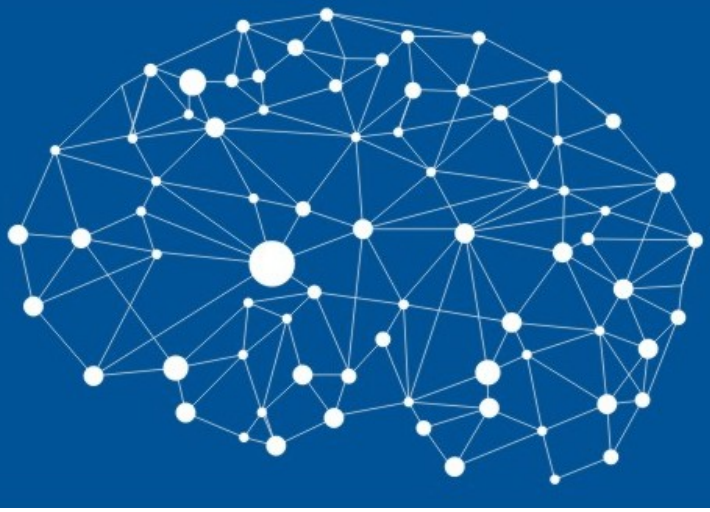

Detecção de outliers nas métricas científicas: estudo preliminar para dados univariados

Luís Fernando Maia Lima[1]; Alexandre Masson Maroldi[2]; Dávilla Vieira Odízio da Silva[3]; Maria Cristina Piumbato Innocentini Hayashi[4]; Carlos Roberto Massao Hayashi[5]

LIMA, L. S. F. M.; et al. Detecção de outliers nas métricas científicas: estudo preliminar para dados univariados In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais**... São Paulo: USP, 2016. p. A21

[1, 2 , 3] Universidade Federal de São Carlos (UFSCAR); [4, 5] Universidade Federal de São Carlos (UFSCAR)

DETECÇÃO DE *OUTLIERS* NAS MÉTRICAS CIENTÍFICAS:
estudo preliminar para dados univariados

Eixo temático: Métodos, Técnicas, e Ferramentos para Estudos Bibliométricos e Cientométricos

Modalidade: Apresentação oral

1 INTRODUÇÃO

O termo *outliers* já é de uso corrente na Estatística e representa valor(es) discrepante(s) no próprio conjunto de dados coletados, ou seja valor(es) que divergem bastante do padrão global dos demais dados observados (BARNETT; LEWIS, 1994; TRIOLA, 2012). Um aspecto de interesse prático é que *outliers* "podem revelar importantes informações" (TRIOLA, 2012, p. 97) sobre o conjunto de dados a ser analisado, inclusive nos estudos bibliométricos (GLÄNZEL; MOED, 2013; LIMA; MAROLDI; SILVA, 2013).

Nos estudos sobre referências de teses e dissertações, por exemplo, a detecção de *outliers* pode auxiliar na identificação dos trabalhos que mais se destacaram (tanto para menores e ou maiores quantidades de referências citadas) em relação ao próprio conjunto de valores observados. É também oportuno esclarecer que *outlier*s influem nos cálculos de média, desvio padrão, histogramas, podendo distorcer conclusões e generalizações sobre o conjunto de dados analisados (TRIOLA, 2012).

Vários métodos de detecção de *outliers* podem ser encontrados nos estudos de Barnett e Lewis (1994). Um destes métodos é cunhado por Barnett e Lewis (1994) de método *ad hoc*, e provém da Análise Exploratória de Dados (AED). A AED (TUKEY, 1977) fornece um método simples e rápido para detecção de *outliers* em dados univariados (a estatística univariada refere-se somente a uma variável). Além disto, a AED pode auxiliar os estudiosos das métricas científicas, conduzindo a análises estatísticas complementares, principalmente se há ocorrência de *outliers*.

Portanto, visto que *outliers* podem contribuir como fonte de informação adicional nas métricas científicas, cabe a seguinte questão: como identificar estes *outliers* para o caso de dados univariados.

2 REVISÃO DA LITERATURA

Ao longo dos anos, vários estudos foram realizados para a detecção de *outliers* via Análise Exploratória de Dados (AED). A seguir, apresentaremos os mesmos juntamente com suas definições:

Fórmula 1: $O.S. > Q_i + K(n,\alpha)*(Q_{ii} - Q_{iii})*F$

Fórmula 2: $O.I. < Q_i - K(n,\alpha)*(Q_{ii} - Q_{iii})*F$

O.S. := *outlier* superior

O. I. := *outlier* inferior

Q_i; Q_{ii}; Q_{iii} := representação genérica para os quartis, podendo ser, não necessariamente na mesma ordem, o primeiro quartil (Q1); o segundo quartil (Q2) e o terceiro quartil (Q3).

K(n,α) := fator que leva em conta o tamanho da amostra coletada ("n") e a probabilidade de ocorrência de *outliers* associada ("α").

F := grandeza que leva em conta outros fatores, como por exemplo a assimetria dos dados.

Um dos primeiros autores a estudar *outliers* via AED foi Tukey (1977), que apresentou a seguinte proposta:

Fórmula 3: $O.I. < Q1 - 1,5*(Q3 - Q1)$

Fórmula 4: $O.S. > Q3 + 1,5*(Q3 - Q1)$

Na teoria de Tukey (1977), temos que $K(n,\alpha) = 1,5$; e $F = 1$; ou seja, a formulação de Tukey (1977) não levava em conta nem o tamanho da amostra, nem a probabilidade associada e nem a assimetria amostral. A proposta de Tukey (1977) aplicava-se melhor para as distribuições gaussianas e com amostra com leve assimetria.

Alguns trabalhos procuraram modificar a ideia original de Tukey (1977), ora alterando-se somente os valores dos quartis (KIMBER, 1990), ora apresentando equações somente para o fator "K(n,α)" (CARLING, 2000; SCHWERTMAN, OWENS, ADNAN, 2004; SIM, GAN, CHANG, 2005; BANERJEE, IGLEWICZ, 2007; DOVOEDO,

CHAKRABORTI, 2015), e outros modificando somente o fator “F”, devida a assimetria dos dados amostrais (HUBERT, VANDERVIEREN, 2008; ADIL, IRSHAD, 2015).

Neste estudo, focaremos na terceira vertente, ou seja, nos modelos que procuram levar em conta somente o fator “F”, pois há possibilidade de serem utilizadas medidas robustas (pouco sensíveis a presença de *outliers*) para o cálculo da assimetria.

Na contribuição de Hubert e Vandervieren (2008), a assimetria dos dados (presente no fator “F”) é dado pelo fator “MC” (denominação de “medcouple”). Na formulação de Hubert e Vandervieren (2008) o valor máximo permissível para “MC” é 0,6. Para dados simétricos (como na distribuição gaussiana), MC = 0.

Portanto, a faixa de valores recomendados para uso do modelo de Hubert e Vandervieren (2008) é $-0{,}6 \leq MC \leq +0{,}6$. Todavia, a quantificação do fator “MC” exige uma complexa rotina computacional, o qual limita o uso da proposta de Hubert e Vandervieren (2008).

Por sua vez, Adil e Irshad (2015) apresentaram uma ligeira modificação da fórmula original de Hubert e Vandervieren (2008), neste sentido a equação da proposta de Adil e Irshad (2015), além de continuar levando em conta o fator “MC”, tal fator é multiplicado pelo fator “SK”, ou seja, o coeficiente momento de assimetria. A formulação final de Adil e Irshad (2015) foi:

Fórmula 5: $O.I. < Q1 - 1{,}5*(Q3 - Q1)*e^{-(SK)*|MC|}$

Fórmula 6: $O.S. > Q3 + 1{,}5*(Q3 - Q1)*e^{(SK)*|MC|}$

Todavia, os autores recomendaram que o valor máximo de “SK” a ser adotado nas fórmulas (5) e (6) seja 3,5; mesmo que na amostra coletada, o coeficiente momento de assimetria seja superior a 3,5.

Outro aspecto da proposta de Adil e Irshad (2015), é que o coeficiente momento de assimetria é justamente influenciado pela presença de *outliers*, ou seja, pode haver um viés inicial no cálculo do valor de “SK”, que por sua vez, irá propagar este viés no cálculo dos próprios *outliers*. Além disto, a proposta dos autores continua utilizando o fator “MC”, que não é de fácil quantificação.

Neste sentido, a proposta de nosso estudo de detecção de *outliers* segue a linha de pensamento de Hubert e Vandervieren (2008) e Adil e Irshad (2015), ou seja, leva em conta a assimetria dos dados amostrais, contudo, sem levar em conta o coeficiente momento de

assimetria (“SK”), que pode ser influenciado pela existência dos próprios *outlier*s, bem como desconsiderando o cálculo complexo do fator “MC”.

A medida de assimetria adotada em nossa formulação é o coeficiente octílico de assimetria (“OC”), que é uma medida de fácil cálculo, além de ser uma medida robusta (mais difícil de ser influenciada pela ocorrência de eventuais *outliers*). Portanto, a proposta deste estudo para detecção de *outliers*, sugeridos para tamanho amostral “n” maior ou igual a 30 ($n \geq 30$) e para dados univariados é:

Fórmula 7: $O.I. < Q1 - 1{,}5*(Q3 - Q1)*e^{-0{,}5*(OC)}$

Fórmula 8: $O.S. > Q3 + 1{,}5*(Q3 - Q1)*e^{0{,}5*(OC)}$

A grandeza “OC” é o coeficiente octílico de assimetria, dado por:

Fórmula 9: $OC = [P_{87,5} - 2*Q2 + P_{12,5}] / [P_{87,5} - P_{12,5}]$

$P_{87,5}$:= representa o 87,5º percentil.

$P_{12,5}$:= representa o 12,5º percentil.

Ressalta-se que na literatura estatística, não há consenso sobre o cálculo dos quartis. Assim, para efeitos deste trabalho, adotaremos para os cálculos dos quartis (Q1, Q2 e Q3) e dos percentis ($P_{87,5}$ e $P_{12,5}$), a metodologia de Triola (2012, p. 93).

3 APLICAÇÃO NAS MÉTRICAS CIENTÍFICAS

A fim de apresentar a aplicação das fórmulas (4); (8) e (9), para efeito de cálculo, usaremos os dados de Silva (2014), representada pelo diagrama de ramo e folha da figura 1:

Figura 1 – Diagrama de ramo e folha sobre a utilização da língua portuguesa nas referências.

0	0 0 1 2 2 2 2 3 4 4 5 5 5 6 6 6 8 8 8 8 8 8 9 9
1	0 1 1 1 2 3 3 3 3 3 4 4 4 5 5 6 6 6 6 6 6 7 7 8 8 8 9
2	0 0 0 0 0 1 1 1 2 3 3 3 5 6 7 7 7 8 8 9
3	0 0 0 0 0 1 1 3 3 3 4 4 5 7 8 9 9
4	0 3 3 3 5 6 6 8 9 9
5	0 1 4 5 6 8 8
6	4
7	0
8	
9	
10	7

Fonte: Elaborado pelos autores com base em Silva (2014)

Pelo diagrama de ramo e folhas, depreende-se que a massa de dados concentra-se nos valores de 0 (zero) a 58 citações, indicando que os valores de 64; 70 e 107 referências citadas

sejam possíveis *outliers* superiores. Também é possível visualizar que esta distribuição é assimétrica à direita, e que não segue a distribuição gaussiana.

Seguindo a recomendação de Triola (2012, p. 93), calculam-se, a partir da figura 1, os seguintes valores:

$P_{12,5}$ = 6 referências; Q1 = 11 referências; Q2 = 20 referências; Q3 = 33,5 referências; $P_{87,5}$ = 46 referências.

Após, utiliza-se a fórmula (9) para o cálculo de "OC":

Fórmula 9: $OC = [P_{87,5} - 2*Q2 + P_{12,5}] / [P_{87,5} - P_{12,5}]$

Substituindo-se os valores encontrados, vem:

OC = [46 – 2*(20) + 6] / [46 – 6]; assim; OC = 0,30 (valor que indica assimetria moderada)

Devido ao fato do diagrama de ramo e folhas (figura 1) esboçar somente a possível presença de *outliers* superiores, utilizaremos as fórmulas (4) e (8). Assim, calculando os *outliers* superiores segundo a fórmula (4) (TUKEY, 1977), vem:

Fórmula 4: O.S. > Q3 + 1,5*(Q3 – Q1)

Substituindo os valores de Q3 e Q1 já calculados, então:

O.S. > 33,5 + 1,5*(33,5 – 11); portanto; O.S. > 67,3 referências.

Verifica-se no diagrama de ramo e folhas (figura 1), que a formulação de Tukey (1977) indica a presença de dois *outliers*: os valores de 70 e 107 referências. É importante ressaltar que a formulação de Tukey presta-se melhor as distribuições gaussianas ou com leve assimetria, todavia, nossos dados indicam assimetria moderada, ou seja, em tese, a proposta de Tukey (1977) não é a mais recomendada para o cálculo de *outliers* para os nossos dados.

Usando a proposta de detecção de *outliers* via fórmula (8), então:

Fórmula 8: $O.S. > Q3 + 1,5*(Q3 - Q1)*e^{0,5*(OC)}$

Substituindo os valores de Q3, Q1 e "OC" já conhecidos, vem:

$O.S. > 33,5 + 1,5*(33,5 - 11)*e^{0,5*(0,30)}$; então, O.S. > 72,7 referências.

Neste caso, a nova proposta (que leva em conta a assimetria dos dados), acusa somente um *outlier*, o valor de 107 referências.

Simulando-se a retirada do *outlier* de 107 citações do conjunto de dados, e reaplicando-se novamente todos os cálculos, novamente a formulação de Tukey (1977) acusa

o valor de 70 referências como *outlier*, todavia a nova proposta não acusa novo *outlier*. Na simulação (com a retirada do *outlier* de 107 citações) não houve alteração significativa do valor do coeficiente octílico de assimetria ("OC"), que permaneceu no valor de 0,30; demonstrando que de fato o coeficiente octílico de assimetria tende a ser robusto, ou seja, pouco influenciável pela presença de *outliers*.

Portanto, a proposta de detecção de *outliers* para dados univariados auxilia os pesquisadores das métricas científicas a obter informações adicionais sobre características notáveis de um conjunto de dados, no nosso caso, a determinação do limite entre valores discrepantes e os valores usuais de uma massa de dados.

CONSIDERAÇÕES FINAIS

Este trabalho apresenta aos estudiosos das métricas científicas uma alternativa para detecção de *outliers* (valores discrepantes) com dados univariados como fonte adicional de informação. Tal formulação procura considerar a assimetria dos dados, fator este que não é contemplado na proposta original de Tukey (1977), que é seguida na maioria dos livros textos de Estatística.

Nos dados utilizados neste estudo, a formulação acusa somente a existência de um *outlier*, ao passo que a proposta de Tukey (1977) indica dois *outliers*. Mas a distribuição dos dados indica uma moderada assimetria dos dados à direita (reiterada pelo valor do coeficiente octílico de assimetria (OC) que é igual a 0,30), bem como que a distribuição não é gaussiana, o que limita a aplicação da formulação de Tukey (1977).

Recomendamos novos estudos e aplicações com outros dados desta formulação, para a validação da mesma junto aos pesquisadores das métricas científicas.

REFERÊNCIAS

ADIL, I. H.; IRSHAD, A. R. A modified approach for detection of outliers. **Pakistan Journal of Statistics and Operation Research**, v. 11, n. 1, p. 91-102, 2015.

BANERJEE, S.; IGLEWICZ, B. A simple univariate outlier identification procedure designed for large samples. **Communications in Statistics: simulation and computation**, v. 36, n. 2, p. 249-263, 2007.

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. 3. ed. New York: John Wiley & Sons, 1994.

CARLING, K. Resistant outlier rules and the non-Gaussian case. **Computational statistics & Data Analysis**, v. 33, n. 3, p. 249-258, may. 2000.

DOVOEDO, Y. H.; CHAKRABORTI, S. Boxplot-based outlier detection for the location-scale family. **Communications in Statistics – Simulation and Computation**, v. 44, n. 6, p. 1492-1513, 2015.

GLÄNZEL, W.; MOED, H. F. Thougts and facts on bibliometric indicators. **Scientometrics**, v. 96, n. 1, p. 381-394, jul. 2013.

HUBERT, M.; VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Computational Statistics & Data Analysis**, v. 52, n. 12, p. 5186-5201, aug. 2008.

KIMBER, A. C. Exploratory data analysis for possibly censored data from skewed distributions. **Journal of the Royal Statistical Society:** series C (applied statistics), v. 39, n. 1, p. 21-30, 1990.

LIMA, L. F. M.; MAROLDI, A. M.; SILVA, D. V. O. Outlier(s) em cálculos bibliométricos: primeiras aproximações. **Liinc em Revista**, v. 9, n. 1, p. 257-268, maio 2013.

SCHWERTMAN, N. C.; OWENS, M. A.; ADNAN, R. A simple more general boxplot method for identifying outliers. **Computational Statistics & Data Analysis**, v. 47, n. 1, p. 165-174, aug. 2004.

SILVA, D. V. O. da. **Elementos bibliométricos das referências nas dissertações defendidas no Programa de Mestrado de Biologia Experimental (PGBIOEXP) na Universidade Federal de Rondônia (UNIR), entre 2003 a 2010**. 2014. 51 f. Trabalho de Conclusão de Curso (Graduação) – Departamento de Ciência da Informação, Universidade Federal de Rondônia, Porto Velho, 2014.

SIM, C. H.; GAN, F. F.; CHANG, T. C. Outlier labeling with boxplot procedures. **Journal of the American Statistical Association**, v. 100, n. 470, p. 642-652, jun. 2005.

TRIOLA, M. F. **Introdução à estatística.** 10. ed. Rio de Janeiro: LTC, 2012.

TUKEY, J. W. **Exploratory data analysis**. Reading, Massachusetts: Addison-Wesley, 1977.