

ESTUDO PRELIMINAR SOBRE A INFLUÊNCIA DE *OUTLIERS* NAS MÉTRICAS CIENTÍFICAS PARA DADOS UNIVARIADOS

Luís Fernando Maia Lima

Universidade Federal de Rondônia, Porto Velho, RO, Brasil
maialima2000@gmail.com

Alexandre Masson Maroldi

Universidade Federal de Rondônia, Porto Velho, RO, Brasil
alexandre@unir.br

Dávilla Vieira Odízio da Silva

Instituto Federal do Amazonas, Manaus, AM, Brasil
davillaodizio@gmail.com

Carlos Roberto Massao Hayashi

Universidade Federal de São Carlos, São Carlos, SP, Brasil
massao@ufscar.br

Maria Cristina Piumbato Innocentini Hayashi

Universidade Federal de São Carlos, São Carlos, SP, Brasil
dmch@ufscar.br

1 INTRODUÇÃO

Os *outliers* são os valores que apresentam um padrão distinto dos demais dados coletados. Para Silva (2011, p. 93), *outlier* é uma “observação aberrante, anormal, atípica, contaminante, dissimilar, estranha, extrema, discordante ou preocupante”. Esclarece Rosado (2006, p. 1) que “uma única observação (não detectada) pode destruir ou contrariar a conclusão de qualquer trabalho”.

Neste sentido, Triola (2012) asseverou que os *outliers* apresentam influência nos cálculos da média, do desvio padrão, na representação gráfica dos dados, bem como “podem revelar importantes informações” (TRIOLA, 2012, p. 97) sobre os dados analisados. Portanto, é impres-



cíndivel nas métricas científicas tanto a detecção quanto o estudo dos efeitos dos *outliers* na análise dos dados.

Assim, os objetivos deste trabalho são: primeiro, apresentar uma nova fórmula para a detecção de *outliers* para dados univariados via Análise Exploratória de Dados (AED) e em segundo, quantificar o efeito dos *outliers* nos cálculos da média e desvio padrão.

2 REVISÃO DA LITERATURA

A sistematização do estudo dos *outliers* ocorreu em meados da segunda metade do século XX; todavia, o assunto *outlier* já era objeto de estudo da Estatística há muito tempo antes (ROSADO, 2006, p. 1).

No caso das métricas científicas, Lima, Maroldi e Silva (2013) alertaram para as conduções de análises com e sem a presença dos *outliers*. Posteriormente, Silva, Maroldi e Lima (2014) propuseram uma fórmula oriunda da AED de detecção de *outliers* para a determinação da elite científica em substituição ao critério da raiz quadrada de Price (PRICE, 1963). Todavia, o critério proposto por Silva, Maroldi e Lima (2014) falha quando o terceiro quartil é igual ao primeiro quartil.

Em extensa revisão sobre as diversas fórmulas com origem na AED para a detecção de *outliers* para dados univariados, Lima et al. (2017) esclareceram que havia três vertentes, sendo que uma destas vertentes procurava levar em conta a assimetria dos dados.

O trabalho pioneiro do uso da assimetria na detecção de *outliers* via AED é devido a Hubert e Vandervieren (2008), sendo a assimetria quantificada pelo “medcouple”. Já Adil e Irshad (2015) propuseram o uso também do coeficiente clássico de assimetria em conjunto com o “medcouple”. Por seu turno, Lima et al. (2017) propuseram o uso do coeficiente octílico de assimetria. Por outro lado, Babura et al. (2017) fizeram uso do coeficiente quartil de assimetria.

As fórmulas propostas neste trabalho para a detecção dos *outliers* são:

$$O.I. < Q_1 - 1,5*(Q_3 - Q_1)* e^{-(As)} \quad \text{Fórmula (1)}$$

$$O.S. > Q_3 + 1,5*(Q_3 - Q_1)* e^{+(As)} \quad \text{Fórmula (2)}$$

A terminologia é a seguinte:



O.I. := *outlier* inferior.

O.S. := *outlier* superior.

Q₁ := primeiro quartil.

Q₃ := terceiro quartil.

e := número de Euler; $e \approx 2,718\dots$

As := coeficiente quartil de assimetria. O valor pode ser positivo ou negativo.

O coeficiente quartil de assimetria é dado por:

$$As = [Q_3 - 2*Q_2 + Q_1] / [Q_3 - Q_1] \quad \text{Fórmula (3)}$$

Q₂ := segundo quartil.

Baseados em Silva (2011, p. 209), podemos afirmar que se:

$|As| = 0$; então a distribuição é simétrica.

$0 < |As| \leq 0,1$; então a distribuição é fracamente assimétrica.

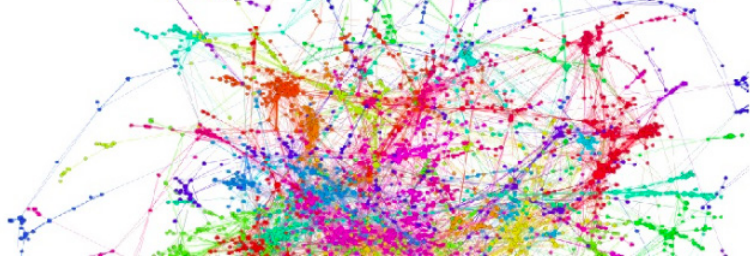
$0,1 < |As| < 0,3$; então a distribuição é moderadamente assimétrica.

$0,3 \leq |As| \leq 1,0$; então a distribuição é fortemente assimétrica.

As fórmulas (1) e (2) são similares à proposta de Babura et al. (2017). Todavia, a contribuição de Babura et al. (2017) utiliza os valores “3” e “5” para multiplicar o coeficiente de assimetria “As”; em nossa proposta, o valor é “1”. A justificativa para o valor “1” de nossa equação é que os coeficientes “3” e “5” conduzem à detecção de poucos *outliers* para o caso de dados com distribuição fracamente assimétrica.

3 METODOLOGIA

Para atingir os fins colimados, faremos uso dos dados de Santos (2010), inicialmente calculando os valores da média e desvio padrão com todos os dados com auxílio do *software* gratuito STATDISK (www.statdisk.org). Para a detecção dos *outliers*, aplicaremos as fórmulas (3); (1) e (2), respectivamente. Para o cálculo dos quartis, usaremos a proposta de Bussab e Morettin (2002, p. 42) de interpolação linear para os dados.



Após, retiramos os *outliers* detectados dos valores, e calcularemos novamente a média e o desvio padrão, a fim de quantificar a influência dos *outliers* nas métricas científicas.

Os dados de Santos (2010, p. 175) para o fator de impacto de 42 periódicos que utilizaremos neste trabalho são: 149; 179; 208; 213; 222; 227; 270; 286; 361; 400; 426; 508; 526; 580; 645; 645; 656; 667; 714; 811; 816; 849; 851; 909; 1000; 1000; 1000; 1000; 1000; 1000; 1000; 1000; 1029; 1053; 1111; 1111; 1333; 1395; 1667; 1786; 1898; 2500. Estes valores devem ser divididos por 10000 ao final.

4 RESULTADOS E DISCUSSÃO

O diagrama de ramo e folhas correspondente aos dados de Santos (2010, p. 175) pode ser encontrado em Lima et al. (2017, p. 266), onde parece haver a ocorrência de 4 a 6 *outliers*. Iniciamos a rotina de cálculo com os quartis; assim encontramos: $Q_1 = 426$ (1º Quartil); $Q_2 = 832,5$ (2º Quartil); $Q_3 = 1000$ (3º Quartil). Com o uso da fórmula (3) calculamos o coeficiente quartil de assimetria:

$$As = [Q_3 - 2*Q_2 + Q_1] / [Q_3 - Q_1] \quad ; \quad As = [1000 - 2*(832,5) + 426] / [1000 - 426]$$

$$As = -0,4164 \text{ (assimetria negativa ou à esquerda forte).}$$

A fórmula (1) identifica os valores para os quais começam os *outliers* inferiores:

$$O.I. < Q_1 - 1,5*(Q_3 - Q_1)*e^{-(As)}; \quad O.I. < 426 - 1,5*(1000 - 426)*e^{-(-0,4164)}$$

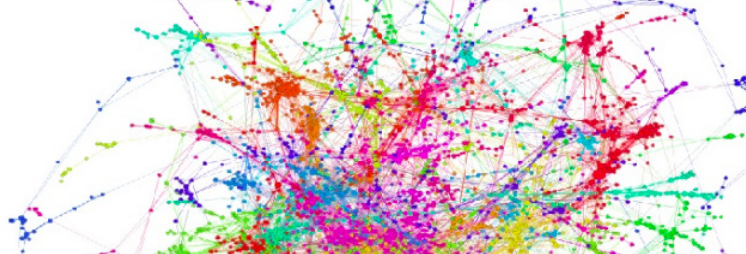
$$O.I. < -879,70 \text{ (não há presença de outliers inferiores)}$$

Já a fórmula (2) quantifica o início dos *outliers* superiores:

$$O.S. > Q_3 + 1,5*(Q_3 - Q_1)*e^{+(As)}; \quad O.S. > 1000 + 1,5*(1000 - 426)*e^{+(-0,4164)}$$

$$O.S. > 1567,76 \text{ (Identificam-se 4 outliers superiores: 1667; 1786; 1898 e 2500)}$$

Lima et al. (2017) calcularam o coeficiente octílico de assimetria, encontrando também uma assimetria negativa ou à esquerda fraca, todavia já no limiar da assimetria moderada. A divergência na assimetria (fraca ou forte) ocorre devido ao uso dos quartis da própria amostra,



contudo, o coeficiente quartil de assimetria é menos influenciado pelos *outliers* do que o coeficiente octílico de assimetria, sendo preferível, pois, o coeficiente quartílico do que o octílico. Em Santos (2010) não houve cálculo do coeficiente de assimetria.

No tocante aos *outliers*, Lima et al. (2017, p. 267) identificaram 3 *outliers* superiores, já Santos (2010, p. 141) visualizou pelo menos 1 *outlier* superior. A título de informação, a formulação clássica de detecção de *outliers* (sem o uso da assimetria) fornece 2 *outliers*, como demonstrado em Lima et al. (2017, p. 267). Como vimos, em nossa pesquisa houve a presença de 4 *outliers*, detectando, portanto, um maior de número deles na amostra.

Finalmente, para quantificar a influência dos *outliers* nas métricas científicas, calculamos para todos os 42 dados no *software* STATDISK a média como 833,36 (ou 0,08, pois os valores ainda devem ser divididos por 1000) e o desvio padrão como 507,63 (ou 0,05, lembrando que os valores necessitam divididos por 1000). Trabalhando somente com 38 dados (ou seja, retirando os 4 *outliers*) encontramos para a média 714,47 (ou 0,07) e para o desvio padrão 349,30 (ou 0,035).

Em seus achados, Santos (2010, p. 141) encontrou para a média (sem retirar nenhum *outlier*) também o valor 0,08 (mesmo valores de nossa pesquisa); e para o desvio padrão 0,04 (em nossa pesquisa encontramos 0,05); a possível divergência talvez seja devida ao uso distinto do *software* estatístico para o cálculo, pois usamos o STATDISK, ao passo que Santos (2010, p. 98) utilizou o SPSS.

Assumindo os cálculos de Santos (2010, p. 141) então a média reduziu de 0,08 para 0,07. Já o desvio padrão apresentou redução de 0,04 para 0,035. Mas vale lembrar que os valores são arredondados para dois dígitos após a vírgula. Então, para visualizarmos melhor o efeito dos *outliers* nas métricas científicas usaremos os achados de nossa pesquisa: média de 833,36 com *outliers* e 714,47 sem *outliers*; para o desvio padrão 507,63 com *outliers* e 349,5 sem *outliers*.

Assim, para a média houve uma redução de 14,3%; em termos de conta: $\{[(714,47 - 833,36)/(833,36)] * 100\}$. Já para o desvio padrão há



uma diminuição de 31,2%; em termos de cálculo: $\{[(349,3 - 507,63)/(507,63)] * 100\}$.

Podemos considerar que a redução dos valores de 14,3% para a média e de 31,2% para o desvio padrão são elevados, e que na condução dos estudos de métricas científicas os *outliers* são objetos a serem levados em consideração na análise dos dados.

Atítulo de curiosidade, a média dos 4 *outliers* é 1962,75. Como a média dos outros 38 dados é 714,47, a média dos *outliers* é 2,75 (1962/714,47) superior à média dos outros dados.

Assim, a fórmula de detecção dos *outliers* serve para identificar dois grupos: um homogêneo (contendo a maioria dos dados e o padrão da distribuição) e um não homogêneo (os *outliers*, valores discrepantes), mas que influem nas conclusões dos dados.

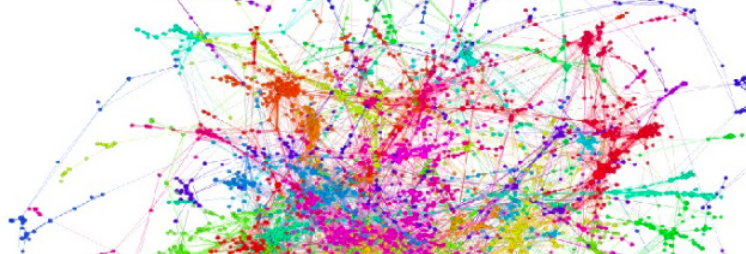
4 CONSIDERAÇÕES FINAIS

A contribuição deste trabalho foi tanto identificar os *outliers* como quantificar os seus efeitos nas métricas científicas em consonância com as contribuições de Lima, Maroldi e Silva (2013) e Triola (2012).

Quanto aos objetivos, apresentamos uma nova fórmula para detecção de *outliers*, levando em consideração a assimetria quartil dos dados. Além disto, usou-se esta nova equação para os dados de Santos (2010), identificando (após a retirada dos *outliers* da amostra) uma redução de 14,3% no cálculo da média e uma redução de 31,2% no cálculo do desvio padrão. Sem dúvida, estas reduções dos valores da média e do desvio padrão são elevadas.

Sugerem-se também estudos com dados que apresentem somente *outliers* inferiores e posteriormente com a existência simultânea de *outliers* inferiores e superiores, e novamente quantificar os efeitos dos *outliers* nas métricas científicas.

Como observação final, alerta-se aos pesquisadores das métricas científicas pelo menos identificar os *outliers* nas suas análises de dados, e se for o caso, quantificar o efeito dos mesmos.



REFERÊNCIAS

- ADIL, I. H.; IRSHAD, A. ur R. A modified approach for detection of outliers. **Pakistan Journal of Statistics and Operation Research**, Lahore, v. 11, n. 1, p. 91-102, Apr. 2015.
- BABURA, B. I. et al. Modified boxplot for extreme data. **AIP Conference Proceedings**, New York, v. 1842, issue 1, May 2017.
- BUSSAB, W.; MORETTIN, P. **Estatística Básica**. 5.ed. São Paulo: Saraiva, 2002.
- HUBERT, M.; VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Computacional Statistics & Data Analysis**, Amsterdam, v. 52, n. 12, p. 5186-5201, Aug. 2008.
- LIMA, L. F. M.; MAROLDI, A. M.; SILVA, D. V. O. da. Outlier(s) nos cálculos bibliométricos: primeiras aproximações. **Liinc em Revista**, Rio de Janeiro, v. 9, n. 1, p. 257-268, maio 2013.
- LIMA, L. F. M. et al. Métricas científicas em estudos bibliométricos: detecção de outliers para dados univariados. **Em Questão**, Porto Alegre, v. 23, Edição Especial 5 EBBC, p. 254-273, jan. 2017.
- PRICE, D. J. de S. **Little Science, Big Science**. New York: Columbia University Press, 1963.
- ROSADO, F. **Outliers em dados estatísticos**. Lisboa: Sociedade Portuguesa de Estatística, 2006.
- SANTOS, S. M. dos. **Perfil dos periódicos científicos de Ciências Sociais e Humanidades**: mapeamento das características extrínsecas. 2010. 176 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2010.
- SILVA, D. V. O. da; MAROLDI, A. M.; LIMA, L. F. M. Outliers na lei do Elitismo. **Em Questão**, Porto Alegre, v. 20, n. 3, p. 43-60, 2014.
- SILVA, D. J. L. da. **Estatística aplicada à investigação científica nas Ciências do Desporto**: análise exploratória de dados com recurso ao SPSS. Medelo, Portugal: Instituto de Estudos Superiores de Fafe, 2011.
- TRIOLA, M. F. **Introdução à Estatística**. 10. ed. Rio de Janeiro: LTC, 2012.