements. Departures from univariate normality are correlated with departures from the multivariate normality.

In the future, more tests will be done in order to have a better characterization of how the distribution of elements influences the multivariate normality, and which elements are the critical ones for the analysis. If necessary, restrictions will be assigned to them in future multivariate data treatment where normality is assumed. Futhermore, it will be investigated the reasons why some groups present higher departures from normality than others, as well as characterize the nature of those departures in terms of skewness and kurtosis coefficients.

Acknowledgments

The present work was realized with the support from “Conselho Nacional de Desenvolvimento Científico e Tecnológico” – CNPq – Brazil. Process Number: 134116/2009-7 and “Fundação de Amparo à Pesquisa do Estado de São Paulo” - FAPESP – Process Number 2010/07659-0

References

BAXTER, M. J. Testing Multivariate Normality, with Applications to Lead Isotope Data Analysis in Archaeology. 1997. Available in: *citeseerx.ist.psu.edu*

BEIER, T.; MOMMSEN, H. Modified Mahalanobis filters for grouping pottery by chemical composition. Archaeometry, 36, pp. 287-306 (1994).

CANTELMO, N. F.; FERREIRA, D. F. Desempenho de Testes de Normalidade Multivariados Analisados por Simulação Monte Carlo. *Ciência e Agrotecnologia*, Lavras, v. 31, n. 6, p. 1630-1636, 2007

DOORNIK, J. A.; HANSEN, H. An Omnibus Test For Univariate and Multivariate Normality. *Oxford Bulletin of Economics and Statistics*, v. 70, issue 1, p. 927-939, 2008.

HAZENFRATZ, R.; MUNITA, C. S.; NEVES, E. G.; OLIVEIRA, P. M. S., TOYOTA, R. G. *Preliminary Characterization of Ceramics from the Lago Grande Archaeological Site in the Central Amazon by INAA*. International Nuclear Atlantic Conference, Rio de Janeiro, September 27 to October 2, 2009

KOCH, G. S.; LINK, R. F. Statistical Analysis of Geological Data. Courier Dove Publications. 2002. 832 p.

MUNITA, C. S.; PAIVA, R. P.; ALVES, M. A.; OLIVEIRA, P. M. S.; MOMOSE, E. F. Provenance study of archaeological ceramic, J. Trace Microprobe Techniques, 21, pp. 697-695. 2003

OLIVEIRA, P. M. S. Influência do Valor Crítico na Detecção de Valores Discrepantes em Arqueometria. 10º SEAGRO, Lavras, July 7 to July 11, 2003

SANTOS, A. C.; FERREIRA, D. F. Definição do tamanho amostral usando simulação Monte Carlo para o teste de normalidade baseado em assimetria e curtose: II. Abordagem multivariada. *Ciência e Agrotecnlogia*, Lavras, v. 24, n. 1, p. 62-69, 2003.

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

SHENTON, L. R.; BOWMAN, K. O. A Bivariate Model for the Distribution of $\sqrt{b_1}$ and b_2. Journal of the American Statistical Association, v. 72, n. 357, p. 206-211, 1977.

APPENDIX: Shenton-Bowman Multivariate Test – Routine for Scilab

```
//Baseado no artigo "An Omnibus Test for Univariate and Multivariate Normality -
Jurgen A. Doornik (Nuffield College) e Henrik Hansen (University of Copenhagen)

//Teste de normalidade multivariável baseado em Shenton & Bowman (1977)
//condição: b2 > 1 + b1
//Hipótese: b2 apresente distribuição gama
//condição: n>=7

//ATENÇÃO: Parâmetros são estimadores -> divisões por n-1 em vez de n.

mode(-1)

X=fscanfMat ("teste_14el_log_g12.sce") //base de dados

[nlin,ncol] = size(X)

//Teste de condição de tamanho mínimo da amostra
printf("O tamanho da amostra deve ser n > 7. Para o presente caso, n = " + string(nlin))
write(%io(2),"")
write(%io(2),"")

//Cálculo do vetor de médias
for i=1:ncol
  media(1,i)=mean(X(:,i)) //é vetor linha para facilitar próximo cálculo
end

//Cálculo alternativo da matriz de covariância
//function [Va]=cov(M)
  //for i=1:ncol
  //M(:,i) = M(:,i) - media(1,i)
  //end
  //Va = (M'*M)/(nlin-1)  //confirmar se é n ou n-1
  //endfunction

//S=cov(X)

//Matriz de desvios relativos
for i=1:nlin
Xrel(i,:)=X(i,:)- media
end

//Matriz de covariância S alternativa
S=(Xrel'*Xrel)/(nlin-1) //no artigo é n!!!

//Matriz diagonal V
for i=1:ncol
        V(i,i) = 1/sqrt(S(i,i))
```

```
end

//Matriz de correlação
C=V*S*V

rankc = rank(C) //Deve ser igual a p
printf("O rank de C é : " + string(rankc) + "; deve ser igual a p = " + string(ncol))
write(%io(2),"")
write(%io(2),"")

//Cálculo de autovetores e autovalores da matriz X
//Lambda é uma matriz diagonal com os autovalores e H é uma matriz modal cujas
colunas são os autovetores
[H,Lambda] = spec(C + %i*0)

//Matriz de observações tranformadas
[H2,Lambda2] = spec(Lambda + %i*0)

  //Fazer A^p = V*D.^p/V para Lambda^(-1/2)

//Rtrans = H*inv((Lambda2*H2.^(1/2)/Lambda2))*H'*V*Xrel' //confirmar
//Utilizado o teorema de decomposição espectral para matrizes simétricas (Jordan):
A^alfa = P*(lambda^alfa)*P^T
Rtrans = H*(H2*Lambda2^(-1/2)*H2')*H'*V*Xrel' //Lambda é simétrica -> Teorema
da Decomposição Espectral
R = Rtrans'

//Cálculo do vetor de médias de R
for i=1:ncol
  mediaR(1,i)=mean(R(:,i)) //aqui usa n-1, mas rever artigo
end

//ATENÇÃO: Utilizando-se parâmetros populacionais, uma distribuição normal
multivariada pode ser transformada em normais padronizadas independentes; utizando-
se estimadores esse procedimento é apenas aproximadamente válido

//Vetor de coeficientes de assimetria

//Momento de ordem 2

Mom2=zeros(ncol,1)
for j=1:ncol
        for i=1:nlin
        Mom2(j,1)=Mom2(j,1) + (R(i,j) - mediaR(1,j))^2
end
end

Mom2=Mom2/nlin
```

```
Mom3=zeros(ncol,1)
//Momento de ordem 3
for j=1:ncol
        for i=1:nlin
        Mom3(j,1)=Mom2(j,1) + (R(i,j)-mediaR(1,j))^3
end
end
Mom3=Mom3/nlin //confirmer

  //Cálculo de raiz(B1)
for i=1:ncol
  B1(i,1)=Mom3(i,1)/(Mom2(i,1)^(3/2))
end

//Vetor de coeficientes de curtose

//Momento de ordem 4
Mom4=zeros(ncol,1)
for j=1:ncol
        for i=1:nlin
        Mom4(j,1)=Mom4(j,1) + (R(i,j)-mediaR(1,j))^4
end
end
Mom4=Mom4/nlin

for i=1:ncol
B2(i,1)=Mom4(i,1)/(Mom2(i,1)^2)
end

//Teste b2 > 1 + b1 - Resíduos devem ser maiores que 0
Residuo = B2 - ones(ncol,1) - B1
printf("Teste b2 -1 - b1 > 0. Vetor de resíduos = ")
write(%io(2),"")
write(%io(2),string(Residuo))
write(%io(2),"")

n=nlin //para simplificação

//Transformação do coeficiente de assimetria segundo D´Agostino (1970)
bet = 3*(n^2+27*n-70)*(n+1)*(n+3)/((n-2)*(n+5)*(n+7)*(n+9))

om2 = -1 + sqrt(2*(bet-1))

del = 1/sqrt(log(sqrt(om2)))

for i=1:ncol
  y(i,1) = B1(i,1)*sqrt((om2-1)/2*(n+1)*(n+3)/(6*(n-2)))
end
```

```
for i=1:ncol
  z1(i,1) = del*log(y(i,1)+sqrt(y(i,1)^2+1))
end

//Transformação do coeficiente de curtose segundo tranformação de raiz cúbica de
Wilson-Hilferty)
del = (n-3)*(n+1)*(n^2+15*n-4)

a = (n-2)*(n+5)*(n+7)*(n^2+27*n-70)/(6*del)

c = (n-7)*(n+5)*(n+7)*(n^2+2*n-5)/(6*del)

k = (n+5)*(n+7)*(n^3+37*(n^2)+11*n-313)/(12*del)

for i=1:ncol
  alfa(i,1) = a + (B1(i,1)^2)*c //confirmar raiz
end

for i=1:ncol
  Qui(i,1) = (B2(i,1)-1-B1(i,1)^2)*2*k //confirmar raiz
end

//Problema de cálculo com raiz cúbica!!! Não pode usar parênteses. Corrigir linha de
Qui.
for i=1:ncol
  cubica=Qui(i,1)/(2.0*alfa(i,1))
  z2(i,1) = (sign(cubica)*(abs(cubica))^(1/3)-1+1/(9*alfa(i,1)))*sqrt(9*alfa(i,1))
end

//Estatística de Teste
Ep = z1'*z1 + z2'*z2

//Valor de Chi-quadrado crítico para comparação
df = 2*ncol
Xcrit1=cdfchi("X",df,0.05,0.95)
Xcrit2=cdfchi("X",df,0.95,0.05)

printf("O valor de EP é: " + string(Ep))
write(%io(2),"")
printf("O intervalo crítico é: [" + string(Xcrit1) + "," + string(Xcrit2) + "]")
```