# Análise Exploratória de Dados

- Resumo de cinco números
- Box-plot (gráfico de caixa)

# Análise exploratória de dados

⇨ A média aritmética e o desvio padrão são medidas muito utilizadas.

⇨ Porém, essas medidas descrevem de forma ótima distribuições de frequências simétricas.

⇨ Numa distribuição assimétrica seus valores são bastante afetados pelos valores discrepantes (não são medidas resistentes).

John Wilder Tukey
(1915 - 2000)

**1970 → John Tukey** propôs técnicas que contornavam esses problemas. O conjunto dessas técnicas recebeu a denominação de **Análise Exploratória de Dados.**

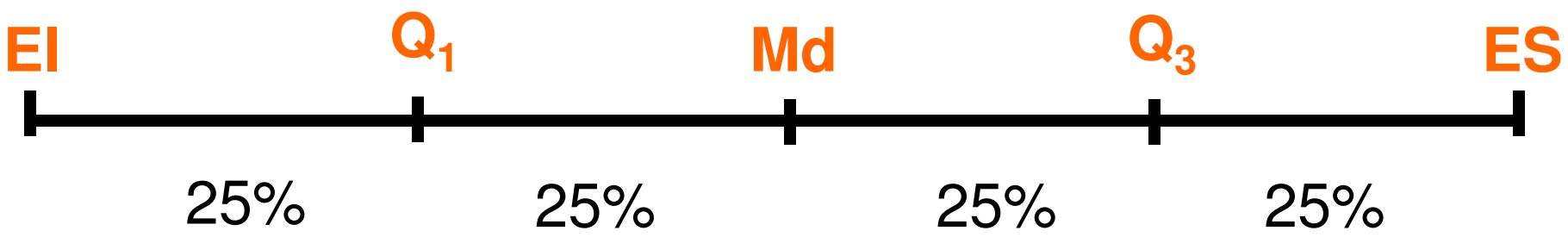
Principais técnicas exploratórias:

- Resumo de cinco números
- Box-plot

# Resumo de cinco números

Descreve o conjunto de dados através de cinco valores:

- mediana (Md)
- primeiro ($Q_1$) e terceiro ($Q_3$) quartis
- extremos inferior (EI) e superior (ES)

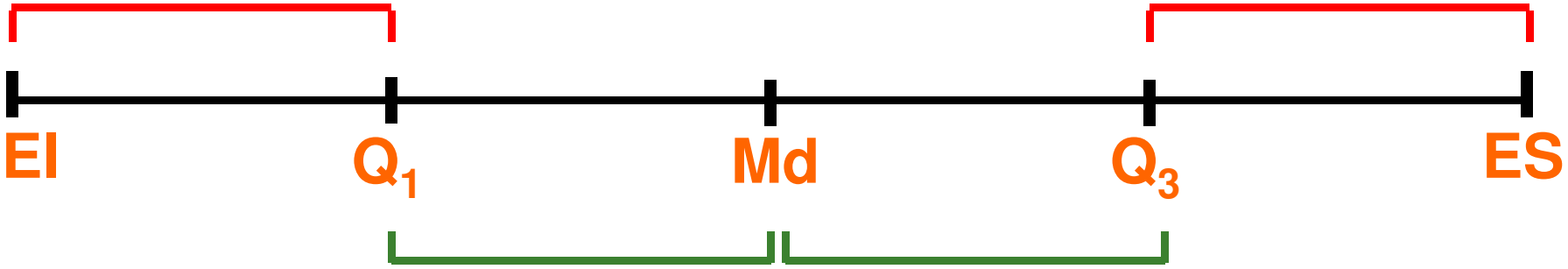



O resumo de cinco números fornece uma ideia da simetria (formato) da distribuição porque o percentual de valores dentro de cada intervalo é conhecido (25%).

# Simetria

A distribuição é considerada simétrica se:

1. A diferença entre o primeiro quartil e extremo inferior é aproximadamente igual à diferença entre o extremo superior e o terceiro quartil ($Q_1 - EI \cong ES - Q_3$)

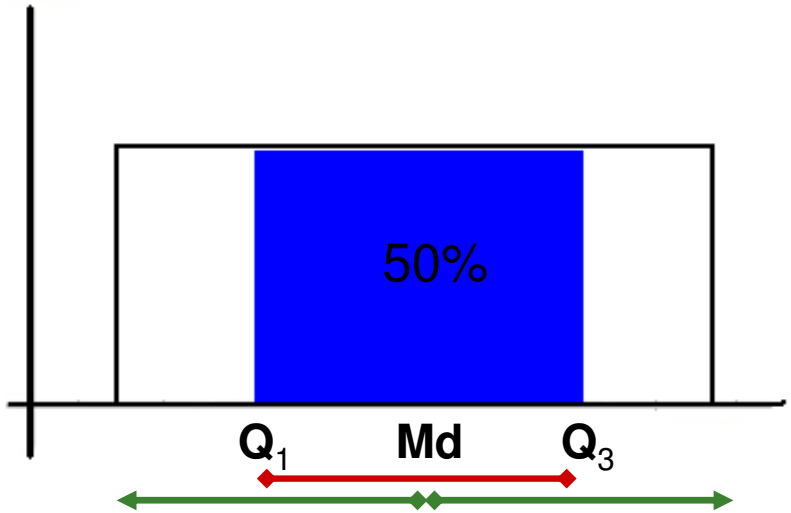



2. A diferença entre a mediana e o primeiro quartil é aproximadamente igual à diferença entre o terceiro quartil e a mediana ($Md - Q_1 \cong Q_3 - Md$)

Condições para a simetria

$$\begin{cases} Q1 - EI \cong ES - Q3 \\ Md - Q1 \cong Q3 - Md \end{cases}$$

## Distribuição Uniforme

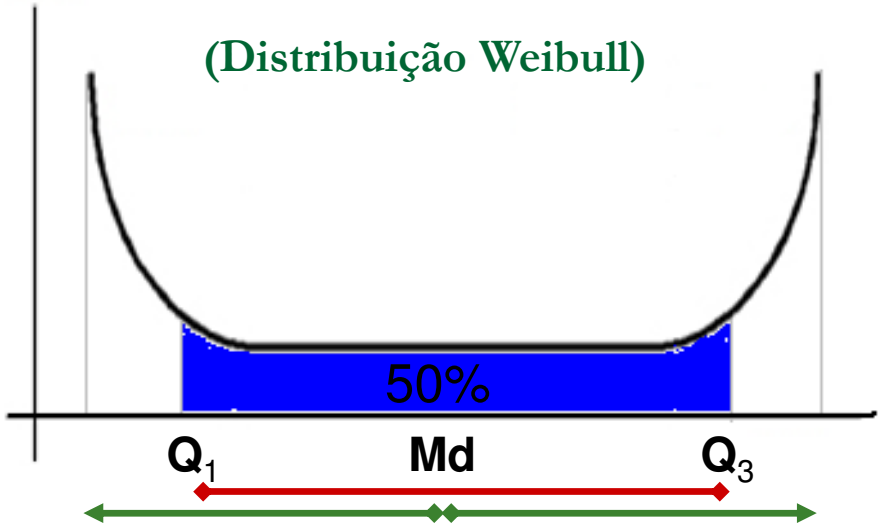



Condições para a simetria

$$\begin{cases} Q1 - EI \cong ES - Q3 \\ Md - Q1 \cong Q3 - Md \end{cases}$$

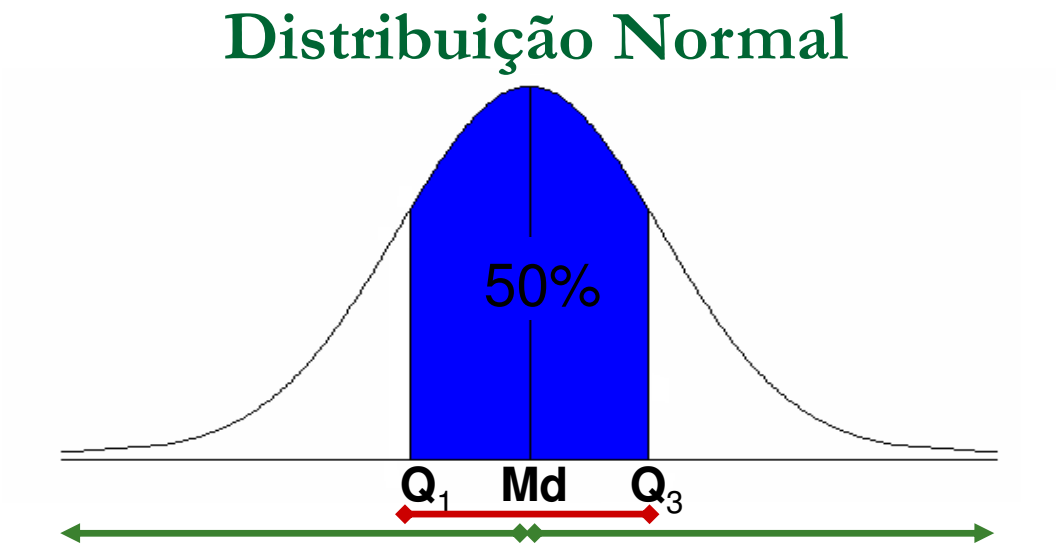
## Curva da Banheira

(Distribuição Weibull)



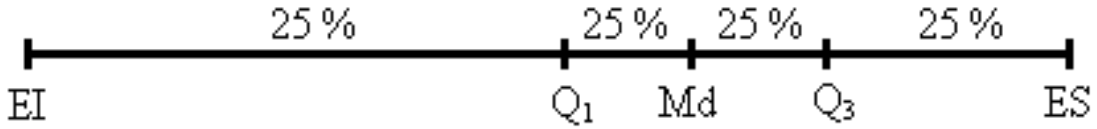

Condições para a simetria

$$\begin{cases} Q1 - EI \cong ES - Q3 \\ Md - Q1 \cong Q3 - Md \end{cases}$$

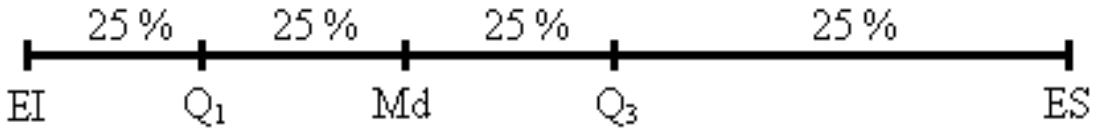



# Casos assimétricos

Assimetria Negativa




Assimetria Positiva



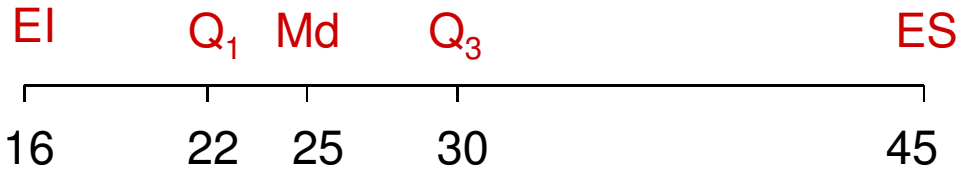
**Exercício proposto:** Os dados abaixo se referem aos pesos ao nascer (em kg) de 61 bovinos machos da raça Ibagé. Encontre o resumo de cinco números e classifique quanto à simetria da distribuição.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 17 | 17 | 18 | 18 | 18 | 19 | 20 | 20 | 20 | 20 | 20 | 21 | 21 | 22 | |
| 22 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 25 | 25 | 25 | 25 | 25 | |
| 25 | 26 | 26 | 27 | 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 30 | 30 | |
| 30 | 30 | 30 | 30 | 30 | 31 | 32 | 33 | 33 | 33 | 34 | 34 | 35 | 36 | 39 | 45 |

$$p_1 = \frac{n+1}{4} = \frac{61+1}{4} = 15{,}5 \qquad p_2 = \frac{2(n+1)}{4} = \frac{2(61+1)}{4} = 31 \qquad p_3 = \frac{3(n+1)}{4} = \frac{3(61+1)}{4} = 46{,}5$$

$$Q_1 = 22 \qquad Md = Q_2 = 25 \qquad Q_3 = 30$$

Resumo de cinco números ➡



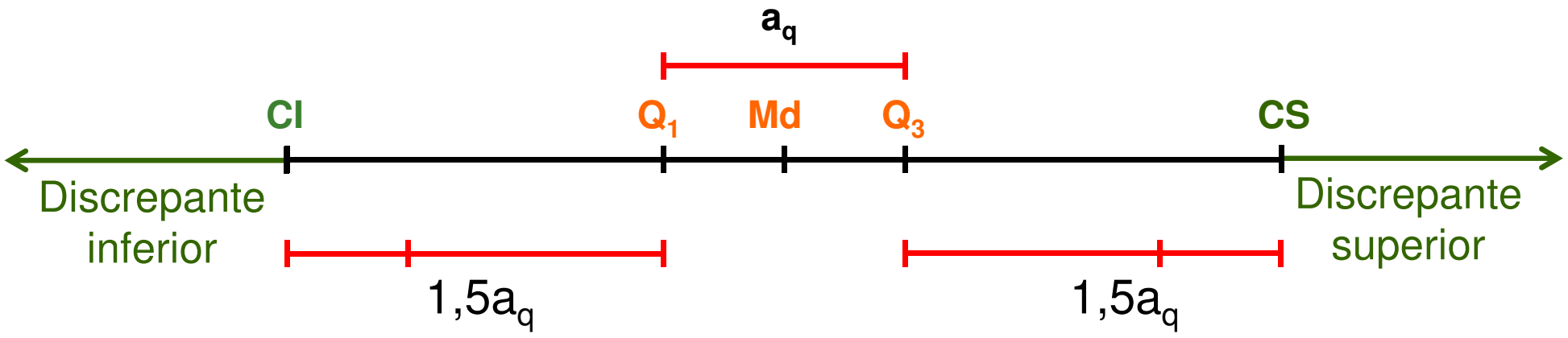

Assimétrica positiva

# Identificação de valores discrepantes (atípicos)

O critério usado para identificar valores discrepantes num conjunto de dados é baseado em duas medidas:

Cerca inferior $\rightarrow$ $CI = Q_1 - 1{,}5a_q$

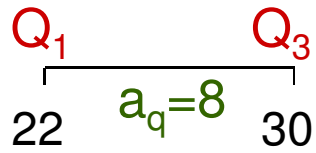
Cerca superior $\rightarrow$ $CS = Q_3 + 1{,}5a_q$



No exemplo referente aos pesos ao nascer de bovinos, serão considerados discrepantes os valores que estiverem fora dos limites da cerca superior e da cerca inferior:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 17 | 17 | 18 | 18 | 18 | 19 | 20 | 20 | 20 | 20 | 20 | 21 | 21 | 22 | |
| 22 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 25 | 25 | 25 | 25 | 25 | |
| 25 | 26 | 26 | 27 | 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 30 | 30 | |
| 30 | 30 | 30 | 30 | 30 | 31 | 32 | 33 | 33 | 33 | 34 | 34 | 35 | 36 | 39 | 45 |

$$CI = Q_1 - 1{,}5\ a_q = 22 - 1{,}5 \times 8 = 10$$




$$CS = Q_3 + 1{,}5\ a_q = 30 + 1{,}5 \times 8 = 42$$

Verificamos que o valor 45 ultrapassa a cerca superior, portanto, é classificado como discrepante superior.

# O que fazer quando identificamos valores discrepantes?

**Investigar a sua origem.**

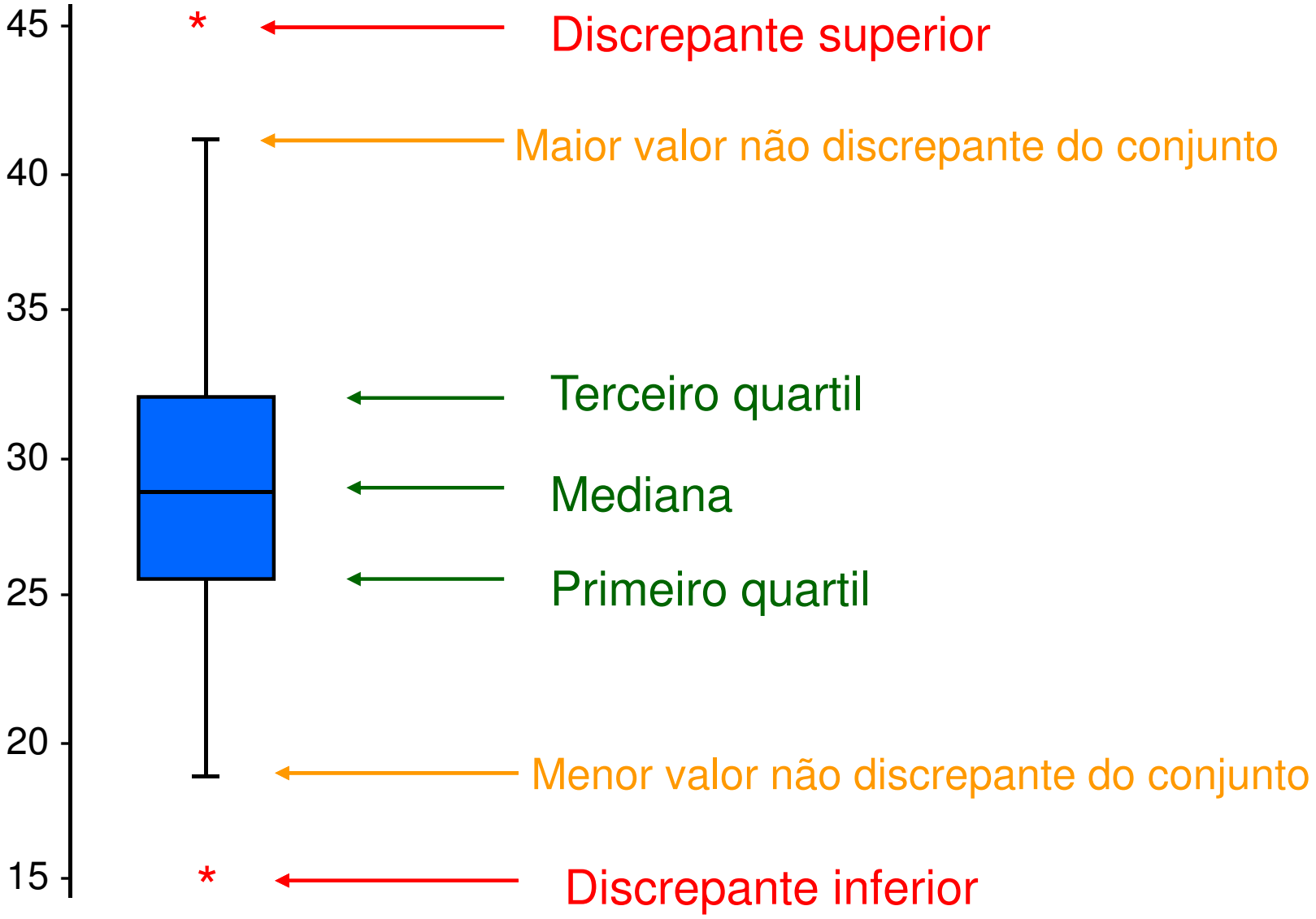
- Eventualmente, esses valores podem ser **oriundos de erros** na aferição ou no registro dos dados.
- Entretanto, valores discrepantes podem, de fato, **fazer parte do conjunto de dados**, reforçando a característica assimétrica da distribuição.

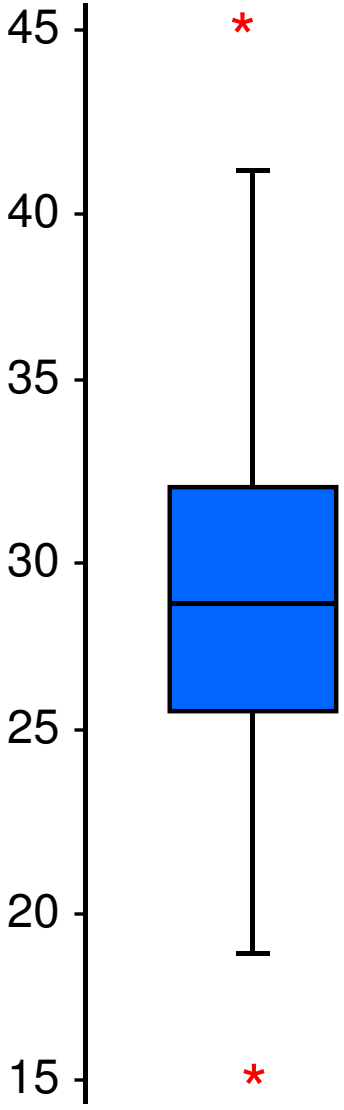
Uma inspeção cuidadosa nos dados e nas eventuais causas da ocorrência de valores discrepantes é sempre uma providência necessária antes que qualquer atitude seja tomada em relação a esses dados.

# Box-plot

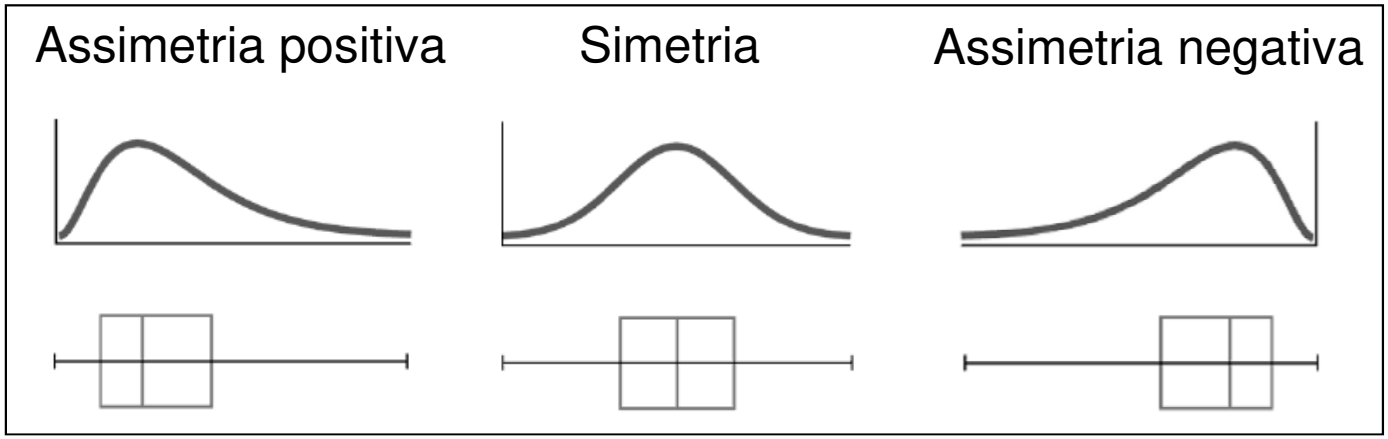
A informação dada pelo resumo de cinco números pode ser apresentada na forma de um box-plot que agrega uma série de informações sobre a distribuição:

- posição
- dispersão
- assimetria
- caudas
- dados discrepantes

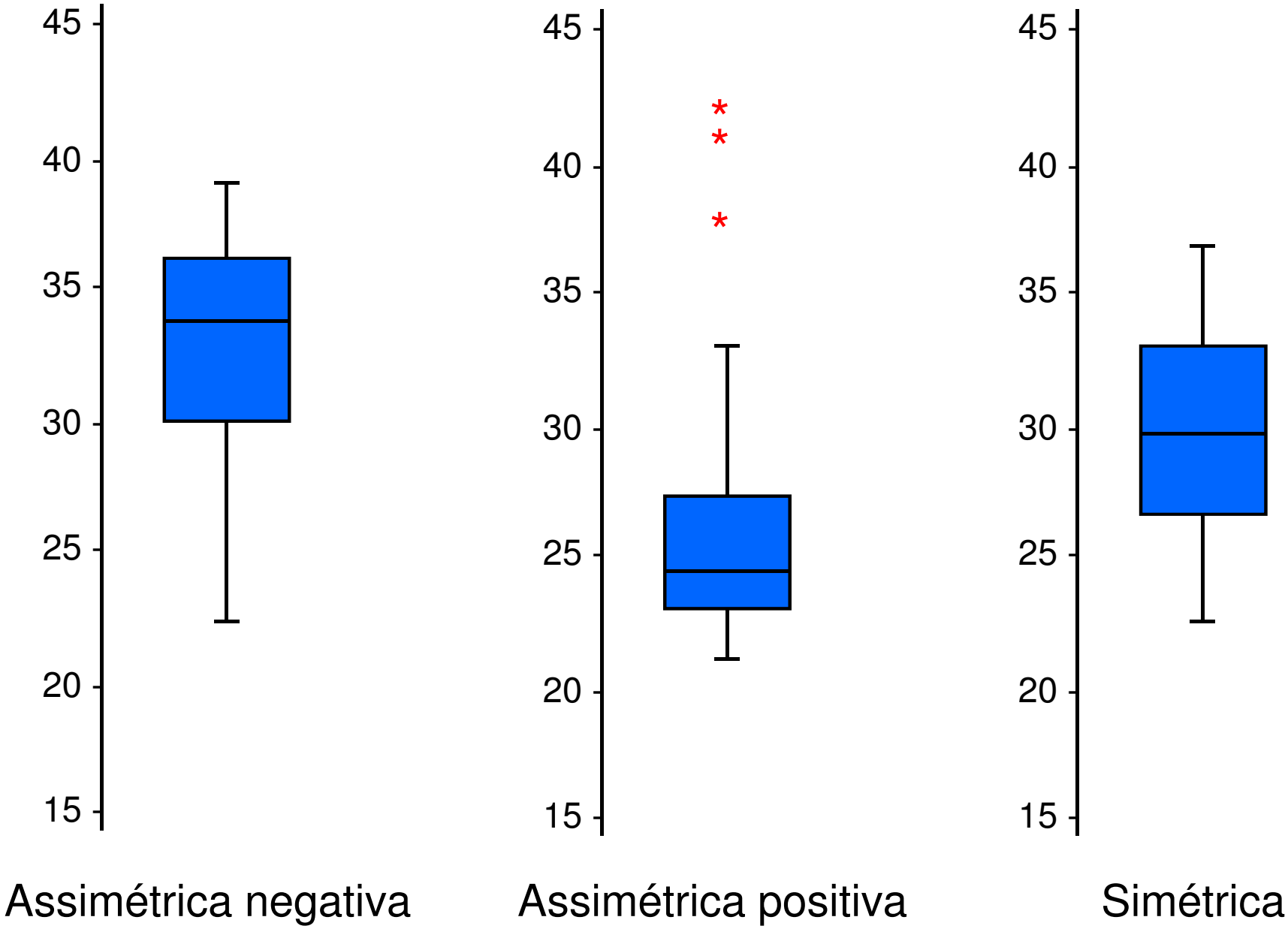

45
40
35
30
25
20
15
*
*
Discrepante superior
Maior valor não discrepante do conjunto
Terceiro quartil
Mediana
Primeiro quartil
Menor valor não discrepante do conjunto
Discrepante inferior




- A posição central dos valores é dada pela mediana e a dispersão pela amplitude interquartílica.
- As posições relativas da mediana e dos quartis e o formato dos bigodes dão uma noção da simetria e do tamanho das caudas da distribuição.

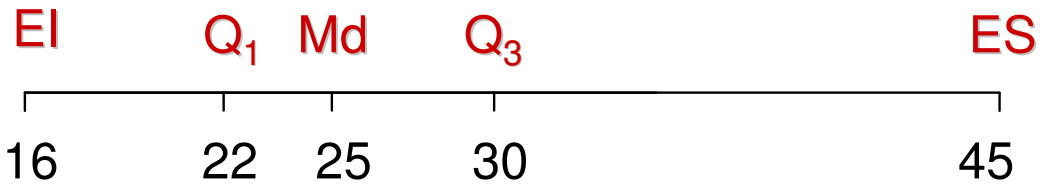


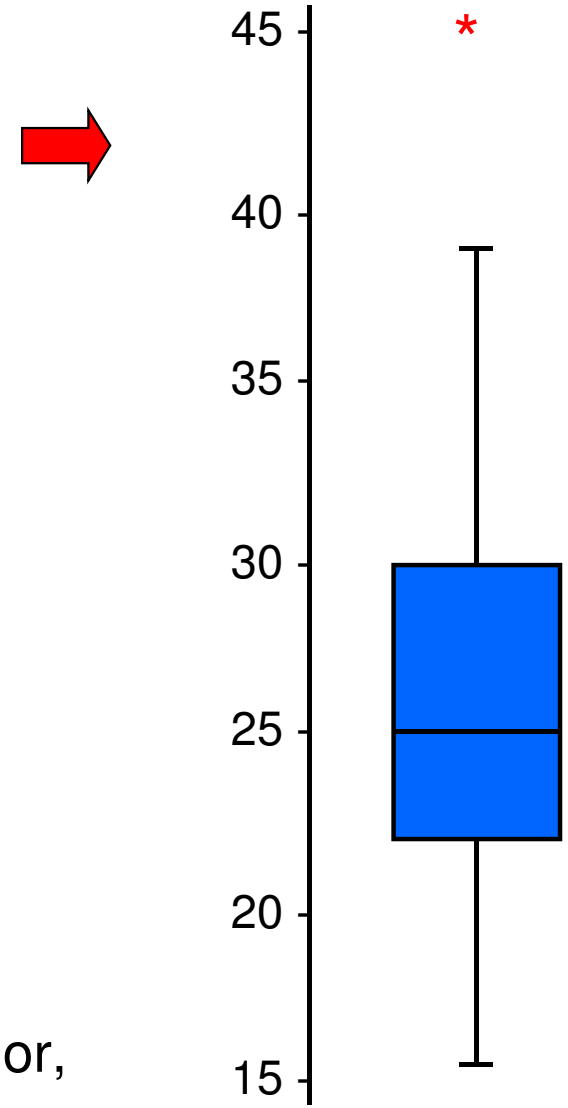
# Exemplos:


45
40
35
30
25
20
15
Assimétrica negativa
45
40
35
30
25
20
15
Assimétrica positiva
45
40
35
30
25
20
15
Simétrica

Consideremos o conjunto de dados referentes ao peso ao nascer (kg) de bovinos machos da raça Ibagé:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 17 | 17 | 18 | 18 | 18 | 19 | 20 | 20 | 20 | 20 |
| 20 | 21 | 21 | 22 | 22 | 23 | 23 | 23 | 23 | 23 | 23 |
| 23 | 23 | 23 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 |
| 27 | 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 30 |
| 30 | 30 | 30 | 30 | 30 | 30 | 31 | 32 | 33 | 33 | 33 |
| 34 | 34 | 35 | 36 | 39 | 45 | | | | | |




$CI = Q1 - 1{,}5\ a_q = 22 - 1{,}5 \times 8 = 10$

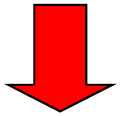
$CS = Q3 + 1{,}5\ a_q = 30 + 1{,}5 \times 8 = 42$

Verificamos que o valor 45 ultrapassa a cerca superior, portanto, é classificado como discrepante superior.

## **Exercício:** Fazer o **resumo de cinco números** e o **box-plot**.

Os dados abaixo se referem aos valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/03/2013.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 9,26 | 10,81 | 3,11 | 85,76 | 70,32 | 82,70 | 18,43 | 19,54 | 23,04 | 24,47 |
| 26,24 | 26,26 | 24,58 | 28,38 | 28,06 | 28,08 | 25,13 | 27,65 | 32,03 | 36,37 |
| 19,27 | 19,50 | 18,36 | 52,75 | 61,22 | 86,37 | 93,34 | 22,22 | 20,16 | 20,59 |
| 54,80 | 59,07 | 50,39 | 45,40 | 44,08 | 44,67 | 38,64 | 42,97 | 46,69 | 48,65 |
| 39,16 | 41,02 | 38,98 | 15,62 | 13,78 | 15,23 | 8,88 | 12,69 | 17,00 | 17,39 |



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3,11 | 8,88 | 9,26 | 10,81 | 12,69 | 13,78 | 15,23 | 15,62 | 17,00 | 17,39 |
| 18,36 | 18,43 | 19,27 | 19,50 | 19,54 | 20,16 | 20,59 | 22,22 | 23,04 | 24,47 |
| 24,58 | 25,13 | 26,24 | 26,26 | 27,65 | 28,06 | 28,08 | 28,38 | 32,03 | 36,37 |
| 38,98 | 38,64 | 39,16 | 41,02 | 42,97 | 44,08 | 44,67 | 45,40 | 46,69 | 48,65 |
| 50,39 | 52,75 | 54,80 | 59,07 | 61,22 | 70,32 | 82,70 | 85,76 | 86,37 | 93,34 |

## Solução:

**p(Q1) = 13**
**p(Q2) = 25,5**
**p(Q3) = 38**

| | |
|---|---|
| **EI =** | **3,1** |
| **Q1 =** | **19,27** |
| **Md =** | **27,855** |
| **Q3 =** | **45,4** |
| **ES =** | **93,3** |

Os valores 85,76 ; 86,37 e 93,34 são considerados discrepantes

**Assimetria positiva**