ISSN 1981-3708

BOXPLOT: UM RECURSO GRÁFICO PARA A ANÁLISE E INTERPRETAÇÃO DE DADOS QUANTITATIVOS

BOXPLOT: A VISUAL RESOURCE FOR ANALYSIS AND INTERPRETATION OF QUANTITATIVE DATA

José VALLADARES NETO[1]; Cristiane Barbosa dos SANTOS[2]; Érica Miranda TORRES[3]; Carlos ESTRELA[4]

1 - Professor Associado de Ortodontia, FO-UFG;

2 - Mestranda do Programa de Pós-Graduação em Odontologia da FO-UFG;

3 - Professora Associada de Metodologia Científica, FO-UFG;

4 - Professor Titular de Endodontia, FO-UFG.

RESUMO

Introdução: O *boxplot* é um recurso gráfico usado regularmente na pesquisa científica para sumarizar e analisar dados quantitativos. Objetivo: Descrever didaticamente a estrutura, interpretação, construção, modificações e aplicações deste recurso visual. Resultados: O *boxplot* tradicional exibe medidas de tendência central não-paramétrica (mediana), de dispersão (quartis), forma de distribuição ou simetria da amostra (valores pontuais mínimo e máximo), valores atípicos (*outliers*) e extremos. Modificações podem ser incorporadas para possibilitar a inserção de parâmetros como média, desvio padrão e intervalo de confiança. As aplicações incluem análise exploratória dos dados, detecção de *outliers e* comparação entre grupos (equivalência). *Conclusão*: O *boxplot* é um recurso gráfico aperfeiçoado que cumpre com a análise exploratória e até mesmo inferencial dos dados e pode substituir o uso de tabelas em casos específicos.

PALAVRAS-CHAVE: Metodologia Científica; Estatística Descritiva; Gráficos; Quartil; Mediana; *Boxplot.*

INTRODUÇÃO

A pesquisa quantitativa lida com dados numéricos e a publicação dos seus resultados é sumarizada por meio de dados agrupados, dispostos em tabelas ou gráficos. A tabela apresenta os dados detalhados, aceita a análise simultânea de múltiplas variáveis e estabelece relações entre elas. Já o gráfico, com diversificada representação visual, permite a análise exploratória (ou descritiva) e a interpretação da tendência conjunta dos dados. Entre os diversos tipos de representação gráfica tem-se o *boxplot*, um sistema alternativo ao tradicional histograma[1].

O *box and whisker plot*, ou simplesmente *boxplot* (ou ainda *box-plot*), foi empregado pela primeira vez pelo matemático estadunidense John W. Tukey (1915–2000) em 1970[2], mas se tornou amplamente divulgado a partir da publicação formal em 1977[3]. O termo *box and whisker* tem origem do inglês e é traduzido *ipsis literis* como "caixa e bigode". O matemático almejou desenvolver um gráfico que sumarizasse a análise exploratória de dados. Pois, o *boxplot* é um recurso visual que resume os dados para exibir a mediana, quartis e os valores pontuais máximos e mínimos. Portanto, apresenta valores de tendência central, dispersão e simetria dos dados agrupados.

O *boxplot* é um tipo de gráfico usado regularmente na pesquisa científica e a sua construção é possível por meio de diversos *softwares* estatísticos. Ao gráfico tradicional podem ser incorporadas modificações, desde a maneira como as hastes são desenhadas, até mesmo a inserção de parâmetros como média, desvio padrão e intervalo de confiança, por exemplo. Por apresentar duas desvantagens peculiares (não são bem compreendidos por não-matemáticos e algumas informações não são transparentes)[1], o *boxplot* pode ser utilizado visando a aplicação e interpretação dos dados quantitativos quando bem indicado[4]. Considerando que em determinados casos o *boxplot* é um substituto aperfeiçoado para as tabelas, o presente artigo se propõe a descrever didaticamente a estrutura, interpretação, modificações e aplicações deste recurso gráfico à pesquisa científica.

BOXPLOT

Estrutura básica

O *boxplot* pode ser configurado em orientação horizontal ou vertical, ambas com o formato de "caixa e haste" (Figura 1). A estrutura básica é constituída por:

a. Caixa (*box*), que assume comumente o formato retangular;
b. Mediana (desenhada como uma linha dentro da caixa e simbolizada por Q2, ou seja, segundo quartil). Caso haja a representação de um conjunto de dados com distribuição normal, a linha é desenhada no centro da caixa, simbolizando a aproximação com a média aritmética;
c. Haste (bigode ou *whisker*), assemelhando-se à letra "T", representativa dos valores compreendidos entre a caixa e os valores limites, inferior e superior, do conjunto de dados. A extremidade da haste é comumente denominada *fence*.

Quais as informações e como interpretá-las?

O *boxplot* exibe a tendência central não-paramétrica (mediana), dispersão (quartis 25% e 75%), forma de distribuição ou simetria da amostra (valores pontuais mínimo e máximo), valores

atípicos (*outliers*) e extremos. Partindo de um *boxplot* com disposição vertical, têm-se as seguintes informações[5] (**Figura 2)**:

a. Eixo vertical: representa dados de valores numéricos;
b. Eixo horizontal: fator de interesse;
c. Primeiro quartil (Q1): onde se localiza ¼ ou 25% dos menores valores. Também chamado de quartil inferior ou 25º percentil. Representado pela linha limite inferior da caixa;
d. Mediana ou segundo quartil (Q2): é o local onde ocorre a divisão da metade superior (ou 50%) da metade inferior da amostra. É o 50º percentil. Representada pela linha dentro da caixa;
e. Terceiro quartil (Q3): onde se localiza ¾ ou 75% dos valores maiores. Também chamado de quartil superior ou 75º percentil. Representado pela linha limite superior da caixa;
f. Intervalo interquartílico (Q3 - Q1 ou IIQ): é definida como a diferença entre Q3 e Q1. No gráfico é representado pela dimensão da caixa. Estende-se do Q1 ao Q3 (percentis 25º a 75º). Representa o intervalo dos 50% dos dados em torno da mediana;
g. Limite inferior (tamanho ou extremidade do *whisker* ou *fence* inferior): valor mínimo do conjunto de dados, até 1,5 vezes o IIQ (uma vez e meia o intervalo interquartílico), excluindo os *outliers* e/ou extremos;
h. Limite superior (tamanho ou extremidade do *whisker* ou *fence* superior): valor máximo do conjunto de dados, até 1,5 vezes o IIQ (uma vez e meia o intervalo interquartílico), excluindo os *outliers* e/ou extremos;
i. *Outliers* (valores atípicos): valores acima e/ou abaixo de 1,5 vezes o IIQ;
j. Extremos: valores acima e/ou abaixo de 2,5 vezes o IIQ (duas vezes e meia o intervalo interquartílico)[6].

Após delimitadas as informações contidas no *boxplot*, é relevante que se informe como cada limite é estabelecido. Para favorecer este entendimento, uma analogia com a curva de distribuição normal (em forma de sino) pode ser realizada, sobrepondo-se a um *boxplot* horizontal (Figura 3).

Cabe ressaltar que a curva de distribuição normal apresenta cauda simétrica e o ponto central representa à média. Cada intervalo do desvio-padrão (1σ, 2σ e 3σ) cumpre com a regra dos 68-95-99, respectivamente, a respeito da dispersão dos dados. O *outlier* (representado na Figura 3 pela estrela cinza à esquerda) é uma circunstância hipotética, pois a distribuição normal não apresenta caudas assimétricas. O *boxplot* cujos dados são representativos de uma distribuição normal deverá apresentar a mediana posicionada com exatidão no meio da caixa (próximo à média) e as hastes distando igualmente ao centro da caixa, ou seja, dispostas simetricamente. A chance de encontrar *outliers* em *boxplot* em amostras com distribuição normal dependerá do tamanho da amostra.

A medida de tendência central, representada graficamente pela linha que dentro da caixa, é a mediana. Por definição, a mediana divide o tamanho da amostra na metade. Portanto, metade da amostra está abaixo da linha e a outra metade acima. É também simbolizada pelo segundo quartil (Q2). A posição simétrica da mediana dentro da caixa a aproxima da média aritmética; e a posição assimétrica simboliza a aproximação com dados não-paramétricos, os quais podem estar mais próximos do quartil inferior (Q1) ou do quartil superior (Q3).

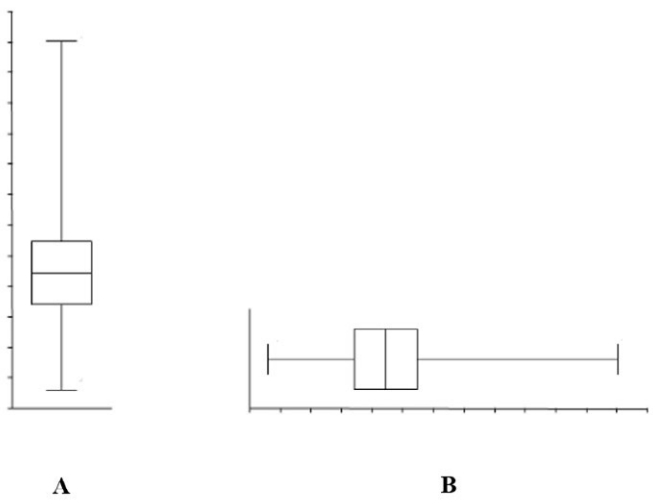

Figura 1 - Estrutura básica do *boxplot* em orientação vertical (A) e horizontal (B).

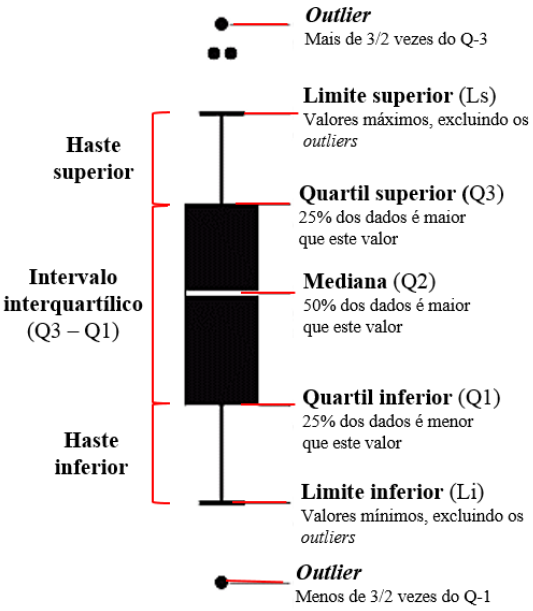

Figura 2 - Elenco de informações contidas no *boxplot*.

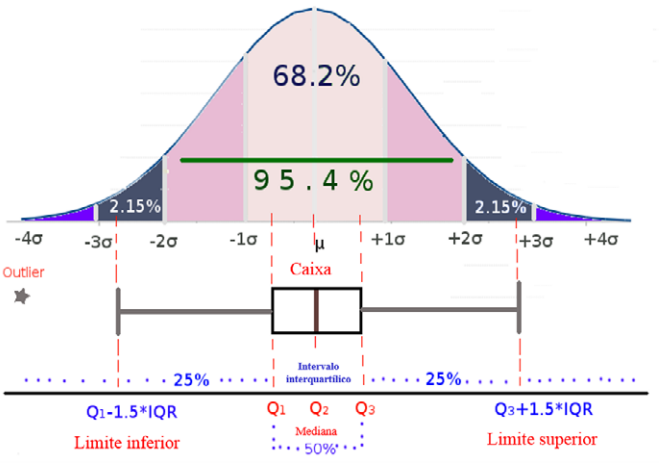

Figura 3 - Sobreposição gráfica de uma curva normal e um *boxplot* horizontal. (Fonte: https://estatistics.eu/what-is-statistics-charts-boxplot/, com modificações).

As medidas de dispersão ou variabilidade são simbolizadas pelas alturas da caixa e da haste. O tamanho da caixa representa a intervalo interquartílico, ou seja, a amostra compreendida entre o quartil inferior (Q1) e quartil superior (Q3). Metade (ou 50%) da amostra total está concentrada neste intervalo, sendo 25% entre Q1 e Q2 e os outros 25% entre Q2 e Q3. O tamanho das hastes é definido pelos valores do intervalo entre o quartil inferior (Q1) ao limite inferior (Li), e do quartil superior (Q3) ao limite superior (Ls), representando respectivamente a haste inferior e superior. Matematicamente, esse intervalo compreendido pelo tamanho das hastes é definido levando-se em conta uma 1,5 vezes o valor interquartílico (Q3-Q1), representado da seguinte forma:

Haste inferior (Li - Q1): Q1 – 1,5 (Q3-Q1)

Haste superior (Ls - Q3): Q3 + 1,5 (Q3-Q1)

O limite estabelecido pelas hastes de 1,5 vezes o IIQ foi determinado subjetivamente por Tukey[3]. Quando representam uma distribuição normal, os limites das hastes representam aproximadamente 2,70 do desvio padrão acima e abaixo da média[7] (Figura 3).

Embora os termos *boxplot* e *box-and-whiskers* sejam usados indistintamente, o segundo foi empregado primeiramente para descrever as hastes que se estendem até o valor mínimo e máximo. Por isso, alguns *boxplots* reconsideram esses valores das hastes com os valores do 2º e 98º percentis, ou seja, as hastes estendem-se até praticamente os valores mínimo e máximo da variável sob análise (Figura 4).

Normalmente, as hastes não englobam os *outliers* e extremos. Os *outliers* são valores individuais atípicos que distorcem os valores de tendência central e de dispersão. Segundo o método de Tukey[3], são considerados *outliers* os valores que ultrapassam os limites inferior (Li) e superior (Ls), designados matematicamente pelo seguinte entendimento:

Outlier inferior < [Q1 – 1,5 (Q3-Q1)] ou

Outlier superior > [Q3 + 1,5 (Q3-Q1)], onde (Q3-Q1) é o intervalo interquartílico.

Eles podem ser identificados graficamente pela representação em formato de "círculos" (º), "cruz"(+) ou asterisco (*). "*Outliers* próximos" se distinguem de "*outliers* distantes". Os *outliers* que ultrapassam os limites inferior ou superior entre 1,5 a 3,0 vezes o IIQ podem ser simbolizados com +, e são considerados próximos; e os "*outliers* distantes" são aqueles além de 3,0 vezes o IIQ, e podem ser marcados distintamente com * (Figura 5). Ou como mencionado anteriormente, outra forma de distinguir *outliers* é chamar de extremos valores acima e/ou abaixo de 2,5 vezes o IIQ[6].

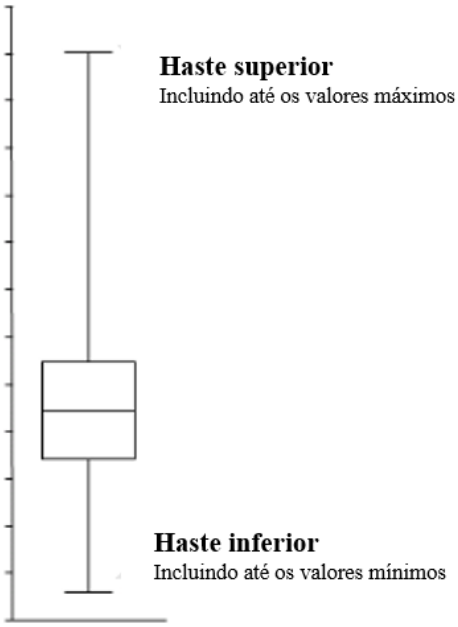

Figura 4 - *Boxplot* tradicional com hastes que englobam os valores mínimos e máximos observados.

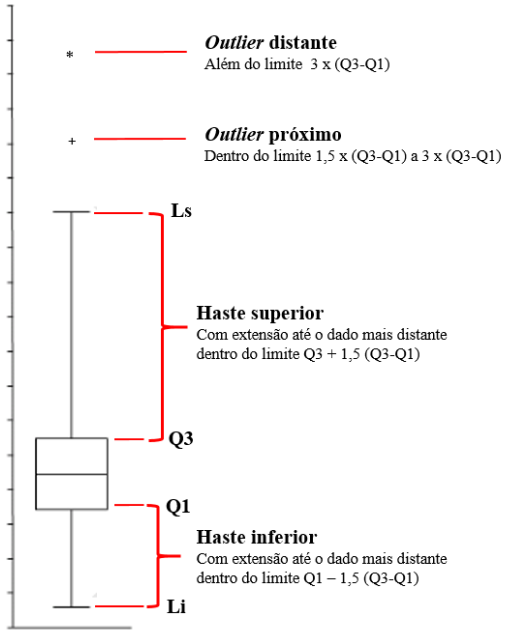

Figura 5 - *Boxplot* com *outliers* próximo e distante.

***Construção do* Boxplot**

A construção gráfica leva em consideração o tipo cartesiano X e Y, onde o eixo X representa a variável de interesse e o eixo Y a mensuração quantitativa (escore) da variável sob análise. Os passos de construção do *boxplot* seguem o método proposto por Tukey, seu idealizador. São eles:

a. Calcular o intervalo interquartílico (a diferença entre o percentil 25º e 75º, ou seja, Q1 e Q3). Denominar de IIQ;
b. Adicionar ao percentil 75º o valor do intervalo interquartílico multiplicado por 1,5. Caso esse valor seja igual ou maior do que o maior dado coletado, desenhar a haste superior equivalente ao maior dado. Caso contrário, interromper a haste superior na altura correspondente ao maior dado abaixo do percentil 75º somado ao valor do IIQ multiplicado por 1,5. Plotar os valores superiores a esse limite como pontos individuais (*outliers* e/ou extremos);
c. Calcular o percentil 25º menos o IIQ. Caso esse valor seja inferior ao menor dado coletado, desenhar a haste inferior coincidente ao menor valor. Caso contrário, interromper a haste inferior na altura correspondente ao menor dado superior ao percentil 25º diminuído ao valor do IIQ multiplicado por 1,5. Plotar os valores inferiores a esse limite como pontos individuais (*outliers* e/ou extremos) (Figura 6).

Boxplot *e suas variações*

O *boxplot* tradicional tem sido modificado visando à inserção de outras propriedades, como a média aritmética[5,8,9]. Entre as variantes, encontram-se (Figura 7):

a. *Boxplot* tradicional, descrito por Tukey;
b. *Boxplot* com largura variável: o tamanho da amostra define o tamanho da largura da caixa;
c. *Notched boxplot*: a fenda ou *notche* enfatiza o tamanho da mediana; e pode incluir o intervalo de confiança (Figura 8);
d. *Violin plot*: o perímetro (em amarelo) exibe a "densidade de probabilidade" dos dados/grupo;
e. *Vase plot*: o perímetro (em amarelo) exibe a presença de dados unimodais ou bimodais;
f. *Beam plot*: as linhas negras representam cada observação individual, e a sua espessura ou largura indicam a presença de dados duplicados. A linha grande exibe a média aritmética. A presença de anomalias nos dados, como distribuição bimodal e medidas duplicadas, são detectadas por esse recurso[1].

***Aplicações do* Boxplot**

Estudos recentes têm exposto a tendência em diversificar a apresentação dos resultados de uma pesquisa[10,11]. Neste contexto, o *boxplot* é um gráfico útil para sumarizar e analisar dados quantitativos, especialmente contínuos[4]. Pode ser utilizado tanto para a análise descritiva como inferencial de dados, independente do tipo de delineamento[11-14]. Como há diferentes possibilidades de inserção dos valores de tendência central (mediana, média) e dispersão (quartil, desvio-padrão, intervalo de confiança), faz-se necessário identificar no corpo do texto ou na legenda da figura qual é a grandeza que está sendo analisada graficamente[14]. Como os demais gráficos, o *boxplot* compõe um recurso para sumarizar tendências e substituir tabelas em casos específicos, especialmente quando os valores de dispersão são mais importantes que os de tendência central[12]. Além disso, ele pode ser construído para dar destaque ao resultado do desfecho primário[13].

Análise exploratória dos dados (estatística descritiva)

A análise exploratória de dados utiliza técnicas estatísticas univariadas para identificar padrões ou tendências que podem estar ocultos em dados agrupados[1]. Essa análise preliminar favorece a avaliação da qualidade dos dados coletados[12].

Gráficos podem transmitir informações múltiplas de forma concisa, e podem ser mais eficazes para a comunicação do que os resultados tabulados. Além disso, os gráficos podem efetivamente reunir os dados e estatísticas apropriadas derivadas deles em uma única exibição, bem como sumarizar as principais características dos dados. Os gráficos podem ser utilizados num cenário exploratório para auxiliar na identificação da qualidade das informações preliminares[12].

A exploração inicial dos dados pode se iniciar pelo *boxplot*, pois é um desses recursos usados para sintetizar visualmente os dados amostrais, e exibir os valores de tendência central (mediana), dispersão (quartis, limites e valores extremos) e verificação da distribuição (simetria) entre as hastes.

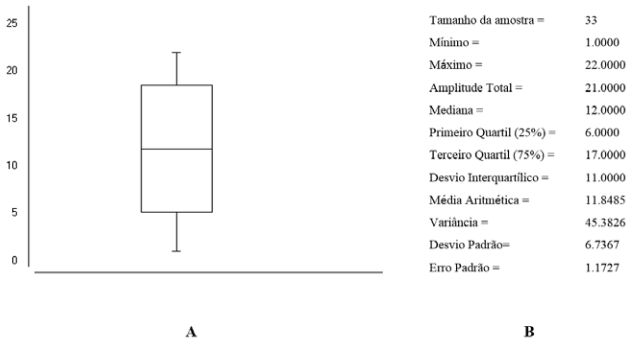

Figura 6 - Construção do boxplot (A) a partir de dados descritivos (B).

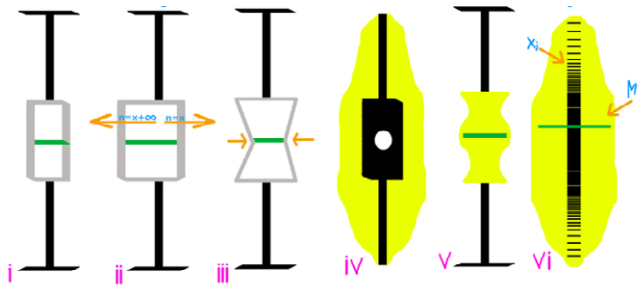

Figura 7 - *Boxplot* tradicional e suas variações: i- tradicional, ii- com largura variável, iii- fenda (*notched*), iv- violino, v- vaso, vi- *beam* (Fonte: https://estatistics.eu/what-is-statistics-charts-boxplot/).

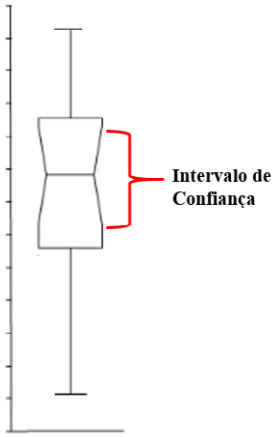

Figura 8 - *Boxplot* tipo fenda (*notched*) exibindo os limites do intervalor de confiança.

Detecção de Outliers e valores extremos

Outliers e valores extremos são valores individuais atípicos ou aberrantes os quais podem ser causados por erros de coleta de dados, incluindo erros de introdução de dados, ou por variações biológicas extremas[15]. Caso a mensuração esteja correta, representa um evento raro. É importante que os *outliers* e valores extremos sejam identificados, pois podem influenciar a análise de dados e conduzir a distorções e conclusões inválidas[7].

Vários métodos são destinados a detectá-los, entre eles o de Tukey[3] (1997) por meio do *boxplot*. Esse método não coincide necessariamente com outros métodos (como por exemplo, o método do desvio padrão e do escore-Z)[15] ou testes estatísticos específicos para identificar *outliers*, como o teste de Grubbs[16].

Comparação entre grupos amostrais

O *boxplot* é também utilizado para comparar a equivalência entre grupos amostrais, tanto para estudo transversal[13] como longitudinal[11,12]. Nestes casos, o gráfico é composto por duas ou mais caixas conforme o número de grupos a serem comparados. Presta-se a fazer comparações diretas de características ou tratamentos entre grupos amostrais não-pareados[11,13] e comparar o efeito de tratamento em estudos pareados (antes e depois)[12] (Figura 9).

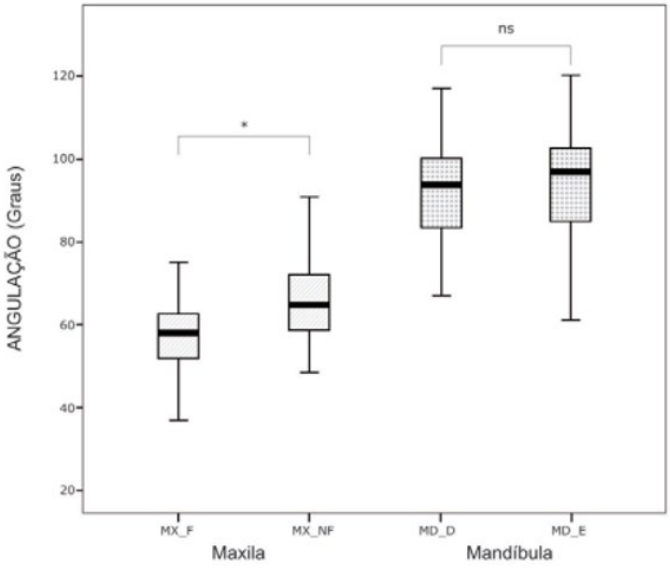

Figura 9 - *Boxplot* podem identificar diferenças entre grupos. Comparação da angulação radicular mesiodistal entre os caninos superiores no lado fissurado (Mx_F) e não-fissurado (Mx_NF); e entre os caninos inferiores no lado direito (MD_D) e esquerdo (MD_E). * = significante (P< 0,05); ns = não significante (P> 0,05) (Fonte: JESUINO[13] *et al*., 2010).

CONCLUSÃO

O *boxplot*, como outros métodos estatísticos visuais, compõe um recurso específico para detectar tendências e substituir tabelas em casos específicos. Quando bem indicado, contribui para melhorar a interpretação de dados, detectar *outliers* e comparar grupos amostrais.

REFERÊNCIAS

01. Kampstra P. Beanplot: A boxplot alternative for visual comparison of distributions. J Statistical Software. 2008; 28: 1-9.
02. Tukey JW. *Exploratory Data Analysis*. Addison-Wesley, preliminary edition, 1970.
03. Tukey JW. Box-and-Whisker Plots. In: *Exploratory Data Analysis.* Reading, MA: Addison-Wesley; 1977. p. 39-43.
04. Pandis N. Statistics for orthodontists. Am J Orthod Dentofacial Orthop. 2015; 147(3): 405-8.
05. Wickham H, Stryjewski L. 40 years of boxplots. Am Statistician, preprint, 2011.
06. Motta VT, Oliveira Filho, PF. SPSS: análise de dados biomédicos. Rio de Janeiro: Medbook; 2009. 334 p.
07. Schwertman NC, Owens MA, Adnan R. A simple more general boxplot method for identifying outliers. Computational Statistics & Data Analysis. 2004; 47(1): 165-174.
08. McGill TJW, Larsen RW. Variations of box plots. The American Statistician, 32: 12-16, 1978.
09. Benjamini Y. Opening the box of a boxplot. Amer Stat. 1998; 42: 257-62.
10. Bailey E, Nelson G, Miller AJ, Andrews L, Johnson E. Predicting tooth-size discrepancy: A new formula utilizing revised landmarks and 3-dimensional laser scanning technology. Am J Orthod Dentofacial Orthop. 2013; 143(4): 574-85
11. LeCornu M, Cevidanes LH, Zhu H, Wu CD, Larson B, Nguyen T. Three-dimensional treatment outcomes in Class II patients treated with the Herbst appliance: a pilot study. Am J Orthod Dentofacial Orthop. 2013; 144(6): 818-30.
12. Amit O, Heiberger RM, Lane PW. Graphical approaches to the analysis of safety data from clinical trials. Pharm Stat. 2008;7(1): 20-35.
13. Jesuino FAZ, Andrade LO, Valladares Neto. Angulação radicular mesiodistal de caninos permanentes em crianças com fissura unilateral completa de lábio e palato. ROBRAC; 19(51): 306-9, 2010.
14. Ponder SN, Benavides E, Kapila S, Hatch NE. Quantification of external root resorption by low- vs high-resolution cone-beam computed tomography and periapical radiography: A volumetric and linear analysis. Am J Orthod Dentofacial Orthop. 2013; 143(1): 77-91.
15. Seo S. A review and comparison of methods for detecting outliers in univarate data sets [Dissertação]. University of Pittsburgh; Pittsburgh, PA; 2006.
16. Grubbs, F.Procedures for detecting outlying observations in samples. Technometrics. 1969; 11(1): 1-21.

ABSTRACT

Introduction: The boxplot is a graphics feature used regularly in this scientific research in various statistical software. Objective: To describe didactically the structure, interpretation, modifications and applications of this graphic feature. Results: The traditional boxplot displays the central trend data nonparametric (median), dispersion (quartiles), the form of distribution or symmetry (point maximum and minimum values) of the sample, and outlier. The boxplot has been modified with a view to inclusion of other properties such as the arithmetic mean and confidence interval. Applications can be compiled for exploratory data analysis, outlier detection, comparison between groups (equivalence) and multivariate data analysis. Conclusion: The boxplot is a graph feature that can replace the use of tables in specific cases.

Keywords: Scientific methodology; Descriptive statistics; Graphics; Quartile; Median; Boxplot.

AUTOR PARA CORRESPONDÊNCIA
Prof. Dr. José Valladares Neto
Faculdade de Odontologia, Universidade Federal de Goiás.
Avenida Universitária esquina com 1ª Avenida, s/ número.
CEP: 74605-220, Goiânia, Goiás, Brasil.
Telefone de contato: +55 62 98231-6000
E-mail: jvalladares@uol.com.br