UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE MATEMÁTICA
CONTEÚDOS DIGITAIS EM MATEMÁTICA E ESTATÍSTICA

O BOXPLOT

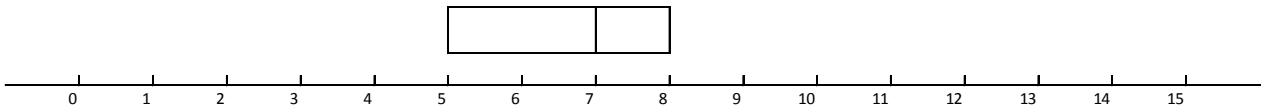
Ana Maria Lima de Farias
Departamento de Estatística (GET/UFF)

Introdução

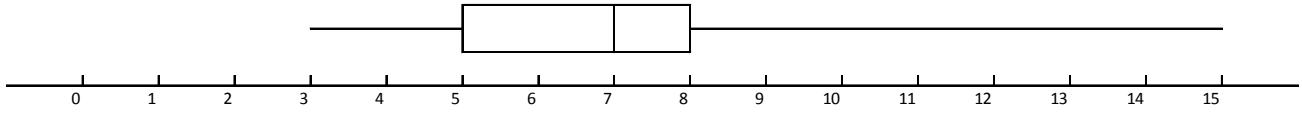
O boxplot é um gráfico construído com base no resumo dos cinco números, constituído por:

- Valor mínimo
- Primeiro quartil (Q1)
- Mediana (segundo quartil Q2)
- Terceiro quartil (Q3)
- Valor máximo

O gráfico é formado por uma caixa construída paralelamente ao eixo da escala dos dados (pode ser horizontal ou vertical). Essa caixa vai desde o primeiro quartil até o terceiro quartil e nela traça-se uma linha na posição da mediana. Essa caixa, que descreve os 50% centrais da distribuição, é comum a todas as variantes do boxplot. Pode-se acrescentar também uma linha, paralela à linha da mediana, para indicar a média. Na figura abaixo, Q1 = 5; Q2 = 7; Q3 = 8.

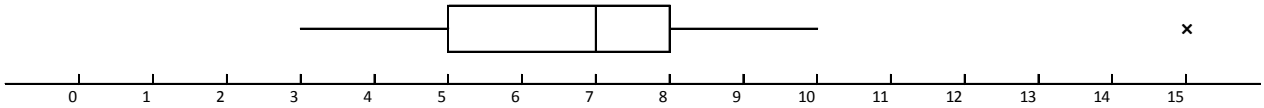


Na variante que usa efetivamente o resumo dos cinco números, continua-se a construção do boxplot traçando-se uma linha paralela à escala que vai de cada extremidade da caixa ao correspondente valor extremo dos dados. É praxe traçarem-se essas linhas pelo meio da caixa. Na figura abaixo, o mínimo é 3 e o máximo é 15.



Outra variante bastante comum, que é a que usamos nas atividades Conhecendo o Boxplot e Construindo Boxplots, trata de forma especial os valores atípicos, que são

valores muito afastados da grande maioria dos dados. Suponhamos que, no nosso exemplo, todas as observações sejam menores que ou iguais a 10, e apenas uma observação seja igual a 15. Então 15 é um valor atípico. A variante do boxplot representaria esses dados da seguinte forma:



Há diferentes opções para se estabelecerem os limites que separam os valores atípicos. Mas estabelecida uma regra, os valores que se encontram entre esses limites são chamados *valores adjacentes* e aqueles fora dos limites são chamados *valores extremos ou atípicos*.

Cálculo dos quartis

Dada a escala de mensuração dos dados, os quartis são valores nessa escala que dividem o conjunto de dados em quatro partes, todas elas com o mesmo número de observações. Isso significa que 25% das observações são menores que o primeiro quartil, 50% são menores que o segundo quartil e 75% são menores que o terceiro quartil. Note que estamos falando de *escala*, de *ordem*. Assim, para calcularmos os quartis, temos que *ordenar* os dados.

O cálculo se inicia com a mediana, ou segundo quartil – ela é o "valor do meio", o valor que deixa metade das observações abaixo e a outra metade acima.

Consideremos o conjunto de dados que gerou o boxplot acima; há 18 observações.

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Valor	3,0	3,5	4,5	5,0	5,0	5,5	6,5	6,5	6,5	7,5	7,6	7,9	8,0	8,0	9,0	9,5	10,0	15,0

A mediana divide o conjunto em duas partes, cada uma com 9 observações.

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Valor	3,0	3,5	4,5	5,0	5,0	5,5	6,5	6,5	6,5	7,5	7,6	7,9	8,0	8,0	9,0	9,5	10,0	15,0

A mediana será, então, a média dos dois valores centrais:

$$Q2 = \frac{6,5 + 7,5}{2} = 7,0$$

Um erro comumente cometido pelos alunos é considerarem a média das posições, e não dos valores! Se assim fosse, todos os conjuntos de dados com 18 observações teriam a mesma mediana 9,5...

O cálculo do primeiro e do terceiro quartis é feito calculando-se as medianas das duas metades – o primeiro quartil é a mediana da metade inferior e o terceiro quartil é a mediana da metade superior. Nesses cálculos despreza-se a mediana.
Para os dados acima, cada metade tem 9 observações. Logo, a mediana deixará 4 observações abaixo e 4 observações acima, ou seja, a mediana de cada uma dessas partes será a quinta observação:

$$Q1 = 5,0$$

$$Q3 = 8,0$$

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Valor	3,0	3,5	4,5	5,0	5,0	5,5	6,5	6,5	6,5	7,5	7,6	7,9	8,0	8,0	9,0	9,5	10,0	15,0

Consideremos outro conjunto de dados como exemplo:

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Valor	15	17	18	19	19	20	25,0	26	26	28	30	32	42

Como há um número ímpar de observações (13), a mediana deixará 6 observações abaixo e 6 observações acima dela. Logo, a mediana é a 7ª observação.

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Valor	15	17	18	19	19	20	25,0	26	26	28	30	32	42

$$Q2 = 25,0$$

Ignorando a mediana, cada metade tem 6 observações e a mediana de cada uma delas será a média das terceira e quarta observações:

$$Q1 = \frac{18+19}{2} = 18,5$$

$$Q3 = \frac{28+30}{2} = 29$$

Determinação de Valores Atípicos

A regra que adotamos para identificação dos valores atípicos se baseia na *amplitude interquartil* AIQ, definida como a distância entre o primeiro e o terceiro quartis:

$$AIQ = Q3 - Q1$$

Note que AIQ é o comprimento da caixa. Quaisquer valores abaixo de Q1 ou acima de Q3 por mais de 1,5×AIQ serão considerados valores atípicos e terão tratamento especial no boxplot. Assim, serão valores atípicos os valores x tais que

$$x < Q1 - 1{,}5 \times AIQ$$

ou

$$x > Q3 + 1{,}5 \times AIQ$$

Os valores que se encontram entre Q1 – 1,5×AIQ e Q3 + 1,5×AIQ são chamados valores adjacentes e sua representação se completa (lembre-se de que já representamos os 50% centrais com a caixa!) traçando uma linha que vai de Q1 até o menor valor adjacente (isto é, o valor mínimo dos dados, excluídos os valores atípicos) e outra que vai de Q3 até o maior valor adjacente (isto é, o valor máximo dos dados, excluídos os valores atípicos).

Possíveis valores atípicos são representados por algum caráter especial.

Para o nosso primeiro exemplo, com 18 observações, obtemos

$$Q1 - 1{,}5 \times AIQ = 5 - 1{,}5 \times (8 - 5) = 0{,}5$$

$$Q3 + 1{,}5 \times AIQ = 8 + 1{,}5 \times (8 - 5) = 12{,}5$$

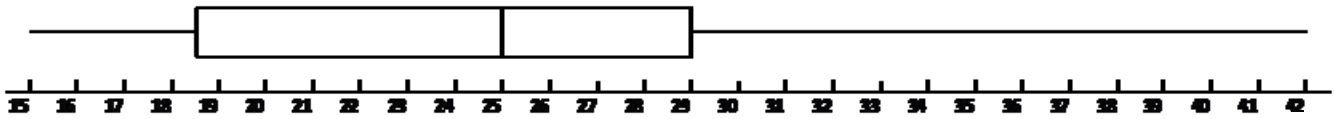
e, portanto, qualquer valor menor que 0,5 ou maior que 12,5 será valor atípico. Resulta que o único valor atípico é 15. Logo, a linha inferior vai de 3 (valor mínimo) até Q1 = 5 e a linha superior vai de Q3 = 8 até 10 (valor máximo, desconsiderando o valor atípico 15).

Para o segundo conjunto com 13 dados, temos

$$Q1 - 1{,}5 \times AIQ = 18{,}5 - 1{,}5 \times (29 - 18{,}5) = 2{,}75$$

$$Q3 + 1{,}5 \times AIQ = 29 + 1{,}5 \times (29 - 18{,}5) = 44{,}75$$

e, portanto, qualquer valor menor que 2,75 ou maior que 44,75 será valor atípico. Resulta que não há valores atípicos. Logo, a linha inferior vai de 15 (valor mínimo) até Q1 = 18,5 e a linha superior vai de Q3 = 29 até 42 (valor máximo).



A atividade “Conhecendo o Boxplot” (http://www.uff.br/cdme/conheceboxplot) ilustra a construção de um boxplot e também apresenta um software interativo que permite explorar suas principais características. Na atividade “Construindo Boxplots” (http://www.uff.br/cdme/constroiboxplot) é fornecido um software para construção de boxplots.

Exemplo: População urbana vs população rural

Na Tabela 1 e na Figura 1 temos os dados sobre a população residente por unidade da federação e pela situação do domicílio (urbano ou rural).

- Posição relativa das caixas – a caixa da população urbana está acima da caixa da população rural, uma vez que a população urbana é, em geral, maior que a rural.
- Dispersão – a caixa da população urbana é mais comprida, assim como as linhas, o que caracteriza maior dispersão]
- Valores atípicos – identifique, a partir da tabela, os estados que são atípicos em termos de população urbana e rural

Exemplo: Notas de 2 Turmas de Estatística Econômica

Na Tabela 2 e na Figura 2 temos as notas de alunos de 2 turmas de Introdução à Estatística Econômica.

- Turma da tarde – menor dispersão, mas notas mais baixas.
- Caixas – 50% centrais da turma da tarde estão com notas entre 40 e 60, enquanto na turma da noite, as notas vão de 45 a 70.

TABELA 1

População Residente (em 1000 hab.)

Grandes Regiões	Situação do domicílio	
e Unidades da Federação	Urbana	Rural
Brasil	137.954	31.845
Região Norte	9.014	3.886
Rondônia	885	495
Acre	370	187
Amazonas	2.107	705
Roraima	247	77
Pará	4.121	2.072
Amapá	425	52
Tocantins	860	297
Região Nordeste	32.975	14.766
Maranhão	3.364	2.287
Piauí	1.789	1.055
Ceará	5.315	2.115
Rio Grande do Norte	2.037	740
Paraíba	2.447	997
Pernambuco	6.058	1.860
Alagoas	1.920	903
Sergipe	1.273	511
Bahia	8.772	4.298
Região Sudeste	65.549	6.863
Minas Gerais	14.672	3.220
Espírito Santo	2.463	634
Rio de Janeiro	13.821	570
São Paulo	34.593	2.440
Região Sul	20.322	4.786
Paraná	7.786	1.777
Santa Catarina	4.218	1.138
Rio Grande do Sul	8.318	1.870
Região Centro-Oeste	10.093	1.544
Mato Grosso do Sul	1.747	331
Mato Grosso	1.988	517
Goiás	4.397	607
Distrito Federal	1.961	90

Fonte: IBGE Censo 2000

FIGURA 1

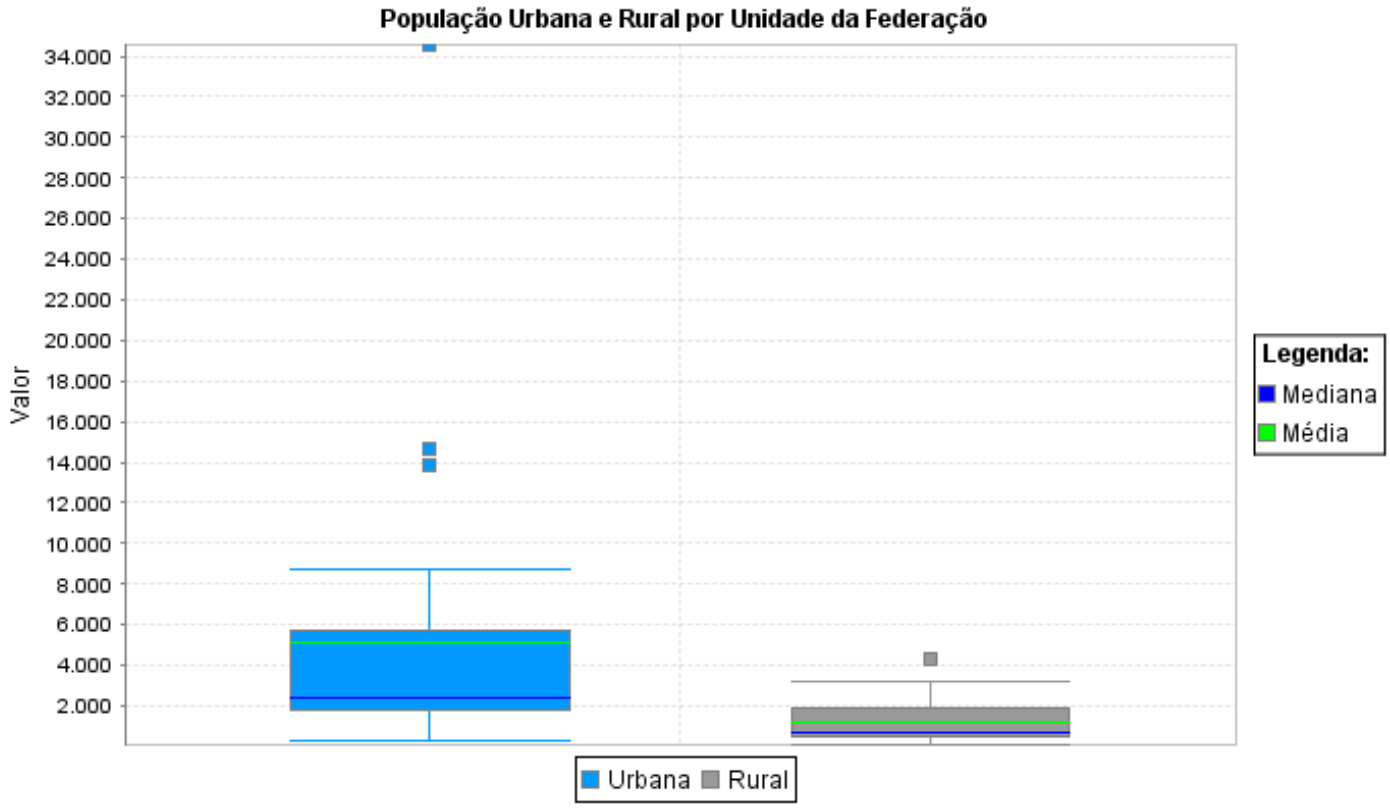
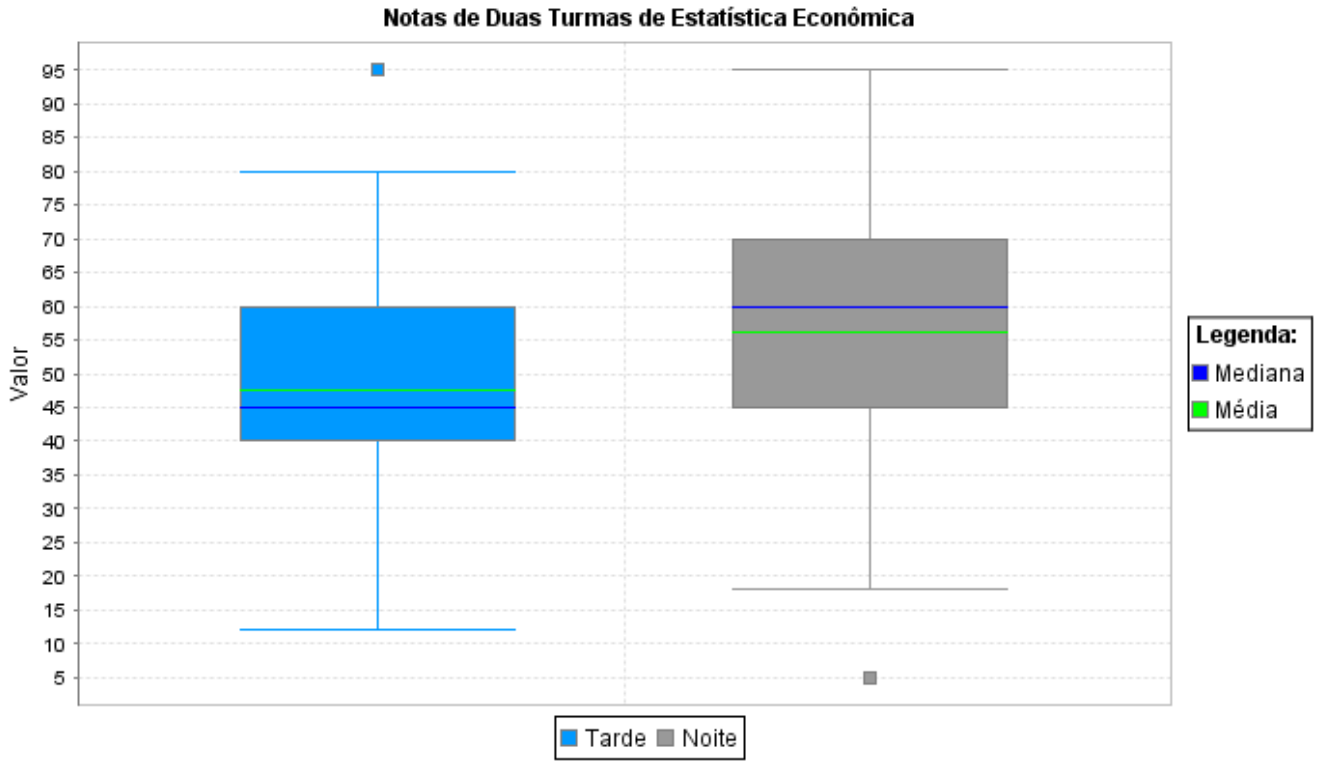

População Urbana e Rural por Unidade da Federação
Valor
34.000
32.000
30.000
28.000
26.000
24.000
22.000
20.000
18.000
16.000
14.000
12.000
10.000
8.000
6.000
4.000
2.000
Urbana
Rural
Legenda:
Mediana
Média

TABELA 2

Turma da Tarde									
12	19	24	25	26	26	26	26	29	30
32	32	33	33	40	40	40	40	40	41
42	42	42	42	43	43	43	43	43	44
45	47	47	48	48	48	50	50	50	52
52	53	53	60	60	60	60	60	60	60
60	61	61	64	68	72	72	72	75	75
80	95								
Turma da Noite									
5	18	18	22	22	23	30	32	40	40
40	41	41	42	42	43	45	45	45	45
47	47	48	48	51	52	52	53	53	60
60	60	60	60	60	60	60	62	62	62
63	63	63	64	65	65	66	68	70	70
70	70	72	72	72	72	74	75	75	80
80	82	83	85	88	95				

FIGURA 2



Bibliografia

Triola, M. F. *Introdução à Estatística*, 10a. edição. Rio de Janeiro: LTC Editora, 2008.

Bussab, W. O. e Morettin, P. A. *Estatística Básica*, 6ª. edição. São Paulo: Editora Saraiva, 2009.

Farias, A. M. L.; Laurencel, L. C. *Estatística Descritiva*, Apostila. Departamento de Estatística. Niterói: UFF 2008 (versão para download em http://www.professores.uff.br/anafarias//estdesc_2008.pdf

Tukey, J. W. *Exploratory Data Analysis*, Addison-Wesley, 1977.