Sobre o Boxplot no GeoGebra

Boxplot in GeoGebra

PÉRICLES CÉSAR DE ARAUJO[1]

CELINA APARECIDA ALMEIDA PEREIRA ABAR [2]

Resumo

O objetivo deste trabalho é apresentar o resultado do uso da ferramenta para construção do gráfico boxplot no ambiente dinâmico do GeoGebra. O boxplot é um gráfico de um conjunto de dados que consiste de uma linha que se estende do valor mínimo ao valor máximo, em uma caixa com linhas verticais, traçadas no primeiro quartil (Q1), na mediana e no terceiro quartil (Q3). Os quartis, isto é, primeiro quartil, a mediana e o terceiro quartil são três valores que dividem os dados ordenados em quatro grupos com aproximadamente 25% dos valores em cada grupo. Na Estatística Descritiva ou na análise exploratória e comparação de dados, o boxplot é um gráfico configurado para poder identificar os outlies (valores discrepantes), valores que são bastante incomuns, no sentido de estarem muito afastados da maioria dos dados. As modificações do boxplot ocorrem nos valores mínimos e máximos que são substituídos pelos valores abaixo do primeiro quartil por uma quantidade que pode ser maior do que o mínimo [Q1-1,5(Q3-Q1)] e uma quantidade que pode ser menor que máximo [Q3+1,5(Q3-Q1)], respectivamente. Por exemplo, a identificação dos valores outlies é importante no cálculo da média aritmética que tem como característica a influência dos valores extremos. Até o momento a ferramenta para construção do gráfico boxplot no ambiente dinâmico do GeoGebra não evidencia os outlies dos dados podendo comprometer, em princípio, o uso do GeoGebra na Estatística Descritiva e na análise exploratória e comparação de dados. Portanto, consideramos importante que nas versões futuras do GeoGebra seja incluída na opção dessa ferramenta uma modificação que permita identificar os outlies.

Palavras chave: *Boxplot, GeoGebra, Estatística Descritiva, Análise exploratória de dados.*

Introdução

O objetivo deste trabalho é apresentar o resultado do uso da ferramenta para construção do gráfico boxplot no ambiente dinâmico do GeoGebra. O GeoGebra é um software de matemática dinâmica gratuito e multiplataforma para todos os níveis de ensino, que combina geometria, álgebra, tabelas, gráficos, estatística e cálculo numa única

[1] Docente da Universidade Estadual de Feira de Santana e doutorando em Educação Matemática da Pontifícia Universidade Católica de São Paulo - e-mail: pericles@uefs.br

[2] Docente do Programa de Estudos Pós-Graduados em Educação Matemática da Pontifícia Universidade Católica de São Paulo - e-mail: abarcaap@gmail.com

aplicação. Tem recebido vários prêmios na Europa e EUA. Observamos que a Estatística, presente no GeoGebra, é mais uma ferramenta para ensino de Matemática. Dessa forma, verificamos que a ferramenta do GeoGebra para construção do gráfico boxplot, um gráfico estatístico de analise exploratória de dados, é apresentado em um formato que pode induzir o estudante a um erro, isto é, o gráfico Boxplot obtido no GeoGebra não destaca os valores discrepantes (*outlier*).

Para identificar o erro de construção do Boxplot no GeoGebra, vamos comparar com Boxplot obtido por meio do R, um programa estatístico no qual são destacados os valores discrepantes (*outlier*).

O R é uma linguagem e ambiente para computação estatística e gráfica. É um projeto GNU - GENERAL PUBLIC LICENSE[3] - que é similar à linguagem e ambiente S-PLUS, que foi desenvolvido no Bell Laboratories (anteriormente AT & T, agora Lucent Technologies) por John Chambers e colegas. O R pode ser considerado como uma implementação diferente de S-PLUS. S-PLUS é um pacote de software comercial de análise estatística e gráfica produzido pela empresa TIBCO[4]. Há algumas diferenças importantes, mas muitos códigos escritos para S-PLUS, funcionam inalterados em R.

O Boxplot

O Boxplot é um gráfico de um conjunto de dados que consiste de uma linha que se estende do valor mínimo ao valor máximo, em uma caixa com linhas verticais, traçadas no primeiro quartil (Q1), na mediana e no terceiro quartil (Q3). Os quartis, isto é, primeiro quartil, a mediana e o terceiro quartil são três valores que dividem os dados ordenados em quatro grupos com aproximadamente 25% dos valores em cada grupo. Na Estatística Descritiva ou na análise exploratória e comparação de dados, o boxplot é um gráfico configurado para poder identificar os *outlies* (valores discrepantes), valores que são bastante incomuns, no sentido de estarem muito afastados da maioria dos dados. Como apresentado por Silva (2011):

[3] http://www.gnu.org/
[4] http://www.tibco.com/

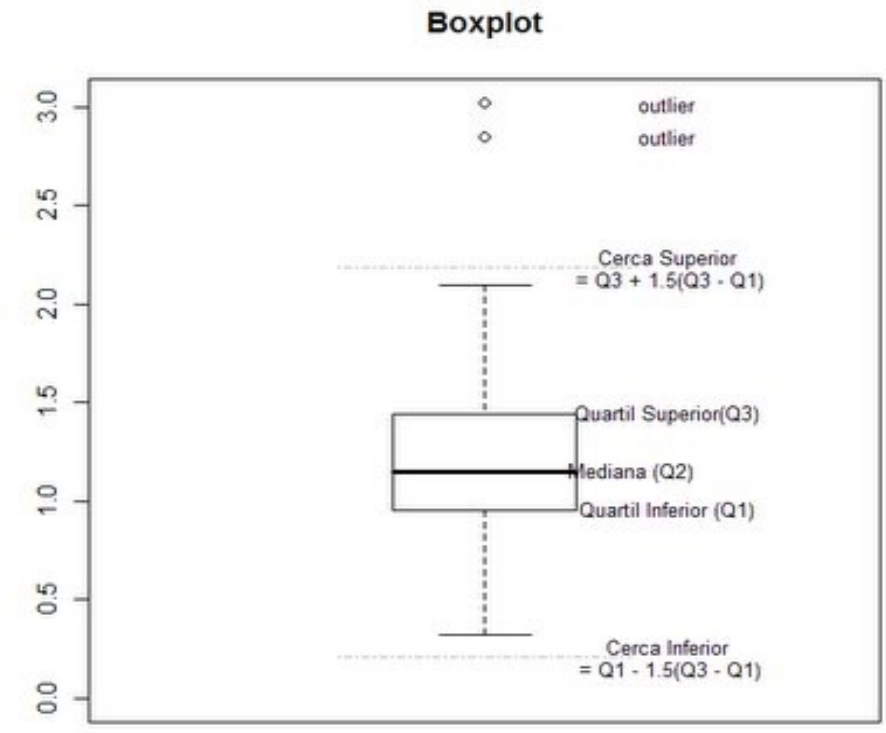

FIGURA 1: Exemplo do uso do BoxPlot
FONTE: Silva (2011)

As modificações do boxplot ocorrem nos valores mínimos e máximos que são substituídos pelos valores abaixo do primeiro quartil por uma quantidade que pode ser maior do que o mínimo [Q1-1,5(Q3-Q1)] e uma quantidade que pode ser menor que máximo [Q3+1,5(Q3-Q1)], respectivamente.

Exemplo explorado

Para comparar o Boxplot obtido no GeoGebra com o Boxplot obtido com R, vamos explorar o seguinte exemplo: Coletaram-se os pesos, em kg, de 40 alunos – 20 rapazes e 20 moças – obtendo-se os dados abaixo. Trace um boxplot para cada sexo. (OLIVEIRA, 2010, p.142)

Primeiro vamos construir o Boxplot no GeoGebra:

BoxPlot[0, 1 {40,49,55,70,40,50,57,75,43,50,60,83,45,52,65,92,47,55,67,105**}] para obter o Box Plot 1 e tecle "Enter"**

BoxPlot[4, 1

{32,40,47,57,33,40,48,58,35,42,50,60,36,43,52,63,38,45,53,65**}] para obter o Box Plot 2 e tecle "Enter"**

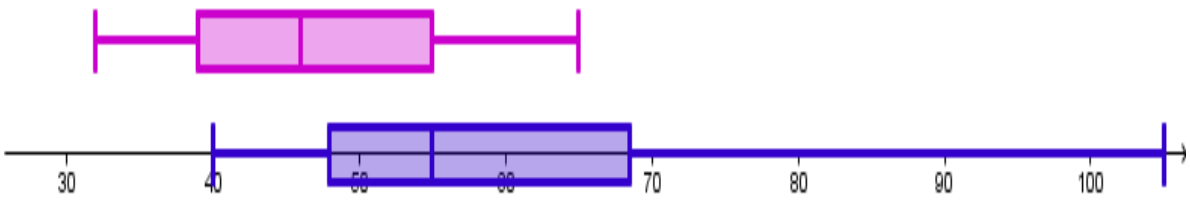

FIGURA 2: Exemplo do uso do BoxPlot no GeoGebra

Agora vamos construir o Boxplot no R:

```
Rapazes=c(40,49,55,70,40,50,57,75,43,50,60,83,45,52,65,92,47,55,67,105)
summary(Rapazes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.00   48.50   55.00   60.00   67.75  105.00
sort(Rapazes)
 [1]  40  40  43  45  47  49  50  50  52  55  55  57  60  65  67  70  75  83  92
[20] 105
```

O valores Min., 1stQu., Median, Mean, 3rdQu., Max. são respectivamente: valor mínimo, primeiro quartil, mediana, média, terceiro quartil e valor máximo.

```
Moças=c(32,40,47,57,33,40,48,58,35,42,50,60,36,43,52,63,38,45,53,65)
summary(Moças)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   39.50   46.00   46.85   54.00   65.00
sort(Moças)
 [1] 32 33 35 36 38 40 40 42 43 45 47 48 50 52 53 57 58 60 63 65
```

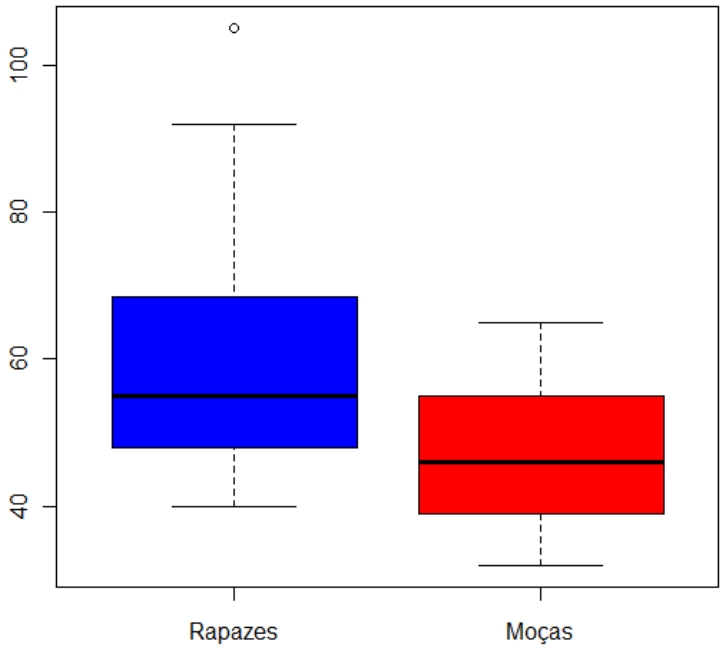

FIGURA 3: Exemplo do uso do BoxPlot no R.

Portanto, como o R é um programa estatístico e gráfico, o Boxplot dos dados *Rapazes* apresenta um valor discrepante (*outlies)* e não existe nos dados das *Moças*. No GeoGebra as duas séries de dados não apresentam valores discrepantes, dessa maneira, poderia levar o aprendiz ao erro, porque os gráficos Boxplot construídos para os dados das *Moças* nos dois programas, não apresentam valores discrepantes, são semelhantes.

Valores discrepantes (*outlies)*

Valores discrepantes (*outlies)* são valores que se localizam muito afastados de quase todos os demais valores. A identificação dos valores discrepantes (*outlies)* é importante no calculo da média aritmética, que tem como característica a influência dos valores extremos. Os valores discrepantes (*outlies)* podem ter efeito sobre o desvio padrão, sobre a escala do histograma e da forma da distribuição de freqüência dos dados (TRIOLA, 2008). O efeito dos valores discrepantes (*outlies)* pode ser observado na sequência de gráficos gerados do programa R:

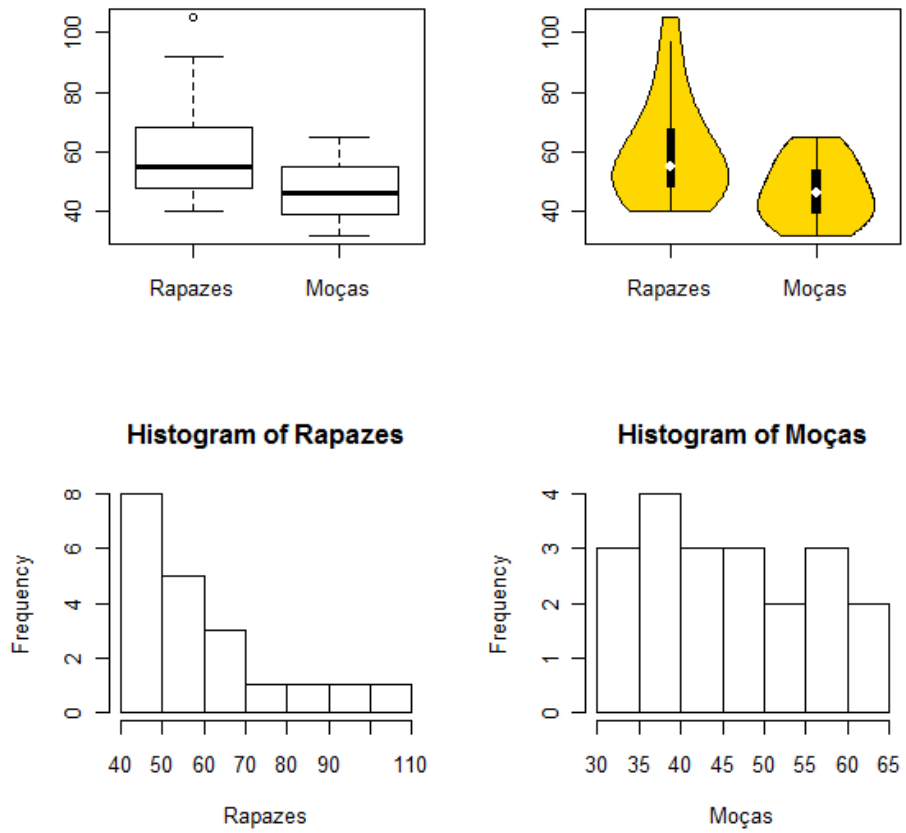

FIGURA 3: Outro exemplo do uso do BoxPlot no R.

Vamos analisar os gráficos acima no sentido horário. O primeiro é do tipo boxplot onde identificamos um valor discrepante, já comentado, o segundo gráfico é um violplot, gráfico em forma de violino no qual observamos o formato da distribuição de freqüência e os outros dois são histogramas. O violplot e o histograma dos dados dos *Rapazes* apresentam uma cauda mais longa porque a distância entre seus extremos é maior do que os dados das *Moças* e também por conta de um peso de 105 kg um valor discrepante. Assim, podemos verificar o efeito de um valor discrepante na dispersão, isto é, os pesos dos Rapazes têm uma distribuição com uma alta dispersão. Observamos que a distribuição dos dados dos *Rapazes* é assimétrica positiva com a mediana menor que a média:

summary(Rapazes)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40.00	48.50	55.00	60.00	67.75	105.00

Enquanto que os pesos das *Moças* têm uma distribuição aproximadamente simétrica, a mediana é igual a média ou aproximadamente:

summary(Moças)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
32.00	39.50	46.00	46.85	54.00	65.00

Os resultados das análises acima apresentadas podem ser resumidos por meio do gráfico qqnorm do R. O gráfico compara os quartis e percentis de uma distribuição teórica Normal com os quartis e percentis dos dados observados. Para o gráfico qqnorm do R, podemos definir essa função por meio de uma adaptação de Triola (2008, p.238) :

> Um gráfico dos quantis normais (ou gráfico de probabilidades normal) é um gráfico de pontos (x,y) onde cada valor y vem do conjunto de dados amostrais e cada valor x é o escore de z correspondente ao valor esperado do quantil da distribuição normal padrão.

Então, para os conjuntos dos dados representados pelos pesos em Kg de *Rapazes* e *Moças* temos os seguintes gráficos dos quantis:

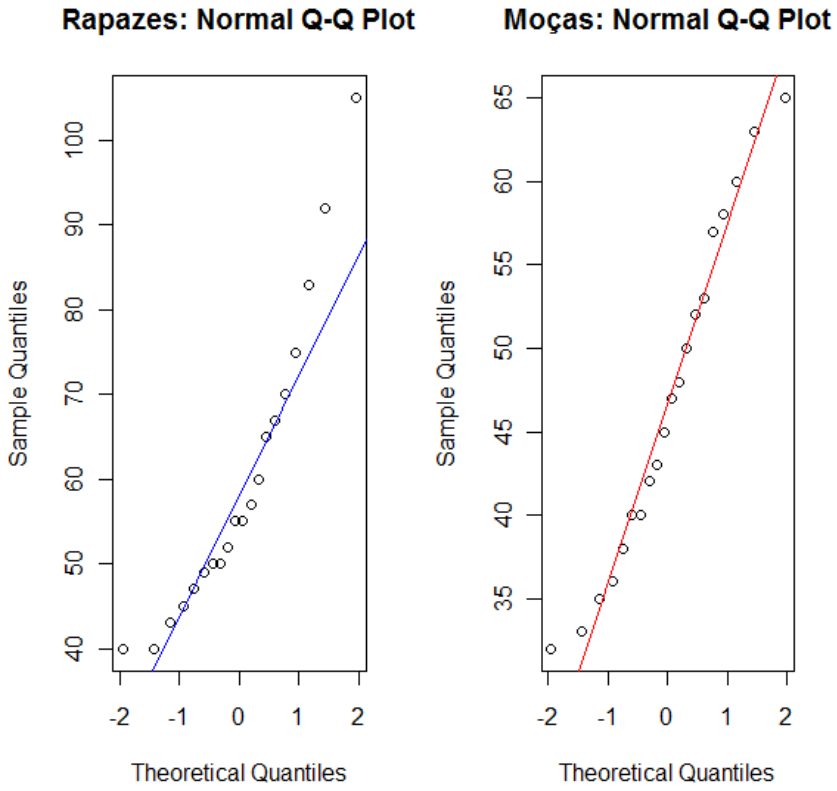

FIGURA 4: Gráficos dos quantis (Triola, 2008, p. 238)

Para fazer a interpretação do gráfico dos quantis normais do R, temos que verificar se os pontos estão razoavelmente próximo de uma reta, uma identidade, não ter um padrão sistemático que não seja de uma reta e se não houver valores discrepantes (*outlies),* como afirma Triola (2008). Alisando os gráficos dos quantis normais dos dados do exemplo em estudo e apresentados pelo programa R, verificamos que os dados dos pesos das Moças por apresentarem uma simetria, regularidade e não há valores discrepantes (*outlies),* estão próximos de uma distribuição Normal, enquanto os dados dos pesos dos *Rapazes* por apresentarem assimetria e valores discrepantes não estão próximos duma distribuição. Essas informações são relevantes no âmbito da Inferência Estatística Clássica e principalmente nos testes de hipótese paramétricos onde a normalidade dos dados é um requisito.

Origem dos valores discrepantes (*outlies)*

Valores discrepantes (*outlies)* podem ter origem em observações, leituras incorretas, ou podem ser valores reais fruto de um país desigual como o Brasil onde há uma grande concentração de recurso e população em algumas capitais, como é apresentado no gráfico abaixo por LANDIM (2011):

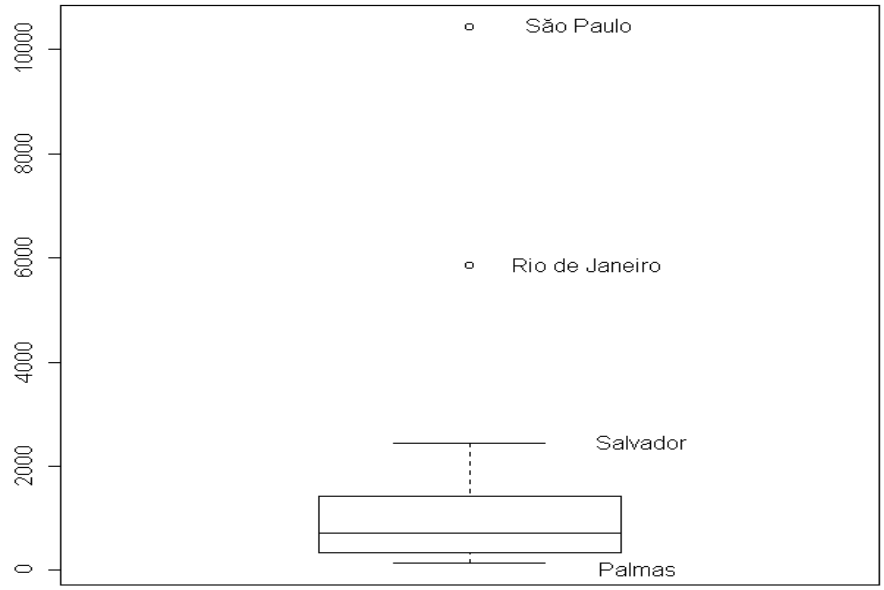

FIGURA 5: Gráfico exemplo de valores discrepantes (Landim, 2011)

Considerações finais

Exploramos exemplos por meio de um programa estatístico para mostrar a importância dos valores discrepantes (*outlies)*. O programa R é mais utilizado no âmbito da pesquisa quantitativa e o GeoGebra é um programa para o ensino de geometria, álgebra e noções de estatística. No entanto é conveniente observar alguns erros do tipo que observamos na ferramenta para construção do gráfico boxplot no GeoGebra, para não comprometer a formação dos estudantes que estão tendo o primeiro contato com uma ferramenta tão importante.

Com ficou evidenciado nesse trabalho, a ferramenta para construção do gráfico boxplot no ambiente dinâmico do GeoGebra não evidencia os *outlies* dos dados comprometendo, desta forma, o uso do GeoGebra na Estatística Descritiva e na análise exploratória e comparação de dados. Portanto, consideramos importante que nas versões futuras do GeoGebra sejam incluídas, na opção da ferramenta para construção do gráfico boxplot, alguma modificação que permita identificar os outlies.

Referências

LANDIM, Flávia M. P. F. **Análise Exploratória de Dados –** 2006 Disponível em *www.dme.ufrj.br/marina/mad114r6.ppt* - acesso em 11/11/2011.

OLIVEIRA, João U.C**. Estatística- Uma nova abordagem**. Rio de Janeiro, Editora Ciência Moderna LTDA, 2010.

R DEVELOPMENT CORE TEAMT. R**: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. 2011.

SILVA, Marcos F**, Aplicação de Métodos Quantitativos em Auditoria: Uso do R em Análise de Dados aplicada à Auditoria.** Disponível em http://sites.google.com/site/marcosfs2006/ acesso em 11/11/2011.

TRIOLA, Mario F. **Introdução à Estatística**, 10ª ed.-Rio de Janeiro, LTC, 2008.