Revista brasileira de iniciação científica
ISSN: 2359-232X

ANÁLISE EXPLORATÓRIA GRÁFICA PARA DADOS ASSIMÉTRICOS COM PRESENÇA DE PONTOS DISCREPANTES

GRAPHICAL EXPLORATORY ANALYSIS FOR ASYMMETRIC DATA WITH THE PRESENCE OF OUTLIERS

ANÁLISIS GRÁFICO EXPLORATORIO PARA DATOS ASIMÉTRICOS CON PRESENCIA DE PUNTOS ATÍPICOS

Ana Flávia Giacondino Soligo Lezcano Tatis[1]
José Eduardo Corrente[2]
Giovana Fumes Ghantous[3]

Resumo: Introdução: Gráficos são ferramentas usuais na análise exploratória de dados contínuos, pois possibilitam visualizar a forma de sua distribuição - simétrica ou assimétrica, e identificar possíveis pontos atípicos. Objetivo: Comparar gráficos para análise exploratória de dados com distribuições assimétricas e presença de pontos *outliers*. Metodologia: Histogramas e construções de *box-plot* com observações discrepantes foram avaliados e uma aplicação foi feita a dados de consumo alimentar. Resultados: Os *box-plot ajustados* evidenciaram com mais precisão os pontos atípicos presentes nas distribuições de consumo, que apresentaram formas assimétricas acentuadas à direita. Conclusões: Para dados assimétricos, os *box-plot ajustados* evidenciam melhor as discrepâncias.
Palavras-chave: *Box-plot*. *Box-plot* ajustado. Histograma. Medidas resumo.

Abstract: Introduction**:** Graphs are usual tools in continuous data exploratory analysis, as they make it possible to visualize the form of its distribution - symmetric or asymmetric, and identify possible atypical points. Objective: To compare graphical tools for exploratory data analysis with asymmetric distributions and presence of outliers. Methodology: Histograms and box-plot constructions with discrepant observations were evaluated and an application was made to food consumption data. Results: The adjusted box-plot showed with more precision the atypical points present in the consumption distributions, which presented asymmetric forms accentuated to the right. Conclusions: For asymmetric data, the adjusted box-plots evidence observations with discrepancies better.
Keywords: Box-plot. Adjusted Box-plot. Histogram. Summary statistics.

[1] Graduanda em Engenharia de Alimentos. Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo. ORCID: https://orcid.org/0000-0002-3974-8214. E-mail: anatatis@usp.br.
[2] Doutor em Estatística e Experimentação Agronômica. Faculdade de Medicina, Universidade Estadual Paulista "Júlio de Mesquita Filho". ORCID: https://orcid.org/0000-0001-5478-4996 . E-mail: jose.corrente@unesp.br.
[3] Doutor em Estatística. Departamento de Ciências Básicas. Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo. ORCID: https://orcid.org/0000-0002-1505-1826 . E-mail: giovana.fumes@usp.br.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Resumen: Introducción: Las representaciones gráficas son herramientas usuales en el análisis exploratorio de datos continuos, pues posibilitan visualizar la forma de su distribución - simétrica o asimétrica, y reconocer posibles puntos atípicos. Objetivo: Comparar herramientas gráficas para el análisis exploratorio de datos con distribuciones asimétricas y la presencia de puntos aislados. Metodología: Histogramas y construcciones de box-plot con observaciones discrepantes fueron evaluados y una aplicación fue hecha a datos de consumo alimentar. Resultados: Los box-plot ajustados evidenciaron con mayor precisión los puntos aislados presentes en las distribuciones de consumo, que presentaron formas asimétricas acentuadas a la derecha. Conclusiones: Para datos asimétricos, los diagramas de caja ajustados evidencian mejor las discrepancias.
Palabras-clave: Diagrama de caja. Diagrama de caja ajustado. Histograma. Medidas de resumen.

Submetido 05/09/2021 **Aceito 12/09/2022** **Publicado 15/09/2022**

Revista brasileira de iniciação científica
ISSN: 2359-232X

Introdução

A análise descritiva de um banco de dados é de extrema importância para caracterizar os dados coletados de uma pesquisa científica, cujo objetivo é resumir esses dados como uma base para posterior realização da inferência estatística. Essa análise pode ser feita por meio da utilização de técnicas gráficas e medidas resumo (MORETTIN e BUSSAB, 2017).

Referente às técnicas gráficas, estas são muito utilizadas para observar a distribuição dos dados e a variabilidade existente nos dados das variáveis de interesse. Ferramentas gráficas como histograma, *box-plot* e *box-plot* ajustado são algumas disponíveis para tal finalidade (MORETTIN e SINGER, 2019).

O histograma, que constitui um gráfico de barras adjacentes, é obtido através da divisão da amplitude dos dados em intervalos menores, chamados de intervalos de classe, é um primeiro procedimento para observar a forma da distribuição de um conjunto de dados (MORETTIN e BUSSAB, 2017).

Além do histograma, outra forma de visualizar a distribuição de dados pode ser feita por meio do *box-plot*, conhecido também como 'gráfico de caixa' ou 'diagrama de caixa'. De modo mais preciso que o histograma, essa ferramenta revela informações sobre posição, dispersão, assimetria e caudas dos dados, bem como pontos discrepantes. Entretanto, para distribuições com assimetria acentuada, em um gráfico de *box-plot*, os pontos que excedem os limites superior e inferior, são considerados distantes do centro da distribuição dos dados, e são muitas vezes declarados como *outliers*, embora não necessariamente o sejam.

Nesse sentido, o *box-plot* ajustado, proposto por Hubert e Vandervierenb (2008) é uma alternativa mais interessante para distribuições assimétricas, uma vez que inclui uma medida robusta de assimetria na determinação dos limites superior e inferior, resultando em uma apresentação mais precisa dos possíveis valores atípicos de um banco de dados.

Uma pesquisa realizada em Dourados no Mato Grosso do Sul (GUIMARÃES, 2019), que teve como objetivo a criação de um índice de eficiência das equipes de fiscalização do corpo de bombeiros no estado em que foi executada, comparou diversos métodos de detecção de *outliers* e a partir de um fluxograma elaborado para orientação da escolha do método que melhor se adequa ao seu projeto (SEO, 2006), optou-se pela utilização do *box-plot* ajustado,

que resultou na identificação e exclusão de pontos considerados discrepantes de maneira equivocada por outros métodos.

Adicionalmente, um estudo realizado em Brasília (DIAS, 2018), apresentou duas aplicações da distribuição de Birnbaum-Saunders (BS), descrita por Leiva (2016), em conjuntos de dados reais com o propósito de averiguar ajustes de modelos de regressão. Nessa pesquisa, foi feita uma comparação entre quatro ferramentas gráficas (histograma, gráfico de dispersão, *box-plot* usual e *box-plot* ajustado), no qual foram observadas distribuições assimétricas à direita. A conclusão foi que, nesses casos, o *box-plot* ajustado fornece representações mais precisas dos dados e de possíveis *outliers*.

Uma conclusão semelhante é feita no trabalho que apresenta uma medida utilizada para avaliar o desempenho de regras para determinação de pontos discrepantes (SILVA, 2019), indicando o uso do *box-plot* ajustado tanto para distribuições assimétricas quanto para aquelas sobre as quais não se tem conhecimento sobre a simetria.

Assim, este estudo teve como objetivo avaliar ferramentas gráficas para análise exploratória de dados contínuos, a fim de caracterizar a forma das distribuições para um conjunto de dados de consumo alimentar de um estudo epidemiológico da cidade do Rio de Janeiro, RJ, Brasil. Este trabalho apresenta a análise exploratória usada para refinamento do referido conjunto de dados, para posterior estudo inferencial com modelos da classe de distribuições Box-Cox Simétricas (FERRARI e FUMES, 2017). Os dados de consumo alimentar tendem a apresentar distribuições assimétricas acentuadas à direita, com presença de pontos atípicos. A geração de dados com tais características se dá pelo fato da maioria das pessoas não manter uma dieta constante durante os diferentes dias da semana, em especial aos finais de semana, nos quais as pessoas participam de atividades de lazer, e, com frequência, extrapolam em seu consumo habitual.

Este artigo está organizado em quatro seções. A presente seção apresenta uma descrição das principais ferramentas gráficas usadas para caracterização de distribuições para dados contínuos, e uma breve descrição e motivação sobre o banco de dados a ser analisado. A próxima seção apresenta a metodologia, na qual foram detalhadas as ferramentas gráficas utilizadas, o conjunto de dados e a ferramenta computacional utilizada - o programa R (RSTUDIO, 2022). A terceira seção apresenta os resultados, e, por fim, as conclusões sobre o estudo realizado são apresentadas.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Metodologia

Para a construção do histograma, a variável resposta é dividida em intervalos de classe e a contagem das frequências de ocorrência em cada classe é realizada (MORETTIN e BUSSAB, 2017).

Já o *box-plot* é construído a partir de um resumo dos dados, que é composto basicamente dos valores mínimo, primeiro quartil (Q1), mediana (Q2), terceiro quartil (Q3) e máximo, constituindo o que se chama de resumo de cinco números. A caixa é desenhada do primeiro quartil (Q1) ao terceiro quartil (Q3), sendo sua amplitude, chamada de amplitude interquartílica (AIQ) obtida através da expressão: $AIQ = Q3 - Q1$. A mediana (Q2) é representada por uma linha grossa, desenhada no interior da caixa em sua altura previamente estabelecida. Os *whiskers* ou bigodes são linhas que partem da extremidade da caixa até os pontos mais remotos da amostra que não ultrapassem os limites inferior e superior, dados por $\mathrm{LI} = \mathrm{Q1} - 1{,}5 \times \mathrm{AIQ}$ e $\mathrm{LS} = \mathrm{Q3} + 1{,}5 \times \mathrm{AIQ}$, , respectivamente. Observações que se encontram fora das demarcações são consideradas atípicas, possuindo comportamento distinto da maior parte dos dados (MORETTIN e BUSSAB, 2017).

A justificativa para usar os limites acima baseia-se em uma curva normal com média zero, e, consequentemente, mediana também igual a zero. Neste caso, para Q1=-0,6745, Q2=0 e Q3=0,6745, e AIQ=1,349, se considerar a medida 1,5 X AIQ, tem-se que os limites inferiores e superiores são dados por LI=-2,698 e LS=2,698. Assim, a área da distribuição normal entre esses dois valores, abaixo da curva, é de 0,993, o que significa que 99,3% da distribuição dos dados está entre esses dois valores. Assim, os valores extremos seriam os que constituiriam os mais distantes do centro da distribuição, cerca de 0,7% da mesma (MORETTIN e BUSSAB, 2017).

E, por fim, o *box-plot* ajustado é construído basicamente da mesma forma, mas com alguns valores de ajustes nos bigodes diferentes. A caixa e a mediana neste caso são representadas da mesma maneira. A diferença entre o *box-plot* e o *box-plot* ajustado está no cálculo dos limites (LI e LS) e consequentemente na determinação dos pontos que de fato estão mais distantes da distribuição dos dados em uma forma assimétrica, ressaltando-os com mais clareza. Para esta ferramenta, a definição dos limites é feita através das expressões $\mathrm{LI} = \mathrm{Q1} - 1{,}5\mathrm{e}^{-4\mathrm{MC}} \times \mathrm{AIQ}$ e $\mathrm{LS} = \mathrm{Q3} + 1{,}5\mathrm{e}^{3\mathrm{MC}} \times \mathrm{AIQ}$, para valores de *medcouple* (MC) iguais ou superiores a zero, ou $\mathrm{LI} = \mathrm{Q1} - 1{,}5\mathrm{e}^{-3\mathrm{MC}} \times \mathrm{AIQ}$ e $\mathrm{LS} = \mathrm{Q3} + 1{,}5\mathrm{e}^{4\mathrm{MC}} \times \mathrm{AIQ}$ para $\mathrm{MC} < 0$. O

Revista brasileira de iniciação científica
ISSN: 2359-232X

conceito de *medcouple* é uma medida estatística robusta, ou seja, que é menos sensível a observações atípicas, e que mede a assimetria de uma distribuição de dados (BRYS; HUBERT; STRUYF, 2004).

O *medcouple* é definido por $MC = med_{x_i \leq Q_2 \leq x_j} \mathrm{h}(x_i, x_j)$, em que Q_2 é a mediana da amostra, e para todo $x_i \neq x_j$, a função kernel h é dada por $h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}$. Em termos de gráfico de caixas, o valor *medcouple=0* resulta no *box-plot* padrão, os valores de *MC<0,* modelam uma distribuição de cauda assimétrica à esquerda, e *MC>0,* modelam uma distribuição com cauda assimétrica à direita. As constantes usadas nos cálculos dos bigodes dos *box-plot* ajustados com o uso do valor *medcouple*, são definidas de forma análoga a tradicional, a qual utiliza a distribuição normal e propõe 0,7% dos dados como valores discrepantes, porém, no caso do *box-plot* ajustado, um modelo exponencial é usado para definir tais constantes (HUBERT e VANDERVIERENB, 2008).

O programa utilizado para obter os gráficos e as medidas resumo foi o R Studio, versão 3.6.2 (RSTUDIO, 2022). Para a construção dos *box-plot* ajustados (HUBERT e VANDERVIERENB, 2008) foi necessária a instalação do pacote *robustbase*, o qual possui a função *adjbox(),* e para a construção dos gráficos de caixa padrão, a função *boxplot()* foi utilizada, e no mesmo gráfico, a função *points*() foi usada para identificar a média. O comando *cut()* foi utilizado para determinar a quantidade de pontos discrepantes de cada variável de acordo com os pontos fornecidos pelos gráficos de caixas e *box-plot* ajustados.

Adicionalmente, usando a mesma ferramenta computacional, os histogramas foram construídos a partir da função *hist(),* e no mesmo gráfico, a função *abline()* foi utilizada duas vezes, o argumento para traçar uma linha vertical, denotado por *v*, foi utilizado com *v=median()* e *v=mean()*, para inclusão de linhas que representam, respectivamente, a posição da mediana e média de cada variável do conjunto de dados. Para obtenção de informações sobre os gráficos de caixas foram utilizados as funções *boxplot()* e *adjbox()*, com o argumento *plot=F* (*False*), para obter-se um resumo das medidas utilizadas para a construção dos gráficos.

Para o estudo da análise gráfica exploratória, foi utilizado um banco de dados referente a uma investigação nutricional de 302 indivíduos, habitantes da cidade do Rio de Janeiro, Rio de Janeiro, Brasil. O objetivo do estudo foi avaliar o consumo alimentar de micro e

Revista brasileira de iniciação científica
ISSN: 2359-232X

macronutrientes da população. Além dos dados de consumo, também foram coletados alguns dados demográficos e antropométricos.

Dados de consumo alimentar, na área nutricional, são importantes para avaliar a adequação de consumo de acordo com as recomendações para uma alimentação saudável. As recomendações nutricionais são distintas para homens e mulheres e, desse modo, todas as análises aqui realizadas foram feitas por gênero.

Para a coleta dos dados fez-se uma amostragem do tipo bola de neve, na qual primeiramente, foram eleitos como entrevistadores, vinte e três graduandos do curso de nutrição, treinados para selecionar indivíduos dispostos a responder detalhes sobre o seu consumo alimentar nas últimas 24 horas por 20 dias não consecutivos. Cada instrumento de coleta de consumo alimentar é chamado de recordatório 24 horas (R24h). Neste caso, para cada participante obteve-se 20 recordatórios.

Os dados de consumo foram coletados entre março de 2013 e abril de 2014, cobrindo tanto dias de semana como finais de semana. Cada indivíduo foi acompanhado por aproximadamente 3 meses, seguindo a recomendação de que o R24h não deveria ser administrado nos mesmos dias da semana para o mesmo entrevistado. Para este estudo foram selecionadas variáveis relativas ao consumo de carboidratos, proteínas, fósforo, magnésio, niacina (vitamina B3) e sódio.

Inicialmente foi feita uma análise descritiva do consumo de carboidratos, proteínas, fósforo, magnésio, niacina (vitamina B3) e sódio por gênero, em que foram calculadas medidas de posição (mínimo, média, mediana e máximo) e de dispersão (variância, desvio padrão e coeficiente de variação). Para análise das ferramentas gráficas foram construídos histogramas, *box-plot* e *box-plot* ajustados.

Para estabelecer critérios de comparação na área nutricional, a Tabela 1 foi consultada.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Tabela 1 – Valores de referências equivalentes à média da distribuição da necessidade estimada (EAR) de um nutriente em um grupo de indivíduos do mesmo sexo com idade superior a 18 anos.

Nutrientes	*EAR*	
	Mulheres	**Homens**
Carboidratos (g)	100	100
Fósforo (mg)	580	580
Niacina (mg)	11	12

Fonte: PADOVANI *et al.*, 2006.

A Tabela 1 apresenta valores de EAR (*Estimated Average Requirement*) para alguns nutrientes estudados. A EAR é uma medida equivalente à média da distribuição da necessidade estimada de um nutriente em um grupo de indivíduos do mesmo sexo e faixa etária (PADOVANI *et al.*, 2006). Os valores de EAR foram consultados para avaliar se os dados coletados eram próximos aos valores recomendados aos indivíduos do estudo.

Resultados

De acordo com as informações dispostas na Tabela 2, o banco foi composto majoritariamente por mulheres (71,2%), com ensino superior (61,4%), eutróficos (82,8%) e faixa etária de 18 a 30 anos (68,4%). Assim, nota-se que os consumos alimentares do estudo são provenientes de pessoas jovens, em sua maioria do sexo feminino, com escolaridade no ensino superior e com uma condição nutricional considerada adequada, uma vez que o grupo foi composto em sua maioria por pessoas com um índice de massa corporal ideal, o que caracteriza uma população jovem adulta com boa instrução e com peso adequado. As análises foram realizadas por gênero, com os participantes que declararam esta informação. A separação por gênero é imprescindível, pois as recomendações nutricionais para pessoas do sexo feminino e masculino são distintas (ASSUMPÇÃO *et al.*, 2017).

Revista brasileira de iniciação científica
ISSN: 2359-232X

Tabela 2 – Caracterização dos participantes da pesquisa - adultos residentes na cidade do Rio de Janeiro em 2014.

Variáveis	Categorias	N	%
Gênero	Feminino	203	71,2
	Masculino	82	28,8
Escolaridade	Até o ensino médio	96	33,7
	Ensino superior	175	61,4
	Não declarado	14	4,9
Índice de massa corporal (IMC)	Eutróficos (IMC ≤ 28)	236	82,8
	Excesso de peso (IMC >28)	49	17,2
Faixa etária	De 18 a 30 anos	195	68,4
	De 31 a 50 anos	64	22,5
	Acima de 50 anos	26	9,1

Fonte: autoria própria (2021).

A Tabela 3 apresenta medidas descritivas (valores mínimo e máximo, média, mediana, desvio padrão e coeficiente de variação em porcentagem) referentes ao consumo de macro e micronutrientes selecionados para este estudo, determinadas de acordo com o gênero dos entrevistados.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Tabela 3 – Medidas resumo sobre o consumo de nutrientes por gênero, relatados por adultos da cidade do Rio de Janeiro em 2014.

Nutrientes	**Homens**						**Mulheres**					
	Mín	Média	Md	DP	CV (%)	Máx	Mín	Média	Md	DP	CV (%)	Máx
Carboidratos (g)	17,0	300,2	273,5	145,1	48,3	1399,3	5,9	224,5	209,6	95,2	42,4	862,0
Proteínas (g)	0,7	109,3	96,4	65,2	59,7	548,5	3,6	81,6	72,4	44,4	54,4	561,6
Fósforo (mg)	95,7	1367,3	1212,3	729,1	53,3	5299,4	81,0	1101,6	968,9	615,5	55,9	7123,3
Magnésio (mg)	0,0	275,0	250,3	139,3	50,6	1075,4	17,1	223,1	202,7	121,2	54,3	1601,8
Niacina (mg)	0,0	24,9	18,5	21,5	86,2	196,6	0,2	18,0	14,6	12,8	71,2	161,8
Sódio (mg)	38,3	2489,9	2093,7	1735,6	69,7	14624,6	29,4	1976,3	1611,7	1520,0	76,9	15297,0

Mín=mínimo, Md=Mediana, DP=desvio padrão, Máx=Máximo, CV(%)=coeficiente de variação em porcentagem, dado por $CV = 100\ x\ \frac{DP}{Media}$.

Fonte: autoria própria (2021).

Pela Tabela 3, notou-se que as médias e medianas, relativas ao consumo de todos os nutrientes, resultaram em valores mais elevados para indivíduos do sexo masculino; para ambos, a maior amplitude de consumo foi na quantidade diária de sódio. Para as medidas de dispersão (desvio padrão e coeficiente de variação), observou-se que tais medidas apresentaram valores maiores para homens, o que indica uma variabilidade de consumo dos macros e micronutrientes considerados em relação à média mais pronunciada, o que aponta uma maior heterogeneidade no consumo.

Por meio do desvio padrão, esta variabilidade pode ser quantificada, e ao calcular o coeficiente de variação, tem-se um percentual do quanto esta variabilidade está ocorrendo em torno da média. Observou-se ainda que a maior parte dos valores médios se encontravam distantes dos valores medianos, destaque para os valores de consumos de fósforo e de sódio para os adultos do sexo masculino, em que as diferenças entre as medidas foram de 155 mg e 396,2 mg respectivamente, o que aponta a existência de pontos aberrantes e de distribuições

Revista brasileira de iniciação científica
ISSN: 2359-232X

assimétricas, uma vez que em distribuições de formato simétrico, as medidas de tendência central média e mediana tendem a ser muito próximas ou coincidentes (MORETTIN e BUSSAB, 2017).

Para as adultas da amostra foi notado um comportamento diferente daquele observado para os homens da amostra, nestes foi possível perceber uma proximidade maior entre as medidas de posição central (média e mediana). Referente ao consumo de proteínas, a distância entre média e a mediana foi pequena, equivalendo a 9,2 g, como visto na Tabela 3, o mesmo foi observado para a ingestão de niacina, que apresentou diferença entre média e mediana menor ainda, apenas 3,4 mg.

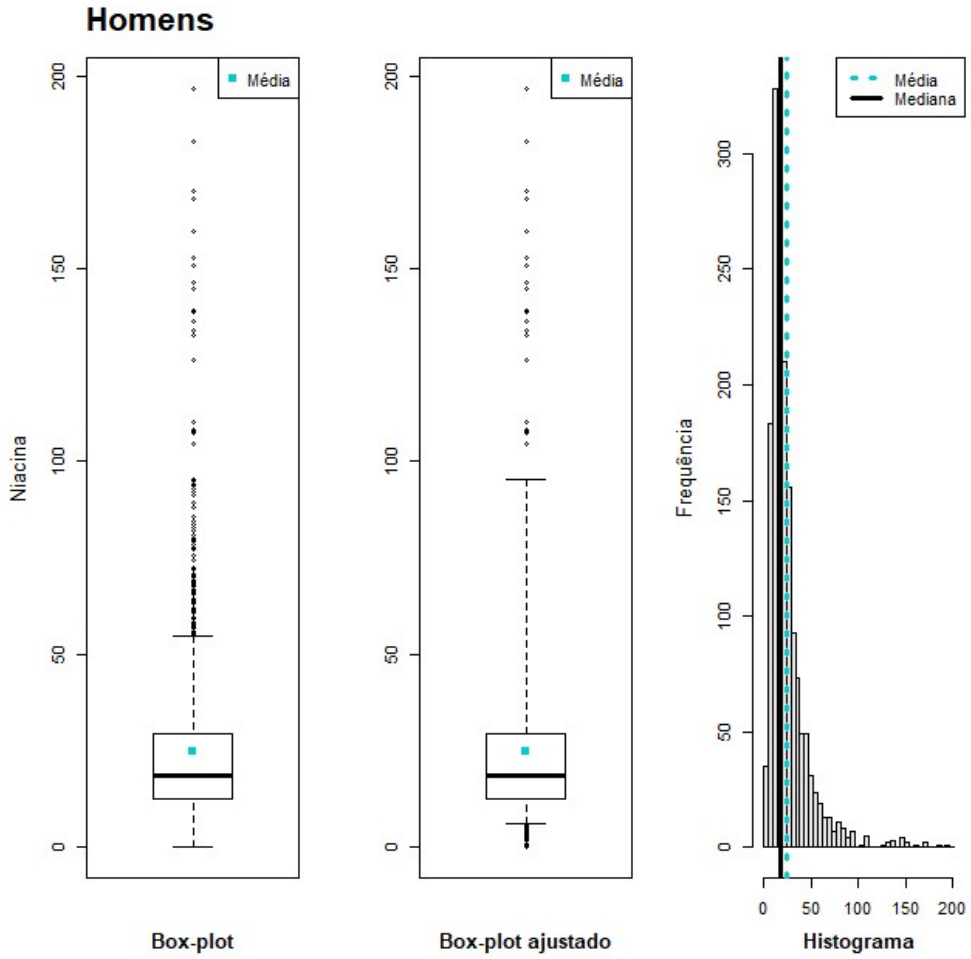
Tanto para os homens quanto para as mulheres, os valores mais altos de desvio padrão encontrados foram referentes ao consumo de sódio, porém, a partir do cálculo do coeficiente de variação, constatou-se que a dispersão dos dados em relação à média de consumo foi maior para a niacina (Vitamina B3) (CV = 86,2%) para entrevistados do sexo masculino, enquanto que para os entrevistados do sexo feminino confirmou-se o consumo de sódio com maior afastamento dos dados em torno da média (CV = 76,9%) (Vide Tabela 3).

Ferramentas gráficas foram utilizadas para representar as distribuições referentes ao consumo de nutrientes pela população estudada, foram construídos histogramas, *box-plot* e *box-plot* ajustados (HUBERT e VANDERVIERENB, 2008), cujo objetivo de aplicação foi comparar as ferramentas gráficas para percepção dos pontos discrepantes em distribuições assimétricas. Os gráficos gerados no programa R foram produzidos e exibidos separadamente para homens e mulheres. Os códigos utilizados e as figuras referentes a todos os nutrientes descritos na Tabela 3 estão alocados na plataforma *GitHub*, acesso disponível em *https://github.com/AnaFTatis/AnaliseDadosEliseu.git*. Aqui, alguns dessas figuras foram exibidas.

No presente estudo observa-se a existência de assimetrias acentuadas à direita nas distribuições de consumo dos nutrientes examinadas através da leitura das ferramentas gráficas, dos nutrientes estudados. Para exemplificar, as distribuições de consumo niacina e sódio para homens e mulheres são explicitadas (Figuras de 1 a 4).

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 1 – Gráficos referentes ao consumo de niacina por homens na cidade do Rio de Janeiro em 2014.

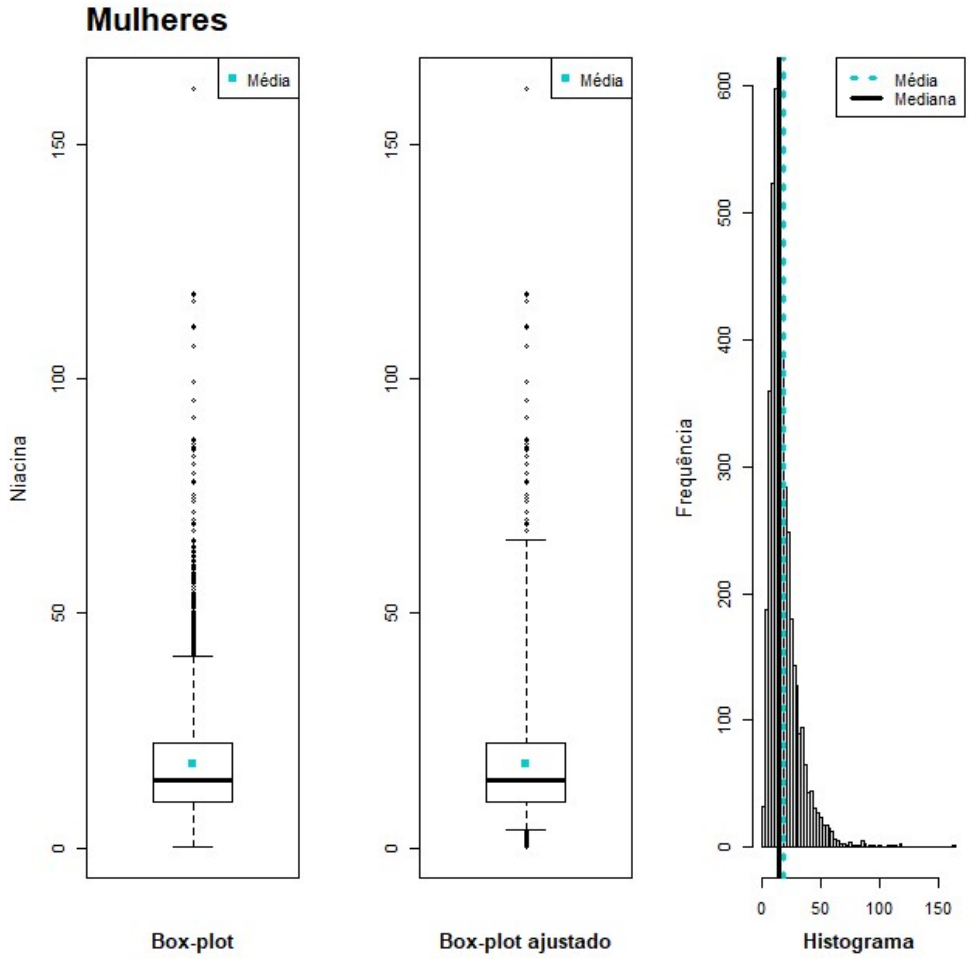


Fonte: autoria própria (2021).

Como esperado, notou-se que as caixas desenhadas para os dois tipos de *box-plot* expostos, o gráfico de caixas padrão e sua variante ajustada (HUBERT e VANDERVIERENB, 2008), assim como a mediana, ilustrada pela linha grossa no interior das caixas e a média, são exatamente iguais nos dois gráficos, enfatizando as diferenças na distorção das distribuições através do tamanho e disposição dos bigodes e consequentemente a quantidade de pontos discrepantes em cada versão da ferramenta gráfica em questão.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 2 – Gráficos referentes ao consumo de niacina por mulheres na cidade do Rio de Janeiro em 2014.

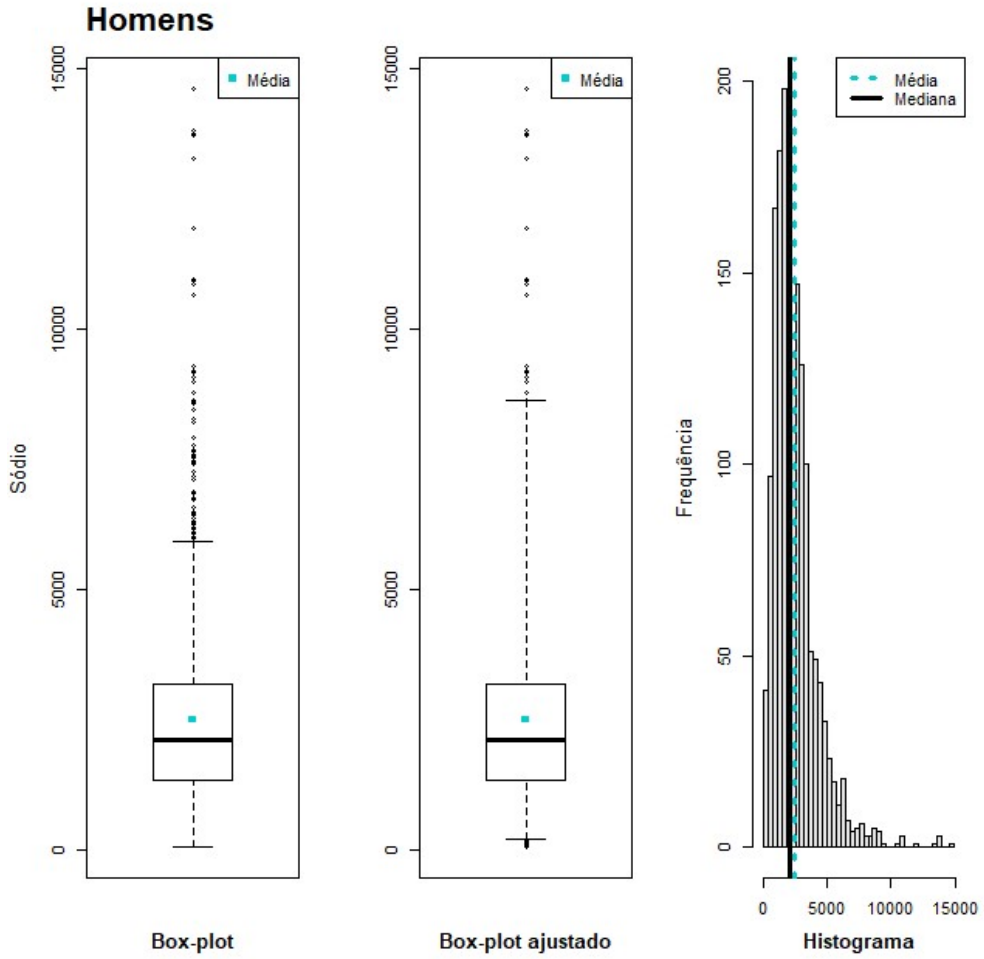


Fonte: autoria própria (2021).

O fato foi notado em todas as figuras expostas no presente estudo (Fig. 1 a 4), em que se percebeu que os bigodes superiores dos *box-plot* ajustados foram mais longos e os inferiores mais curtos quando comparados aos bigodes dos *box-plot* padrão. Desse modo, os gráficos de caixa ajustados evidenciaram a distorção da distribuição, refletindo mais claramente a cauda esquerda mais curta.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 3 – Gráficos referentes ao consumo de sódio por homens na cidade do Rio de Janeiro em 2014.

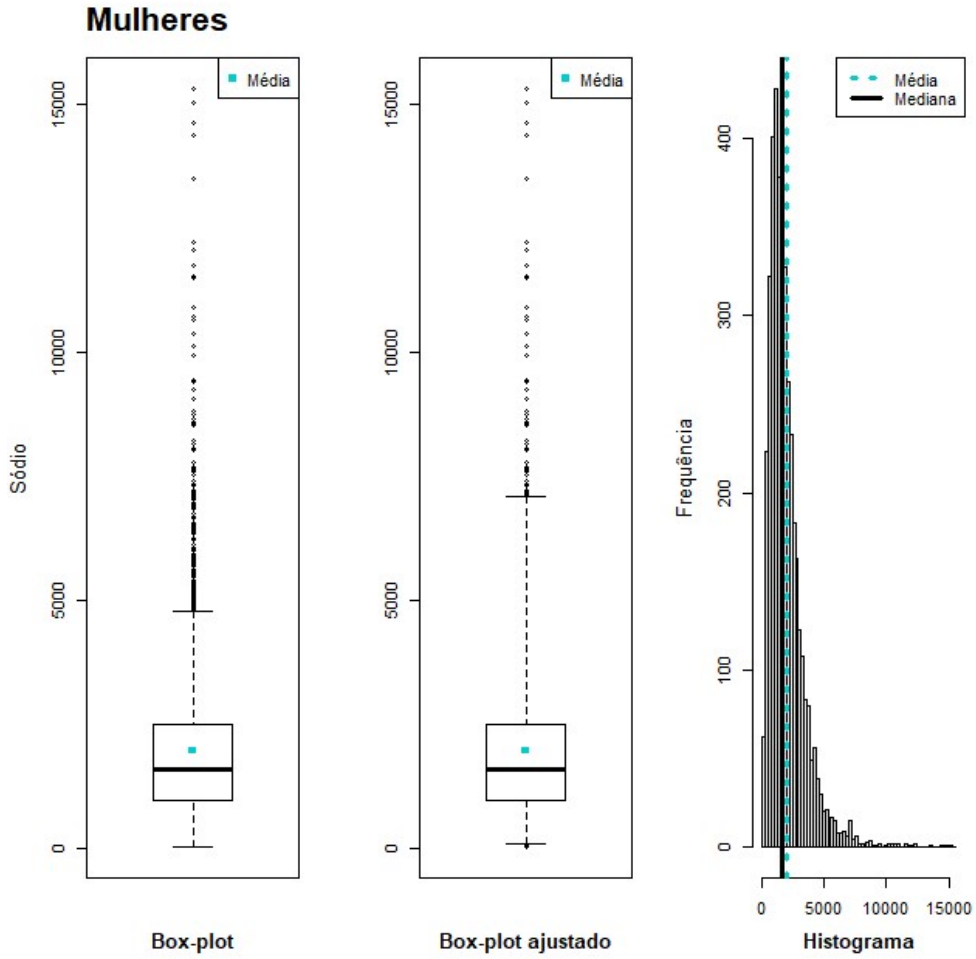


Fonte: autoria própria (2021).

Como consequência do alongamento dos bigodes superiores e redução dos bigodes inferiores, menos observações ultrapassaram os limites superiores nos *box-plot* ajustados, enquanto observações abaixo dos limites inferiores nestes mesmos gráficos foram consideradas pontos discrepantes, tornando a conferência de tais pontos menos trabalhosos, sendo que tais pontos de fato são considerados discrepantes, uma vez que a classificação dos mesmos é feita em relação a um conjunto de dados assimétrico.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 4 – Gráficos referentes ao consumo de sódio por mulheres na cidade do Rio de Janeiro em 2014.



Fonte: autoria própria (2021).

As ocorrências anteriormente descritas foram notadas nos gráficos construídos tanto para o gênero masculino, quanto para o gênero feminino. A Tabela 4 apresenta a quantidade de pontos discrepantes encontradas nos *box-plot* e nos *box-plot* ajustados. As informações estão dispostas na tabela de acordo com o gênero dos entrevistados. Por meio dela, uma comparação pode ser feita em relação a classificação de pontos considerados atípicos por uma técnica gráfica em relação a outra.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Tabela 4 – Número de *Outliers* detectados nos gráficos de caixa padrão e na sua versão ajustada.

Variáveis	***Número de Outliers***							
	Homens				**Mulheres**			
	Box-plot	*Box-plot* ajustado			*Box-plot*	*Box-plot* ajustado		
	Total (>LS)	Total	>LS	<LI	Total (>LS)	Total	>LS	<LI
Carboidratos	70	46	23	23	109	88	41	47
Proteínas	65	36	32	4	174	126	53	73
Fósforo	76	46	21	25	161	109	53	56
Magnésio	51	35	15	20	183	110	85	25
Niacina	117	100	24	76	225	131	33	98
Sódio	70	30	18	12	184	60	54	6

LS: Limite Superior e LI: Limite Inferior

Fonte: autoria própria (2022).

Neste aspecto os gráficos construídos sobre os consumos de niacina (Fig. 1) e sódio (Fig. 3) para os homens destacam aspectos interessantes, em que, no caso do consumo de niacina, o limite superior era de 55 mg no *box-plot* padrão, contendo 117 *outliers* e passou para 95 mg na sua versão ajustada, com apenas 24 observações acima deste valor e 76 observações abaixo do limite inferior definido como 6 mg; já no caso sódio, das 1640 observações, 70 foram determinadas como aberrantes no *box-plot* padrão, já no gráfico de caixas ajustado apenas 30 pontos foram considerados discrepantes, sendo 12 abaixo do limite inferior (191 mg) e 18 acima do limite superior (8625 mg).

Para as mulheres da amostra os gráficos construídos para os nutrientes sódio (Fig. 4) e niacina (Fig. 2) também se destacaram, os dados obtidos dessas figuras apontaram uma diferença de 124 pontos discrepantes entre os dois gráficos de caixas plotados para o consumo de sódio, uma diferença de 192 valores excedentes do limite superior entre os *box-plot* (padrão e ajustado) gerados para o consumo de niacina e o surgimento de 98 observações discrepantes abaixo do limite inferior de consumo de 4 mg de niacina no gráfico de caixas ajustado.

A partir da observação das medidas descritivas (Tabela 3), e dos gráficos construídos (*box-plot, box-plot* ajustado e histograma) juntamente com o levantamento de seus principais pontos, foi possível observar um padrão de comportamento claro e distinto para entrevistados dos gêneros feminino e masculino.

Em geral os homens da amostra consumiram mais macro e micronutrientes e de maneira mais irregular que as mulheres, revelando distribuições de dados com assimetrias mais

Revista brasileira de iniciação científica
ISSN: 2359-232X

acentuadas, reforçando a variabilidade observada nos histogramas construídos para a ingestão dos nutrientes estudados. Esse padrão de comportamento foi observado em todos os gráficos neste trabalho apresentados (Fig. 1 a 4).

Os consumos medianos de carboidratos, fósforo e niacina de homens e mulheres, foram superiores aos valores recomendados pelas EARs – *Estimated Averages Requirement*, que são valores correspondentes às médias das distribuições das necessidades dos nutrientes em um grupo de indivíduos saudáveis, segundo cada gênero e adultos. Desse modo, pode-se afirmar que os consumos medianos diários de tais nutrientes pelos indivíduos da população em estudo foram superiores as recomendações estabelecidas - Vide Tabela 1 (PADOVANI *et al.*, 2006)).

É importante ressaltar que apenas foram feitas comparações diretas com os valores encontrados para o consumo mediano dos nutrientes e seus respectivos valores de EAR, visando a complementação da interpretação dos resultados obtidos por meio do uso das ferramentas gráficas, sendo suficiente para atingir os objetivos pretendidos com este estudo.

Para uma análise mais detalhada sobre a avaliação dietética seria necessário o corrigir os dados de consumo pela variabilidade intrapessoal do grupo, que elimina a variabilidade da dieta e da ingestão de nutrientes dos indivíduos ao longo dos dias em que foram aplicados os recordatórios. Esta avaliação permitiria a verificação da prevalência de inadequação de consumo desses nutrientes (MORIMOTO e*t al.*, 2011). O presente estudo trata-se de uma análise descritiva e preliminar dos dados, no estudo inferencial a ser realizado posteriormente, um modelo de regressão adequado deve ser sugerido para que esta correção seja realizada.

Conclusões

As ferramentas gráficas são de grande valia para análise descritiva de um conjunto de dados, explorá-las de forma adequada favorecem resultados mais interessantes para posterior análise inferencial, como exemplo, distribuições da classe Box-Cox simétricas (FERRARI e FUMES, 2017) são candidatas a ajustar dados com tais características.

A partir do presente estudo conclui-se que os gráficos histograma, *box-plot* e *box-plot* ajustado são excelentes ferramentas para investigação sobre a forma da distribuição de um conjunto de dados, uma análise inferencial foi proposta posteriormente a este estudo, e tais ferramentas foram essenciais para identificação de possíveis problemas nos dados, e assim por meio delas, um refinamento pode ser realizado, e de forma especial, quando se trata de

distribuições assimétricas, os *box-plot* ajustados podem expressar com mais precisão a presença de pontos que se destacam de um conjunto de dados, uma vez que os pontos identificados por ele como aberrantes, foram de fato aqueles que não expressavam um consumo alimentar verdadeiro.

Adicionalmente, o *box-plot* usual, apresenta uma forma de classificação de pontos como sendo discrepantes diferente do *box-plot* ajustado, baseado na distribuição normal, sendo desta forma mais interessante para identificação de pontos atípicos em dados cuja distribuição é mais próxima da simetria, o *box-plot* ajustado apresenta uma forma de especificação que evidencia os pontos que estão de fato mais extremos dos valores considerados centrais da distribuição de um banco de dados para distribuições assimétricas, sendo por isso, é uma alternativa interessante para casos nos quais se deseja detectar pontos que estão muito distantes do conjunto de dados e que são dignos de melhor investigação.

No contexto de dados de nutrição, tais gráficos auxiliam os pesquisadores a investigarem se tais valores extremos têm significado prático, e como eles podem evidenciar a forma da distribuição usual de consumo, muito importante em estudos epidemiológicos para averiguar a inadequação de um grupo alimentar; de uma maneira geral, para os nutrientes aqui apresentados, é em geral muito assimétrica e apresenta diversas observações que necessitam de um cuidado especial para serem investigadas.

Agradecimentos

Os autores agradecem ao prof. Dr. Eliseu Verley Junior, do Departamento de Epidemiologia do Instituto de Medicina Social da UERJ – Rio de Janeiro por ceder os dados para este estudo e a Fundação de Amparo à pesquisa do Estado de São Paulo pelo apoio financeiro (processo FAPESP nº2020/03228-6 e processo FAPESP nº2019/02231-6).

Referências

ASSUMPÇÃO, D. D. *et al*. Diferenças entre homens e mulheres na qualidade da dieta: estudo de base populacional em Campinas, São Paulo, **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 22, n. 2, p. 347-358, fev. 2017.

BRYS, G.; HUBERT, M..; STRUYF, A. A robust measure of skewness, **Journal Computational and Graphical Statistics**, v. 13, p. 996–1017, 2004.

DIAS, R. V. B. **Modelagem baseada na distribuição.** 2018. Trabalho de Conclusão de Curso (Bacharelado em Estatística) - Universidade de Brasília, Brasília, 2018.

FERRARI, S. L. P.; FUMES, G. Box-Cox symmetric distributions and applications to nutritional data. **AStA-Advances in Statistical Analysis,** v. 101, p. 321-344, 2017.

GUIMARÃES, F. P. **Proposta de Criação de um Índice de Eficiência das Equipes de Fiscalização do Corpo de Bombeiros Militar do Estado de Mato Grosso Do Sul**. 2019. Dissertação (Mestrado em Administração Pública) - Universidade Federal da Grande Dourados, Dourados, 2019.

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

LEIVA, V. The Birnbaum-Saunders Distribution, **Academic Press**, Londres, v.4, n.16, p. 996-1017, 2016.

MORETTIN, P. A.; BUSSAB, W. D. O. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

MORETTIN, P. A.; SINGER, J. M. **Introdução à Ciência de Dados:** fundamentos e aplicações. São Paulo: Departamento de Estatística, USP, 2019.

MORIMOTO, J. M. *et al.* Variância intrapessoal para ajuste da distribuição de nutrientes em estudos epidemiológicos. **Rev. Saúde Pública**, São Paulo, v. 45, n. 3, p. 621-625, jun. 2011.

PADOVANI, R. M. et al. Dietary reference intakes: aplicabilidade das tabelas em estudos nutricionais. **Revista de Nutrição**, Campinas, v. 19, n. 6, p. 741-760, nov/dez 2006.

RSTUDIO. Disponível em: < https://www.rstudio.com>. Acesso em: 03 maio. 2022.

SEO, S. **A Review and Comparison of Methods for Detecting Outliers**. 2006. Dissertação (Mestrado em Saúde Pública) - Universidade de Pittsburgh Graduate School of Public Health, Pittsburgh, 2006.

SILVA, K. C. R. D **Robust Outilier Labeling Rules for Light-tailed and Heavy-tailed dat**. Tese (Doutorado em Ciências Matematicas e Cmputação) - Universidade de São Paulo: São Paulo, 2019.