Determinação de *outliers* para Pequenas Amostras

Anderson da Silva Soares[1], Clarimar José Coelho[2], Gustavo Teodoro Laureano[3], Daniel Vitor Lucena[4], e Roberto Kawakami Harrop Galvão[5]

[1] Universidade Católica de Goiás, Departamento de Computação, Brasil
barnesucg@hotmail.com
[2] Universidade Católica de Goiás, Departamento de Computação, Brasil
clarimar@brturbo.com
[3] Universidade Católica de Goiás, Departamento de Computação, Brasil
gustavoeng@hotmail.com
[4] Universidade Católica de Goiás, Departamento de Computação, Brasil
danielvitor@brturbo.com
[5] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, Brasil
kawakami@ita.cta.br

Resumo Determinação de *outliers* em pequenas amostras empregando método de Bonferroni e gráficos de controle. É feito um breve resumo da teoria e o conceito de *outlier* também é introduzido. A título de ilustração, é considerado o problema de monitorar variáveis que influenciam a qualidade das amostras multivariadas analisadas. A análise é feita com base em dados de espectrometria de emissão atômica em plasma. Os resultados obtidos demonstram que os métodos empregados são eficientes para estimar os parâmetros populacionais de pequenas amostras e obtenção de limites de controle estatístico LCS e LCI.

Palavras chaves: Detecção de *outliers*, Controle de qualidade, inferência estatística.

1 Introducão

Inferência estatística obtém conclusões válidas para uma população baseada em populações amostrais [3]. O problema de verificar se um determinado valor μ_0 é plausivel para a verdadeira média populacional desconhecida pode ser resolvido através do teste

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0 \tag{1}$$

onde, H_0 é a hipótese nula e H_1 é a hipótese alternativa.

A presença de valores moderados em uma população normal é mais provável que a presença de valores extremos [6]. Assim, a suposição de normalidade de uma população qualquer é devida à alta probabilidade dos dados serem normalmente distribuídos [4][13].

Seja $\{X_1, X_2, \ldots, X_n\}$ uma amostra aleatória extraída de uma população normal para o caso univariado. O teste estatístico para esta hipótese, quando $p = 1$ é:

$$t = \frac{(\bar{X}^{-}\mu_0)}{\frac{S}{\sqrt{n}}} \tag{2}$$

onde, $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$, $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$ e p é o número de colunas da matriz de dados.

O teste mostrado na equação (2) segue a distribuição de t-student com $n-1$ graus de liberadade. A hipótese H_0 é rejeitada se o valor observado para $|t|$ exceder um valor (crítico) específico da distribuíção de t-student.

A distância quadrada da média amostral $\bar{X}$ também pode ser considerada para o valor a ser testado. A hipótese H_0 pode ser rejeitada a um nível de significância α, se

$$t^2 = n(\bar{X} - \mu_0)(S^2)^{-1} \geq t^2_{n-1}(\alpha/2) \tag{3}$$

onde, $t^2_{n-1}(\alpha/2)$ é o quantil quadrático superior $100(\alpha/2)$ da distribuição de t-student com $n-1$ graus de liberdade.

Se H_0 não é rejeitada, então μ_0 é um valor plausível para representar a média populacional normal ou existem outros valores de μ consistentes com os dados. A partir da correspondência entre a região de aceitação dos testes de hipóteses e o intervalo de confiança para μ, tem-se:

$$\left| \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \right| < t_{n-1}(\alpha/2) \tag{4}$$

a não rejeição de H_0 equivale a:

$$\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \leq \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \tag{5}$$

O intervalo de confiança $100(1-\alpha)\%$ é aleatório e depende das variáveis aleatórias $\bar{X}$ e S. A probabilidade desse intervalo conter μ é $100(1-\alpha)\%$.

Para o caso multivariado o problema consiste na determinação de um vetor $\mu_0(p \times 1)$ plausível para a média de uma distribuição normal multivariada. A generalização da distância quadrada mostrada na equação (3) é dada por

$$T^2 = n(\bar{\mathbf{X}} - \mu)^T S^{-1}(\bar{\mathbf{X}} - \mu_0) \tag{6}$$

onde,

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j, S = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^t$$

A equação (6) é conhecida como distribuição de Hotelling. A distribuição de Hotelling não necessita de tabelas com pontos percentuais para a realização dos testes de hipótese devido T^2 ser distribuída como:

$$\frac{(n-1)p}{n-p}F_{p,n-p} \tag{7}$$

onde $F_{p,n-p}$ é uma variável com distribuição F com p e $n-p$ graus de liberdade. A distribuição de Hotelling pode ser generalizada para pequenas observações de modo que a análise de todo o grupo é dada por

$$T_j^2 = (X_j - \bar{X})^T S^{-1} (X_j - \bar{X}) \tag{8}$$

O teste de hipótese $h_0 : \mu = \mu_0$ geralmente não satisfaz o analista no caso multiviariado. A estimação de uma região de confiança envolve a quantificação do valor de um determinado parâmetro populacional desconhecido. O teste de hipótese indica decisão a ser tomada sobre o valor específico do parâmetro populacional. Assim, é preferível encontrar regiões de valores μ plausíveis para representar a média populacional para os dados observados [7].

A região de confiança μ para uma distribuição normal p variada serão todos os valores de μ dados por

$$P\left[n(\bar{\mathbf{X}} - \mu)^T S^{-1} (\bar{\mathbf{X}} - \mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right] \tag{9}$$

O cálculo do valor μ_0 plausível para μ é dado pela distância quadrada generalizada $n(\bar{\mathbf{X}} - \mu)^t S^{-1} (\bar{\mathbf{X}} - \mu)$. O resultado é comparado com $(n-1)pF_{p,n-p}(\alpha)/(n-p)$. Se a distância quadrada for maior que $(n-1)pF_{p,n-p}(\alpha)/(n-p)$, então μ_0 não pertence a região de confiança. O teste da hipótese $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ permite afirmar que a região de confiança é constituída por todos os valores de μ_0 cujo teste T^2 não rejeita a hipótese nula a favor da hipótese alternativa a um nível de significância α.

2 Gráficos de Controle

Gráficos de controle são registros de dados mensurados em pontos críticos para um processo estatístico e construídos num sistema de coordenadas cartesianas [8]. O eixo das ordenadas é representado por mensurações para uma determinada característica. O eixo das abscissas é representado por subgrupos da amostra analisada e definidos com uma divisão racional da amostra coletada [11]. Paralelo ao eixo das abscissas são definidas duas linhas de controle obtidas a partir da expressão:

$$\mu \pm 3 * \sigma \tag{10}$$

com $\sigma = \sqrt{S}$.

As linhas paralelas ao eixo das abscissas são definidas como Limite de Controle Superior (LCS) e Limite de Controle Inferior (LCI). A Figura 1, mostra o gráfico de Shewhart [12] construído a partir do vetor $\mathbf{x}_2$ [10]. As medições obtidas são representadas na ordem do tempo e comparadas com os limites de controle. Se alguma medição ultrapassar os limites de controle o processo é considerado fora dos limites de controle estatístico e o valor identificado é definido como *outlier*.

Concluir que determinado valor de um conjunto de dados é *outlier*, é subjetivo. A definição de *outliers* é sujeita à análise e interpretação de resultados. Decisões a respeito da identificação de *outliers* devem ser tomadas individualmente e dependem de um experimento específico [4]. Os valores de $\mathbf{x}_2$, mostrados na Figura 1, estão localizados dentro dos limites LCI e LCS. Neste caso, assume-se que nenhum valor do conjunto de dados é considerado *outlier*.

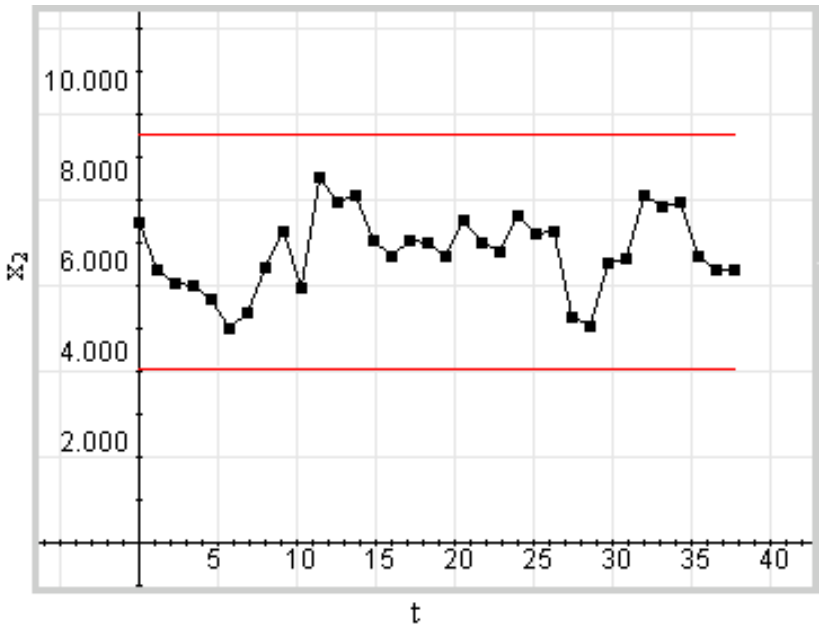

Figura 1. Gráfico de controle de Shewhart com limites de controle LCS e LCI.

Os limites LCS e LCI mostrados na Figura 1 são obtidos, respectivamente, pela soma e subtração da expressão (10).

O trabalho de Shewhart é baseado na classificação de pequenas variações aleatórias inerentes ao processo estatístico que prejudicam a inferência [12]. Assim, a variável de $\mathbf{x}_2$ escolhida para análise é monitorada individualmente por meio de sucessivas amostras espaçadas no tempo, desprezando-se possíveis correlações entre as variáveis.

Jonhson [4] descreve dois fatores discriminantes no uso do método T^2 para controle estatístico que leva o analista a conclusões erradas sobre a detecção de *outliers*. O primeiro, diz respeito ao uso direto da distância quadrada definida na equação (9) para definição dos limites de controle. Quando o método T^2 sinaliza a j-ésima variável fora de controle deverão ser determinadas quais observações são responsáveis pelo *outlier*. O segundo, diz respeito a utilização da expressão

(10) para pequenas amostras, onde o parâmetro σ da população verdadeira não pode ser conhecido e requer o uso de um estimador $\bar{X}$ para μ.

Pequeno número de intevalos de confiança μ é requerido em muitas situações da análise multivariada. Assim, intervalos de confiança constituem uma boa alternativa para estimar $\bar{X}$. Esta alternativa é conhecida como método de Bonferroni:

$$\bar{\mathbf{X}}_i \pm \mathbf{t}_{n-1}\Big(\frac{\alpha}{2p}\Big)\sqrt{\frac{S_{ii}}{n}} \quad i = 1, 2, \ldots, p = m \tag{11}$$

O método de Bonferroni deriva da distribuição de t-student sendo indicado para um conjunto pequeno de dados multivariados [5]. Estima-se com $\alpha = 95\%$, o vetor de parâmetros populacionais μ_0 da equação (10) para pequenas amostras para a obter os limites de controle estatístico LCS e LCI, respectivamente.

3 Exemplo de Aplicação

Apresenta-se os resultados obtidos para pequenas amostras. Pequenas amostras,

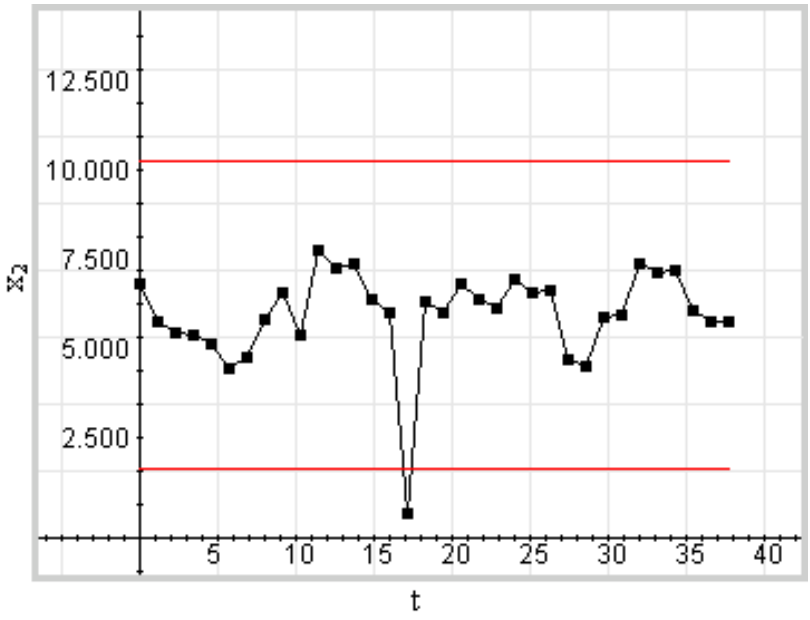

Figura 2. *Outlier* na variável $\mathbf{x}_2$ detectado com os limites de controle LCS e LCI.

possuem número de observações menores que 35 [9]. Dados espectroscópicos

foram extraídos da análise de amostras de aço-ligas contendo Manganês (Mn), Molibdênio (Mo), Cromo (Cr), Níquel (Ni) e Ferro (Fe) conforme descrito em Pimentel [10]. Os dados são apresentados na forma matricial: linhas representam observações amostrais; colunas representam as variáveis analisadas.

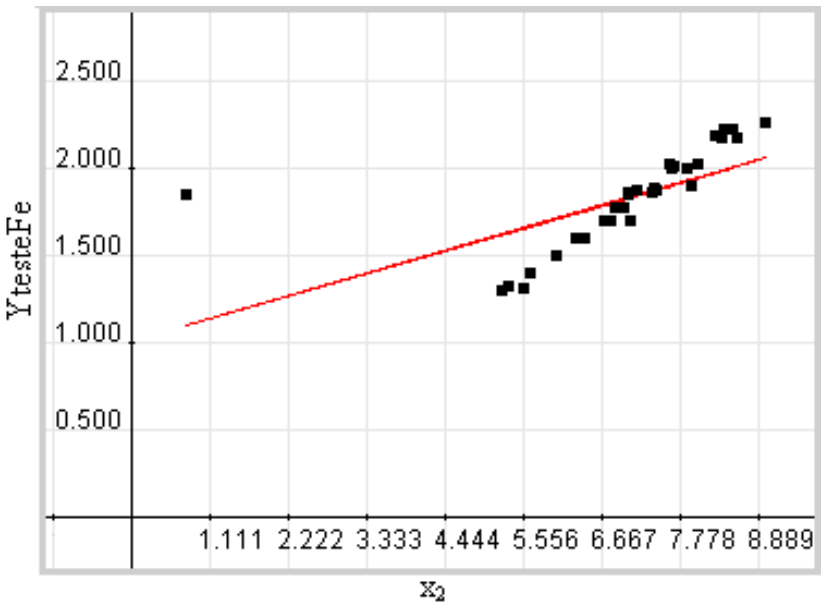

Figura 3. Reta de Regressão Linear sobre $\mathbf{x}_2$ e **YtesteFe** com o *outlier*

O vetor coluna $\mathbf{x}_2$ é extraído da matriz **XtesteFe**. De modo que, o vetor $\mathbf{x}_2$ contém 34 linhas da segunda coluna de **XtesteFe**. Todas as variáveis de **XtesteFe** são analisadas e nenhum valor é considerado *outlier* considerando os critérios utilizados. Como no exemplo da seção 2, devido o fato da matriz de dados **XtesteFe** ter sido pré-processada. Assim, o valor da linha 16 do vetor $\mathbf{x}_2$ é alterado de 6.7 para 0.7 para demonstrar o método.

A Figura 2 ilustra a presença de *outlier* após emprego do método para o vetor $\mathbf{x}_2$ alterado. Constata-se que algum valor da observação número 16 extrapola o limite de controle inferior. Quando isto ocorre, diz-se que *outliers* são detectados, segundo critérios estabelecidos pelo método empregado. A Figura 3, mostra a reta de regressão linear [1][2] para valores de $\mathbf{x}_2$ e **YtesteFe** e as alterações provocadas pelo *outlier* em $\mathbf{x}_2$. Isto significa que o *outlier* interfere bruscamente na interpretação de resultados. Neste exemplo, a reta de regressão se inclina a um ponto irreal para a base de dados analisada.

A interpretação é modificada quando o *outlier* é excluído de $\mathbf{x}_2$. A Figura 4, mostra a reta de regressão para $\mathbf{x}_2$ e **YtesteFe** sem a presença do *outlier*.

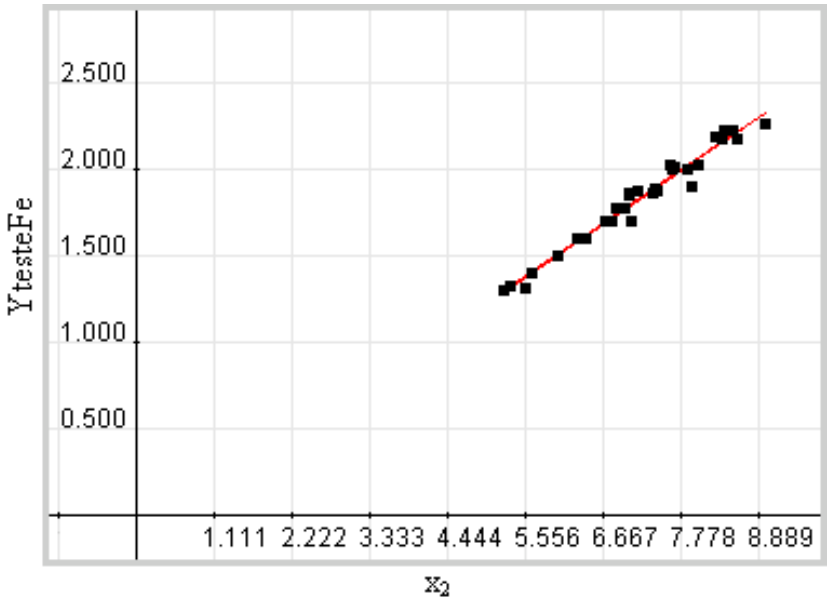

Figura 4. Reta de Regressão Linear sobre as variáveis $\mathbf{x}_2$ e **YtesteFe** sem o *outlier*

Observe que a reta de regressão tem outra inclinação em relação a anterior. Altera-se a inferência quanto aos possíveis valores a serem preditos pela reta de regressão na relação das variáveis independentes e dependentes.

4 Conclusão

Este trabalho apresenta os resultados obtidos com o método de Bonferroni e gráficos de controle para análise multivariada. O método de Bonferroni e gráficos de controle foram empregados para a determinação de *outliers* estabelecendo limites de controle inferior e superior. Neste exemplo particular, introduziu-se um *outlier* num conjunto de dados pré-processados para o teste do método proposto. Trabalhos futuros incluem o estudo e o desenvolvimento de software para a análise multivariada empregando técnicas tradicionais em quimiometria como regressão linear múltipla, regressão em componentes principais e regressão em mínimos quadrados parciais.

Agradecimentos

Pesquisa suportada pela PROPE-UCG através do projeto número 577 e fundos BIC-UCG. Detalhes do projeto podem ser encontrados na página de administração de projetos da UCG (www.ucg.br/pesquisa).

À Prof^a Dr^a Maria Fernanda Pimentel (Univerisidade Federal de Pernambuco) pela cessão dos dados de espectrometria de emissão em plasma.

Referências

1. Coelho, C. J. Calibração Multivariada Empregando Transformada Wavelet Adaptativa, Tese de Doutorado, ITA, São José dos Campos, (2002).
2. Chatterjee, S., Hadi, A. S. and Price, B. Regression Analysis By Example, v. 1, John Wiley (2000).
3. da Fonseca,J. S. Estatística Aplicada, v. 1, Atlas (2002).
4. Johnson, R. A. and Wichern, D. W. Applied Multivariate Statistical Analysis, v. 1, Prentice Hall (2002).
5. Draper, N. R. and Smith, H. Applied Regresion analysis, v. 1, 3, Wiley-Interscience (1998).
6. Downing, D. and Clark, J. Estatística Aplicada, v. 1, Saraiva (2000).
7. Kleinbaum, D. G., anda Keith E. Muller, L. L. K. and Nizm, A. Applied Regresion Analysis and Other Multivariable Methods, v. 1, (1998).
8. Konrath, A. C. Decomposição da estatística do gráfico de controle multivariado t de hotelling por meio de um algorítmo computacional, Universidade Federal de Santa Catarina (2002).
9. Levine, D. M., Berenson, M. L. and Stephan, D. Estatística: Teoria e aplicações,v. B, n. 52, p. 2151-2161 Spectrochimica Acta (1997).
10. Pimentel, M. F., de Barros Neto, B., de Araújo, M. C. U. and Pasquini, C. Simultaneous multielemental determination using a low-resoluiton inductively coupled plasma spectrometer/diode array detection system, Spectrochimica Acta **52** p. 2151–2161 (1997).
11. Ross M., Sheldon. Introduction to Probability Models - Seventh Edition, Berkeley, Califórnia, (2000).
12. Shewhart, W. A. The applications of statistics as an aid in maintaining quality of manufactured products., Journal of the American Statistical Association 20: **20** p.546–548 (1925).
13. Spiegel, Murray R. Theory and Problems of Probability and Statistics, Schaum's Outline Series, Mc Graw-Hill, New York, (1992).