em questão
E-ISSN 1808-5245

Métricas científicas em estudos bibliométricos: detecção de *outliers* para dados univariados

Luís Fernando Maia Lima
Doutor; Fundação Universidade Federal de Rondônia, Porto Velho, RO, Brasil;
maialima2000@gmail.com

Alexandre Masson Maroldi
Doutorando; Fundação Universidade Federal de Rondônia, Porto Velho, RO, Brasil;
alexandre@unir.br

Dávilla Vieira Odízio da Silva
Bacharela; Fundação Universidade Federal de Rondônia, Porto Velho, RO, Brasil;
davilla_jp@hotmail.com

Carlos Roberto Massao Hayashi
Doutor; Universidade Federal de São Carlos, São Carlos, SP, Brasil;
massao@ufscar.br

Maria Cristina Piumbato Innocentini Hayashi
Doutora; Universidade Federal de São Carlos, São Carlos, SP, Brasil;
dmch@ufscar.br

Resumo: Apresenta fórmulas, para dados univariados, de detecção de *outliers* que levem em conta a assimetria dos dados, tanto positiva como negativa. A nova formulação, proveniente da Análise Exploratória de Dados, é simulada comparando os resultados com a proposta oriunda da Análise Exploratória de Dados, presente na maioria dos livros-textos de estatística e *softwares* estatísticos, mas que se aplica somente para distribuições normais ou gaussianas, ou seja, simétricas ou com leve assimetria. Para a simulação, são utilizados dados reais publicados por dois trabalhos na área de métricas científicas. Para assimetrias positivas (negativas) moderadas ou fortes, a nova formulação detecta menor (maior) quantidade de *outliers* superiores que a proposta clássica. É importante levar em conta a existência de *outliers* nos dados bibliométricos, pois recomenda-se quantificar a influência dos mesmos nos cálculos estatísticos, tais como média e desvio padrão.

Palavras-chave: *Outliers*. Análise Exploratória de Dados. Assimetria. Bibliometria. Univariado.

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

diverge(m) bastante do padrão global dos demais dados observados (BARNETT; LEWIS, 1994; TRIOLA, 2012).

Um aspecto relevante é que *outliers* "podem revelar importantes informações" (TRIOLA, 2012, p. 97) sobre o conjunto de dados a ser analisado, inclusive nos estudos bibliométricos (GLÄNZEL; LIMA; MAROLDI; SILVA, 2013; MOED, 2013). Outro detalhe também muito importante é que *outliers* influem nos cálculos de média, desvio padrão, histogramas, podendo distorcer conclusões e generalizações sobre o conjunto de dados analisados (TRIOLA, 2012).

Um dos estudos a apontar como *outliers* podem influir nos cálculos bibliométricos e a conduzir a interpretações errôneas foi o de Bensman, Smolinsky e Pudovkin (2010), que detectaram e ilustraram a presença de um *outlier* extremo na distribuição de frequência do fator de impacto da área físico-matemática, e como esse *outlier* extremo influía tanto na interpretação dos dados como nas conclusões de seus resultados.

Por sua vez, Mutz e Daniel (2012) também afirmam que o fator de impacto de periódicos, proposto inicialmente por Garfield em 1955, é altamente sensível quando há *outliers*, gerando valores elevados que podem conduzir a conclusões errôneas.

Nesse sentido, 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, como ocorreu no trabalho de Silva (2014).

Já Santos (2010), em seus achados sobre periódicos científicos indexados na SciELO Brasil nas áreas de Ciências Sociais e Humanidades, detectou *outliers* para os seguintes dados: número de artigos publicados, quantidade de fascículos publicados, citações concedidas, citações recebidas, vida média dos periódicos, fator de impacto e número de acessos.

Existem vários métodos para detecção de *outliers*, os quais podem ser encontrados nos estudos de Barnett e Lewis (1994). Um desses métodos foi

cunhado por Barnett e Lewis (1994) de método *ad hoc*, e provém da Análise Exploratória de Dados (AED).

A AED fornece um método simples e rápido para detecção de *outliers* em dados univariados (a estatística univariada se refere somente a uma variável). (TUKEY, 1977) Ademais, a AED pode auxiliar os estudiosos das métricas científicas, conduzindo a análises estatísticas complementares, principalmente se há ocorrência de *outliers*.

Bornmann et al. (2008) reforçaram que a AED deve ser utilizada em cálculos bibliométricos com o uso de gráficos e também a detecção de *outliers*, em virtude de poderem suprir informações relevantes.

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

Fórmula 1: O.I. < Q1 – 1,5*(Q3 – Q1)

Fórmula 2: O.S. > Q3 + 1,5*(Q3 – Q1)

O.S. := *outlier* superior

O.I. := *outlier* inferior

Q1 := primeiro quartil

Q2 := segundo quartil

Q3 := terceiro quartil

As fórmulas (1) e (2) são usadas na maioria dos *softwares* estatísticos, como os que foram utilizados nos trabalhos de Silva (2014) e Santos (2010). Todavia, as fórmulas acima propostas por Tukey (1977) se aplicam para as distribuições normais ou gaussianas, ou seja, distribuições simétricas ou com leve (fraca) assimetria.

No entanto, outros trabalhos apontam que as Fórmulas (1) e (2) devem também levar em consideração: o tamanho da amostra (“n”), a probabilidade de ocorrência de *outliers* (“α”) ou assimetria da distribuição de frequência dos dados, fatores não contemplados nas Fórmulas (1) e (2). (ADIL; IRSHAD, 2015; BANERJEE; IGLEWICZ, 2007; BRANT, 1990; BRUFFAERTS; VERARDI; VERMANDELE, 2014; CARLING, 2000; CARTER; SCHWERTMAN; KISER, 2009; DOVOEDO; CHAKRABORTI, 2015; HOAGLIN; IGLEWICZ; TUKEY, 1986; HOAGLIN; IGLEWICZ, 1987; HUBERT; VANDERVIEREN, 2008;

SCHWERTMAN; OWENS; ADNAN, 2004; SCHWERTMAN; SILVA, 2007; SIM; GAN; CHANG, 2005).

Portanto, visto que *outliers* podem contribuir como fonte de informação adicional nas métricas científicas, cabe a seguinte questão: como identificar os *outliers*, via AED, para o caso de dados univariados, levando-se em conta a assimetria dos dados?

2 Revisão da literatura

Em termos gerais, as diversas fórmulas para a detecção de *outliers* via Análise Exploratória de Dados (AED) possuem a seguinte estrutura:

Fórmula 3: O.S. > $Q_i + K(n,\alpha)*(Q_{ii} - Q_{iii})*F$

Fórmula 4: 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).

$K(n,\alpha)$:= 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.

Na teoria de Tukey (1977), fórmulas (1) e (2), temos que $K(n,\alpha) = 1,5$; e $F = 1$; ou seja, a formulação de Tukey (1977) não considera nem o tamanho da amostra, nem a probabilidade associada e nem a assimetria amostral. A proposta de Tukey (1977) se aplicava melhor para as distribuições gaussianas e com amostra com leve assimetria.

Alguns trabalhos procuraram modificar a ideia original de Tukey (1977), podendo ser identificadas três vertentes: a primeira vertente procura alterar somente a posição dos valores dos quartis (KIMBER, 1990); já a segunda vertente apresenta equações somente para o fator "K(n,α)" (BANERJEE; IGLEWICZ, 2007; BRANT, 1990; CARLING, 2000; CARTER; SCHWERTMAN; KISER,

2009; DOVOEDO; CHAKRABORTI, 2015; HOAGLIN; IGLEWICZ; TUKEY, 1986; HOAGLIN; IGLEWICZ, 1987; SCHWERTMAN; OWENS; ADNAN, 2004; SCHWERTMAN; SILVA, 2007; SIM; GAN; CHANG, 2005); finalmente, a terceira vertente visa modificar somente o fator "F", devido à assimetria dos dados amostrais (ADIL; IRSHAD, 2015; BRUFFAERTS; VERARDI; VERMANDELE, 2014; HUBERT; VANDERVIEREN, 2008).

Neste estudo, focaremos a terceira vertente, ou seja, os modelos que procuram levar em conta somente o fator "F" (ADIL; IRSHAD, 2015; BRUFFAERTS; VERARDI; VERMANDELE, 2014; HUBERT; VANDERVIEREN, 2008), pois há possibilidade de serem utilizadas medidas robustas (pouco sensíveis à presença de *outliers*) para o cálculo da assimetria.

Na contribuição de Hubert e Vandervieren (2008), a assimetria dos dados (presente no fator "F") é dada 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, a qual limita o uso da proposta de Hubert e Vandervieren (2008).

Já Bruffaerts, Verardi e Vermandele (2014) utilizam uma transformação dos dados originais para as denominadas distribuições "g" e "h" de Tukey, de modo a levar em conta assimetrias fortes (fator "MC" $< -0,6$ ou fator "MC" $> 0,6$), bem como a cauda íngreme da distribuição considerada, constituindo-se, pois, em um aperfeiçoamento das ideias de Hubert e Vandervieren (2008). Contudo, permanece o problema da complexa rotina computacional necessária para a transformação dos dados.

Por sua vez, Adil e Irshad (2015) apresentaram uma ligeira modificação da fórmula original de Hubert e Vandervieren (2008). De acordo com a proposta de Adil e Irshad (2015), continua-se a levar em conta o fator "MC", contudo, 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. > $Q3 + 1{,}5*(Q3 - Q1)*e^{(SK)*|MC|}$

e := número de Euler; e ≈ 2,718...

Todavia, Adil e Irshad (2015) 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 esse viés no cálculo dos próprios *outliers*. Além disso, a proposta dos autores continua utilizando o fator "MC", que não é de fácil quantificação.

Nesse sentido, a proposta de nosso estudo de detecção de *outliers* segue a linha de cálculo 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 *outliers*, bem como desconsiderando o cálculo complexo do fator "MC".

Então, para substituir o coeficiente momento de assimetria ("SK") e o fator "MC", faremos uso das denominadas medidas robustas de assimetria (mais difíceis de serem influenciadas pela ocorrência de eventuais *outliers*). O exemplo mais clássico, presente na maioria dos livros-textos da área, é o coeficiente quartil de assimetria, que leva em conta os quartis. Mas para a nossa proposta, utilizaremos o coeficiente octílico de assimetria ("OC"), que também foi citado nos estudos de Hubert e Vandervieren (2008).

A grandeza "OC" é dada por:

Fórmula 7: $OC = [P_{87,5} - 2*Q2 + P_{12,5}] / [P_{87,5} - P_{12,5}]$

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

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

A grandeza "OC" é adimensional. Uma interpretação para os valores do coeficiente octílico de assimetria ("OC"), **adaptada por nós** (grifo nosso), com base em Silva et al. (1996, p. 125) é, neste caso:

$|OC| = 0$; distribuição simétrica.

$0 < |OC| \leq 0,1$; assimetria fraca.

$0,1 < |OC| < 0,3$; assimetria moderada.

$0,3 \leq |OC| \leq 1,0$; assimetria forte.

Portanto, a proposta deste estudo para detecção de *outliers* sugere as seguintes hipóteses: tamanho amostral "n" maior ou igual a 30 ($n \geq 30$); dados univariados; o valor do terceiro quartil diferente do valor do primeiro quartil ($Q3 \neq Q1$) é:

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

Fórmula 9: $O.S. > Q3 + 1,5*(Q3 - Q1)*e^{0,5*(OC)}$

Por fim, 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 proposta por Triola (2012, p. 93).

3 Metodologia

A fim de apresentar a aplicação das fórmulas (1); (2); (7); (8) e (9) nas métricas científicas, usaremos os dados de Santos (2010) e Silva (2014). Justifica-se o uso das Fórmulas (1) e (2), pois são as mais utilizadas nos softwares estatísticos. As Fórmulas (8) e (9) correspondem à nossa proposta, sendo ambas precedidas pelo cálculo do coeficiente octílico de assimetria, fórmula (7).

Também elegemos a escolha dos trabalhos de Santos (2010) e Silva (2014), pois são muito poucos os trabalhos na área de métricas científicas os quais apresentam a presença de outliers. Ademais, os trabalhos citados de alguma forma calcularam ou detectaram a ocorrência de outliers em seus dados. Além disso, ambos os trabalhos apresentam os dados brutos, o que permite replicar os resultados.

4 Resultados e discussão

Começaremos pelo trabalho de Silva (2014), que estudou 108 referências de dissertações em um programa de pós-graduação na Universidade Federal de Rondônia. Um dos achados foi a utilização da língua portuguesa nas referências de cada dissertação. Esses dados são apresentados pelo diagrama de ramo e folhas da Figura 1:

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

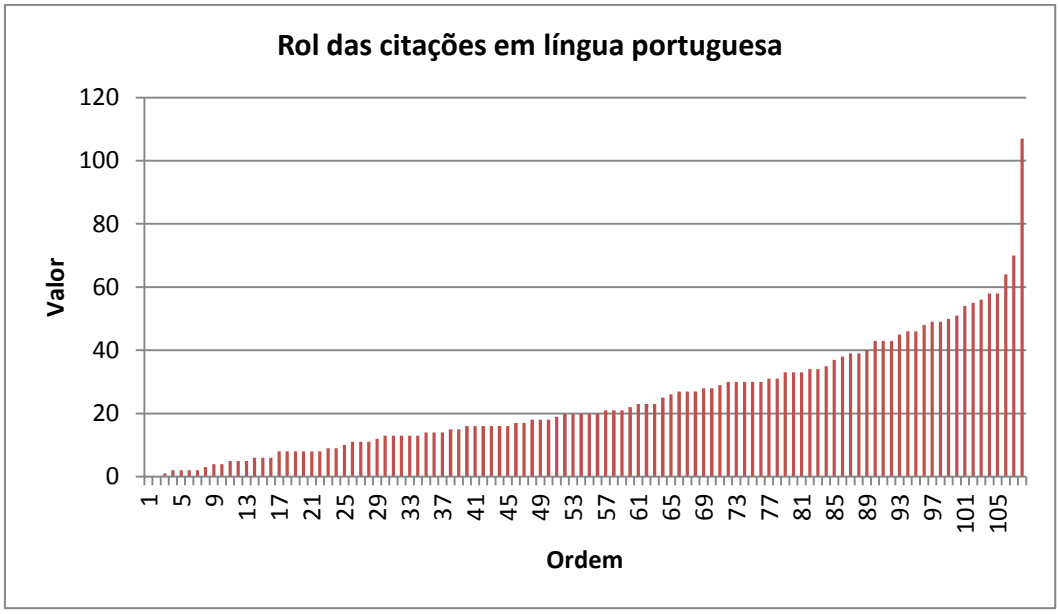
0	0 0 1 2 2 2 2 3 4 4 5 5 5 6 6 6 8 8 8 8 8 8 9 9
1	0 1 1 1 2 3 3 3 3 3 4 4 4 5 5 6 6 6 6 6 6 7 7 8 8 8 9
2	0 0 0 0 0 1 1 1 2 3 3 3 5 6 7 7 7 8 8 9
3	0 0 0 0 0 1 1 3 3 3 4 4 5 7 8 9 9
4	0 3 3 3 5 6 6 8 9 9
5	0 1 4 5 6 8 8
6	4
7	0
8	
9	
10	7

Fonte: Elaborado pelos autores com base em Silva (2014, anexo em Excel).

Pelo diagrama de ramo e folhas, depreende-se que a massa de dados se concentra 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 essa distribuição é assimétrica à direita e que não segue a distribuição gaussiana.

A Figura 2 reforça que os valores 64; 70 e 107 sejam candidatos a *outliers*, pois a disposição dos dados segue uma reta crescente que justamente nos três últimos valores citados (64; 70 e 107) apresenta um aumento (mudança) na declividade da reta.

Figura 2 – Rol com os dados da utilização da língua portuguesa nas referências.



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

Seguindo a recomendação de Triola (2012, p. 93), calculam-se, a partir da Figura 1, as seguintes estatísticas para os 108 dados de Silva (2014):

a) localização e valores dos elementos $P_{12,5}$; Q1; Q2; Q3 e $P_{87,5}$:

a.1.) Elemento $P_{12,5}$

108*(0,125) = 13,5º elemento. Portanto, para $P_{12,5}$, devemos selecionar o 14º elemento. Assim: $P_{12,5}$ = 6 referências.

a.2.) Elemento Q1

108*(0,25) = 27º elemento. Portanto, para Q1, devemos calcular a média aritmética entre o 27º elemento (valor 11) e o 28º elemento (valor 11 também, nesse caso). Logo: Q1 = 11 referências.

a.3.) Elemento Q2

108*(0,50) = 54º elemento. Portanto, para Q2, novamente devemos calcular a média aritmética, agora entre o 54º elemento (valor 20) e o 55º elemento (valor 20). Portanto: Q2 = 20 referências.

a.4.) Elemento Q3

108*(0,75) = 81º elemento. Portanto, para Q3, calculamos a média aritmética entre o 81º elemento (valor 33) e o 82º elemento (valor 34). Logo: Q3 = 33,5 referências.

a.5.) Elemento $P_{87,5}$

108*(0,875) = 94,5º elemento. Nesse caso, para $P_{87,5}$, selecionamos o 95º elemento. Então: $P_{87,5}$ = 46 referências.

De posse dos valores de $P_{12,5}$; Q1; Q2; Q3 e $P_{87,5}$; calcula-se o coeficiente octílico "OC":

Fórmula 7: $OC = [P_{87,5} - 2*Q2 + P_{12,5}] / [P_{87,5} - P_{12,5}]$

Substituindo-se os valores encontrados, vem:

OC = [46 – 2*(20) + 6] / [46 – 6]; assim; OC = 0,300; valor que indica assimetria positiva (ou à direita) forte (SILVA et al., 1996).

Depois, passamos ao cálculo de *outliers*:

Fórmula 1: O.I. < Q1 – 1,5*(Q3 – Q1); O.I. < 11 – 1,5*(33,5 – 11);

O. I. < – 22,75 referências. Não há *outlier* inferior, pois o valor mínimo é zero referência.

Fórmula 2: O.S. > Q3 + 1,5*(Q3 – Q1); O.S. > 33,5 + 1,5*(33,5 – 11);

O.S. > 67,25 referências.

Verifica-se no diagrama de ramo e folhas (Figura 1) que a formulação de Tukey (1977), Fórmulas (1) e (2), indica a presença de dois *outliers* superiores: os valores de 70 e 107 referências. Esses dois *outliers* também foram os mesmos encontrados por Silva (2014, p. 33), pois houve a utilização da metodologia de Triola (2008) para o cálculo dos quartis.

É importante ressaltar que a formulação de Tukey (1977) se presta melhor às distribuições gaussianas ou com leve assimetria. Todavia, nossos dados indicam assimetria forte, ou seja, em tese, a proposta de Tukey (1977) não é a mais recomendada para o cálculo de *outliers* para os nossos dados.

Fazendo-se uso da fórmula (8) vem:

Fórmula 8: $O.I. < Q1 - 1,5*(Q3 - Q1)*e^{-0,5*(OC)}$;

$O.I. < 11 - 1,5*(33,5 - 11)*e^{-0,5*(0,30)}$; O.I. < – 18,05 referências. Novamente não há *outlier* inferior, pois o valor mínimo é nenhuma referência.

Para a detecção de *outliers* superiores:

Fórmula 9: $O.S. > Q3 + 1,5*(Q3 - Q1)*e^{0,5*(OC)}$

$O.S. > 33,5 + 1,5*(33,5 - 11)*e^{0,5*(0,30)}$; O.S. > 72,71 referências.

Portanto, pela nova proposta de detecção de *outliers*, que leva em conta a assimetria dos dados, há a presença de somente um *outlier* superior, o valor de 107 referências. O efeito da assimetria positiva é aumentar o valor de O.S., portanto, há maior possibilidade de ser detectada menor quantidade de *outliers* do que a proposta clássica de Tukey (1977).

Já a dissertação de Santos (2010) estudou 70 periódicos científicos nas áreas de Ciências Sociais e Humanidades indexados na base SciELO Brasil. Uma análise preliminar usando estatística multivariada detectou três grupos de periódicos com características similares. A nossa análise será limitada ao grupo um, que englobou 42 periódicos, e também limitada aos seguintes dados univariados: citações recebidas, fator de impacto e número de acessos.

Para as citações recebidas, a Figura 3 apresenta os dados de Santos (2010, p. 175):

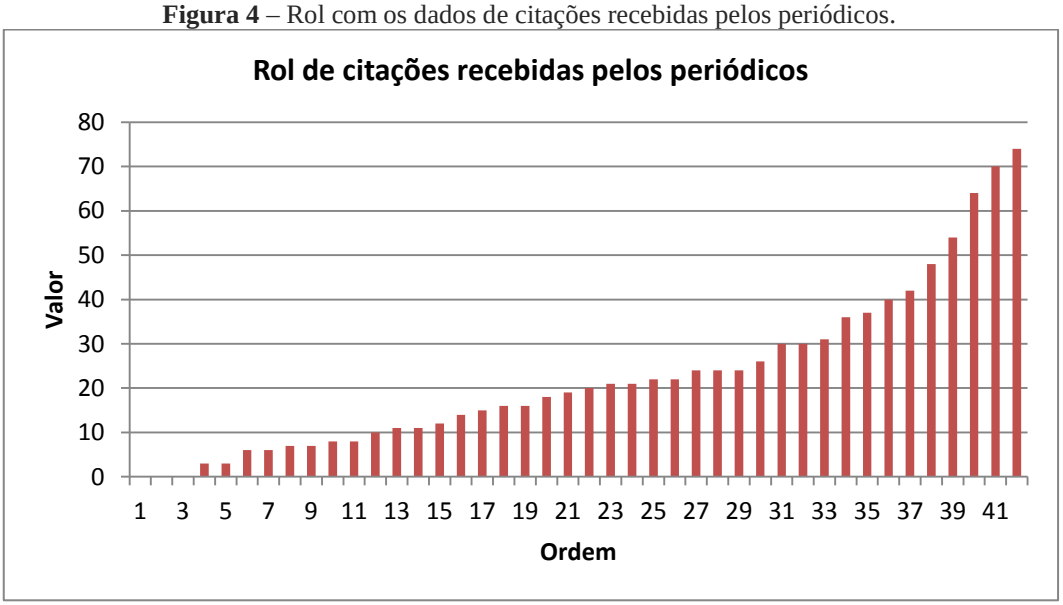
Figura 3 – Diagrama de ramo e folhas sobre as citações recebidas pelos periódicos.

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

Fonte: Elaborado pelos autores com base em Santos (2010, p. 175).

Pela análise da Figura 3, há quatro possíveis candidatos a *outliers* superiores: os valores 54; 64; 70 e 74. Também se observa que a distribuição dos dados não segue a forma normal ou gaussiana. Já pela Figura 4, aparentemente há três *outliers* superiores: 64; 70 e 74 (apresentam declividade de reta diferente dos demais dados).

Figura 4 – Rol com os dados de citações recebidas pelos periódicos.



Fonte: Elaborado pelos autores com base em Santos (2010).

Seguindo-se o roteiro de cálculo explanado anteriormente, encontram-se as seguintes grandezas:

$P_{12,5}$ = 6 referências; Q1 = 8 referências; Q2 = 19,5 referências; Q3 = 30 referências; $P_{87,5}$ = 42 referências. O valor de "OC" fica 0,25 (assimetria positiva moderada).

Pela formulação de Tukey (1977), fórmulas (1) e (2), achamos: O.I. < – 25 referências (não há *outliers* inferiores) e O.S. > 63 referências. Portanto, detecta-se a presença de três *outliers* (valores 64, 70 e 74 referências). Nossos resultados estão em concordância qualitativa com os achados de Santos (2010, p. 138), que também detectou mais de um *outlier*. É importante rememorar que possíveis divergências de cálculo ocorram devido ao cálculo dos quartis, que variam de acordo com o *software* estatístico utilizado.

Todavia, pela proposta nova de detecção de *outliers*, encontramos O.I. < – 17,70 referências (não há *outliers* inferiores) e O.S. > 72,37 referências. Assim, há somente um *outlier* superior, o valor de 74 referências.

Já para o fator de impacto, os dados de Santos (2010, p. 175) são representados pela Figura 5:

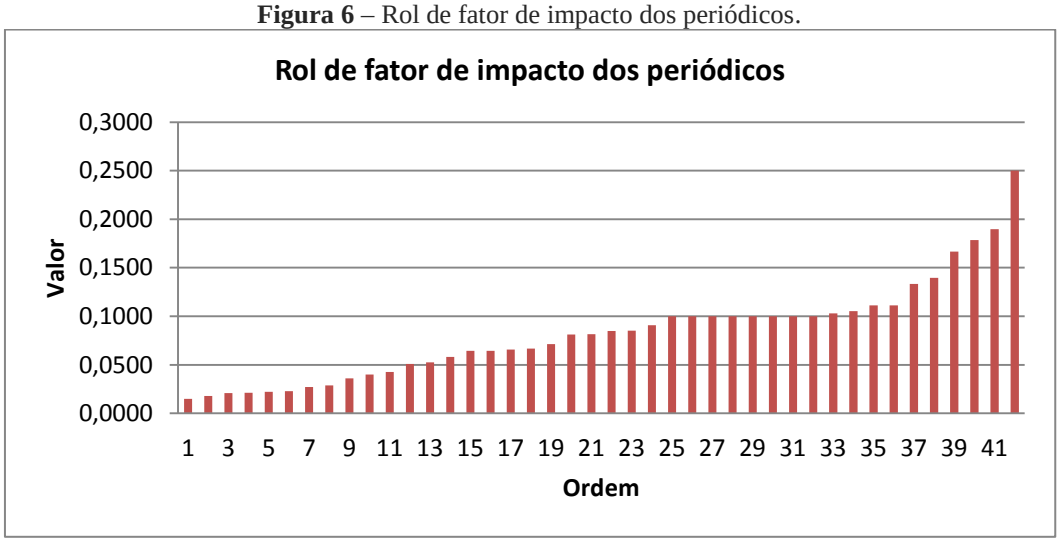
Figura 5 – Diagrama de ramo e folhas para o fator de impacto. Os valores devem ser divididos por 10.000.

1	49; 79;
2	08; 13; 22; 27; 70; 86;
3	61;
4	00; 26;
5	08; 26; 80;
6	45; 45; 56; 67;
7	14;
8	11; 16; 49; 51;
9	09;
10	00; 00; 00; 00; 00; 00; 00; 00; 29; 53;
11	11; 11;
12	
13	33; 95;
14	
15	
16	67;
17	86;
18	98;
19	
20	
21	
22	
23	
24	
25	00;

Fonte: Elaborado pelos autores com base em Santos (2010, p. 175).

O diagrama de ramo e folhas da Figura 5 parece sugerir de quatro a seis *outliers* superiores: 0,1333; 0,1395; 0,1667; 0,1786; 0,1898; 0,2500. Observar novamente que a distribuição dos dados não segue a curva normal ou gaussiana. A Figura 6 também corrobora a possível presença dos mesmos seis *outliers* superiores (observar a declividade distinta dos demais dados).

Figura 6 – Rol de fator de impacto dos periódicos.



Fonte: Elaborado pelos autores com base em Santos (2010).

Novamente usando-se o roteiro de cálculo para os quartis e percentis, vem: $P_{12,5}$ = 0,0227; Q1 = 0,0426; Q2 = 0,08325; Q3 = 0,1000; $P_{87,5}$ = 0,1333. O valor de "OC" fica – 0,095; assimetria negativa fraca, mas muito próximo da assimetria negativa moderada, conforme Silva et al. (1996, p. 125). É importante reiterar que agora os dados fornecem uma assimetria negativa.

Pelas Fórmulas (1) e (2) devidas a Tukey (1977) achamos: O.I. < – 0,0435 (não há *outliers* inferiores) e O.S. > 0,1861. Portanto, detecta-se a presença de dois *outliers* superiores (valores 0,1898 e 0,2500).

De novo, os resultados acima estão em concordância qualitativa com os achados de Santos (2010, p. 141), que detectou pelo menos um *outlier*. A possível divergência é devida ao cálculo dos quartis, que, reforçamos, varia de acordo com o *software* estatístico usado.

Com a utilização das Fórmulas (7) e (8), achamos O.I. < – 0,0521 (não há *outliers* inferiores) e O.S. > 0,1783. Agora, há a detecção de três *outliers* superiores (valores 0,1786; 0,1898 e 0,2500).

É importante neste ponto observar que para assimetrias positivas, quanto maior o valor de "OC", menor a quantidade de *outliers* superiores detectados, como ocorreu nos dados de Silva (2014) e Santos (2010) para citações recebidas. O inverso ocorre para *outliers* inferiores: há a possibilidade de se encontrar maior quantidade de *outliers* pela nova formulação.

Todavia, para assimetrias negativas, quanto menor o valor de "OC", maior a quantidade de *outliers* superiores encontrados, como vimos agora para o fator de impacto com os dados de Santos (2010). Novamente o inverso ocorre para *outliers* inferiores: provavelmente menor a quantidade de *outliers*.

Para os dados de números de acesso por periódico de Santos (2010), optamos por apresentar os dados brutos e realizar um diagrama de ramos e folhas simplificado. Os dados são (SANTOS, 2010, p. 175): 30.861; 33.591; 40.989; 49.446; 51.665; 54.562; 65.728; 66.866; 78.300; 90.487; 101.393; 105.368; 106.806; 111.850; 113.221; 114.990; 118.032; 123.354; 127.364; 134.513; 138.608; 160.254; 164.039; 173.140; 179.777; 185.617; 193.276; 206.831; 211.992; 222.685; 230.255; 245.697; 250.092; 300.635; 305.645; 322.341; 351.242; 413.582; 466.984; 469.225; 665.384; 756.163.

Com base nos dados acima, realiza-se o diagrama de ramo e folhas simplificado da Figura 7.

Figura 7 – Diagrama de ramo e folhas simplificado dos números de acessos de periódicos.

0	X X X X
0	X X X X X X
1	X X X X X X X X X X X
1	X X X X X X
2	X X X X X
2	X
3	X X X
3	X
4	X
4	X X
5	
5	
6	
6	X
7	
7	X

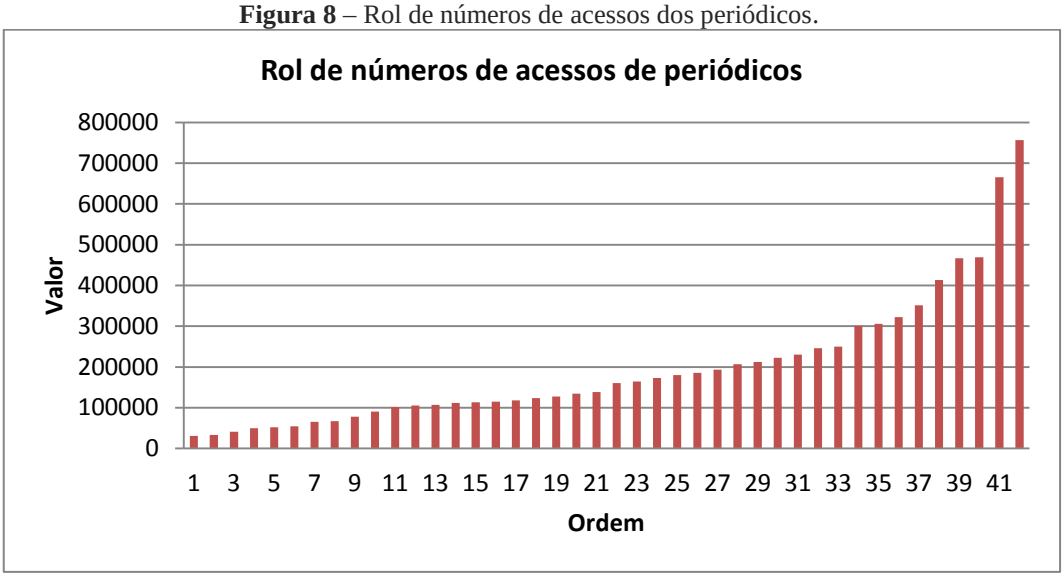
Fonte: Elaborado pelos autores com base em Santos (2010, p. 175).

A Figura 7 ilustra a possibilidade de dois *outliers* (valores 665.384 e 756.163), além dos valores superiores a 350.000 serem também possíveis *outliers*

(valores 351.242; 413.582; 466.984; 469.225); portanto, há a possibilidade de haver até seis *outliers*. Outro aspecto que deve ser visualizado na Figura 7 é a assimetria existente dos dados.

Já pela Figura 8, a suspeita dos dois valores mais elevados serem *outliers* é reforçada; além disso, a declividade dos três valores imediatamente anteriores é diferente do padrão dos demais dados. Desse modo, pela Figura 8, parece haver a existência de dois a cinco *outliers*.

Figura 8 – Rol de números de acessos dos periódicos.



Fonte: Elaborado pelos autores com base em Santos (2010).

A análise estatística dos dados para os valores de quartis e percentis fornece: $P_{12,5}$ = 54.562; Q1 = 101.393; Q2 = 149.431; Q3 = 245.697; $P_{87,5}$ = 351.242. O valor de "OC" fica 0,360; assimetria positiva forte (SILVA et al., 1996, p. 125).

Pela contribuição de Tukey (1977), as Fórmulas (1) e (2) geram: O.I. < – 115.063 (não há *outliers* inferiores) e O.S. > 462.153. Portanto, há a detecção de quatro *outliers* superiores (valores 466.984; 469.225; 665.384; 756.163). Nossos resultados estão em concordância qualitativa com os resultados de Santos (2010, p. 142), que também visualizou quatro *outliers*.

Considerando-se a assimetria dos dados, pelas Fórmulas (7) e (8), encontramos O.I. < – 49.623,2 (não há *outliers* inferiores) e O.S. > 661.494,8. Agora, há a presença de somente dois *outliers* superiores (valores 665.384; 756.163).

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.

5 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 (assimetria tanto positiva como negativa), fator este que não é contemplado na proposta original de Tukey (1977), que é seguida na maioria dos livros-textos de Estatística e ou também nos *softwares* estatísticos.

É importante atentar para o efeito que a assimetria exerce na detecção dos *outliers*: para assimetrias positivas (negativas), há maior chance de ser encontrado menor (maior) número de *outliers* superiores. Já para os *outliers* inferiores, a assimetria positiva (negativa) possibilita encontrar maior (menor) número de *outliers*.

O efeito da assimetria na detecção dos *outliers* é mais dramático (mais sentido) quanto maior for a assimetria dos dados, ou seja, para assimetrias moderadas ou fortes (e, portanto, distribuições que se afastam da distribuição normal ou gaussiana), quando a hipótese de utilização da formulação clássica de Tukey (1977) é violada.

Sugerimos aos bibliometristas verificarem a possibilidade de existência de *outliers* em seus dados univariados, tanto para detecção como para realizar análises alternativas (por exemplo, para média e desvio padrão) com e sem a presença dos *outliers* achados, pois os mesmos podem conter informações relevantes para análises e interpretações de seus dados de pesquisa. Isso vai ao encontro das observações de Bornmann et al. (2008); Bensman, Smolinsky e Pudovkin (2010); Mutz e Daniel (2012); Glänzel e Moed (2013); Lima, Maroldi e Silva (2013).

Por fim, consideramos relevante alertar a comunidade científica dos estudos métricos que durante nossas buscas bibliográficas, encontramos dificuldade em alguns trabalhos que não disponibilizam os dados brutos de seus estudos, essenciais para a replicação de resultados, como também é escasso o número de pesquisas que se preocupam com a existência de *outliers* em seus achados, o que pode comprometer suas análises quantitativas.

Referências

ADIL, Iftikhar Hussain; IRSHAD, Ateeq ur Rehman. A modified approach for detection of outliers. **Pakistan Journal of Statistics and Operation Research**, Lahore, v. 11, n. 1, p. 91-102, Apr. 2015.

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

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

BENSMAN, Stephen J.; SMOLINSKY, Lawrence J.; PUDOVKIN, Alexander I. Mean citation rate per article in Mathematics journals: differences from the scientific model. **Journal of the American Society for Information Science and Technology**, New York, v. 61, n. 7, p. 1440-1463, July 2010.

BORNMANN, Lutz et al. Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. **Ethics in Science and Environmental Politics**, Oldendorf/Luhe, v. 8, p. 93-102, 2008. Disponível em: <http://www.int-res.com/articles/esep2008/8/e008p093.pdf>. Acesso em: 5 set. 2016.

BRANT, Rollin. Comparing classical and resistant outlier rules. **Journal of the American Statistical Association**, Boston, v. 85, n. 412, p. 1083-1090, Dec. 1990.

BRUFFAERTS, Christopher; VERARDI, Vincenzo; VERMANDELE, Catherine. A generalized boxplot for skewed and heavy-tailed distributions. **Statistics and Probability Letters**, Amsterdam, v. 95, p. 110-117, Dec. 2014.

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

CARTER, Nancy; SCHWERTMAN, Neil C.; KISER, Terry L. A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. **Statistical Methodology**, Amsterdam, v. 6, n. 6, p. 604-621, Nov. 2009.

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

GLÄNZEL, Wolfgang; MOED, Henk. F. Thougts and facts on bibliometric indicators. **Scientometrics**, Dordrecht, v. 96, n. 1, p. 381-394, Jul. 2013.

HOAGLIN, David C.; IGLEWICZ, Boris. Fine-tuning some resistant rules for outlier labeling. **Journal of the American Statistical Association**, Boston, v. 82, n. 400, p. 1147-1149, Dec. 1987.

HOAGLIN, David C.; IGLEWICZ, Boris; TUKEY, John W. Performance of some resistant rules for outlier labeling. **Journal of the American Statistical Association**, Boston, v. 81, n. 396, p. 991-999, Dec. 1986.

HUBERT, M.; VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Computational Statistics & Data Analysis**, Amsterdam, 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)**, London, v. 39, n. 1, p. 21-30, Jan. 1990.

LIMA, Luís Fernando Maia Lima; MAROLDI, Alexandre Masson; SILVA, Dávilla Vieira Odízio da. Outlier(s) em cálculos bibliométricos: primeiras aproximações. **Liinc em Revista**, Rio de Janeiro, v. 9, n. 1, p. 257-268, maio 2013.

MUTZ, Rüdiger; DANIEL, Hans-Dieter. Skewed citation distributions and bias factors: solutions to two core problems with the journal impact factor. **Journal of Informetrics**, Amsterdam, v. 6, n. 2, p. 169-176, Apr. 2012.

SANTOS, Solange Maria 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.

SCHWERTMAN, Neil C.; OWENS, Margaret Ann; ADNAN, Robiah. A simple more general boxplot method for identifying outliers. **Computational Statistics & Data Analysis**, Amsterdam, v. 47, n. 1, p. 165-174, Aug. 2004.

SCHWERTMAN, Neil C.; SILVA, Rapti de. Identifying outliers with sequencial fences. **Computational Statistics & Data Analysis**, Amsterdam, v. 51, n. 8, p. 3800-3810, May 2007.

SILVA, Dávilla Vieira Odízio 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.

SILVA, Ermes Medeiros da; et al. **Estatística para os cursos de Economia, Administração, Ciências Contábeis**. 2. ed. São Paulo: Saraiva, 1996. v. 1.

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

TRIOLA, Mario F. **Introdução à Estatística.** 10. ed. Rio de Janeiro: LTC, 2012.

TUKEY, John Wilder. **Exploratory data analysis**. Reading, Massachusetts: Addison-Wesley, 1977.

Scientific metrics on bibliometric studies: detection of outliers for univariate data

Abstract: This study presents formulas for detection of outliers for univariate data, taking into consideration the positive as well as the negative asymmetry of data. This new formula is based on the Exploratory Data Analysis and is simulated through the comparison of the outcome of the Exploratory Data Analysis found in statistical text books and statistical software. However, only normal or Gaussian distribution, i.e., symmetric or slightly asymmetric values, are applied. Real data published in two scientific papers on metrics are used for the simulation. For moderate or strong positive (negative) asymmetries, the new formulation detects a lower (higher) quantity of superior outliers. It is important to take into account the existence of outliers in bibliometric data; it is recommended to quantify the influence of outliers in statistical calculation, such as mean and standard deviation.

Keywords: Outliers. Exploratory Data Analysis. Asymmetry. Bibliometry. Univariate.

Recebido em: 17/09/2016

Aceito em: 08/11/2016