# Estudo comparativo de métodos de normalização em resultados experimentais.

P.M.S. Oliveira, C.S. Munita

Instituto de Pesquisas Energéticas e Nucleares - IPEN-CNEN/SP, Av. Prof. Lineu Prestes 2242. CEP 05508-000, São Paulo, SP, Brasil

## Introdução

A caracterização físico-química dos fragmentos cerâmicos é de grande interesse em estudos arqueológicos, uma vez que permitem evidenciar aspectos da vida de grupos ancestrais, de modo a inferir aspectos da cultura, comércio e o desenvolvimento tecnológico. As características macroscópicas das cerâmicas, tais como, decoração da superfície e forma, são, freqüentemente, utilizadas como indicadores culturais e cronológicos. Entretanto, as propriedades microscópicas como: textura, composição química e mineralógica, podem ser usadas para estudo da tecnologia de fabricação e proveniência desses materiais.

Tendo em vista o crescente avanço das técnicas físico-químicas em estudos arqueométricos, a quantidade de dados gerados tem aumentado significativamente. Para a interpretação desses resultados, faz-se necessário o uso de métodos estatísticos cada vez mais sofisticados, tais como as técnicas multivariadas. Estas técnicas, de uma forma geral, consideram que cada amostra analisada pode ser representada como um ponto no espaço multidimensional, onde cada dimensão do hiper-espaço corresponde a eixos determinados pela composição físico-química das amostras. Com o objetivo de agrupar as amostras, conforme sua similaridade/dissimilaridade, devem ser formados grupos de amostras de acordo com alguns critérios estatísticos. Os resultados podem ser organizados dentro de uma matriz de dados $X_{np}$, sendo o $n$ o número de amostras variando de 1 até $n$ e $p$ número de variáveis variando de 1 até $p$.

O objetivo deste trabalho é estudar e avaliar modelos estatísticos com o propósito de encontrar as condições mais adequadas para estudo da normalidade dos dados para serem aplicados a resultados experimentais de análise de cerâmica de sítios arqueológicos em diferentes regiões do país. Os estudos serão realizados usando uma base de dados de cerâmicas.

## Estudo de normalização dos resultados.

Quando se faz a interpretação de um conjunto de resultados arqueométricos, uma das questões que se coloca, quase sempre, é saber a distribuição da população de onde este subconjunto de amostras foi obtido. Inicialmente se faz a verificação da normalidade das variáveis por meio das concentrações elementares. Se os resultados não atendem as condições de uma distribuição normal, para normalizar esses dados, aplicam-se algumas transformações, como: log na base 10, log na base *e* outras.

Como na análise de dados arqueométricos as amostras são constituídas por um conjunto de vetores, que correspondem às concentrações elementares, existe a suposição de que cada vetor da amostra venha de uma distribuição normal (Santos et al., 2007). Portanto, torna-se necessário que os resultados sigam uma distribuição normal multivariada, especialmente quando são usados métodos estatísticos multivariados, tais como: análise de componentes principais, análise discriminante, análise de conglomerados (cluster analysis), entre outros (Baxter and Freestone, 2006).

Por outro lado, nas situações em que há um grande número de amostras, as técnicas de normalização dependem, apenas, do comportamento da média, ou da distância envolvendo a média na forma

$$n\left(\bar{\underset{\sim}{X}}-\underset{\sim}{\mu}\right)' S^{-1}\left(\bar{\underset{\sim}{X}}-\underset{\sim}{\mu}\right) \qquad (1)$$

onde, *n,* é o número de amostras; $\bar{X} - \mu$, é vetor diferença entre os vetores da média das amostras e a amostra; e por fim, $S^{-1}$, é o inverso da matriz variância-covariância das amostras.

A suposição da normalidade das observações individuais de uma amostra *n* é menos significativa porque a distribuição normal é assintótica em relação às principais estimativas estatísticas. Entretanto, a qualidade da inferência na normalização será melhor, quanto mais se aproximam as variáveis das amostras à distribuição normal multivariada representada pela matriz $X_{np}$ (Ferreira, 1996).

Na literatura são apresentados vários métodos para a normalização dos resultados (Jonhson and Wichern, 1998; Baxter and Gale, 1998). Contudo, nesta proposta serão estudados três métodos, a saber:

***a) Método de Anderson***

Trata-se de um teste válido para verificar, não somente o ajuste da normalidade, como também o ajuste para outras distribuições de probabilidade. Este teste é utilizado para medir a proximidade dos pontos à reta estimada no gráfico de probabilidade; apresenta mais peso nos pontos mais próximos da cauda da distribuição. Assim, para pequenos valores da estatística de Anderson (valores que faz com que o nível de significância obtido seja menor que 0,05), obtida por meio da expressão (2), indica que a distribuição normal é melhor estimada.

O teste de Anderson estabelece um critério de aceitação ou rejeição da distribuição normal (distribuição de probabilidade normal), considerando o seguinte teste de hipótese:

$H_0$: o conjunto das amostras segue uma distribuição normal

$H_1$: o conjunto das amostras não segue uma distribuição normal

A estatística do teste é dada por:

$$A^2 = n - \sum_{j=1}^{n} \frac{(2j-1)}{n} \ln\left[F(x_j) + \ln\left(1 - F(x_{n+1-j})\right)\right] \quad (2)$$

onde, *F,* é a função de distribuição acumulada da distribuição de probabilidade normal; *j,* é o valor da *j*-ésima amostra ordenada; e pó fim, $x_{j,}$ são os valores, ordenados, das variáveis das amostras.

Para o teste de Anderson, os valores críticos, são valores tabelados. É um teste unicaudal e a hipótese nula ($H_0$) é rejeitada se o valor calculado é maior que o valor crítico.

***b) Método de skewness e kurtosis***

É um procedimento formal para a verificação da normalidade por meio dos dados que utilizam o vetor de média. Os coeficientes de assimetria e curtose multivariada (Martinez-Espinoza et al., 2004) são calculados por meio das expressões:

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \left(X_i - \bar{X}\right)' S^{-1} \left(X_j - \bar{X}\right)^3 \right] \quad (3)$$

e

$$b_{2,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \left(X_i - \bar{X}\right)' S^{-1} \left(X_j - \bar{X}\right)^2 \right] \quad (4)$$

onde, $b_{1,P,}$ é o valor do coeficiente de assimetria; $b_{2,P,}$ é o valor do coeficiente de curtose; $P = 2$ ou 3 para pequenas quantidades de amostras; $X_i$ e $X_j$, valores da variável na *i*-ésima e *j*-ésima amostra; $\bar{X}$, é o vetor de media; *n,* é o número de amostras; e por fim,S, é a matriz, de variância-covariância.

Os valores calculados $b_{1,P}$ e $b_{2,P}$ são comparados com o valor crítico para verificar os desvios da distribuição normal (Mardia and Kent, 1991).

***c) Método de Q-Q plot***

Método que consiste em colocar, em um gráfico, os percentis ou z-score esperados pelo ajuste de uma distribuição normal. Se os pontos pertencem a uma linha reta a suposição de normalidade deve ser aceita.

Suponhamos que $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ são os valores de cada variável, separadamente e ordenada crescentemente, para as *n´s* amostras, calcula-se a média das concentrações para cada variável. Dividindo o valor da diferença entre o resultado de cada amostra (de cada variável) e a média da variável das amostras pelo desvio padrão, encontram-se os valores dos z-score ou quantil para cada amostra. Para pequenas quantidades de amostras ($n < 30$), a diferença entre os valores da variância populacional e amostral é significativa. Nesse caso, o valor da variância utilizada para calcular o desvio padrão divide-se por $n - 1$, enquanto que, para grandes quantidades de amostras ($n > 30$), a diferença entre os valores da variância populacional e da variância amostral torna-se pouco significativa, bastando para isso, que a variância seja dividida por $n$.

Os percentis $q_{(i)}$ de $X_{(i)}$ ($1 \leq i \leq n$) são plotados em um sistema cartesiano com $q_{(i)}$ na abscissa e $X_{(i)}$ na ordenada. Desvios de normalidade podem ser observados pela inspeção deste tipo de gráfico, cujos pontos, quando existir normalidade deve pertencer a uma reta de mínimos quadrados.

Este processo gráfico, embora bastante poderoso para se verificar desvios da normalidade, não constitui um teste formal para este propósito por ser subjetivo. Para contornar esta limitação, (Jonhson and Wichern, 1998) apresentaram um teste complementar a este processo gráfico, o qual mede o ajuste dos pontos do Q-Q plot a linha reta de mínimos quadrados por meio de uma medida de coeficiente de correlação calculado por:

$$r_Q = \frac{\sum_{i=1}^{n}(X_{(i)} - \bar{X})(q_{(i)} - \bar{q})}{\sqrt{\sum_{i=1}^{n}(X_{(i)} - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(q_{(i)} - \bar{q})^2}} \quad (5)$$

A hipótese de normalidade não é satisfeita se o valor calculado for menor que o valor crítico para um determinado nível de significância.

Em geral, deve verificar para o caso da normalidade multivariada, se ela obedece a seguinte desigualdade:

$$\left(\bar{X} - \underset{\sim}{\mu}\right)' \Sigma^{-1} \left(\bar{X} - \underset{\sim}{\mu}\right) \leq \chi_p^2(\alpha) \quad (6)$$

onde, $\bar{X} - \underset{\sim}{\mu}$, é a matriz transposta da diferença entre a média das variáveis e o valor de cada variável para cada amostra; $\bar{X}$, é a media das concentrações de cada variável; $\underset{\sim}{\mu}$, é o vetor das concentrações das variáveis para cada amostra; $\Sigma^{-1}$, é o inverso da matriz de variância co-variância amostral; e por fim, $\chi_p^2(\alpha)$, é o valor crítico obtido pelo valor de uma distribuição quiquadrado com $p$ graus de liberdade associado a uma distribuição normal multivariada representada por $N_p \sim (\mu, \Sigma)$ para um nível de confiança $\alpha$.

**Resultados e Discussão**

Neste trabalho, com o objetivo de avaliar a normalidade, foram calculados os valores para a média, desvio padrão, coeficiente de variação, mínimo e máximo, e, valores calculados por meio dos procedimentos de Anderson, skewness, kurtosis e Q-Q plot para uma base de dados constituída de 89 amostras de fragmentos cerâmicos onde foram determinadas as concentrações de As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th e U.

Os critérios considerados para esta avaliação foram: métodos de Anderson e Q-Q Plot, para nível descritivo (p-value) maior que 0.05, e por fim, kurtosis e skewness para valores entre –1 e +1 e –0,5 e +0,5, respectivamente.

A Tabela 1 apresenta os resultados para os valores brutos das concentrações elementares. Pode-se ver que As apresenta o maior valor de coeficiente de variação (0,47).

Os resultados para cada método estão indicados pelos símbolos + ou – que significam que os ajustes foram considerados compatíveis (+) e não compatíveis (-) a uma distribuição normal.

Tabela 1. Resultados para o conjunto de dados, n=89, $\alpha = 0,05$

| Variável | As | Ce | Cr | Eu | Fe | Hf | La | Na | Nd | Sc | Sm | Th | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mínimo | 0.20 | 84.90 | 82.00 | 1.78 | 1.81 | 5.40 | 43.60 | 217.00 | 37.00 | 10.16 | 6.33 | 9.06 | 0.80 |
| Máximo | 5.80 | 227.00 | 275.00 | 3.92 | 5.61 | 15.50 | 116.60 | 4068.00 | 102.00 | 23.31 | 14.35 | 19.10 | 2.36 |
| Média | 2.17 | 123.14 | 158.23 | 2.51 | 3.34 | 8.51 | 71.68 | 1860.42 | 58.63 | 15.87 | 9.68 | 12.72 | 1.37 |
| CV | 0.47 | 0.19 | 0.22 | 0.18 | 0.23 | 0.16 | 0.17 | 0.36 | 0.20 | 0.16 | 0.16 | 0.15 | 0.21 |
| Desvio Padrão | 1.02 | 23.21 | 34.79 | 0.45 | 0.75 | 1.35 | 12.25 | 672.96 | 11.87 | 2.59 | 1.56 | 1.96 | 0.29 |
| Q-Q plot | 0.91 | 0.85 | 0.91 | 0.91 | 0.99 | 0.90 | 0.96 | 0.98 | 0.94 | 0.98 | 0.97 | 0.97 | 0.95 |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.713 | 0.000 | 0.005 | 0.288 | 0.000 | 0.127 | 0.018 | 0.077 | 0.001 |
| resultado | - | - | - | - | + | - | - | + | - | + | - | + | - |
| Anderson | 2.459 | 2.399 | 2.161 | 1.966 | 0.204 | 1.289 | 1.066 | 0.375 | 1.391 | 0.586 | 0.849 | 0.448 | 1.690 |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.871 | 0.002 | 0.008 | 0.407 | 0.001 | 0.124 | 0.028 | 0.273 | 0.000 |
| resultado | - | - | - | - | + | - | - | + | - | + | - | + | - |
| kurtosis | 1.776 | 6.755 | 2.942 | 1.685 | 0.002 | 7.543 | 1.616 | 0.903 | 1.740 | 0.191 | 0.760 | 0.735 | 0.865 |
| skewness | 1.179 | 1.999 | 1.169 | 1.193 | 0.305 | 1.597 | 0.876 | -0.008 | 1.013 | 0.462 | 0.705 | 0.632 | 0.845 |
| Resultado | - | - | - | - | + | - | - | + | - | + | - | - | - |

Verifica-se na Tabela 1 que 4 variáveis (Fe, Na, Sc e Th) tiveram ajustes considerados normais pelo método de Anderson e Q-Q plot, e apenas, 3 variáveis (Fe, Na e Sc) pelo método de skewness e kurtosis.

A Tabela 2 mostra os resultados, após aplicar a transformação logaritmo na base *e* para os valores das concentrações elementares e, pode-se ver, que o As apresenta maior valor de coeficiente de variação (0,77).

Tabela 2. Resultados obtidos após aplicar transformação log base *e* ao conjunto de dados, n=89, $\alpha = 0,05$.

| Variável | As | Ce | Cr | Eu | Fe | Hf | La | Na | Nd | Sc | Sm | Th | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mínimo | -0.70 | 1.93 | 1.91 | 0.25 | 0.26 | 0.73 | 1.64 | 2.34 | 1.57 | 1.01 | 0.80 | 0.96 | -0.10 |
| Máximo | 0.76 | 2.36 | 2.44 | 0.59 | 0.75 | 1.19 | 2.07 | 3.61 | 2.01 | 1.37 | 1.16 | 1.28 | 0.37 |
| Média | 0.29 | 2.08 | 2.19 | 0.39 | 0.51 | 0.93 | 1.85 | 3.23 | 1.76 | 1.19 | 0.98 | 1.10 | 0.13 |
| CV | 0.77 | 0.04 | 0.04 | 0.19 | 0.20 | 0.07 | 0.04 | 0.07 | 0.05 | 0.06 | 0.07 | 0.06 | 0.69 |
| Desvio Padrão | 0.22 | 0.07 | 0.09 | 0.07 | 0.10 | 0.07 | 0.07 | 0.22 | 0.08 | 0.07 | 0.07 | 0.07 | 0.09 |
| Q-Q plot | 0.94 | 0.94 | 0.96 | 0.96 | 0.99 | 0.96 | 0.99 | 0.81 | 0.98 | 0.99 | 0.99 | 0.99 | 0.98 |
| p-value | 0.000 | 0.001 | 0.004 | 0.014 | 0.658 | 0.007 | 0.499 | 0.000 | 0.254 | 0.704 | 0.714 | 0.928 | 0.175 |
| resultado | - | - | - | - | + | - | + | - | + | + | + | + | + |
| Anderson | 1.12 | 0.89 | 1.17 | 0.79 | 0.30 | 0.66 | 0.39 | 4.26 | 0.55 | 0.37 | 0.33 | 0.15 | 0.87 |
| p-value | 0.006 | 0.022 | 0.004 | 0.039 | 0.589 | 0.080 | 0.378 | 0.000 | 0.155 | 0.420 | 0.516 | 0.965 | 0.025 |
| resultado | - | - | - | - | + | + | + | - | + | + | + | + | - |
| kurtosis | 3.91 | 2.64 | 1.84 | 0.56 | -0.12 | 3.03 | 0.83 | 5.89 | 0.47 | -0.07 | 0.26 | 0.06 | 0.09 |
| skewness | -1.03 | 1.02 | 0.05 | 0.67 | -0.30 | 0.40 | 0.21 | -2.11 | 0.32 | 0.00 | 0.19 | 0.16 | 0.24 |
| Resultado | - | - | - | - | + | - | + | - | + | + | + | + | + |

Por outro lado, na Tabela 2, 7 variáveis tiveram ajustes considerados normais, sendo que, pelo método de Anderson e de skewness e kurtosis foram as variáveis Fe, La, Nd, Sc, Sm, Th e U, e, para o Q-Q plot, Fe, Hf, La, Nd, Sc, Sm e Th.

Finalmente, na Tabela 3 contem os resultados, após aplicar a transformação logaritmo na base 10 para os valores das concentrações elementares e, pode-se ver, que As apresenta maior valor de coeficiente de variação (0,77).

Tabela 3. Valores obtidos após aplicar transformação log base 10 ao conjunto de dados, n=89, $\alpha = 0,05$

| Variável | As | Ce | Cr | Eu | Fe | Hf | La | Na | Nd | Sc | Sm | Th | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mínimo | -0.70 | 1.93 | 1.91 | 0.25 | 0.26 | 0.73 | 1.64 | 2.34 | 1.57 | 1.01 | 0.80 | 0.96 | -0.10 |
| Máximo | 0.76 | 2.36 | 2.44 | 0.59 | 0.75 | 1.19 | 2.07 | 3.61 | 2.01 | 1.37 | 1.16 | 1.28 | 0.37 |
| Média | 0.29 | 2.08 | 2.19 | 0.39 | 0.51 | 0.93 | 1.85 | 3.23 | 1.76 | 1.19 | 0.98 | 1.10 | 0.13 |
| CV | 0.77 | 0.04 | 0.04 | 0.19 | 0.20 | 0.07 | 0.04 | 0.07 | 0.05 | 0.06 | 0.07 | 0.06 | 0.69 |
| Desvio Padrão | 0.22 | 0.07 | 0.09 | 0.07 | 0.10 | 0.07 | 0.07 | 0.22 | 0.08 | 0.07 | 0.07 | 0.07 | 0.09 |
| Q-Q plot | 0.94 | 0.94 | 0.96 | 0.96 | 0.99 | 0.96 | 0.99 | 0.81 | 0.98 | 0.99 | 0.99 | 0.99 | 0.98 |
| p-value | 0.000 | 0.001 | 0.004 | 0.014 | 0.658 | 0.007 | 0.499 | 0.000 | 0.254 | 0.704 | 0.714 | 0.928 | 0.175 |
| Resultado | - | - | - | - | + | - | + | - | + | + | + | + | + |
| Anderson | 1.12 | 0.89 | 1.17 | 0.79 | 0.30 | 0.66 | 0.39 | 4.26 | 0.55 | 0.37 | 0.33 | 0.15 | 0.87 |
| p-value | 0.006 | 0.022 | 0.004 | 0.039 | 0.589 | 0.080 | 0.378 | 0.000 | 0.155 | 0.420 | 0.516 | 0.965 | 0.025 |
| Resultado | - | - | - | - | + | + | + | - | + | + | + | + | - |
| kurtosis | 3.91 | 2.64 | 1.84 | 0.56 | -0.12 | 3.03 | 0.83 | 5.89 | 0.47 | -0.07 | 0.26 | 0.06 | 0.09 |
| skewness | -1.03 | 1.02 | 0.05 | 0.67 | -0.30 | 0.40 | 0.21 | -2.11 | 0.32 | 0.00 | 0.19 | 0.16 | 0.24 |
| Resultado | - | - | - | - | + | - | + | - | + | + | + | + | + |

Por outra parte, na Tabela 3, foram encontrados os mesmos resultados para o ajuste de normalidade mostrado na Tabela 2 para os diferentes métodos.

**Conclusões**

Finalmente, comparando os resultados sem transformação com os resultados com transformação por log base *e* e por log base 10, nas Tabelas 1, 2 e 3, pode-se concluir que há diferença entre os estudos de normalidade aplicado nas 89 amostras de objetos cerâmicos.

Por outra parte, não há diferença significativa ao comparar os dados usando a transformação por log nas bases *e* e 10.

**Referencias Bibliográficas**


BAXTER, M.J.; FREESTONE, I.C. Log-ratio compositional data analysis in archaeometry. Archaeometry, 48(3): 511—531, 2006.

BAXTER, M.J.; GALE, N. H. Testing for multivariate normally via univariate tests: a case study using lead isotope ratio data. Journal of Applied Statistics, 25:671—683, 1998.

FERREIRA, D.F. Análise multivariada, UFLA, Lavras – MG, 1996.

JONHSON, R.A.; WICHERN, R.A. Applied multivariate statistical analysis. Fifth Edition, Practice Hall, New Jersey, 1998.

MARDIA, K.V.; KENT, J. T. Rao score tests for goodness of fit and independence. Biometrika, 78(2):355—363,1991.

MARTINEZ-ESPINOZA, M.; JÚNIOR, C.C.; LAHR, F.A.R. Parametric and non-parametric methods to determine the characteristic value in wood tests results. Scientia Forestalis, 66:76—83, 2004.

SANTOS, J.O.; MUNITA, C.S.; VERGUE, C.; OLIVEIRA, P.M.S. Normalização e padronização por meio da transformação logarítmica em estudos arqueométricos de cerâmicas. 52ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria - RBRAS, e 12º Simpósio de Estatística Aplicada à Experimentação Agronômica– SEAGRO, Santa Maria RS, 23 a 27 de Julho, 2007.