Em Questão
ISSN: 1807-8893
ISSN: 1808-5245
emquestao@ufrgs.br
Universidade Federal do Rio Grande do Sul
Brasil

A influência de outliers nos estudos métricos da informação: uma análise de dados univariados

Lima, Luís Fernando Maia; Maroldi, Alexandre Masson; Silva, Dávilla Vieira Odízio da; Hayashi, Carlos Roberto Massao; Hayashi, Maria Cristina Piumbato Innocentini

A influência de outliers nos estudos métricos da informação: uma análise de dados univariados
Em Questão, vol. 24, 2018
Universidade Federal do Rio Grande do Sul, Brasil
Disponível em: https://www.redalyc.org/articulo.oa?id=465658737012
DOI: https://doi.org/10.19132/1808-5245240.216-235

Este trabalho está sob uma Licença Creative Commons Atribuição-NãoComercial 3.0 Internacional.

Artigos

A influência de outliers nos estudos métricos da informação: uma análise de dados univariados

The influence of outliers on metric studies of information: an analysis of univariate data

Luís Fernando Maia Lima 1
Universidade Federal de Rondônia, Brasil
luis.fernando@unir.br

Alexandre Masson Maroldi 2
Universidade Federal de Rondônia, Brasil
alexandre@unir.br

Dávilla Vieira Odízio da Silva 3
Instituto Federal do Amazonas, Lábrea, Brasil
davilla.odizio@gmail.com

Carlos Roberto Massao Hayashi 4
Universidade Federal de São Carlos, São Carlos, Brasil
massao@ufscar.br

Maria Cristina Piumbato Innocentini Hayashi 5
Universidade Federal de São Carlos, São Carlos, Brasil
dmch@ufscar.br

DOI: https://doi.org/10.19132/1808-5245240.216-235
Redalyc: https://www.redalyc.org/articulo.oa?id=465658737012

Recepção: 14 Setembro 2018
Aprovação: 03 Dezembro 2018

Resumo:

Este artigo apresenta uma nova fórmula de detecção de *outliers* via Análise Exploratória de Dados, levando em conta a assimetria dos dados, e também estuda o efeito da remoção dos *outliers* dos dados originais. Aplica-se a fórmula para três conjuntos de dados publicados na literatura de estudos métricos da informação. O primeiro conjunto de dados apresenta cinco *outliers* inferiores. A média, dos dados agregados, conduz à falsa impressão de que 40 universidades, de um total de 49, estão acima da média. A remoção dos cinco *outliers* inferiores conduz a uma nova média em que somente 22 universidades estão acima da média. No segundo conjunto de dados há a presença de cinco *outliers* inferiores e um *outlier* superior. Neste caso, o outlier superior ameniza o efeito dos *outliers* inferiores. No terceiro conjunto de dados, detectam-se cinco *outliers* superiores e um outlier inferior. A média, dos dados agregados, aponta que dez universidades estão acima da média. Removendo-se os seis *outliers* dos dados originais, encontra-se que

Autor notes

1 Doutor; Universidade Federal de Rondônia, Porto Velho, RO, Brasil
luis.fernando@unir.br

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

3 Especialista; Instituto Federal do Amazonas, Lábrea, AM, Brasil
davilla.odizio@gmail.com

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

5 Doutora; Universidade Federal de São Carlos, São Carlos, SP, Brasi
dmch@ufscar.br

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

28 universidades estão acima do novo valor da média. Para os três conjuntos de dados analisados o trabalho também demonstra o efeito dos *outliers* na estimativa intervalar (inferência estatística): a remoção dos *outliers* gera valores mais representativos tanto para a média como para o desvio padrão da amostra analisada. Portanto, evidencia-se como *outliers* podem afetar resultados e conclusões nos estudos métricos da informação. Todavia, a fórmula para a detecção de outliers apresenta-se aberta para futuras pesquisas.

Palavras-chave: Outliers, Univariados, Bibliometria, Assimetria, Análise Exploratória de Dados.

Abstract:

This paper presents a new formula for detecting outliers through Exploratory Data Analysis, while taking data asymmetry into account. The effect of removing outliers from the original dataset was also assessed. The new formula was applied on three datasets published in the literature on metric studies of information. The first dataset presented five lower outliers. The average of aggregate data conveyed a false impression that 40 universities, from a total of 49, were above average. The removal of the five lower outliers leads to a new average in which only 22 universities were above average. In the second dataset, there were five lower outliers and one upper outlier. In this case, the upper outlier eventually weaken the effect of the lower outliers. In the third dataset, five upper outliers and one lower outlier are detected. The average of aggregate data revealed that 10 universities were above average. Removing the six outliers from the original dataset, it was found that 28 universities were above the new average score. For the three datasets analyzed, the assessment demonstrated the effect of the outliers on the interval estimation (statistical inference): the removal of outliers generated a mean and standard deviation that were more representative of the sample analyzed. Therefore, became evident how outliers could influence results and conclusions in metric studies of the information. However, the formula for outliers' detection is open for future research.

Keywords: Outliers, Univariate, Bibliometry, Assimetry, Exploratory Data Analysis.

1 Introdução

A Estatística (TRIOLA, 2012), entre as suas diversas partes, apresenta a produção de dados (amostragem), a estatística descritiva (resumo e descrição dos dados coletados), a probabilidade (que faz a conexão entre a estatística descritiva e a inferencial) e a estatística inferencial (generalização de características da população com base nos dados da amostra).

Em que pese todo o cuidado para que a coleta dos dados seja de uma amostra aleatória e representativa (TRIOLA, 2012), há a possibilidade da ocorrência de *outliers* na amostra.

Os *outliers* são os valores que apresentam um padrão distinto dos demais dados coletados. De outra maneira, acabam sendo valores que não são representativos da população estudada. Como exemplo hipotético, caso seja feita uma pesquisa sobre peso da população humana, mesmo a amostragem sendo aleatória, há a possibilidade de ocorrência de *outlier* inferior (pessoas com peso muito baixo, sintoma da anorexia), bem como ser coletado *outlier* superior (pessoas com peso elevado, sintoma de obesidade).

Silva (2011, p. 93) define *outlier* como uma "observação aberrante, anormal, atípica, contaminante, dissimilar, estranha, extrema, discordante ou preocupante". Além disto, esclarece Rosado (2006, p. 1) que "uma única observação (não detectada) pode destruir ou contrariar a conclusão de qualquer trabalho".

Um excelente exemplo sobre a influência dos *outliers* nos estudos das métricas da informação encontra-se em Silva e Schulz (2018). Os autores comentam que o Chile aparece bem posicionado no *ranking* de "eficiência financeira" entre os artigos publicados e o gasto em pesquisa no país. Contudo, Silva e Schulz (2018) observam que a simples análise superficial dos dados escondia um verdadeiro outlier superior: a produção científica na área de Astronomia e Astrofísica (em colaboração com pesquisadores de outros países). Este único *outlier* superior gera um viés nos resultados dos indicadores do Chile, dando a falsa impressão de que há uma "eficiência financeira" digna de nota e em consonância com o esclarecimento de Rosado (2006).

Outro exemplo de presença de *outliers* nos estudos métricos da informação é dado no trabalho de Alvarez e Caregnato (2018), que ao estudarem os agradecimentos devido ao financiamento de pesquisa citados nos artigos brasileiros na Web of Science (WoS), entre 2009 a 2016, apresentam em sua Figura 1 diversos diagramas de caixa (*boxplot*) para a porcentagem de artigos com agradecimentos em relação ao total publicado por área de conhecimento.

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

Nos achados de Alvarez e Caregnato (2018), há a presença de um *outlier* inferior nas seguintes quatro áreas de conhecimento: primeiro em Agricultura, Biologia e Meio Ambiente (o *outlier* inferior é a subárea Política e Economia Agrícola); segundo em Biomedicina (o *outlier* inferior é a subárea Anatomia e Morfologia); terceiro na área de Engenharia e Tecnologia (o *outlier* inferior é a subárea Transporte) e quarto na área de Física (o *outlier* inferior é a subárea Termodinâmica). Para cada área de conhecimento, o *outlier* inferior representa o valor mínimo destoante (discrepante) dos demais valores.

Porém, na área de Humanidades ocorre a presença de um outlier superior (Arqueologia) nos resultados de Alvarez e Caregnato (2018). Esse *outlier* superior significa que a presença de agradecimentos na subárea de Arqueologia é um valor máximo destoante (discrepante) das demais subáreas de Humanidades.

Já a ocorrência de múltiplos *outliers* superiores pode ser encontrada na Figura 1 do trabalho de Silva, Almeida e Grácio (2018), em que o *boxplot* dos 25% maiores periódicos com Fator de Impacto apresenta três *outliers* superiores. Já o *boxplot* dos 25% maiores periódicos com índice "*h*" apresenta seis *outliers* superiores.

É importante reforçar que a detecção dos *outliers* "podem revelar importantes informações" (TRIOLA, 2012, p. 97) sobre os dados analisados. Além disto, outro aspecto, além da detecção, é verificar como os *outliers* influenciam os cálculos da média, do desvio padrão e do histograma (TRIOLA, 2012) dos valores coletados.

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 ou a influência dos *outliers* nos cálculos: a) da média e desvio padrão, tanto da estatística descritiva (estimativa pontual) como da estatística inferencial (estimativa intervalar), e b) na conclusão dos resultados. Trata-se de um estudo ampliado da contribuição de Lima et al. (2018) nos Anais do 6º Encontro Brasileiro de Bibliometria e Cientometria (6º EBBC). Entre os principais aspectos da ampliação há a inclusão de outros trabalhos apresentados no 6º EBBC que apresentam *outliers* em suas análises (ALVAREZ; CAREGNATO, 2018; SILVA; ALMEIDA; GRÁCIO, 2018; SILVA; SCHULZ, 2018) para demonstrar a atualidade e a importância dos *outliers* nos estudos bibliométricos e cientométricos. Outra modificação na ampliação é a substituição do único conjunto de dados de Lima et al. (2018) por três outros conjuntos de dados. Por fim, procura-se mostrar os efeitos dos *outliers* nas inferências (generalizações) estatísticas.

2 Revisão da literatura

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

No caso dos estudos métricos da informação, Lima, Maroldi e Silva (2013) alertam para as conduções de análises com e sem a presença dos outliers. Posteriormente, Silva, Maroldi e Lima (2014) propõem uma fórmula oriunda da AED (TUKEY, 1977) 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), com base em Tukey (1977), falha quando o terceiro quartil é igual ao primeiro quartil.

A conexão entre *outliers* e a elite de Price (raiz quadrada do total de autores) é dada no trabalho de Lima et al. (2017a) que citam os trabalhos de Barnett e Lewis (1978) e Chhikara e Feiveson (1980) sobre o número máximo de *outliers* em uma amostra, este valor é dado também pela raiz quadrada do total da amostra, ou seja, estrutura idêntica ao enunciado de Price (1963).

Em extensa revisão sobre as diversas fórmulas com origem na AED para a detecção de *outliers* para dados univariados, Lima et al. (2017b) esclarecem que há três vertentes.

A primeira vertente altera somente a utilização dos valores dos quartis, sendo representado somente pelo trabalho de Kimber (1990), segundo Lima et al. (2017b). Todavia, vale ressaltar que os trabalhos de Tambay (1988) e Dümbgen e Riedwyl (2007) também se enquadram nesta vertente.

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

Já a segunda vertente altera somente o fator que leva em conta o tamanho amostral e a probabilidade de ocorrência de outliers (LIMA et al., 2017b).

Apesar das tentativas da primeira e da segunda vertente em aperfeiçoarem a contribuição original de Tukey (1977), é imperioso reforçar que se aplicam primordialmente para distribuições simétricas ou com fraca assimetria, gerando resultados insatisfatórios quando utilizadas em distribuições com assimetria moderada ou forte.

Finalmente, a terceira vertente inclui o uso da assimetria dos dados para a detecção dos dados, que, no trabalho pioneiro de Hubert e Vandervieren (2008), é a assimetria quantificada pelo "medcouple". O "medcouple" é a mediana da função de Kernel sobre todos os pares de valores inferiores e superiores ao valor da mediana da amostra. O cálculo do "medcouple" exige extensa rotina computacional.

Já Adil e Irshad (2015) propõem o uso também do coeficiente clássico de assimetria em conjunto com o "medcouple". Por seu turno, Lima et al. (2017b) propõem o uso do coeficiente octílico de assimetria. Em outro trabalho recente, Babura et al. (2017) fazem uso do coeficiente quartil de assimetria.

O coeficiente octílico de assimetria leva em conta a mediana, o 12,5o percentil e o 87,5o percentil. Já o coeficiente quartil de assimetria leva em consideração a mediana, o primeiro quartil e o terceiro quartil.

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

O.I. < Q1 – 1,5*(Q3 – Q1)*e–(As) Fórmula (1)

O.S. > Q3 + 1,5*(Q3 – Q1)*e+(As) 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 ≈ 2,718....

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

O coeficiente quartil de assimetria é dado por:

As = [Q3 – 2*Q2 + Q1]/[Q3 – Q1] Fórmula (3)

Q2 := segundo quartil.

Segundo Silva (2011, p. 209), se:

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

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

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

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

As fórmulas (1) e (2) são recomendadas para tamanhos amostrais ("n") superiores a 30 (n > 30); todavia é desejável que n > 40. Ambas as fórmulas são similares à proposta de Babura et al. (2017). Todavia, a contribuição de Babura et al. (2017) utiliza os valores "3" (três) e "5" (cinco) 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.

Deve também ser mencionado que no trabalho de Lima et al. (2017b) o coeficiente octílico é multiplicado por "0,5" (meio). Nesta nova contribuição, o valor adotado é "1" (um) para multiplicar o coeficiente quartil de assimetria. As divergências entre os valores multiplicadores (meio ou um ou outro valor) devem ser objeto de futuras pesquisas.

Por fim, complementando o trabalho de Lima et al. (2017b), pode-se identificar uma quarta vertente, representada pelas contribuições de Walker e Chakraborti (2013) e Barnett e Cohen (2000), que utilizam as razões interquartílicas ou razões quartílicas, respectivamente, como fator de modificação da fórmula de Tukey (1977).

3 Metodologia

Como este trabalho é sobre modelação matemática/estatística de detecção de *outliers* e seus efeitos nos estudos métricos da informação, é mister que os dados a serem modelados devem, portanto, possuir outliers. Para atingir os fins colimados, fazemos uso dos dados de Vanti e Casado (2015), conforme a Tabela 1, pois como se pode ver na seção de resultados e discussão, os três conjuntos de dados apontam para a existência de outliers.

A primeira coluna apresenta o número total de páginas *web*. A segunda coluna representa as citações URL. Já a terceira coluna é o Fator de Impacto Web (FIW) revisado, que é obtido pela divisão dos valores da segunda coluna pela primeira coluna. Os dados foram encontrados no *webcrawler* acadêmico *Webometric Analyst (WeboAnaly)*, segundo Vanti e Casado (2015).

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

TABELA 1
Avaliação de um periódico de comunicação

Número de páginas WeboAnaly	Citações URL WeboAnaly	FIW Revisado
949	605	0,6375
932	612	0,6567
898	620	0,6904
940	536	0,5702
956	448	0,4686
934	486	0,5203
889	280	0,3150
971	539	0,5551
918	537	0,5850
902	602	0,6674
972	490	0,5041
932	477	0,5118
937	488	0,5208
917	555	0,6052
977	488	0,4995
983	570	0,5799
598	682	1,1405
986	564	0,5720
793	479	0,6040
927	566	0,6106
966	544	0,5631
921	547	0,5939
943	458	0,4857
947	504	0,5322
838	298	0,3556
872	497	0,5700
942	384	0,4076
979	540	0,5516
933	431	0,4620
726	513	0,7066
861	537	0,6237
955	567	0,5937
409	499	1,2200
977	428	0,4381
944	491	0,5201
968	477	0,4928
900	715	0,7944
988	373	0,3775
929	499	0,5371
962	543	0,5644
966	537	0,5559
945	348	0,3683
936	590	0,6303
915	800	0,8743
933	484	0,5188
936	463	0,4947
909	511	0,5622
149	495	3,3221
929	505	0,5436

Fonte: Adaptado de Vanti e Casado (2015).

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

Para cada coluna, inicialmente calculamos os valores da média e desvio padrão (estatística descritiva) com todos os dados com auxílio do *software* gratuito STATDISK (<www.statdisk.org>). O *software* também fornece os valores do intervalo de confiança (estatística inferencial) dos valores da média e desvio padrão.

Para a detecção dos *outliers*, foram aplicadas as fórmulas (3); (1) e (2), respectivamente. Para o cálculo dos quartis, usamos a proposta de Bussab e Morettin (2002) de interpolação linear para os dados.

Em seguida, retiramos os *outliers* detectados dos valores e calculamos novamente a média, o desvio padrão (estatística descritiva) e os intervalos de confiança (estatística inferencial) da média e do desvio padrão, a fim de quantificar a influência dos *outliers* nos estudos métricos da informação.

4 Resultados e discussão

Para melhor compreensão deste trabalho, para cada coluna da Tabela 1, realiza-se a estatística dos dados univariados.

.1 Análise dos resultados de "Número de Páginas Webo Analy"

Os 49 dados geram o seguinte histograma no *software*:

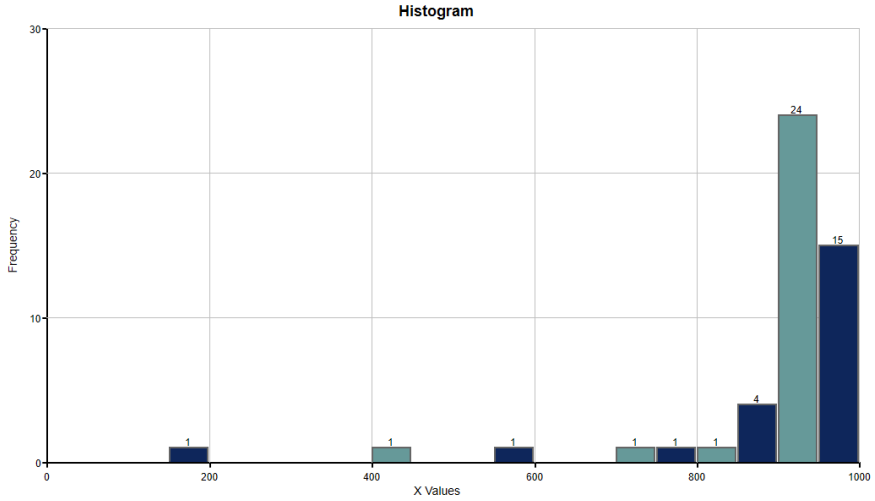

FIGURA 1
Histograma do número de páginas weboanaly de 49 universidades espanholas
Fonte: Elaborado pelos autores com base nos dados de Vanti e Casado (2015).

É importante observar que a concentração mais elevada de valores está na faixa de 900 a 999 páginas, com 39 dados dos 49 possíveis. Na faixa de 800 a 899 páginas encontram-se mais cinco dados. Os outros cinco dados restantes apresentam valores inferiores a 799, denotando subjetivamente a presença de possíveis *outliers* inferiores.

A estatística descritiva para os 49 dados fornece uma média de 895,7 páginas. Já o desvio padrão é 147,5 páginas, o que fornece um Coeficiente de Variação (CV) de 16,5%, valor considerado moderado.

Observar-se que a média (895,7 páginas) é inferior a 899 páginas, ou seja, a média não se encontra na faixa de 900 a 999 páginas, que é o padrão dos dados. Isto é reflexo da influência dos possíveis *outliers* inferiores no cálculo da média. Além disto, há a falsa impressão de que, das 49 universidades espanholas, 40 delas estão acima da média.

A estatística inferencial para os 49 dados apresenta que o intervalo de confiança (IC) de 95% para a média populacional deve variar entre 853,3 a 938,1 as páginas e o desvio padrão variam entre 123,0 a 184,2 páginas (IC de 95%).

O cálculo dos quartis, segundo o método de Bussab e Morettin (2002, p. 42), fornece para o primeiro quartil o valor de 907,25 páginas; segundo quartil (mediana) de 934,00 páginas; e terceiro quartil de 957,50 páginas.

Deve-se atentar que a média calculada (895,7 páginas) também é inferior ao valor do primeiro quartil (907,25 páginas), reforçando novamente a influência dos *outliers* inferiores no cálculo da média.

Por meio da aplicação da fórmula (3) com os valores dos quartis para o coeficiente quartil de assimetria encontra-se o valor de (– 0,06467), ou seja, assimetria negativa fraca.

O uso da fórmula (1) para a detecção dos *outliers* inferiores gera:

O.I. < Q1 – 1,5*(Q3 – Q1)*e–(As);

O.I. < 907,25 – 1,5*(957,5 – 907,25)*e–(– 0,06467)

O.I. < 826,8 páginas.

Há então a presença de cinco *outliers* inferiores: 149; 409; 598; 726; 793. Estes valores correspondem às cinco universidades espanholas com bem menor número de páginas em relação às outras 44 universidades. Deve-se observar que no histograma da Figura 1 os *outliers* inferiores detectados correspondem exatamente a cauda à esquerda da distribuição.

Para a detecção dos outliers superiores, utiliza-se a fórmula (2):

O.S. > Q3 + 1,5*(Q3 – Q1)*e+(As)

O.S. > 957,50 + 1,5*(957,50 – 907,25)*e+(– 0,06467)

O.S. > 1028,2 páginas. Não há presença de outliers superiores.

A remoção dos cinco *outliers* inferiores fornece uma nova média de 936,7 páginas. Nota-se que a nova média encontra-se agora entre os 39 dados que variam de 900 a 999 páginas no histograma da Figura 1. Além disto, verifica-se que somente 22 das 49 universidades espanholas encontram-se acima da média.

Já o novo desvio padrão é de 33,3 páginas para os 44 dados. A redução aqui é substancial, pois o desvio padrão original (49 dados) foi de 147,5 páginas. O CV original foi de 16,5% (moderado) e agora com a remoção dos *outliers* inferiores é de 3,6% (baixo).

Para a estatística inferencial, o novo intervalo de confiança (IC) de 95% para a média fornece valores entre 926,6 páginas e 946,8 páginas (diferença de 20,2 páginas). Para os dados completos (sem remoção dos *outliers*), os valores eram de 853,3 a 938,1 páginas (diferença de 84,8 páginas).

Apesar de haver sobreposição dos valores do intervalo de confiança (IC), é nítido que a diferença do intervalo de confiança sem a presença dos *outliers* é bem menor (20,2 contra 84,4 páginas), denotando que os dados sem *outliers* estão mais agrupados ou dentro do padrão representativo dos dados.

Outro detalhe a ser apontado é que o IC da média sem *outliers* (926,6 a 946,8 páginas) encontra-se na massa de dados de 900 a 999 páginas. Já o IC da média com todos os dados (853,3 a 938,1 páginas) apresenta valores inferiores a 900 páginas.

Ainda dentro da estatística inferencial, a faixa do intervalo de confiança (IC) de 95% para o desvio padrão sem os *outliers* inferiores é de 27,5 a 42,1 páginas. Inicialmente, a faixa de variação com todos os dados era de 123,0 a 184,2 páginas, ou seja, há bastante influência dos cinco *outliers* inferiores no cálculo do desvio padrão; de outro modo, não há sobreposição dos valores das faixas de variação do desvio padrão. Além disso, como a faixa de variação do desvio padrão é bem menor sem a presença dos *outliers*, novamente representa que os dados estão mais agrupados, mais representativos.

Corroborando a remoção dos *outliers* conduz a dados mais representativos e mais agrupados (Figura 2).

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

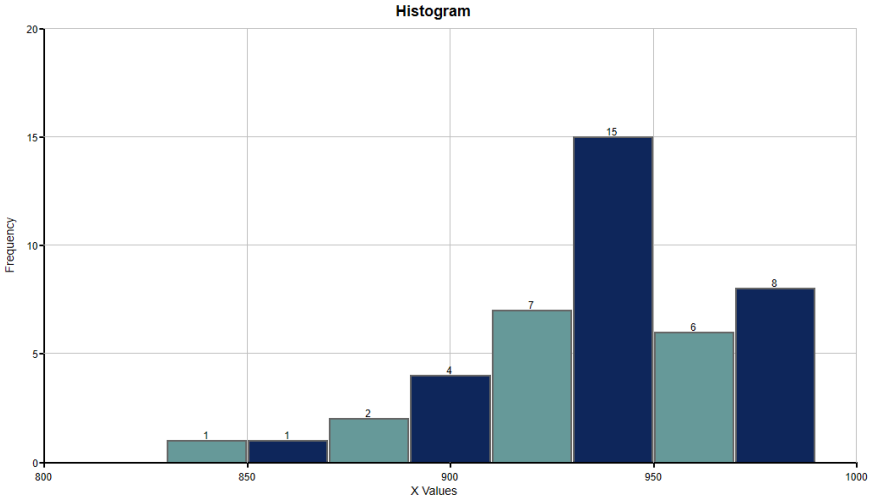

FIGURA 2
Histograma do número de páginas weboanaly de 44
universidades espanholas, sem a presença dos outliers inferiores
Fonte: Elaborado pelos autores com base nos dados de Vanti e Casado (2015).

Em síntese, esse exemplo dos dados de Vanti e Casado (2015) ilustra perfeitamente como os outliers influem nos cálculos de média e desvio padrão (tanto na estatística descritiva como na inferencial), bem como na forma do histograma, em consonância com as informações de Triola (2012).

Além disto, vale rememorar a falsa impressão que 40 universidades espanholas estavam acima da média (com a presença de todos os dados), quando na realidade a remoção dos *outliers* dos cálculos conduz somente a 22 universidades espanholas acima da média.

4.2 Análise dos resultados de "Citações URL WeboAnaly"

Realizando a mesma rotina de cálculo do item 4.1 anterior, temos para os 49 dados originais da citação:

Média = 514,3 citações I.C. de 95%: 487,7 < média < 540,9 citações.

Desvio padrão (dp) = 92,6 citações IC de 95%: 77,2 < dp < 115,7 citações.

Coeficiente de Variação (CV) = 18,0%.

Q1 = 478,50 citações; Q2 = 505,00 citações; Q3 = 557,25 citações.

As = + 0,32698 (assimetria positiva forte).

Para a detecção dos outliers inferiores:

O.I. < Q1 – 1,5*(Q3 – Q1)*e–(As);

O.I. < 478,50 – 1,5*(557,25 – 478,50)*e–(+ 0,32698)

O.I. < 393,3 citações.

Há a presença de cinco *outliers* inferiores: 280; 298; 348; 373 e 384 citações.

Já para a detecção dos *outliers* superiores:

O.S. > Q3 + 1,5*(Q3 – Q1)*e+(As)

O.S. > 557,25 + 1,5*(557,25 – 478,50)*e+(+ 0,32698)

O.S. > 721,1 citações.

Agora há a presença de um *outlier* superior: o valor de 800 citações. Vale mencionar também que o valor de 715 citações (2o maior valor dos dados) está muito próximo do limite de 721,1 citações. Ao valor de 715 citações é dado o nome de valor adjacente e pode ser visto como um indicativo da necessidade de aperfeiçoamento da fórmula de detecção deste trabalho.

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

Por fim, o histograma da Figura 3 ilustra a distribuição dos dados, no qual se pode observar os cinco *outliers* inferiores (menores que 400 citações) e o único *outlier* superior (igual a 800 citações).

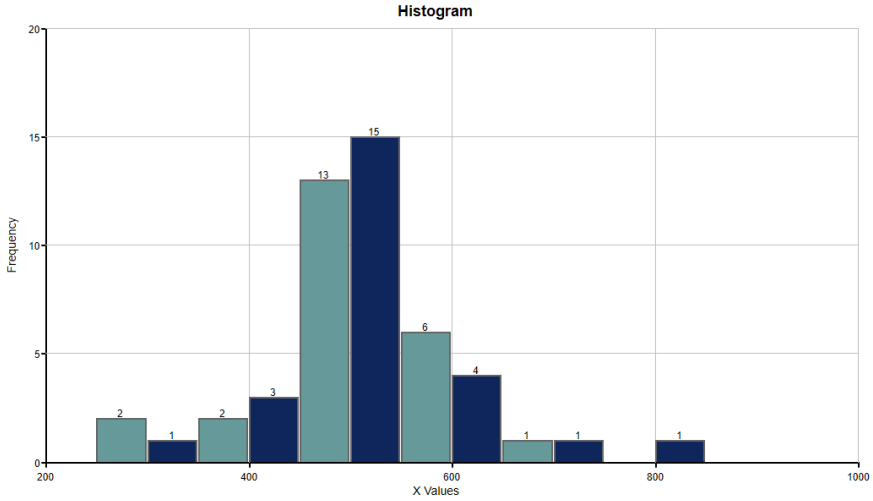

FIGURA 3
Histograma do número de citações URL webo analy de 49 universidades espanholas
Fonte: Elaborado pelos autores com base nos dados de Vanti e Casado (2015).

As remoções dos *outliers* conduzem a novos valores:

Média = 528,3 citações I.C. de 95%: 509,6 < média < 547,1 citações.

Desvio padrão (dp) = 60,8 citações IC de 95%: 50,1 < dp < 77,3 citações.

Coeficiente de Variação (CV) = 11,5%

Em comparação aos dados originais, não há tanta redução na média, pois o outlier superior de certa forma contrabalança em algum grau o efeito dos *outliers* inferiores. Além disto, há pouca diminuição no CV, diminuindo de 18,0% para 11,5%. Outro detalhe é que não há alteração do número de universidades acima da média.

Novamente, para o desvio padrão, a redução é substancial, passando de 92,6 citações para 60,8 citações; apesar de no IC ainda haver sobreposição de valores no limiar (77,2 citações mínimas do desvio padrão com todos os dados, contra 77,3 citações máximas do desvio padrão sem os *outliers*).

4.3 Análise dos resultados do Fator de Impacto Web Revisado "FIW Revisado"

Seguindo o roteiro de cálculo, temos para os 49 dados originais da citação:

Média = 0,6342 I.C. de 95%: 0,5124 < média < 0,7560.

Desvio padrão (dp) = 0,4240 IC de 95%: 0,3536 < dp < 0,5297.

Coeficiente de Variação (CV) = 66,9% (muito elevado)

Q1 = 0,50245; Q2 = 0,5622; Q3 = 0,613875.

As = – 0,07247 (assimetria negativa fraca).

Para a detecção dos *outliers* inferiores:

O.I. < Q1 – 1,5*(Q3 – Q1)*e–(As);

O.I. < 0,50245 – 1,5*(0,613875 – 0,50245)* e–(– 0,07247)

O.I. < 0,32275. Há a presença de um *outlier* inferior: 0,3150.

Já para a detecção dos *outliers* superiores:

O.S. > Q3 + 1,5*(Q3 – Q1)*e+(As)

O.S. > 0,613875 + 1,5*(0,613875 – 0,50245)*e+(– 0,07247)

O.S. > 0,76933. Agora há a presença de cinco *outliers* superiores: 0,7944; 0,8743; 1,1405; 1,2200; 3,3221.

O histograma da Figura 4 indica tanto o *outlier* inferior como os *outliers* superiores, bem como ilustra que a massa de dados situa-se entre 0,35 e 0,75.

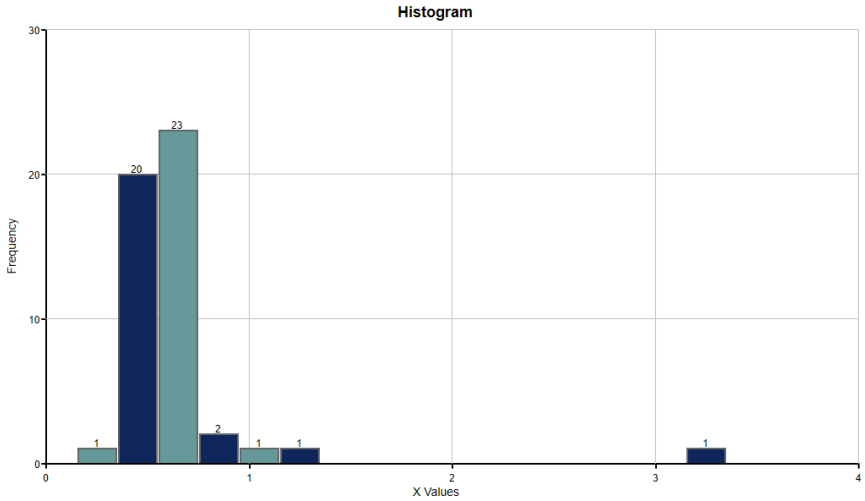

FIGURA 4
Histograma do FIW Revisado de 49 universidades espanholas
Fonte: Elaborado pelos autores com base nos dados de Vanti e Casado (2015).

A remoção do *outlier* inferior junto com os cinco outliers superiores geram os seguintes novos valores:

Média = 0,5444 I.C. de 95%: 0,5196< média < 0,5692.

Desvio padrão (dp) = 0,0807 IC de 95%: 0,0665 < dp < 0,1026.

Coeficiente de Variação (CV) = 14,8% (moderado)

Para a média, há diminuição do valor de 0,6342 para 0,5444. Já o IC variou da faixa [0,5124; 0,7560] (para todos os dados), para a faixa [0,5196; 0,5692] (sem a presença dos *outliers*). Apesar de ainda haver sobreposição de valores no IC, observar que o valor máximo do IC da média diminui substancialmente de 0,7560 para 0,5692 (ou seja, um conjunto mais compacto de dados, mais representativos); novamente ilustrando como os *outliers* influem nos cálculos e, portanto, nas conclusões.

Outro aspecto da influência dos *outliers* é que, para o valor original de 0,6342, há somente 10 universidades espanholas (inclusas as cinco que são outliers superiores) acima da média do FIW revisado. Como a remoção dos *outliers* gera uma nova média de 0,5444 para o FIW revisado, agora há 28 universidades espanholas (inclusos os cinco *outliers*) que estão acima da média.

Para o CV, a redução de valores é drástica, passando de 66,9% (muito elevado) de todos os dados, para 14,8% (moderado) com a remoção dos *outliers*, o que denota maior representatividade dos dados sem os *outliers*.

5 Considerações finais

Este trabalho apresenta uma nova fórmula para detecção de *outliers* via AED. Para o estudo da influência dos *outliers* nos estudos métricos da informação, usam-se três conjuntos de dados publicados na literatura.

No primeiro conjunto de dados, há a detecção de cinco *outliers* inferiores. A média calculada com todos os dados é de 895,7 páginas, valor inferior ao primeiro quartil dos dados e dando a falsa impressão que 40 universidades (de um total de 49 universidades) estão acima da média.

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

A remoção dos cinco *outliers* inferiores do primeiro conjunto de dados conduz a uma nova média de 936,7 páginas. Este valor encontra-se próximo da mediana original de 934 páginas. Além disto, a nova média encontra-se na massa de dados de 39 universidades com 900 a 999 páginas.

Em relação ao intervalo de confiança da média do primeiro conjunto de dados, há sobreposição de valores (853,3 a 938,1 páginas para todos os dados; contra 926,6 a 947,0 páginas removendo os *outliers*); entretanto, é nítido que a nova faixa de valores (sem os *outliers*) é mais representativa do primeiro conjunto de dados.

Em relação ao desvio padrão do primeiro conjunto de dados, o valor original é de 147,5 páginas, com um intervalo de confiança entre 123,0 e 184,2 páginas. A remoção dos cinco *outliers* inferiores gera um novo desvio padrão bem inferior, de 33,3 páginas, sendo o novo intervalo de confiança entre 27,5 a 42,2 páginas. Agora não há sobreposição dos valores do intervalo de confiança do desvio padrão; ademais, a faixa de variação do novo desvio padrão [27,5 a 42,2] páginas também é bem inferior à faixa original dos valores [123,0 a 184,2] páginas.

Para o segundo conjunto de dados, há a detecção de cinco *outliers* inferiores e um *outlier* superior. Neste caso específico, o *outlier* superior, de certa forma, contrabalançou o(s) possível(is) efeito(s) do(s) *outlier*(s) inferiores. Todavia, alerta-se aos estudiosos da área que nem sempre haverá este evento de minoração dos efeitos na existência simultânea de *outliers* superiores e inferiores.

O terceiro conjunto de dados apresenta um *outlier* inferior e cinco outlierssuperiores. A média original é de 0,6342; conduzindo a conclusão que somente 10 universidades (do total de 49 universidades) estão acima da média. A remoção dos seis *outliers* conduz a uma nova média de 0,5444; neste caso, na realidade, 28 universidades encontram-se acima da média.

Ainda para o terceiro conjunto de dados, observa-se que há sobreposição dos valores do intervalo de confiança da média; todavia, o novo intervalo de confiança (sem os *outliers*) apresenta valores mais compactos e, portanto, mais representativos dos dados analisados.

Já para o desvio padrão do terceiro conjunto de dados, não há sobreposição do intervalo de confiança; assim, o novo intervalo de confiança do desvio padrão (sem os *outliers*) é mais representativo dos dados.

Portanto, este trabalho demonstra, para os três conjuntos de dados analisados, como *outliers* podem influir tanto nos cálculos de média, desvio padrão, inferência estatística (estimativa intervalar), bem como nas conclusões de estudos. Neste sentido, é mister nos estudos métricos da informação que *outliers* sejam levados em consideração, a fim de evitar vieses tanto nos resultados como na interpretação dos dados.

Todavia, deve-se mencionar que as fórmulas (1) e (2) deste trabalho ainda merecem atenção dos pesquisadores da área para aperfeiçoamento:

O.I. < Q1 – 1,5*(Q3 – Q1)*e–(As) Fórmula (1)

O.S. > Q3 + 1,5*(Q3 – Q1)*e+(As) Fórmula (2)

As possíveis melhorias podem envolver a junção das vertentes um (alteração dos valores dos quartis) e vertente dois (alteração da constante "1,5"), combinadas com a definição de qual deva ser o valor do multiplicador (visto que na literatura os multiplicadores variam entre 0,5 a 5,0) para o coeficiente de assimetria (vertente três).

Referências

ADIL, Iftikhar Hussain; IRSHAD, Ateequr 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.

ALVAREZ, Gonzalo Rubén; CAREGNATO, Sônia Elisa. Presença de agradecimentos por financiamento nos artigos brasileiros indexados na Web of Science (2009-2016). In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. Anais... Rio de Janeiro: UFRJ, 2018. p. 172-180.

BABURA, Babangida Ibrahim et al. Modified boxplot for extreme data. AIP Conference Proceedings, New York, v. 1842, issue 1, May 2017.

BARNETT, Ofra; COHEN, Ayala. The histogram and boxplot for the display of lifetime data. Journal of Computational and Graphical Statistics, England, v. 9, n. 4, p. 759-778, Dec. 2000.

BARNETT, Vic; LEWIS, Toby. Outliers in statistical data. New York: John Wiley & Sons, 1978.

BUSSAB, Wilton; MORETTIN, Pedro. Estatística Básica. 5. ed. São Paulo: Saraiva, 2002.

CHHIKARA, R. S.; FEIVESON, A. L. Extended critical values of extreme studentized deviate test statistics for detecting multiple outliers. Communications in statistics: simulation and computation, England, v. B9, n. 2, p. 155-166, [s.d.], 1980.

DÜMBGEN, Lutz; RIEDWYL, Hans. On fences and asymmetric in box-and-whiskers plots. The American Statistician, Alexandria, VA, v. 61, n. 4, p. 356-359, Nov. 2007.

HUBERT, Mia; VANDERVIEREN, Ellen. An adjusted boxplot for skewed distributions. Computacional 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.1980.

LIMA, Luís Fernando Maia et al. Estudo preliminar sobre a influência de outliers nas métricas científicas para dados univariados. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. Anais... Rio de Janeiro: UFRJ, 2018. p. 446-452.

LIMA, Luís Fernando Maia et al. Proposta de um critério auxiliar para a determinação da elite científica. In: BORGES, Maria Manuel; CASADO, Elias Sanz (Coord.). A ciência aberta: o contributo da Ciência da Informação: atas do VIII Encontro Ibérico EDICIC. Coimbra: Universidade de Coimbra, 2017a. p. 301-310. Disponível em: https://purl.org/sci/atas/edicic2017. Acesso em: 26 ago. 2018.

LIMA, Luís Fernando Maia 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. 2017b.

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

PRICE, John Derek de Solla. Little science, big science. New York: Columbia University Press, 1963.

ROSADO, Fernando. Outliers em dados estatísticos. Lisboa: Sociedade Portuguesa de Estatística, 2006.

SILVA, Dávilla Vieira Odizio da; MAROLDI, Alexandre Masson; LIMA, Luís Fernando Maia. Outliers na Lei do Elitismo. Em Questão, Porto Alegre, v. 20, n. 3, Edição Especial, p. 43-59, dez. 2014.

SILVA, Deise Deolindo; ALMEIDA, Cátia Cândida de; GRÁCIO, Maria Claúdia Cabrini. Avaliação científica de periódico em Ciências Sociais: junção dos indicadores Fator de Impacto e índice h. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. Anais... Rio de Janeiro: UFRJ, 2018, p. 264-271.

SILVA, Domingos J. Lopes 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.

SILVA, Fábio Salomão Vinco; SCHULZ, Peter Alexander. Impacto de uma única área de conhecimento sobre os indicadores de um país: a astronomia no Chile. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. Anais... Rio de Janeiro: UFRJ, 2018, p. 181-189.

TAMBAY, J. L. An integrated approach for the treatment of outliers in sub-annual economic surveys. American Statistical Association Proceedings of the Survey Research Methods. Alexandria, VA: American Statistical Association, 1988, p. 229-234.

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

TUKEY, John Wilder. Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley, 1977.

VANTI, Nadia; CASADO, Elías Sanz. O uso do fator de impacto web alternativo para avaliar as universidades públicas espanholas. In: ARAÚLO, Ronaldo Ferreira de (Org.). Estudos métricos da informação na web: atores, ações e dispositivos. Maceió: EDUFAL, 2015. p. 109-127.

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto

WALKER, Michael; CHAKRABORTI, Subha. An asymmetrically modified boxplot for Exploratory Data Analysis. [S.I.], 2013. Disponível em: https://louisville.edu/sphis/bb/src-2013/student-poster-competition/Abstract_WalkerM.pdf. Acesso em: 25 ago. 2018.

PDF gerado a partir de XML Redalyc JATS4R
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa acesso aberto