ISSN: 2317-0840

Rotulaçao de outliers: método de Faleschini para dados quantitativos univariados

Luís Fernando M. Lima[1†]; João Marcelo B. Protázio[2]

[1]*Universidade Federal de Rondônia (UNIR).*
[2]*Universidade Federal do Pará (UFPA).*

Resumo: *O objetivo deste trabalho é apresentar o método de Faleschini para a rotulação de outliers para dados quantitativos univariados. O método de Faleschini utiliza a média, desvio padrão, coeficientes momentos de assimetria e curtose, resultando em uma equação do quarto grau, onde a menor raiz e a maior raiz podem ser adotados como rotuladores de outliers. O método de Faleschini é comparado com o método de Tukey e com o método recente de Adil e Zaman. Para as distribuições teóricas contínuas que não apresentem parâmetros ou de média, ou de desvio padrão, ou de assimetria ou de curtose, propõe-se "pseudos" parâmetros baseados em percentis e quartis, e também estes "pseudos" parâmetros são utilizados para distribuições discretas e para dados amostrais, bem como uma proposta para o cálculo destes percentis e quartis. O artigo finaliza apontando que o método de Faleschini apresenta vantagem conceitual, pois além de levar os parâmetros de localização, dispersão, assimetria e curtose, não faz distinção entre distribuições com cauda leve ou pesada, contudo, são necessárias novas pesquisas visando tanto o aperfeiçoamento dos "pseudo" parâmetros como ampliação para outros campos, como análise multivariada, série temporais e outros.*

Palavras-chave: *Outliers; Rotulação de Outliers; Faleschini.*

Outliers labeling: Faleschini method for univariate quantitative data

Abstract: *The objective of this work is to present Faleschini's method for labeling outliers for univariate quantitative data. Faleschini's method uses the mean, standard deviation, moment coefficients of asymmetry and kurtosis, resulting in a fourth degree equation, where the smallest root and the largest root can be adopted as outlier labelers. Faleschini's method is compared with Tukey's method and the recent method of Adil and Zaman. For continuous theoretical distributions that do not present parameters or mean, standard deviation, asymmetry or kurtosis, "pseudo" parameters based on percentiles and quartiles are proposed, and these "pseudo" parameters are also used for discrete distributions. and for sample data, as well as a proposal for calculating these percentiles and quartiles. The article ends by pointing out that Faleschini's method has a conceptual advantage, as in addition to taking the parameters of location, dispersion, asymmetry and kurtosis, it does not distinguish between distributions with a light or heavy tail. "pseudo" parameters as extension to other fields, such as multivariate analysis, time series and others.*

Keywords: *Outliers; Outliers Labeling; Faleschini.*

Introdução

Introduz-se este trabalho expondo uma contextualização do seu tema até se chegar ao seu objetivo (conforme já figura no título). Depois disso, nos demais itens, apresentam-se Revisão da Literatura, o Método de Faleschini como Proposta para a Rotulação de Outliers e Comparação com Tukey (1977) e Adil e Zaman (2020), Considerações Finais e Referências.

Os outliers já atraíam a atenção dos pesquisadores pelo menos desde a metade do século XVIII, conforme Barnett e Lewis (1994); e neste diapasão, Rosado (2006) esclarece que muitas pesquisas sistemáticas ou de iniciação ao tema sobre outliers foram escritas a partir da metade do século passado, e tal assunto continua sendo objeto de pesquisa contemporânea (SILVA; OLIVEIRA; CARVALHO, 2021; RODRIGUES; ALMEIDA; MUSTAFA, 2020; BARBOSA; DUARTE; MARTINS, 2020; SILVA, 2019; BARBOSA;

[1†]Autor correspondente: luis.fernando@unir.br.

PEREIRA; OLIVEIRA, 2018; VELOSO; CIRILLO, 2016; ANDRADE; CIRILLO; BEIJO, 2014; PEREIRA; CIRILLO; OLIVEIRA, 2014; VISSOTTO JUNIOR; DIAS, 2013).

A definição de Barnett e Lewis (1994) para outlier é aquele valor que aparenta ser inconsistente com os restantes dos dados. Rosado (2006) lembra que em uma análise de dados, é mister a pesquisa de outliers, pois a sua não detecção pode arruinar a conclusão sobre os dados.

Para dados quantitativos univariados, Tukey (1977) apresentou uma proposta de rotulação de *outlier* (*outlier labeling*) (SILVA, 2019) que leva em conta uma medida de localização (primeiro ou terceiro quartil) e uma medida de dispersão (desvio interquartil). O método de Tukey (1977) é citado por Barnett e Lewis (1994) como um método *ad hoc*.

Aperfeiçoamentos na proposta de Tukey (1997), modificando a medida de localização ou a de dispersão ou ambas, foram feitos por Tambay (1988); Kimber (1990) e Adil e Zaman (2020).

A inclusão da assimetria na contribuição de Tukey (1977) ocorreu em Hubert e Vandervieren (2008). Desde então outros trabalhos seguiram a filosofia de Hubert e Vandervieren (2008) com uso da assimetria, em conjunto com uma medida de localização e dispersão, notadamente: Adil e Irshad (2015); Babura *et al.* (2017); Lima *et al*. (2017); Lima *et al*. (2018); Walker *et al.* (2018) e Silva (2019).

As caudas pesadas, dentro do ponto de vista de Hubert e Vandervieren (2008), foram trabalhadas por Bruffaerts, Verardi e Vermandele (2014) e Silva (2019).

Por seu turno, Carling (2000) procurou trabalhar tanto a questão da assimetria quanto a da curtose em conjunto com a contribuição de Tukey (1977); todavia, seu trabalho não é utilizado na prática do dia a dia.

Outro aspecto a ser mencionado é que, para o caso de assimetria nula, as contribuições de Adil e Irshad (2015); Babura *et al.* (2017); Lima *et al*. (2017); Lima *et al*. (2018); Walker *et al.* (2018) e Silva (2019) recaem na formulação original de Tukey (1977).

Assim, na fórmula de Tukey (1977), para a distribuição teórica normal, a taxa total de rotulação é de 0,70%; mas para a distribuição teórica de Laplace, a proposta de Tukey alcança uma taxa total de rotulação de 6,26%. Como as distribuições são simétricas (assimetria nula), um efeito não quantificado na fórmula de Tukey (1977) é justamente a curtose.

Pelo parágrafo anterior, as contribuições de Adil e Irshad (2015); Babura *et al.* (2017); Lima *et al*. (2017); Lima *et al*. (2018); Walker *et al.* (2018) e Silva (2019) igualmente rotulam 6,26% de taxa de *outliers* para a distribuição teórica de Laplace; pois também não levam em conta o efeito da curtose.

Outro aspecto é sobre a divisão de distribuições de cauda leve (exemplo, a normal) das distribuições de cauda pesada (exemplo, a de Laplace). Pois seria desejável uma formulação para rotulação de *outliers* que levasse em conta – medida de localização, medida de dispersão, medida de assimetria e medida de curtose – independente se distribuição de cauda leve ou cauda pesada.

O objetivo deste trabalho é propor uma nova formulação – levando-se em conta medida de localização, dispersão, assimetria e curtose –, tanto para uso nas distribuições teóricas, sejam simétricas ou assimétricas, sejam de cauda leve ou pesada, seja distribuição contínua ou discreta, como também para o caso de dados amostrais. Esta nova modelação será comparada com os resultados de Tukey (1977) (presente na maiores dos *softwares* estatísticos) e Adil e Zaman (2020).

Revisão da literatura

A estrutura geral de muitos modelos de rotulação de *outliers* é a seguinte:

$$x \geq (\leq)\ P_x \pm K(n) * (P_y - P_z) * e^{f(A)} \quad (1)$$

x := valor numérico que define *outliers* superiores (O.S), com uso dos sinais "≥" e "+" ou então, o valor numérico que caracteriza os outliers inferiores (O.I), neste caso, com uso dos sinais "≤" e "–".

P_x := percentil como medida de localização, em geral, a maioria das pesquisas da área usam quartis, contudo, Adil e Zaman (2020) utilizaram de fato o percentil.

K(n) := constante, em geral "1,5" como Tukey (1977) e outros ou função do tamanho amostral, como em Carling (2000).

P_y ; P_z := percentis como medida de dispersão, onde em geral $P_y > P_z$, novamente a maioria dos estudos da área utilizam quartis, todavia, Adil e Zaman (2020) usaram de fato o percentil.

e := número de Euler.
f(A) := constante ou função da assimetria dos dados ou distribuição.

A contribuição de Tukey (1977) foi:

$$O.S. \geq Q3 + 1{,}5 * (Q3 - Q1) \quad (2)$$

$$O.I. \leq Q1 - 1{,}5 * (Q3 - Q1) \quad (3)$$

Onde:
Q1 := primeiro quartil
Q3 := terceiro quartil

Como asseverado na introdução, muitas pesquisas remontam à ideia original de Tukey (1977); que, como vimos, para a distribuição de Laplace rotula uma taxa de *outliers* de 6,26%, valor este que consideramos excessivamente alto (acima de 5%).

Lado outro, a contribuição de Walker *et al.* (2018) para o caso do coeficiente de assimetria de Bowley ser o valor máximo de 1,0; a fórmula gera divisão por zero, daí, sem generalidade matemática.

Conceitualmente, a contribuição de Silva (2019) de contraste quartílico, no caso de distribuição de caudas leves, conduz a mesma fórmula matemática de Walker *et al.* (2018), e portanto, para assimetria nula, retorna a Tukey (1977); assim, também rotulando uma taxa de 6,26% de *outliers* para a distribuição de Laplace.

Para caudas pesadas, existem as contribuições de Bruffaerts, Verardi e Vermandele (2014) e Silva (2019); porém, estas pesquisas exigem rotina computacional, sem a simplicidade matemática da equação (1).

Retomando-se o aspecto da distribuição de Laplace (simétrica e cauda pesada), a contribuição de Adil e Zaman (2020) rotula uma taxa de 2,41%, bem inferior (e abaixo de 5%) dos 6,26% de Tukey (1977).

A proposta de Adil e Zaman (2020) é:

$$O.I. \leq P12{,}5 - 1{,}5 * (P37{,}5 - P12{,}5) \quad (4)$$

$$O.S. \geq P87{,}5 + 1{,}5 * (P87{,}5 - P62{,}5) \quad (5)$$

Onde:
P12,5 := décimo segundo e meio percentil.
P37,5 := trigésimo sétimo e meio percentil.
P62,5 := sexagésimo segundo e meio percentil.
P87,5 := octagésimo sétimo e meio percentil.

A situação torna-se mais dramática do uso da fórmula de Tukey (1977) para distribuições com assimetria e curtose elevadas: a distribuição log-normal [LN (0,1)], com média zero e desvio padrão, apresenta uma taxa total de rotulação de 7,76%; para a distribuição qui-quadrado de grau um [X^2_1], uma taxa total de rotulação de 7,56%; para a distribuição de Weibull [W (1; 0,5)] com fator de escala igual a um e fator de forma igual a meio rotula 11,49%. Ou seja, todas as rotulações citadas foram novamente acima de 5%.

Para as mesmas distribuições LN (0,1); X^2_1 e W (1; 0,5), a fórmula de Adil e Zaman (2020) rotulam respectivamente 3,89%; 3,01% e 4,69%; portanto todas as rotulações são inferiores a 5%.

Uma possível explicação para a menor taxa de rotulação da fórmula de Adil e Zaman (2020) em relação à Tukey (1977) – seja para distribuições simétricas ou assimétricas –, talvez seja pelo uso de percentis mais próximos do efeito da cauda da distribuição (P12,5 e P87,5), ao passo de que Tukey (1977) usa praticamente valores centrais (primeiro quartil e terceiro quartil).

Outro aspecto do aperfeiçoamento trazido em Adil e Zaman (2020) é o seu uso tanto para distribuições de cauda leve como para distribuições de cauda pesada, como para distribuições simétricas ou assimétricas; conforme vimos nos parágrafos anteriores, mesmo sem uso da assimetria e da curtose.

Mas, em que pese o avanço conceitual de Adil e Zaman (2020), reitera-se a necessidade de uma modelação matemática que englobe as medidas de localização, dispersão, assimetria e curtose.

Método de Faleschini como proposta para a rotulação de *outliers* e comparação com Tukey (1977) e Adil e Zaman (2020)

Faleschini (1948) estudou, para uma sequência discreta de valores positivos e suas respectivas frequências, qual seria o efeito do acréscimo de um novo valor (que também denominaremos de "x") nos momentos da sequência discreta, notadamente nos momentos de segunda ordem (associado à variância e desvio padrão), momento de terceira ordem (associado à assimetria) e de quarta ordem (associado à curtose).

A seguinte notação é útil:
μ := média aritmética.
σ := desvio padrão.
x := novo valor a ser inserido na sequência discreta.

Para o caso do momento de segunda ordem, Faleschini (1948) obteve as seguintes expressões, oriundas da solução de uma equação de segundo grau:

Se $\mu - \sigma < x < \mu + \sigma$; então o segundo momento diminui. Caso contrário, se $x \leq \mu - \sigma$; ou se $x \geq \mu + \sigma$; então o segundo momento aumenta.

Já para o caso particular de assimetria nula (ou terceiro momento nulo), Faleschini (1948) obteve as seguintes soluções de uma equação do terceiro grau:

Se $\mu - (\sqrt{3})^{\star}\sigma < x \leq 0$ ou $x \geq \mu + (\sqrt{3})^{\star}\sigma$; então o terceiro momento aumenta. Caso contrário, se $x \leq \mu - (\sqrt{3})^{\star}\sigma$ ou se $0 < x < \mu + (\sqrt{3})^{\star}\sigma$; então o terceiro momento diminui.

Importa aqui reconhecer a estrutura (ou similaridade): uma medida de posição (representada por μ) e uma medida de dispersão (representada por σ) que conecta a contribuição original de Faleschini (1948) com a proposta de rotulação de *outliers* devida a Tukey (1977).

Por fim, ao chegar ao quarto momento (associado à curtose, denominada nas equações de "B"), Faleschini (1948) obtém uma equação do quarto grau, e para o caso particular novamente de uma distribuição simétrica, obtém como quatro raízes, sendo de nosso interesse a menor e a maior raiz:

$$x \leq \mu - \sigma * \{B + [B * (B-1)]^{1/2}\}^{1/2} \ (curtose\ aumenta) \quad (6)$$

$$x \geq \mu + \sigma * \{B + [B * (B-1)]^{1/2}\}^{1/2} \ (curtose\ aumenta) \quad (7)$$

As fórmulas (6) e (7) obtidas originalmente por Faleschini (1948) representam justamente a influência da inclusão do novo valor ("x") nas extremidades: em ambos os casos a curtose aumenta! Outro detalhe é a presença da medida de posição (μ) e a medida de variabilidade (σ), sendo esta multiplicada pelo fator "$\{B + [B*(B-1)]^{1/2}\}^{1/2}$" que leva em conta a curtose ("B") da sequência original, novamente fazendo a ponte entre a contribuição de Faleschini (1948) e a ideia de Tukey (1977).

Por seu turno, a extensão do estudo de Faleschini (1948) de sequência discreta para as distribuições contínuas e conexão para a denominada função de influência de Hampel é dada no artigo de Fiori e Zenga (2005).

Fiori e Zenga (2005) retrabalham a equação do quarto grau encontrada por Faleschini (1948), que usava os momentos, para utilizar a variável normalizada "Z", o coeficiente momento de assimetria "A" e o coeficiente momento de curtose "B". A equação do quarto grau de Fiori e Zenga (2005) é:

$$(Z^2 - B)^2 - B * (B - 1) - 4 * A * Z = 0 \quad (8)$$

Ou:

$$Z^4 - 2 * B * Z^2 - 4 * A * Z + B = 0 \quad (9)$$

Onde:

$$Z = (x - \mu) / (\sigma) \quad (10)$$

A fórmula (9) relaciona a média aritmética "μ" e o desvio padrão "σ" (inclusas na variável normalizada "Z") com a assimetria "A" e a curtose "B"; desta forma, admitida a conexão entre os estudos de Faleschini (1948) e Tukey (1977) com o aperfeiçoamento trazido por Fiori e Zenga (2005); então a solução da equação (9) fornece uma fórmula genérica para a rotulação de *outliers*, seja para distribuições discretas (e conjecturamos também para dados amostrais) por conta do estudo de Faleschini (1948), seja para distribuições contínuas devido a Fiori e Zenga (2005), além de servir como equação única para o caso de caudas leves ou pesadas, daí o avanço conceitual desta proposta.

As diferentes distribuições simétricas podem ser diferenciadas pela curtose, apenas para citar exemplos, distribuição uniforme (B = 1,8), distribuição normal (B = 3), distribuição logística (B = 4,2) e distribuição de Laplace (B = 6,0).

Por exemplo, para a distribuição teórica normal padrão, temos: μ = 0; σ = 1; A = 0 e B = 3. Substituindo os valores numéricos da distribuição normal padrão em (9) e (10) obtemos:

$$Z = x \leq -2{,}3344 \ (a\ curtose\ aumenta) \quad (11)$$

$$Z = x \geq +2{,}3344\ (a\ curtose\ aumenta) \quad (12)$$

A área (probabilidade) à esquerda de (10) é 0,0098 ou 0,98%; e devido à simetria, obviamente a área (probabilidade) à direita de (11) é também 0,0098 ou 0,98%. Portanto, para a distribuição normal, a taxa total de rotulação é de 1,96%.

Assim, podemos conjecturar que as equações (9) e (10) servem como indicadores de *outliers*, de modo que então rotulamos como *outliers* os valores extremos em ambas as caudas que ocasionam o aumento da curtose; deste modo, denominamos como método de Faleschini, em justa homenagem ao seu trabalho pioneiro de 1948, o uso das equações (9) e (10) como rotuladores de *outliers*.

A aplicabilidade do método de Faleschini será ilustrada para o caso geral de distribuições contínuas, distribuições discretas e dados amostrais nos próximos tópicos; e também já comparando a taxa total de rotulação com os métodos de Tukey (1977), presente na maioria dos *softwares* estatísticos e também o método de Adil e Zaman (2020).

Método de Faleschini para distribuições contínuas

Para as distribuições que possuem média, desvio padrão, coeficientes momentos de assimetria e curtose, o uso das equações (9) e (10) é direta, como foi feito com a distribuição normal da secção anterior. Seguem algumas distribuições teóricas com comparações de taxa de rotulação de *outliers* entre os métodos de Tukey (1977), Adil e Zaman (2020) e Faleschini.

a) Distribuição Uniforme com média igual a zero e desvio padrão igual a um; assimetria nula e curtose igual a 1,8. Neste caso específico, as três formulações fornecem como taxa total de rotulação zero por cento.

b) Distribuição Logística com média igual a zero; fator de escala igual a um e desvio padrão igual a 1,8138; assimetria nula e curtose igual a 4,2. Aqui, Tukey (1977) rotula um total de 2,44% de *outliers,* Adil e Zaman (2020) rotula 3,26% e o método de Faleschini rotula 1,23%.

c) Distribuição de Laplace com média igual a zero; fator de escala igual a um e desvio padrão igual a 1,4142; assimetria nula e curtose igual a 6,0. A fórmula de Tukey (1977) rotula um total de 6,26% de *outliers,* Adil e Zaman (2020) rotula 4,82% e o método de Faleschini rotula 0,83%. É importante atentar, como já asseverado anterior, que mesmo uma distribuição simétrica, Tukey (1977) rotula acima de 5% os *outliers,* contudo Adil e Zaman (2020) e principalmente o método de Faleschini apresentam tais taxa abaixo de 5%.

d) Distribuição de Student com cinco graus de liberdade (t_5), com média igual a zero; desvio padrão igual a 1,291; assimetria nula e curtose igual a 9,0. Nesta situação, Tukey (1977) rotula um total de 3,35% de *outliers*, Adil e Zaman (2020) rotula 4,04% e o método de Faleschini rotula 0,30%.

e) Distribuição Lognormal com parâmetros $\mu = 0$ e $\sigma = 1$ [LN(0;1)]; com média igual a 1,6487, desvio padrão igual a 2,1612; assimetria de 6,185 e curtose igual a 113,94. Tukey (1977) rotula um total de 7,76% de *outliers,* novamente acima de 5%, Adil e Zaman (2020) rotula 3,89% e o método de Faleschini rotula 0,02%.

f) Distribuição Qui-Quadrado com um grau de liberdade [X^2_1] com média igual a um, desvio padrão igual a 1,4142; assimetria de 2,8284 e curtose igual a 15. As taxas totais de rotulação são 7,56%; 3,01% e 0,38% respectivamente para Tukey (1977); Adil e Zaman (2020) e o método de Faleschini. Observe-se mais uma vez que Tukey (1977) rotula mais de 5% de *outliers*.

g) Distribuição de Weibull com fator de escala igual a um e fator de forma igual a meio [W(1;0,5)], com média dois, desvio padrão de 4,4721, assimetria de 6,6188 e curtose de 87,72. Agora, as taxas totais de rotulação são 11,49%; 4,69% e 0,04%, respectivamente, para Tukey (1977); Adil e Zaman e o método de Faleschini. Verifica-se de novo que Tukey (1977) rotula mais de 5% de *outliers,* ao passo que Adil e Zaman (2020) e o método de Faleschini rotulam abaixo de 5% de *outliers*.

h) Distribuição de Weibull com fator de escala igual a um e fator de forma igual a um [W(1;1)], com média um, desvio padrão um, assimetria de dois e curtose de nove (função exponencial com parâmetro igual a um). As taxas de rotulação são de 4,81% (TUKEY, 1977); 2,41% (ADIL; ZAMAN, 2020) e 0,73% pelo método de Faleschini. Kimber (1990) informa também que a taxa de rotulação de outliers por Tukey é de 4,8%; e que seu método rotula 3,1%. Pelo método de Walker et al. (2018) e por extensão, Silva (2019) para distribuições de cauda leve, encontra-se uma taxa total de rotulação de 1,59% para um coeficiente de assimetria de Bowley de 0,262. Importante atentar que a contribuição de Adil e Zaman (2020) apresenta taxa de rotulação

inferior ao proposto por Kimber (1990), mas superior ao de Walker *et al*. (2018) e Silva (2019); contudo, o método de Faleschini é a que rotula a menor taxa de *outliers* e até mesmo inferior a 1%.

i) Distribuição de Gumbel com fator de locação nula e fator de escala igual a um [G(0;1)], com média 0,5772, desvio padrão 1,2825, assimetria de 1,140 e curtose de 5,4.

O aspecto interessante da distribuição de Gumbel é que igualmente as fórmulas de Tukey (1977) e Faleschini rotulam menos de 0,01% para os *outliers* inferiores; ao passo que Adil e Zaman (2020) rotulam 0,16%. Para os *outliers* superiores; as taxas de rotulação são 2,68% para Tukey (1977); 2,0% para Adil e Zaman (2020) e 1,29% pelo método de Faleschini.

j) Distribuição de Weibull com fator de escala igual a um e fator de forma igual a dois [W(1;2)], com média 0,8862, desvio padrão 0,4633, assimetria de 0,6311 e curtose de 3,2455. As taxas de rotulação são 1,03%; 1,38% e 2,91%, respectivamente, para Tukey (1977), Adil e Zaman (2020) e Faleschini. Chama a atenção neste caso específico que o método de Faleschini foi o de maior taxa de rotulação e muito próximo de 3%, de modo que nem sempre o método de Faleschini rotula menos *outliers* que Tukey (1977) e Adil e Zaman (2020).

Uma questão importante aparece quando a distribuição não possui seus parâmetros (média, desvio padrão, assimetria e curtose) definidos, como a distribuição padrão de Cauchy. Este trabalho propõe, como uma aproximação inicial, sendo passível e também necessário futuros aperfeiçoamentos, os seguintes "pseudo" parâmetros, tanto para distribuições teóricas como para amostras:

$$\mu = (P85 + P15) / 2 \quad (13)$$

$$\sigma = (P85 - P15) / 2 \quad (14)$$

$$A = [(P90 - \mu) / (Q2 - P15)] + [(P10 - \mu) / (P85 - Q2)] \quad (15)$$

$$B = [(P90 - \mu) / (Q2 - P15)]^2 + [(P10 - \mu) / (P85 - Q2)]^2 \quad (16)$$

Onde:

P85:= octogésimo quinto percentil.
P10:= décimo quinto percentil.
P90:= nonagésimo percentil.
P10:= décimo percentil.
Q2 := segundo quartil.

As aplicações das equações (13) a (16) para a distribuição padrão de Cauchy retornam: $\mu = 0$; $\sigma = 1,96$; A = 0 e B = 4,92. O baixo valor da curtose de Cauchy, em relação à distribuição de Student "t_5"; reforça que as equações (13) a (16) necessitam de mais pesquisas e aprimoramentos, portanto, neste momento, devem ser vistas como aproximação inicial.

O uso dos "pseudo" valores encontrados para a distribuição de Cauchy retornam uma rotulação via Faleschini de 10,53% de *outliers,* ao passo que o modelo de Tukey (1977) rotula 15,60% e a de Adil e Zaman (2020) rotula 11,63%. A elevada rotulação do método de Faleschini com uso dos "pseudo" parâmetros, novamente reforça novas contribuições ou visões à área.

Método de Faleschini para distribuições discretas

A fim de comparar o método de Faleschini com o método de Tukey (1977) e Adil e Zaman (2020), para o caso das distribuições discretas (e posteriormente para dados amostrais), é mister um novo modo de cálculo dos percentis. Vale lembrar que a literatura da área já aponta diversos métodos para o cálculo dos percentis, conforme se depreende de Triola (2012). Por exemplo, uma parte da distribuição de probabilidade acumulada para Poisson de parâmetro um é: P(0) = 0,36788; P(1) = 0,73576; P(2) = 0,91970; P(3) = 0,98101

Adota-se, antes da interpolação numérica (linear), a seguinte distribuição de valores:
X = 0 e área = 0,0; X = 0,5 e área = 0,36788; X = 1,5 e área = 0,73576; X = 2,5 e área = 0,91970.

A interpolação linear dos valores acima retorna os seguintes percentis:
P90 = 2,39290; P87,5 = 2,25699; P85 = 2,12107; P75 = 1,57742; P62,5 = 1,19892; P50 = 0,85914; P37,5 = 0,51935; P25 = 0,33978; P15 = 0,20387; P12,5 = 0,16989; P10 = 0,13591.

A distribuição de Poisson possui os seguintes parâmetros: média, desvio padrão e assimetria, todas iguais a um, e curtose igual a quatro. Para este caso, a modelação de Faleschini retorna como *outlier* superior X = 3,36; Tukey (1977) X = 3,43; Adil e Zaman (2020) X = 3,84.

Neste caso específico, todas as três modelações são arredondadas para o maior inteiro, no caso X = 4; e a distribuição teórica de Poisson retorna uma taxa de rotulação de 1,90% para as três modelações.

Método de Faleschini para dados amostrais

Para dados amostrais, não há evidentemente os parâmetros média, desvio padrão, assimetria e curtose. Neste caso, recomendam-se inicialmente os cálculos dos percentis via interpolação linear, como ilustrado na distribuição de Poisson de parâmetro um da seção 3.2, e posterior uso das equações (13), (14), (15) e (16) em conjunto com o método de Faleschini das fórmulas (9) e (10).

Os percentis calculados via interpolação linear proposta também serão usados nas fórmulas (2) e (3) de Tukey (1977) e equações (4) e (5) de Adil e Zaman (2020); como foi feita na seção 3.2 com a distribuição de Poisson com parâmetro um.

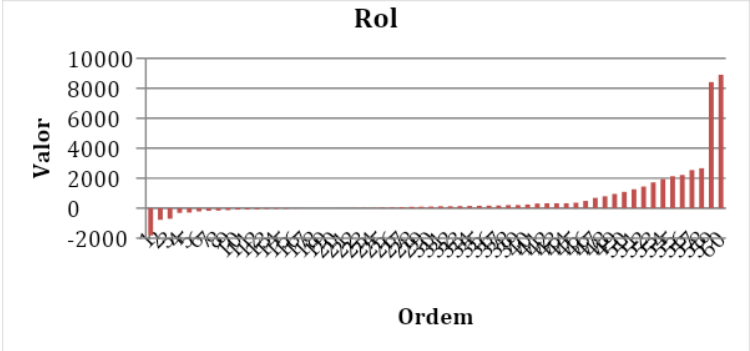
Por fim, um último aspecto necessita ser abordado ou comentado: o tamanho mínimo amostral. O método de Faleschini, para as distribuições teóricas, em geral apresentou valores de taxa de rotulação inferiores a 5% (exceto em Cauchy, quando faz uso de "pseudo" parâmetros).

Este trabalho recomenda inicialmente o uso do método de Faleschini para amostras com tamanho igual ou superior a 50; todavia, recomendam-se pesquisas e novas contribuições para tamanhos inferiores a 50. Outro aspecto que cabe novas visões é levar em conta o tamanho amostral nas equações (13) a (16).

Usaremos os dados reais de Brys, Hubert e Struyf (2004); Velleman e Hoaglin (1981); Barnett e Lewis (1994) e Chambers *et al*. (1983), cujos valores podem ser consultados nos próprios originais ou na compilação realizada em arquivo suplementar a este trabalho.

A Figura 1 apresenta os dados reais de Brys, Hubert e Struyf (2004).

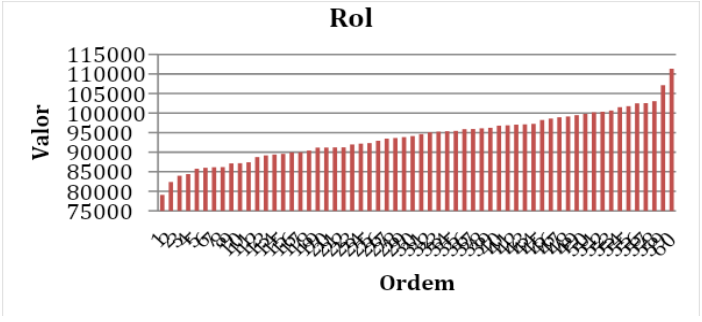
Figura 1 – Dados reais de Brys, Hubert e Struyf (2004).



Fonte: Elaborado pelos autores com base em Brys, Hubert e Struyf (2004).

A inspeção visual da Figura 1 evidencia os dois *outliers* superiores, contudo, não há tanta evidência para os *outliers* inferiores nos dados de Brys, Hubert e Struyf (2004). Já a Figura 2 apresenta os dados reais de Velleman e Hoaglin (1981).

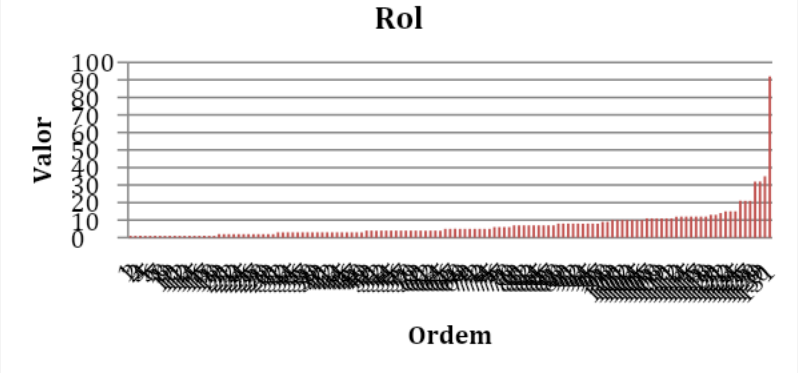
Figura 2 – Dados reais de Velleman e Hoaglin (1981).



Fonte: Elaborado pelos autores com base em Velleman e Hoaglin (1981).

Uma inspeção expedita da Figura 2 revela a possibilidade de dois *outliers* superiores, contudo, e provavelmente um *outlier* inferior nos dados de Velleman e Hoaglin (1981). Por seu turno, a Figura 3 apresenta os dados reais de Barnett e Lewis (1994).

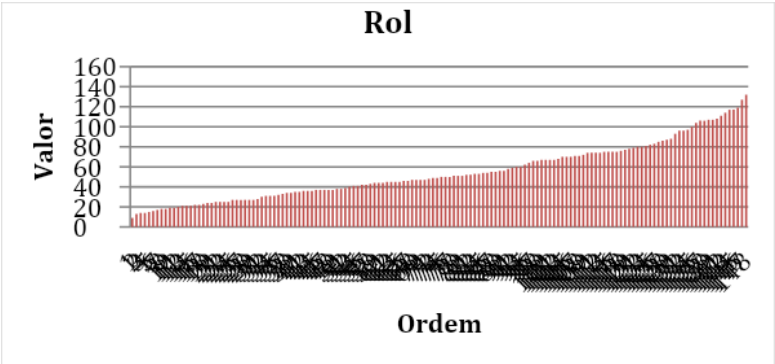
Figura 3 – Dados reais de Barnett e Lewis (1994).



Fonte: Elaborado pelos autores com base em Barnett e Lewis (1994).

A Figura 3 mostra a possibilidade de um *outlier* superior, todavia, há também seis outros valores que se destacam dos demais e, em tese, são candidatos a *outliers* superiores. Por sua vez, a Figura 4 apresenta os dados reais de Chambers *et al*. (1983).

Figura 4 – Dados reais de Chambers *et al*. (1983).



Fonte: Elaborado pelos autores com base em Chambers *et al*. (1983).

Finalmente, a inspeção visual da Figura 4 deixa evidente os dois *outliers* superiores, entretanto, não há tanta evidencia para o *outlier* inferior nos dados de Chambers *et al*. (1983). A Tabela 1 resume os elementos e cálculos efetuados para os dados reais selecionados para este artigo.

Tabela 1 – Cálculos para a rotulação de *outliers*

(Continua)

Cálculo	Brys, Hubert e Struyf (2004)	Velleman e Hoaglin (1981)	Barnett e Lewis (1994)	Chambers *et al*. (1983)
Tamanho amostral	60	60	131	148
P90	2036,50	101631,50	12,486	96,900
P87,5	1707,75	100787,50	12,018	89,000
P85	1352,50	100270,00	11,550	83,700

Tabela 1 – Cálculos para a rotulação de *outliers*

(Continuação)

Cálculo	Brys, Hubert e Struyf (2004)	Velleman e Hoaglin (1981)	Barnett e Lewis (1994)	Chambers *et al*. (1983)
P75	426,00	98412,00	9,536	73,750
P62,5	199,50	96089,75	6,931	59,875
P50	119,00	94368,50	4,650	49,500
P37,5	18,50	91842,00	3,570	41,750
P25	-25,00	89748,00	2,653	33,500
P15	-114,00	87155,50	1,638	25,325
P12,5	-159,25	86411,00	1,455	23,750
P10	-199,50	86077,00	1,364	21,433
μ (Equação (13))	619,25	93712,75	6,59	54,51
σ (Equação (14))	733,25	6557,25	4,96	29,19
A (Equação (15))	5,42	-0,20	1,20	0,79
B (Equação (16))	37,44	2,88	4,40	4,01
O.I. Tukey (1977)	-701,5 (3)*	76752,0 (0)*	-7,7 (0)*	-26,9 (0)*
O.S. Tukey (1977)	1102,5 (10)*	111408,0 (0)*	19,9 (7)*	134,1 (0)*
O.I. Adil, Zaman (2020)	-425,9 (3)*	78264,5 (0)*	-1,7 (0)*	-3,2 (0)*
O.S. Adil, Zaman (2020)	3970,1 (2)*	107834,1 (1)*	19,6 (7)*	132,7 (0)*
O.I. Faleschini	-5809,3 (0)*	79344,8 (1)*	-8,9 (0)*	-31,2 (0)*
O.S. Faleschini	6832,4 (2)*	109194,3 (1)*	18,9 (7)*	126,5 (2)*

Fonte: Elaborado pelos autores com base na literatura citada.
Nota: * O valor entre parênteses indica o número de observações que são rotuladas como *outliers* pelo respectivo método.

Para os dados de Brys, Hubert e Struyf (2004), a rotulação de Tukey (1977) aponta a presença de três *outliers* inferiores e dez *outliers* superiores; já o método de Adil e Zaman (2020) indica também três *outliers* inferiores e dois *outliers* superiores, e finalmente, o método de Faleschini acusa somente os dois *outliers* superiores. Pela inspeção visual da Figura 1, de fato os dois *outliers* superiores eram bem evidentes, ao passo que os *outliers* inferiores suscitavam dúvidas, nos quais tanto Adil e Zaman (2020) quanto Tukey (1977) apontam três *outliers* inferiores; o método de Faleschini não declara nenhum *outlier* inferior.

A inspeção visual da Figura 2 com os dados de Velleman e Hoaglin (1981) indica a possibilidade de dois *outliers* superiores e um *outlie*r inferior. A rotulação de Tukey (1977) não aponta presença de *outliers*, já o método de Adil e Zaman (2020) indica um *outlier* superior, e por fim, o método de Faleschini acusa um *outlier* inferior e um *outlier* superior. Este exemplo é importante para ilustrar que o método de Faleschini pode rotular mais *outliers* que os métodos de Tukey (1977) e Adil e Zaman (2020), além da possibilidade de indicar *outliers* em cauda onde os dois métodos anteriores não apontam existência.

Já para os dados de Barnett e Lewis (1994), os três métodos rotulam sete *outliers* superiores, em concordância com a inspeção visual da Figura 3. O aspecto a ser salientado aqui é que os sete *outliers* rotulados representam 5,34% dos dados, ou seja, o método de Faleschini para amostras pode também rotular mais de 5%

dos dados, o que demonstra de fato mais uma vez a necessidade de aperfeiçoamento das fórmulas (13), (14), (15) e (16), como já havia sido citado na distribuição teórica padrão de Cauchy, onde a taxa de rotulação de Faleschini foi superior a 10% com uso das equações (13) a (16).

A inspeção expedita da Figura 4 com os dados de Chambers *et al*. (1983) ilustra a possibilidade de dois *outliers* superiores e talvez um *outlie*r inferior. As rotulações de Tukey (1977) e Adil e Zaman (2020) não apontam presença de *outliers*, já o método de de Faleschini aponta dois *outliers* superiores. Este exemplo reitera que o método de Faleschini pode rotular mais *outliers* que os métodos de Tukey (1977) e Adil e Zaman (2020), além de novamente a possível indicação de *outliers* em cauda onde os dois métodos anteriores não apontam existência.

Em resumo, o método de Faleschini não segue um padrão definido, ou seja, ora pode apresentar rotulação de dados *outliers* em quantidade inferior, igual ou superior aos métodos de Tukey (1977) e Adil e Zaman (2020), bem como rotular *outliers* em cauda onde os dois métodos anteriores não indicam a presença de *outliers.*

Considerações finais

Este trabalho traz uma nova contribuição ao campo da rotulação de *outliers* para dados quantitativos univariados, mostrando que o método de Faleschini possui vantagem conceitual – pois leva em conta todas as medidas descritivas estatística de média, desvio padrão, assimetria e curtose –, e ademais tem o aspecto inovador da não necessidade de se diferenciar distribuições de cauda leve e pesada.

Os resultados apontam que o método de Faleschini pode rotular ora menos, ora igual ou até mesmo ora a mais *outliers* que os métodos de Tukey (1977) e Adil e Zaman (2020), ou seja, não há um padrão definido.

Recomenda-se também a modelação de Adil e Zaman (2020) para rotulação de *outliers* em conjunto com o método de Faleschini, ampliando o leque de opção do analista para decidir se determinado dado pode ser rotulado como *outlier*.

Em que pesem os avanços conceituais elencados, é mister mais pesquisas visando quantificar os parâmetros de média, desvio padrão, assimetria e curtose pelo uso de medidas robustas (estatísticas de ordem) tanto para distribuições teóricas como para o caso de amostra para uso no método de Faleschini.

A proposta deste artigo para média, desvio padrão, assimetria e curtose, deve ser encarada como uma aproximação inicial, pois, como visto tanto na distribuição teórica padrão de Cauchy como no caso de dados reais de Barnett e Lewis (1994), a taxa de rotulação foi superior a 5%. Outro possível aperfeiçoamento mediante novas pesquisas é referente à inclusão da influência do tamanho amostral para os cálculos dos parâmetros citados via estatísticas de ordem.

Outra sugestão para futuros trabalhos sobre o assunto é a necessidade de se verificar se a equação (9) sempre apresenta quatro raízes reais, dada a conhecida inequação da estatística (Jones; Rosco; Pewsey, 2011): $B \geq A^2 + 1$.

Sugere-se ainda o estudo e pesquisa para a adaptação do método de Faleschini para a rotulação de *outliers* para outros campos, como por exemplo, para dados bivariados, séries temporais, dados multivariados e outros.

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.

ADIL, Iftikhar Hussain; ZAMAN, Asad. Outliers detection in skewed distributions: split sample skewness based boxplot. **Economic Computation and Economic Cybernetics Studies and Research**, Bucharest, v. 54, n. 3, p. 279-296, 2020.

ANDRADE, Larissa Ribeiro de; CIRILLO, Marcelo Angelo; BEIJO, Luiz Alberto. Proposal of a bootstrap procedure using measures of influence in non-linear regression models with outliers; doi: 10.4025/actascitechnol.v36i1.17564. **Acta Scientiarum. Technology**, v. 36, n. 1, p. 93-99, 7 jan. 2014.

BABURA, Babangida Ibrahim; ADAM, Mohd Bakri; FRITIANTO, Anwar; SAMAD, Abdul Rahim Abdul. Modified boxplot for extreme data. **AIP Conference Proceedings**, New York, v. 1842, issue 1, May 2017.

BARBOSA, Josino José; PEREIRA, Tiago Martins; OLIVEIRA, Fernando Luiz Pereira de. Uma proposta para identificação de outliers multivariados. **Ciência e Natura**, *[S. l.]*, v. 40, p. e40, 2018. DOI: 10.5902/2179460X29535. Disponível em: https://periodicos.ufsm.br/cienciaenatura/article/view/29535 Acesso em: 21 ago. 2023.

BARBOSA, Josino José; DUARTE, Anderson Ribeiro; MARTINS, Helgem Souza Ribeiro. A performance evaluation in multivariate outliers identification methods. **Ciência e Natura**, *[S. l.]*, v. 42, p. e16, 2020. DOI: 10.5902/2179460X41662. Disponível em: https://periodicos.ufsm.br/cienciaenatura/article/view/41662 Acesso em: 21 ago. 2023.

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

BRYS, Guy; HUBERT, Mia; STRUYF, Anja. A robust measure of skewness. **Journal of computacional and graphical statistics**, v. 13, n. 4, p. 996-1017, December 2004.

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. Resistance outlier rules and the non-Gaussian case. **Computational Statistics & Data Analysis**. V. 33, n. 3, p. 249-258, 2000.

CHAMBERS, J. M. et al. **Graphical methods for data analysis**. USA: Wadsworth, 1983.

FALESCHINI, Luigi. Su alcune proprietà dei momenti impegati nello studio dela variabilità, assimetria e curtosi. **Statistica**, anno VIII, n. 4, p. 503-513, Ottobre-Dicembre 1948.

FIORI, Anna Maria; ZENGA, Michele. The meaning of kurtosis, the influence function and an early intuition by L. Faleschini. **Statistica**, anno LXV, n. 2, 2005, p. 135-144.

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

JONES, M. C.; ROSCO, J. F.; PEWSEY, Arthur. Skewness-Invariant Measures of Kurtosis. **The American Statistician**, v. 65, n. 2 , p. 89-95, May 2011.

KIMBER, A. C. Exploratory data analysis for possibly censored data from skewed distributions. **Applied Statistics**. V. 39, n. 1, p. 21-30, 1990.

LIMA, Luís Fernando Maia; MAROLDI, Alexandre Masson; SILVA, Dávilla Vieira Odízio da; HAYASHI, Carlos Roberto Massao; HAYASHI, Maria Cristina Piumbato Innocentini. 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.

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**, Porto Alegre, v. 24, Edição Especial 6 EBBC, p. 216-235, 2018.

PEREIRA, Tiago Martins; CIRILLO, Marcelo Ângelo; OLIVEIRA, Fernando Luiz Pereira de. Chisquaremax rotation criterion in factor analysis: a Monte Carlo assessment of the effect of outliers. **Acta Scientiarum. Technology**, v. 36, n. 4, p. 643-649, 12 set. 2014.

RODRIGUES, Paulo Jorge Canas; ALMEIDA, Rafael; MUSTAFA, Kézia. The usefulness of robust multivariate methods: A case study with the menu items of a fast food restaurant chain. **Ciência e Natura**, *[S. l.]*, v. 42, p. e17, 2020. DOI: 10.5902/2179460X39892. Disponível em: https://periodicos.ufsm.br/cienciaenatura/article/view/e18%27 Acesso em: 21 ago. 2023.

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

SILVA, Kelly C. Ramos da; OLIVEIRA, Helder L. Costa de; CARVALHO, André C.P.L.F. de. Performance evaluation of outlier rules for labelling outliers in multidimensional dataset. **International Journal of Business and Data Mining**, v. 19, n. 2, p. 135-152, 21 July 2021.

SILVA, Kelly Cristina Ramos da. Regras robustas para rotular *outliers* em dados de caudas leves e caudas pesadas. **Tese de Doutorado**. 2019. Disponível em: https://teses.usp.br/teses/disponiveis/55/55134/tde-29042019-145141/pt-br.php Acesso em 12 ago. 2023.

TAMBAY, Jean-Louis. 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, p. 229-234, 1988.

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

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

VELLEMAN, P. F.; HOAGLIN, D.C. **Applications, basics, and computings of exploratory data analysis**. Boston: Duxbury, 1981.

VELOSO, Manoel Vitor de; CIRILLO, Marcelo Angelo. Principal components in the discrimination of outliers: A study in simulation sample data corrected by Pearson's and Yates´s chi-square distance. **Acta Scientiarum. Technology**, v. 38, n. 2, p. 193-200, 1 Apr. 2016.

VISSOTTO JUNIOR, Dornelles; DIAS, Nelson Luís. Método Empírico para Determinação de outliers em Séries de Fluxos de dados Micrometeorológicos Pós-processados. **Ciência e Natura**, *[S. l.]*, v. 35, p. 169–171, 2013. DOI: 10.5902/2179460X11585. Disponível em: https://periodicos.ufsm.br/cienciaenatura/article/view/11585 Acesso em: 21 ago. 2023.

WALKER, M. L; DOVOEDO, Y. H.; CHAKRABORTI, S.; HILTON, C. W. An improved boxplot for univariate data. **The American Statistician**, v. 72, n. 4, p. 348-353, November 2018.