Medidas de tendência central,dispersão, posição, associação, boxplot e proporção

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

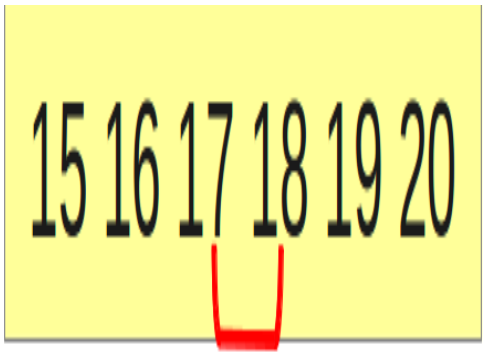
Introdução

- Uma vez entendido qual o comportamento dos dados, como eles estão distribuídos na população, se faz necessário outras medidas que nos digam pontuamente alguma característica da população.

MEDIDAS DE TENDÊNCIA CENTRAL

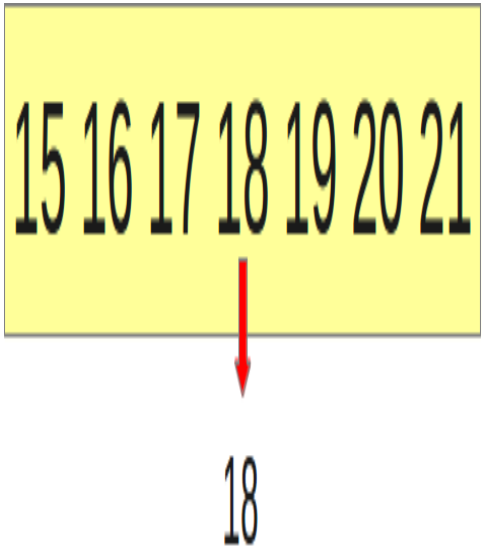
Mediana

- É o valor que divide os dados ao meio, ou seja, 50% dos valores estarão a esquerda e 50% dos valores estarão a direita da mediana.
- Para **n** par, a mediana é calculada como a média dos dois valores centrais.



$$(17 + 18)/2 = 17{,}5$$

- Para **n** ímpar, a mediana é o valor central.



Aplicação

1. De acordo com a revista *Chemical Engineering* uma propriedade importante da fibra é a sua absorção de água. Uma amostra aleatória de 20 pedaços de fibra foi coletada e sua absorvência medida em cada peça. A seguir estão os valores coletados.

Rol de dados

1-5	6-10	11-15	16-20
18.71	22.43	18.92	21.77
21.41	20.17	20.33	22.11
20.72	23.71	23.00	19.77
21.81	19.44	22.85	18.04
19.29	20.50	19.25	21.12

a. Calcule a mediana.

Ordenando os dados tem-se:

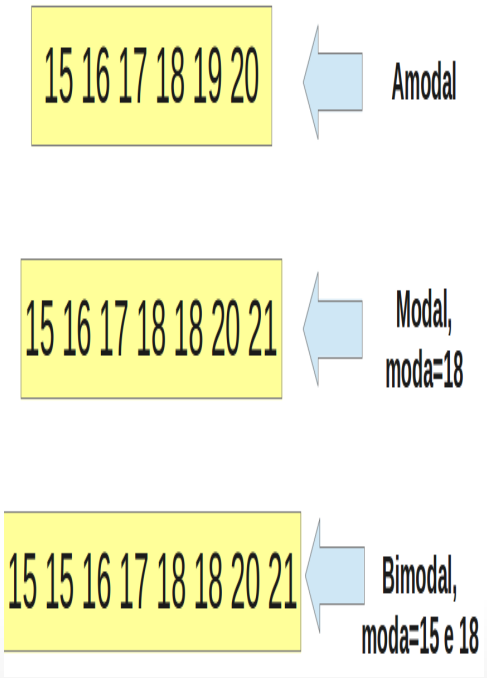
Dados ordenados			
1-5	6-10	11-15	16-20
18.04	19.44	20.72	22.11
18.71	19.77	21.12	22.43
18.92	20.17	21.41	22.85
19.25	20.33	21.77	23.00
19.29	20.50	21.81	23.71

Como os dados são pares, a mediana será a média dos valores que estão na posição 10 e 11. Logo, tem-se:

$$mediana = \frac{20,5 + 20,72}{2} = 20.61$$

Moda

- É o valor que ocorre com maior frequência.
- É possível que os dados não apresentem moda (denominando amodal), apresentem uma moda (modal), duas modas (bimodal) e assim por diante.
- No caso de variáveis contínuas o cálculo da moda só é possível utilizando-se de técnicas para **dados agrupados** por meio da **tabela de distribuição de frequências.**



Aplicação

2. Aproveitando os dados do exemplo 1, calcule a moda.

   Percebam que as variáveis são contínuas. Então, ficará como dever avaliativo o aluno procurar e calcular a moda do referido exemplo. Dica: Procure material sobre **dados agrupados**.

Média

- É a soma de todos os valores da série dividida pela quantidade de elementos na série.
- Se **n** observações são tomadas de uma amostra cujo as observações podem ser denotadas como $x_1, x_2, \cdots, x_n$, então a **média amostral** é:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Se **N** observações são tomadas de uma população finita cujo as observações podem ser denotadas como $x_1, x_2, \cdots, x_n$ a **média populacional** é:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$
$$= \frac{\sum_{i=1}^{N} x_i}{N}$$

Aplicação

3. Aproveitando os dados do exemplo 1, tem-se:

   Pegando os dados brutos, temos:

$$\bar{x} = \frac{18,04 + 18,71 + 18,92 + \cdots + 23 + 23,71}{20} = 20,77$$

Propriedades e características da média

- A soma dos desvios em relação à média é igual a zero para qualquer amostra.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- A soma ou subtração de uma constante (k) aos dados altera a média de tal forma que a nova média fica adicionada ou subtraída pela constante.

$$\frac{\sum_{i=1}^{n}(k \pm x_i)}{n} = k \pm \frac{\sum_{i=1}^{n} x_i}{n}$$

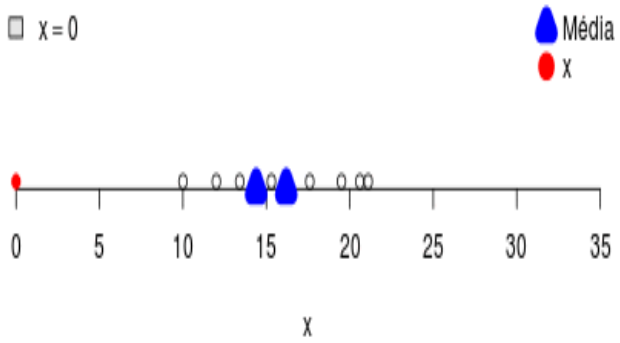
- A multiplicação ou divisão de uma constante (k) aos dados altera a média de tal forma que a nova média fica multiplicada ou dividida pela constante.

$$\frac{\sum_{i=1}^{n}(k \cdot x_i)}{n} = k \cdot \frac{\sum_{i=1}^{n} x_i}{n}$$

ou

$$\frac{\sum_{i=1}^{n} x_i / k}{n} = \frac{\sum_{i=1}^{n} x_i}{n} / k$$

- Embora a média amostral seja uma medida preferida por todos por uma série de propriedades que será vista na introdução a inferência, ela é influenciada por valores extremos, sejam eles baixos ou altos.



MEDIDAS DE DISPERSÃO

Amplitude

- É a medida mais simples de dispersão.
- Uma vez que os dados estejam ordenados de modo crescente, basta subtrair o maior valor do menor valor da série.

$$A = x_n - x_1$$

Aplicação

4. Aproveitando os dados do exemplo 1, tem-se:

$$A = 23.71 - 18.04 = 5.67$$

Variância

- É a medida mais utilizada dentre as medidas de dispersão, pois dentre vários aspectos positivos, está o fato de contemplar todos os valores da amostra.

- A variância é uma distância média de cada observação em relação a média. No entanto, em si tratando de amostras e por motivos que serão vistos mais adiante, esta distância precisa ser elevada ao quadrado e dividida pelo o que nós chamamos de **graus de liberdade** da amostra, de acordo com a seguinte definição:

- Sejam $x_1, x_2, \cdots, x_N$ observações provenientes de uma população, a variância é calculada como:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- No caso de uma amostra:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- No link http://nbcgib.uesc.br/lec/avale-es/amb-virtual/estimadores/var há uma simulação para um melhor entendimento do porquê que o denominador é **n-1** e não **n**.

- Vamos entender um pouquinho mais do porque do $n-1$?

- Vamos imaginar que você tenha 7 camisas diferentes de acordo com a figura abaixo:

- Suponha que você queira usar uma camisa diferente por dia ao longo de uma semana. Quantos dias da semana você poderia escolher **livremente** uma camisa?

- Agora imaginemos um conjunto de dados. Imagine então, que coletamos uma amostra de tamanho 5 cujo a média foi 4. Quantos elementos podemos atribuir **livremente** um valor para que a média seja 4?

Aplicação

5. Aproveitando os dados da aplicação 1, tem-se:

$$S^2 = \frac{(18,04 - 20,77)^2 + (18,71 - 20,77)^2 + \cdots + (23,71 - 20,77)^2}{20 - 1}$$
$$= 2,53$$

- Deve-se ter uma atenção especial. A unidade de medida estará elevada ao quadrado. No referido exemplo, a variância foi de $2,53\%^2$.

Desvio padrão

- O desvio padrão é uma medida utilizada para contornar o inconveniente de unidade de medida apresentada pela variância.
- A unidade de medida do desvio padrão é igual a unidade de medida mensurada na variável.
- Sejam $x_1, x_2, \cdots, x_n$ observações provenientes de uma amostra, o desvio padrão é dado por:

$$s = \sqrt{s^2}$$

- Se as observações forem provenientes de uma população, então:

$$\sigma = \sqrt{\sigma^2}$$

Aplicação

6. Aproveitando os dados da aplicação 1, tem-se:

$$S = \sqrt{S^2} = \sqrt{2,53} = 1,59$$

- Logo, com a unidade de medida na escala linear, tem-se 1,59%.

Coeficiente de variação

- Tanto a variância como o desvio padrão são medidas dependentes da grandeza, escala ou unidade de medida da variável.
- Conjunto de dados com diferentes unidades de medida não podem ter suas dispersões comparadas pela variância ou pelo desvio padrão, e até mesmo dados com uma mesma unidade não podem ser comparados se possuem médias de diferentes magnitudes.
- Logo, um estimador que não seja dependente desses fatores se faz necessário.

- Sejam $x_1, x_2, \cdots, x_n$ observações provenientes de uma amostra, o coeficiente de variação é dado por:

$$cv = \frac{S}{\bar{x}} * 100$$

- No caso de uma população:

$$CV = \frac{\sigma}{\mu} * 100$$

Aplicação

7. Vamos considerar uma amostra proveniente do peso (em kg) de alunos do curso A e outra amostra provenientes de alunos do curso B.

Curso A	Curso B
52	82
55	90
51	88
60	81
54	78

- O desvio padrão de ambas as amostras são:

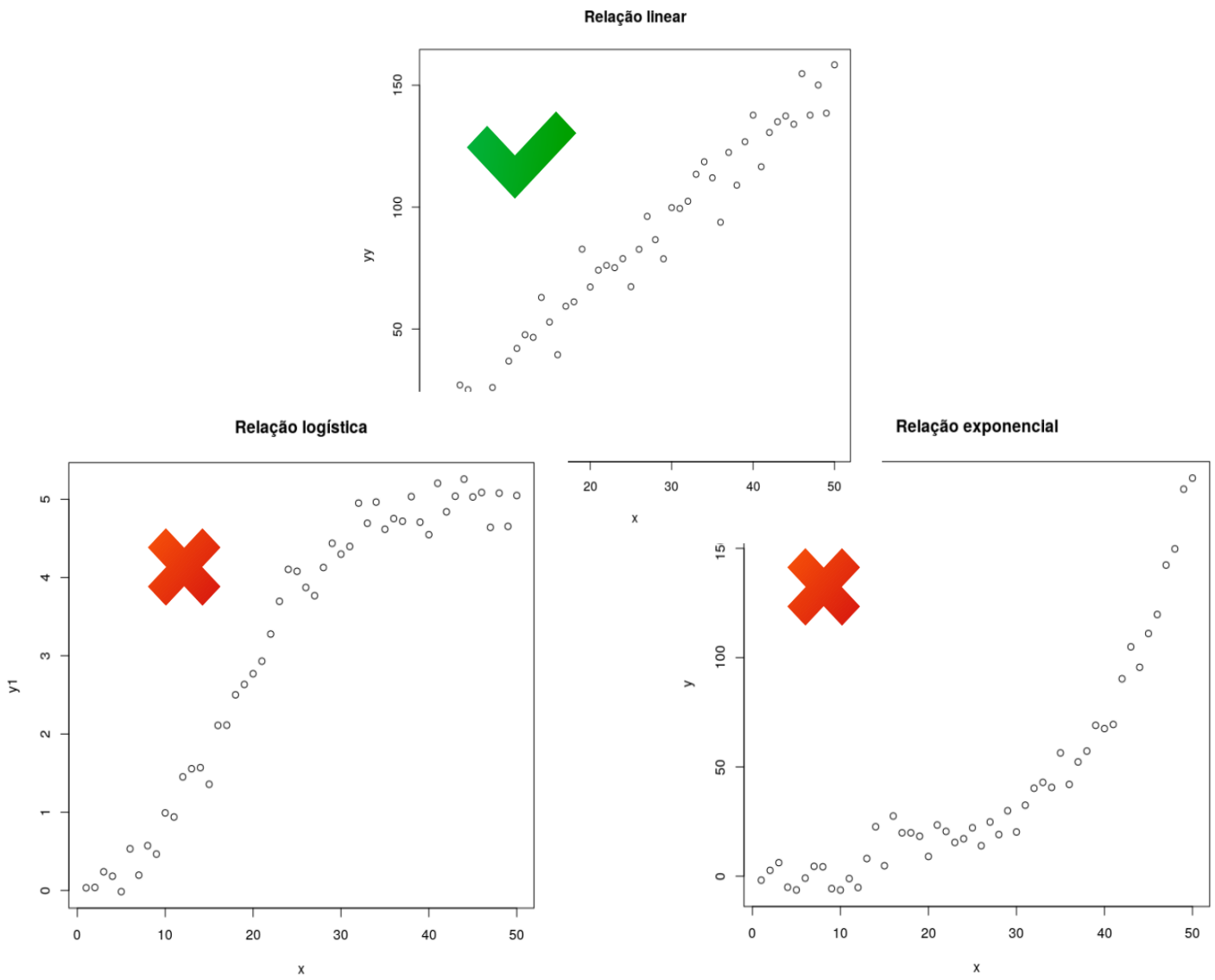
$$S_A = 3,51$$
$$S_B = 5,02$$

- Se perguntássemos qual das duas amostras foi a mais homogênea, certamente irias dizer que foi o curso A. Esta conclusão seria equivocada, pois proporcionalmente os dados do curso B estão mais próximos da média quando comparado com os dados do curso A. Vamos a prova!

$$CV_A = \frac{3,51}{54,4} * 100 = 6,45\%$$
$$CV_B = \frac{5,02}{83,8} * 100 = 5,99\%$$

MEDIDAS DE ASSOCIAÇÃO

- São medidas utilizadas para avaliar a relação entre duas variáveis.
- Serão abordados duas medidas: a covariância e a correlação.
- É importante ter em mente que as medidas que serão abordadas empressam a relação linear entre duas variáveis.


Relação linear
yy
x
Relação logística
y1
x
Relação exponencial
y
x

Covariância

- Mede a associação linear entre duas variáveis.
- No entanto, é impossível saber qual o grau de associação entre as variáveis pois os valores podem variar de menos infinito a mais infinito.
- Ainda, a covariância é influenciada pela unidade de medida das variáveis.
- Sua fórmula é dada por:

$$cov = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Aplicação

8. Considerem os seguintes dados sobre o número de comerciais (NC) e o volume de vendas (VV).

NC	VV	NC	VV
2	50	1	38
5	57	5	63
1	41	3	48
3	54	4	59
4	54	2	46

Qual a relação entre as variáveis NC e VV?

- Avaliando a covariância temos:

$$\bar{x}_{NC} = 3$$
$$\bar{y}_{VV} = 51$$
$$cov = \frac{(2-3)\cdot(50-51) + (5-3)\cdot(57-51) + \cdots + (2-3)\cdot(46-51)}{10-1}$$
$$= 11$$

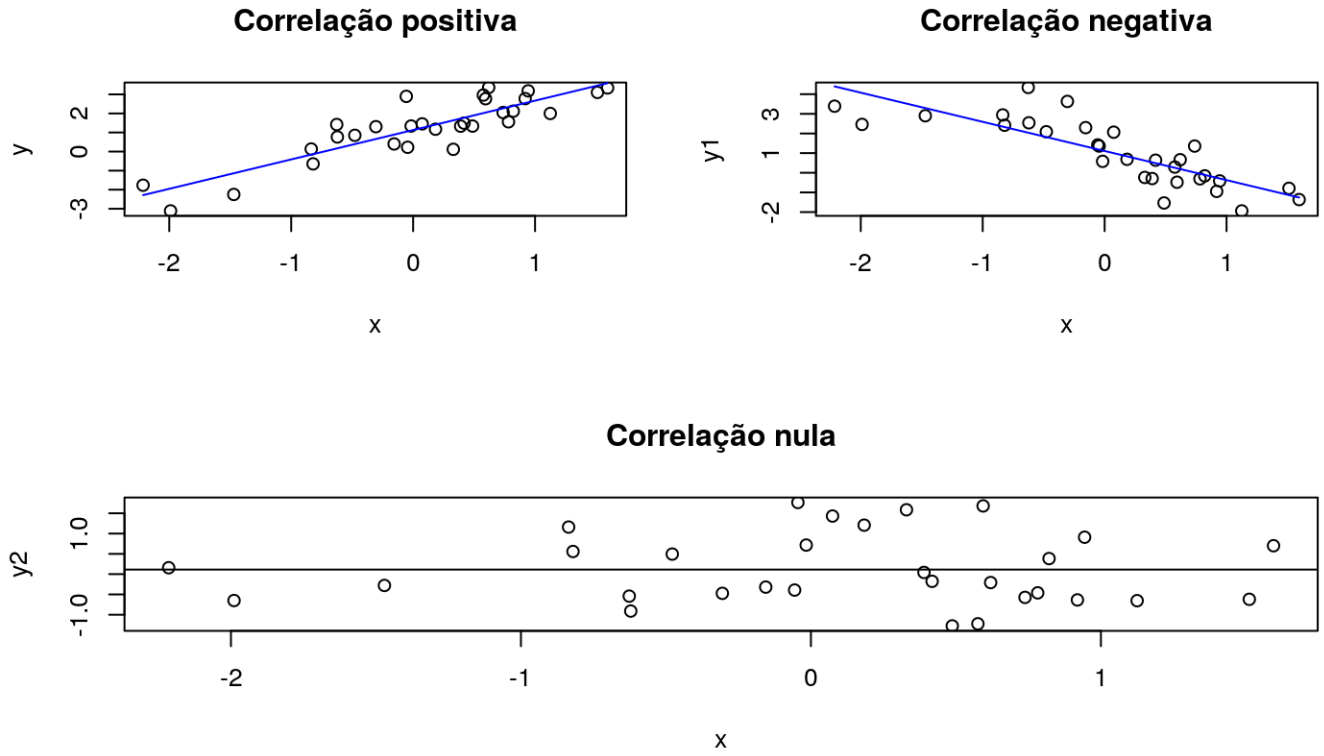
Percebam que o valor 11 não nos dá uma idéia do grau (força) de associação entre as duas variáveis em estudo. Só é possível afirmar que a relação é positiva.

Coeficiente de correlação de Pearson

- Também mede a associação linear entre duas variáveis quantitativas, mais é preferida por que os resultados ficam entre -1 e 1, valores estes que nos permite avaliar qual o grau de associação entre as variáveis estudada.
- Ainda, a correlação de Pearson não é influenciada pela unidade de medida das variáveis.
- A fórmula é a seguinte:

$$r_{xy} = \frac{cov_{xy}}{s_x s_y}$$

- Existem as seguintes possibilidades de relação entre duas variáveis:



9. Aproveitando os dados da aplicação 8 tem-se:

$$
\begin{aligned}
S_{NC} &= 1,49 \\
S_{VV} &= 7,93 \\
cov &= 11 \\
r_{NC-VV} &= \frac{11}{1,49 * 7,93} \\
&= 0,93
\end{aligned}
$$

Vejam que com o coeficiente de correlação é possível dizer a direção da associação e também o grau, a força de associação. Neste exemplo, temos uma forte associação positiva entre NC e VV, ou seja, o aumento no número de comerciais acarreta em maior volume de vendas.

MEDIDAS DE POSIÇÃO

- Também denominadas de separatrizes, são medidas utilizadas para dividir os dados em partes iguais e orientar quanto a posição da observação nos dados.

- Dentre as medidas utilizadas será abordado os percentis e os quartis.

Percentil

- p-ésimo percentil é um valor tal que pelo menos p por cento das observações são menores ou iguais a esse valor e pelo menos (100-p) por cento das observações são maiores ou iguais a esse valor.
- Existem vários métodos na literatura para encontrar o percentil no roll de dados. Será abordado aqui aquele que não faz nenhuma pressuposição a cerca da distribuição dos dados.

- Para calcular o *p*-ésimo percentil tem-se os seguintes passos:
  - Organize os dados em ordem crescente,
  - Calcule um índice *i*,

$$i = \left(\frac{p}{100}\right) n$$

  - em que *p* é o percentil procurado e *n*, o número de observações.
  - Se *i* não for um número inteiro, arredonde-o para cima. O número inteiro seguinte maior que *i* denota a posição do *p*-ésimo percentil.
  - Se *i* for um número inteiro, o *p*-ésimo percentil será a média dos valores nas posições *i* e *i+1*.

Exemplo

10. Considere os seguintes dados (parte dos dados) sobre a renda familiar mensal (em quantidade de salários mínimos) em duas localidades de Florianópolis - SC .

Monte verde	Pq da figueira	Monte verde	Pq da figueira
10.3	5.4	2.4	14.0
15.4	6.4	4.1	8.5
9.6	4.4	8.4	7.7
5.5	2.5	10.3	5.8
9.0	5.5	4.6	5.0

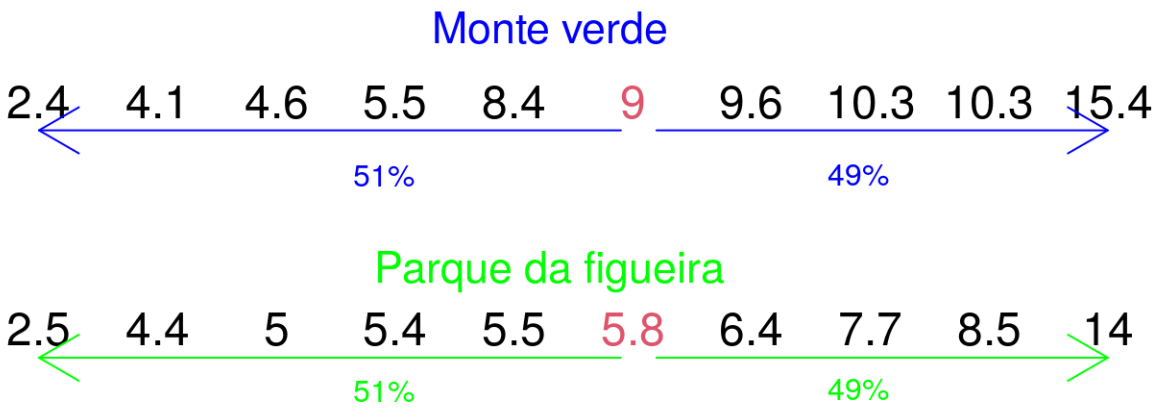
Segundo do dados publicados na seguinte página: http://g1.globo.com/economia/seu-dinheiro/noticia/2013/08/veja-diferencas-entre-conceitos-que-definem-classes-sociais-no-brasil.html, poderíamos classificar os bairros monte verde e parque da figueira em qual classe?

Imaginemos o seguinte: Sejam X e Y duas categorias distintas. Todos concordam que se a grande maioria (51% ou mais) dos integrantes de um grupo são classificados com a categoria X, então esse grupo pode ser classificado como X.

Como base no raciocínio acima, iremos calcular o percentil 51% e classificarmos os bairros de acordo com a tabela já exposta.

$$i = \frac{51}{100} \cdot 10 = 5,1 \approx 6$$

Vamos ordenar os dados para avaliarmos qual é o valor que está na posição 6.



Consirando o salário mínimo em 2014, temos que para a localidade *monte verde*, 51% das famílias ganham até 9 salários mínimos, ou seja, $9 \cdot 724 = R\$6516$. Segundo a tabela já exposta, podemos classificar esta localidade como famílias de **baixa classe alta**.

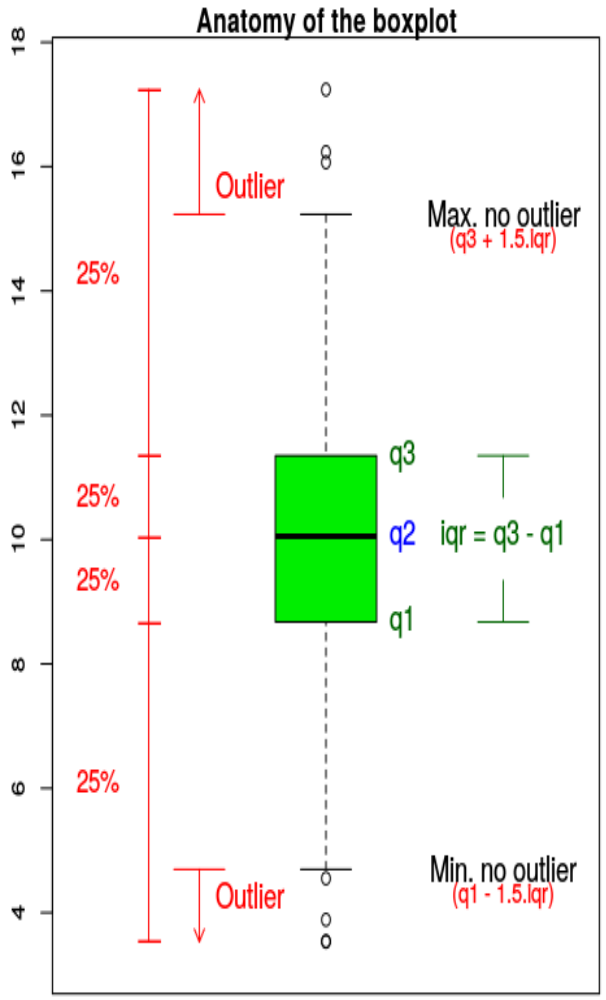
Quanto ao *parque da figueira*, 51% das famílias ganham até 5,8 salários mínimos, ou seja, $5,8 \cdot 724 = R\$4199,2$. Segundo a tabela já exposta, podemos classificar esta localidade como famílias de **baixa classe alta** também.

Quartil

- Muitas vezes é desejável dividir os dados em quatro partes, tendo cada parte aproximadamente um quarto, ou 25% das observações.
- Os quartis são muito utilizados para elaboração do box-plot.
- Tem-se os seguintes quartis:
  - 1° Quartil - corresponde ao 25° percentil.
  - 2° Quartil - corresponde ao 50° percentil. Coincide com a mediana.
  - 3° Quartil - corresponde ao 75° percentil.

BOXPLOT

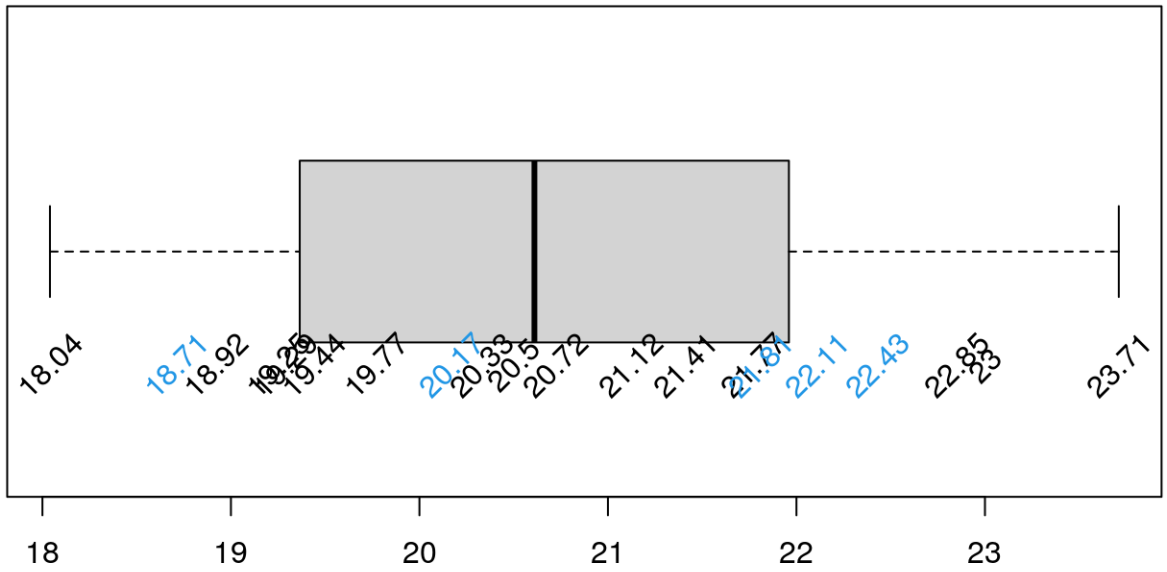
- O box-plot ou gráfico de caixa é um desenho esquemático utilizado para descrever as características mais proeminentes de conjuntos de dados. Essas características incluem:
  - centro
  - dispersão
  - extensão e a natureza de qualquer desvio em relação à simetria
  - identificação de *outliers*.


Anatomy of the boxplot
18
16
14
12
10
8
6
4
Outlier
Max. no outlier
(q3 + 1.5.iqr)
25%
25%
25%
25%
q3
q2
q1
iqr = q3 - q1
Outlier
Min. no outlier
(q1 - 1.5.iqr)

Aplicação

11. Aproveitando os dados da aplicação 1 tem-se:

$$q1 = 78$$
$$q2 = 82$$
$$q3 = 87$$
$$iqr = 87 - 78 = 9$$


18.04
18.71
18.92
19.44
19.77
20.17
20.33
20.5
20.72
21.12
21.41
22.11
22.43
22.85
23
23.71
18
19
20
21
22
23

PROPORÇÃO

- É uma medida estatística utilizada para avaliar o número de sucessos em um total de elementos estudados. A unidade de medida estará em decimal ou percentual, tanto faz.
- A variável de interesse pode ser tanto qualitativa quanto quantitativa.
- Se a variável for qualitativa, então a proporção será calculada pela razão do número de vezes que a característica de interesse foi observada pelo total de características.
- Suponha que o interesse seja em avaliar a proporção de solteiros no rol de dados a seguir. Então:

solteiro casado solteiro viúvo casado solteiro viúvo solteiro

$$\frac{solteiro(4)}{total(8)} = 0,5 \quad \textbf{ou} \quad 50\%$$

- Se a variável for quantitativa, então é provável que o interesse estará em um determinado intervalo de valores, principalmente se a variável for do tipo contínua.
- Suponhamos que estejamos interessados em avaliar a proporção de pessoas abaixo de 1,78m. Então:

1,89 1,56 1,68 1,75 1,61 1,77 1,56 1,80

$$\frac{< 1,78(6)}{total(8)} = 0,75 \quad \textbf{ou} \quad 75\%$$

- Se **N** observações são tomadas de uma população, então iremos calcular e donatar a **proporção** como:

$$\pi = \frac{n^o \quad sucessos}{N}$$

- Se **n** observações são tomadas de uma amostra, então a **proporção amostral** será:

$$p = \frac{n^o \quad sucessos}{n}$$