

D1EAD – Análise Estatística para Ciência de Dados 2021.1



Análise Exploratória de Dados

Prof. Ricardo Sovat

sovat@ifsp.edu.br

Prof. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br



Estadística Descriptiva

Estatística Descritiva

É o ramo da estatística que visa **sumarizar e descrever** o conjunto de dados.

A disponibilidade de uma **grande quantidade de dados** e de **métodos computacionais muito eficientes** revigorou esta área da estatística.

Junto com **visualizações (gráficos)**, elas formam a **base** para a análise exploratória de dados.

Estatística Descritiva

É o ramo da estatística que visa **sumarizar e descrever** o conjunto de dados.

A disponibilidade de uma **grande quantidade de dados** e de **métodos computacionais muito eficientes** revigorou esta área da estatística.

Junto com **visualizações (gráficos)**, elas formam a **base** para a análise exploratória de dados.

Qual a nota média dos alunos na última prova?

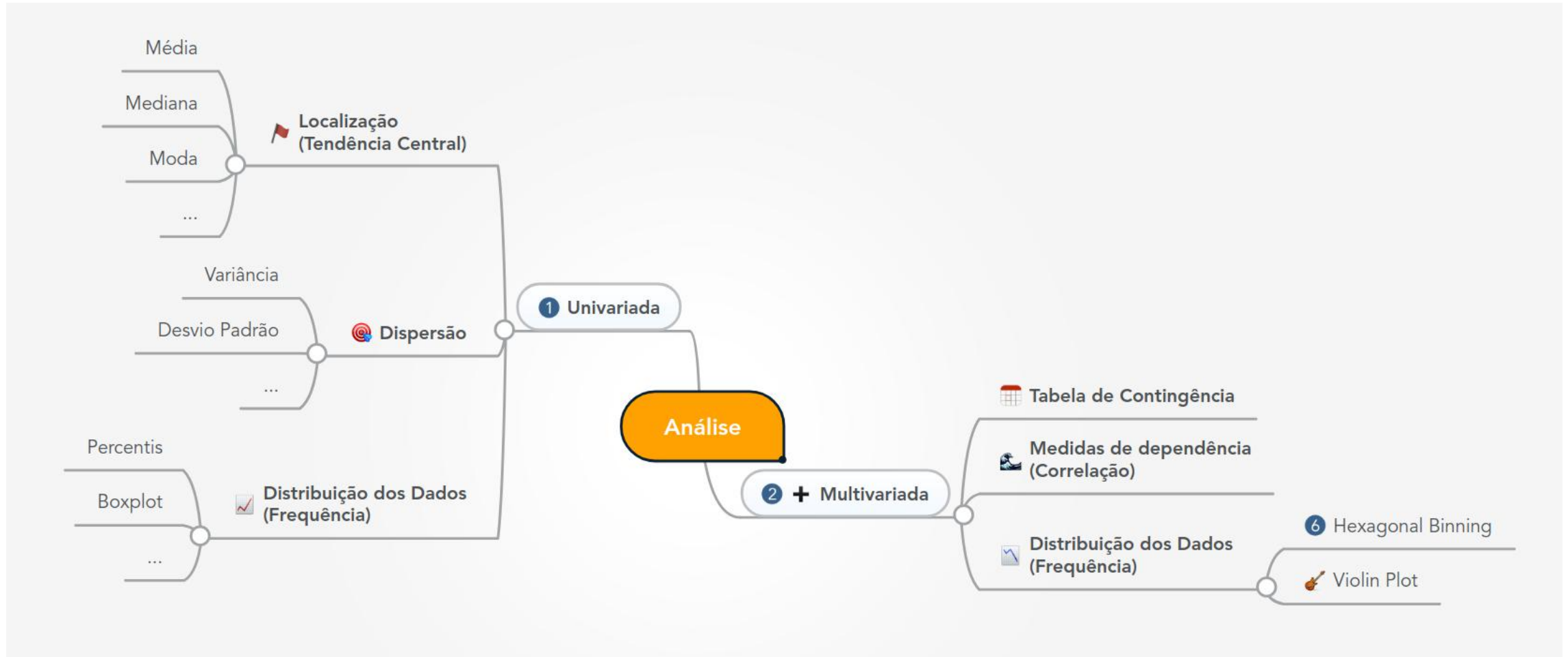
Qual é o preço aproximado da gasolina no estado de SP?

Como a riqueza do Brasil está distribuída?

Qual é a eficácia da vacina contra a doença X?

...

Estatística Descritiva



Tendência Central

Média

Média

Qual o **preço médio** da **Gasolina Comum** no estado de **São Paulo** ao longo dos anos?

| ESTADO | PREÇO MÉDIO REVENDA | ANO |
|-----------|---------------------|------|
| SAO PAULO | 1.891 | 2004 |
| SAO PAULO | 1.888 | 2004 |
| SAO PAULO | 1.894 | 2004 |
| SAO PAULO | 1.912 | 2004 |
| SAO PAULO | 1.919 | 2004 |
| ... | ... | ... |

785 registros

Média

Qual o **preço médio** da **Gasolina Comum** no estado de **São Paulo** ao longo dos anos?

| ESTADO | PREÇO MÉDIO REVENDA | ANO |
|-----------|---------------------|------|
| SAO PAULO | 1.891 | 2004 |
| SAO PAULO | 1.888 | 2004 |
| SAO PAULO | 1.894 | 2004 |
| SAO PAULO | 1.912 | 2004 |
| SAO PAULO | 1.919 | 2004 |
| ... | ... | ... |

785 registros

Fórmula

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

x_i : iésimo valor/registro

n : número de valores/registros

Todos os registros **contribuem igualmente (mesmo peso)** no cálculo da média

Média

Qual o **preço médio** da **Gasolina Comum** no estado de **São Paulo** ao longo dos anos?

| ESTADO | PREÇO MÉDIO REVENDA | ANO |
|-----------|---------------------|------|
| SAO PAULO | 1.891 | 2004 |
| SAO PAULO | 1.888 | 2004 |
| SAO PAULO | 1.894 | 2004 |
| SAO PAULO | 1.912 | 2004 |
| SAO PAULO | 1.919 | 2004 |
| ... | ... | ... |

785 registros

Fórmula

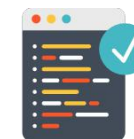
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

x_i : iésimo valor/registro

n : número de valores/registros

Todos os registros **contribuem igualmente (mesmo peso)** no cálculo da média

$$\bar{x} = \frac{1.891 + 1.888 + 1.894 + \dots}{785} = \mathbf{R\$ 2.846}$$



Média Ponderada

Qual a nota média final do João?

P1: 10

P2: 8

T: 7.5

A: 8.5

$$NotaFinal = \frac{0.2 * P1 + 0.2 * P2 + 0.5 * T + 0.1 * A}{(0.2 + 0.2 + 0.5 + 0.1)}$$

Fórmula

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

x_i : iésimo valor/registro

w_i : peso do iésimo valor/registro

n : número de valores/registros

Cada registro contribui com
um **peso diferente**
no cálculo da média

Média Ponderada

Qual a nota média final do João?

P1: 10

P2: 8

T: 7.5

A: 8.5

$$NotaFinal = \frac{0.2 * P1 + 0.2 * P2 + 0.5 * T + 0.1 * A}{(0.2 + 0.2 + 0.5 + 0.1)}$$

$$NotaFinal = \frac{0.2 * 10 + 0.2 * 8 + 0.5 * 7.5 + 0.1 * 8.5}{1.0} = 8.2$$

Fórmula

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

x_i : iésimo valor/registo

w_i : peso do iésimo valor/registo

n : número de valores/registos

Cada registo **contribui com**
um peso diferente
no cálculo da média

Um Problema da Média

Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

Um Problema da Média

Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |


$$\bar{x}=4000.0$$

Um Problema da Média

Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

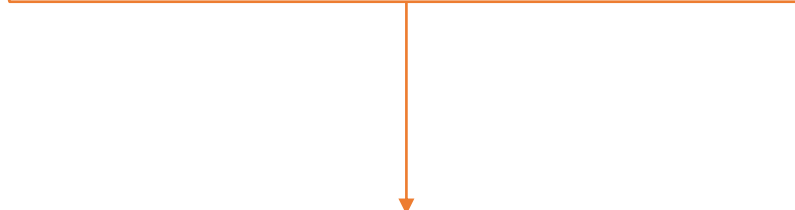

$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|---------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |

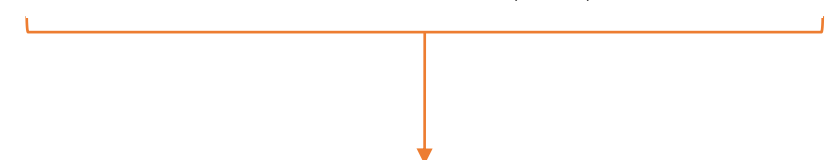
Um Problema da Média

Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |


$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|---------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |


$$\bar{x}=203,200.0$$

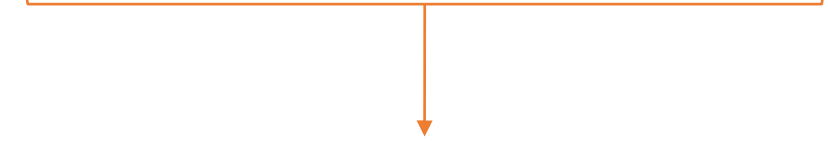
Um Problema da Média

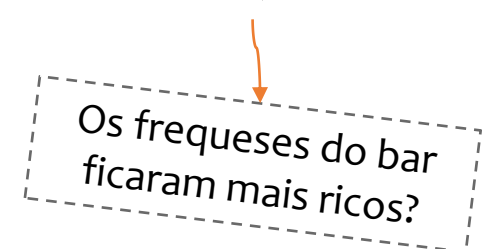
Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |


$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|---------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |


$$\bar{x}=203,200.0$$



Os fregueses do bar
ficaram mais ricos?

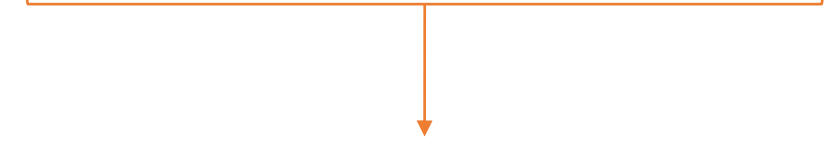
Um Problema da Média


Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |


$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|---------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |


$$\bar{x}=203,200.0$$



Os fregueses do bar ficaram mais ricos? **NÃO!**

Um Problema da Média

Qual o **salário médio** dos fregueses do Bar do Juca?


| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |


$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

outlier

| | |
|--------------|-------------|
| Fausto Silva | 1,000,000.0 |
|--------------|-------------|


$$\bar{x}=203,200.0$$

Os fregueses do bar
ficaram mais ricos?

NÃO!

Um Problema da Média

Qual o **salário médio** dos fregueses do Bar do Juca?

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

$$\bar{x}=4000.0$$

Soluções

- Remover os outliers
- *Trimmed Mean*
- Mediana

outlier

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |

$$\bar{x}=203,200.0$$

Os fregueses do bar ficaram mais ricos?

NÃO!

Mediana

Número central de uma lista de elementos **ordenados** (ranqueados).

Se o número de elementos é **par**, a **mediana** é igual a **média dos dois valores centrais** que dividem a lista ao meio.

Qual a **mediana** dos fregueses do Bar do Juca?

$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

$$\bar{x}=203,2000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |

outlier

Mediana

Número central de uma lista de elementos **ordenados** (ranqueados).

Se o número de elementos é **par**, a **mediana** é igual a **média dos dois valores centrais** que dividem a lista ao meio.

Qual a **mediana** dos fregueses do Bar do Juca?

$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

3000; 3500; 4500; 5000

$$mediana = \frac{3500 + 4500}{2} = 4000.0$$

$$\bar{x}=203,2000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |

outlier

3000; 3500; 4500; 5000; 1,000,000

$$mediana = 4500.0$$

Mediana

Número central de uma lista de elementos **ordenados** (ranqueados).

Se o número de elementos é **par**, a **mediana** é igual a **média dos dois valores centrais** que dividem a lista ao meio.

Qual a **mediana** dos fregueses do Bar do Juca?

$$\bar{x}=4000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |

3000; 3500; 4500; 5000

$$mediana = \frac{3500 + 4500}{2} = 4000.0$$

outlier

Em distribuições sem
sérios problemas de
outliers, a **média e a
mediana são similares**

$$\bar{x}=203,2000.0$$

| Nome | Salário Mensal (R\$) |
|-------------------|----------------------|
| João das Neves | 5000.0 |
| Daineres da Silva | 4500.0 |
| Luke Escaiuolker | 3000.0 |
| Leia Morgana | 3500.0 |
| Fausto Silva | 1,000,000.0 |

3000; 3500; 4500; 5000; 1,000,000

$$mediana = 4500.0$$

Mediana

Qual é a **mediana do preço** a **Gasolina Comum** no estado de **São Paulo** ao longo dos anos?

| ESTADO | PREÇO MÉDIO REVENDA | ANO |
|-----------|---------------------|------|
| SAO PAULO | 1.891 | 2004 |
| SAO PAULO | 1.888 | 2004 |
| SAO PAULO | 1.894 | 2004 |
| SAO PAULO | 1.912 | 2004 |
| SAO PAULO | 1.919 | 2004 |
| ... | ... | ... |

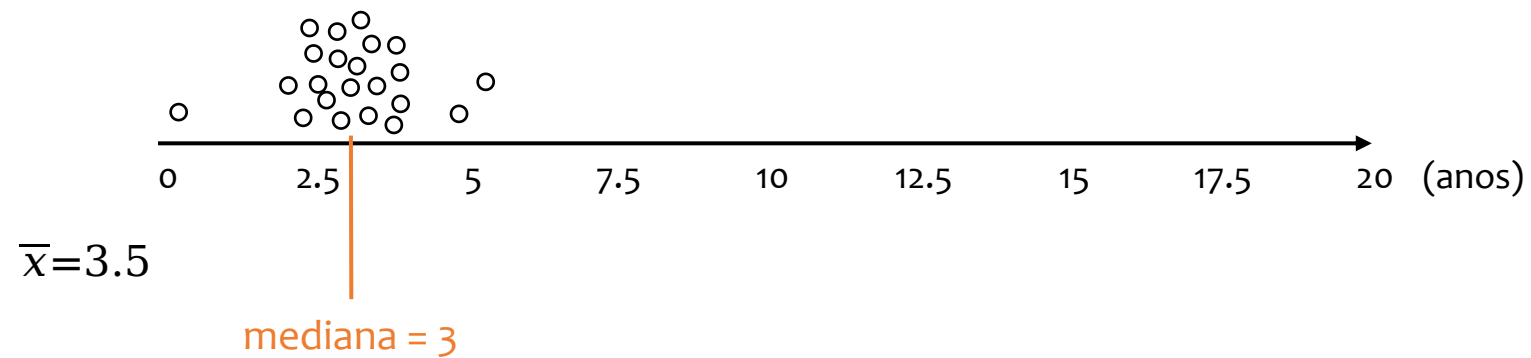
785 registros

$$\bar{x} = \text{R\$ } 2.846$$

$$\textit{mediana} = \text{R\$ } 2.638$$

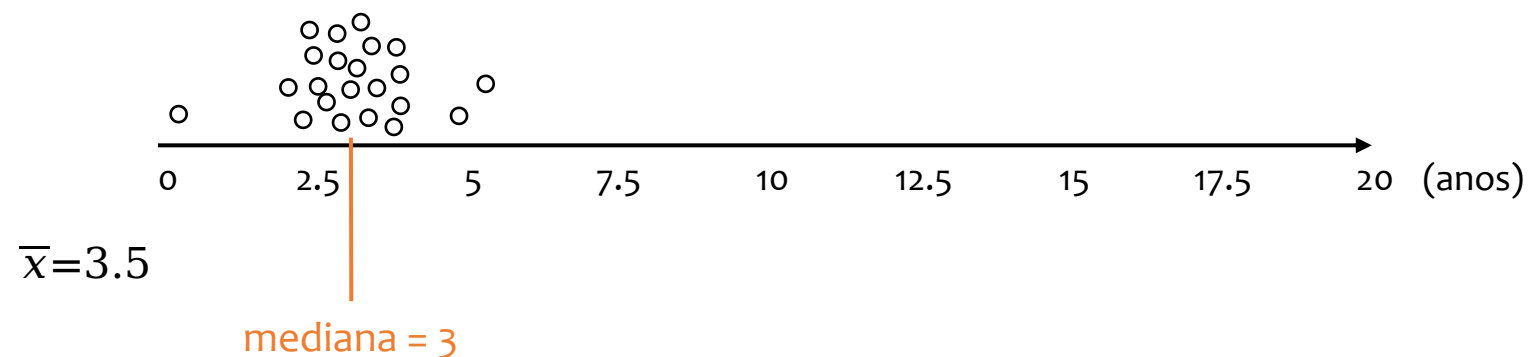
Outliers podem ser importantes

Suponha que os pacientes com uma dada **doença fatal** possuem a seguinte **expectativa de vida**:

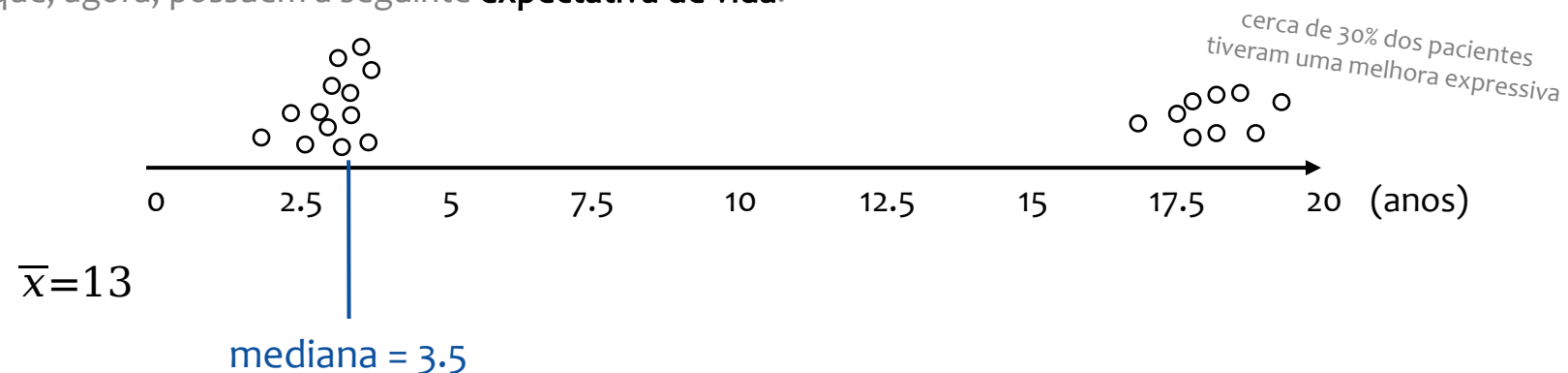


Outliers podem ser importantes

Suponha que os pacientes com uma dada **doença fatal** possuem a seguinte **expectativa de vida**:

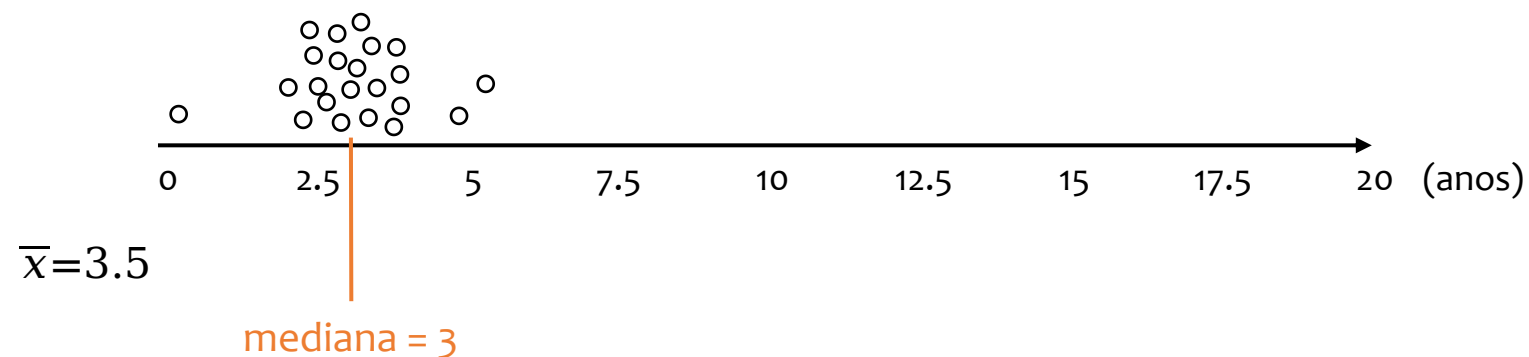


Uma nova droga, muito **cara** e que apresenta efeitos colaterais, foi aplicada em todos os pacientes que, agora, possuem a seguinte **expectativa de vida**:

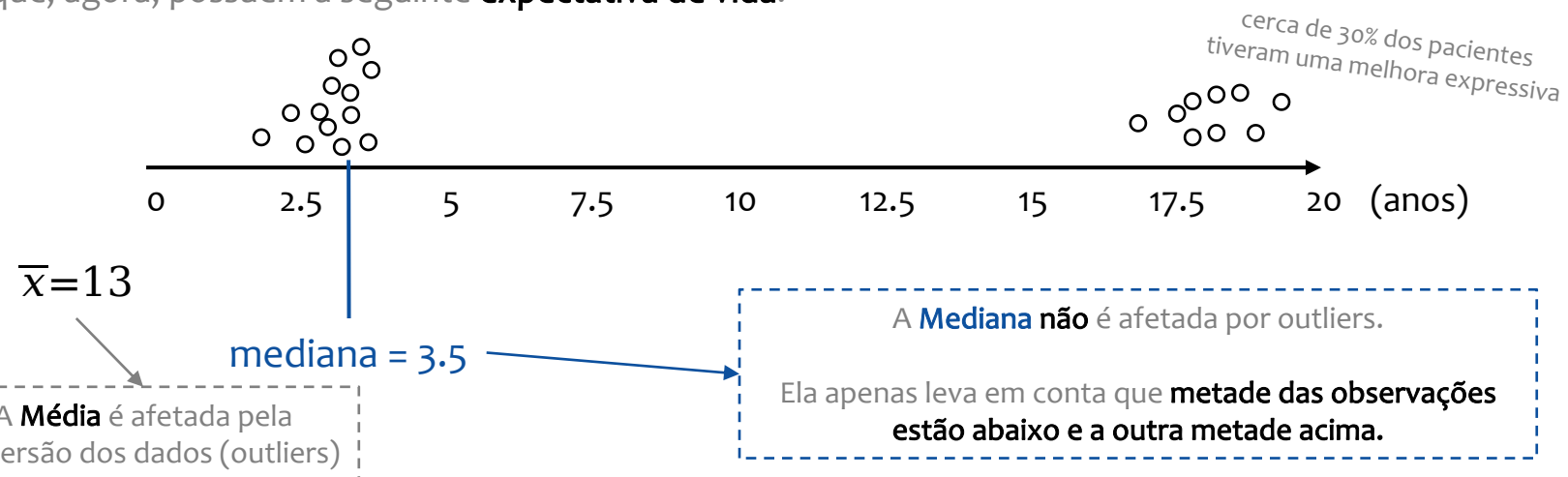


Outliers podem ser importantes

Suponha que os pacientes com uma dada **doença fatal** possuem a seguinte **expectativa de vida**:

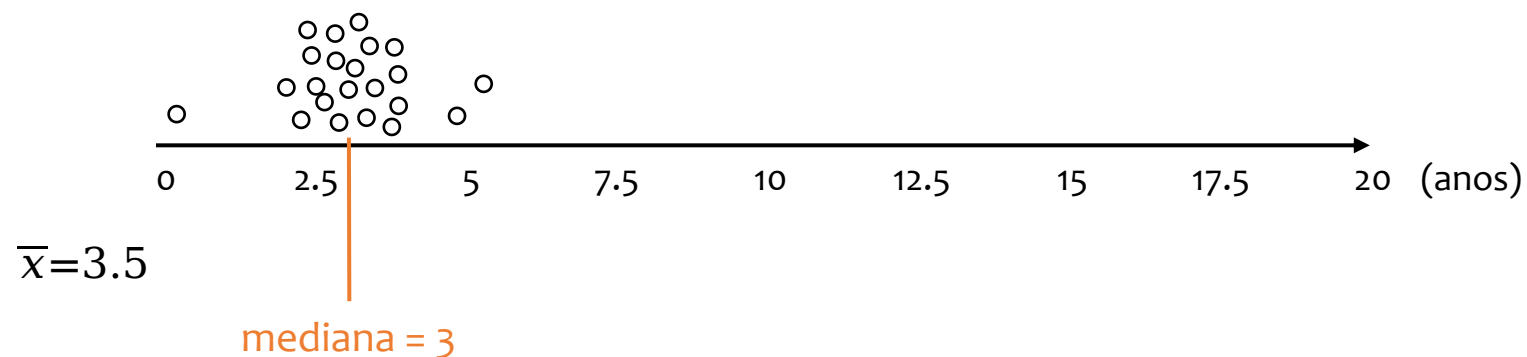


Uma nova droga, muito **cara** e que apresenta efeitos colaterais, foi aplicada em todos os pacientes que, agora, possuem a seguinte **expectativa de vida**:



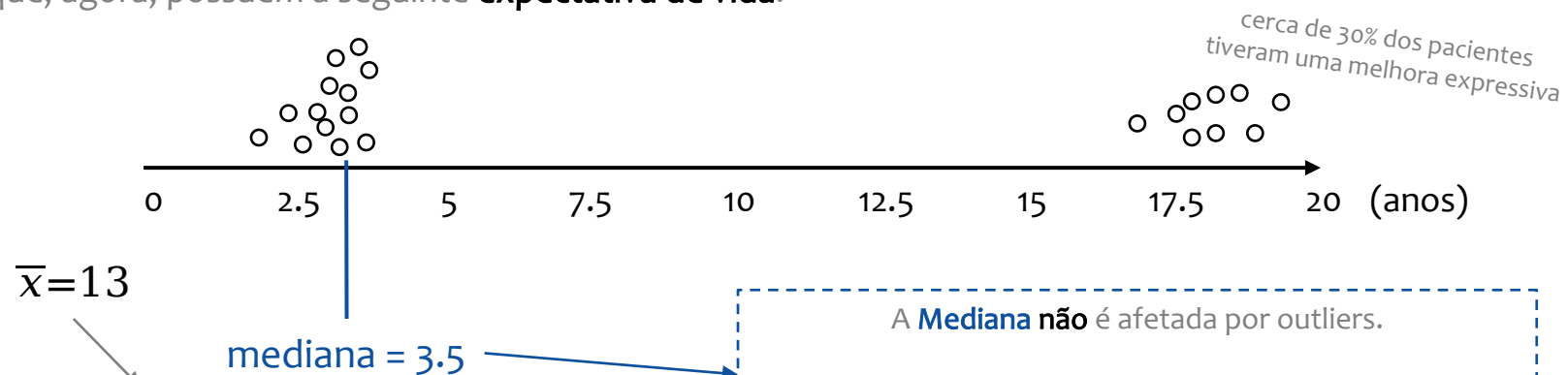
Outliers podem ser importantes

Suponha que os pacientes com uma dada **doença fatal** possuem a seguinte **expectativa de vida**:



A droga vale a pena?

Uma nova droga, muito **cara** e que apresenta efeitos colaterais, foi aplicada em todos os pacientes que, agora, possuem a seguinte **expectativa de vida**:

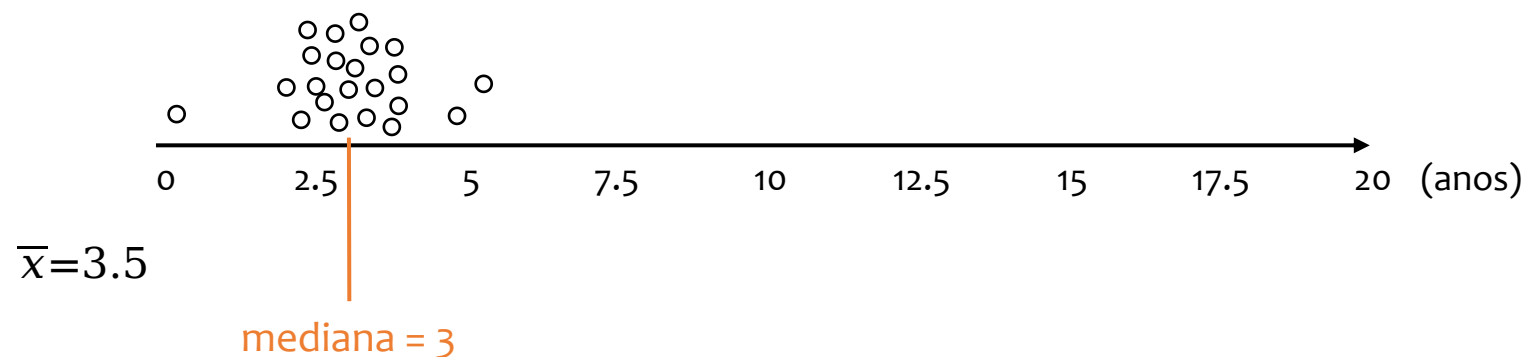


A **Média** é afetada pela dispersão dos dados (outliers)

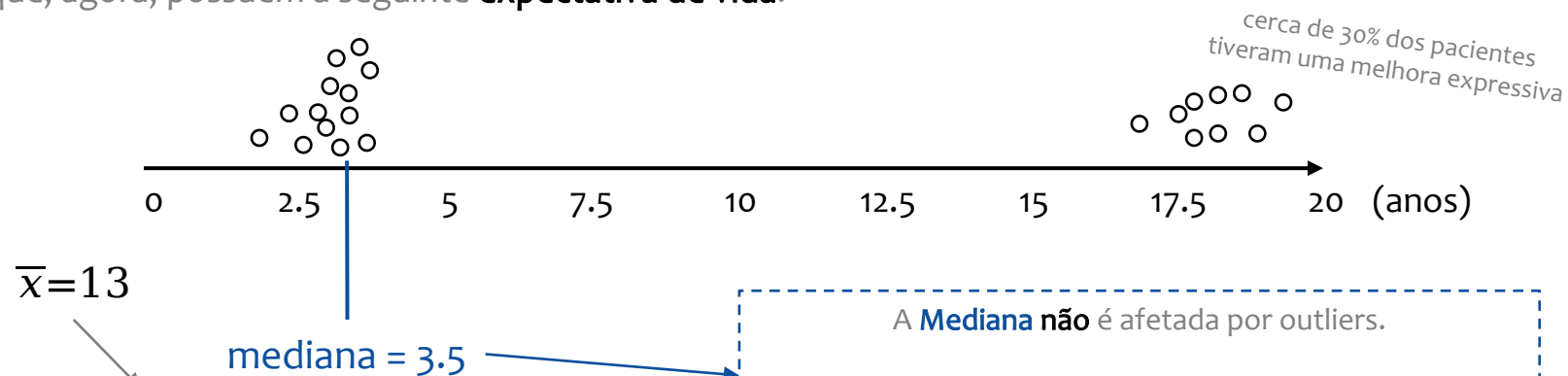
A **Mediana** não é afetada por outliers.
Ela apenas leva em conta que metade das observações estão abaixo e a outra metade acima.

Outliers podem ser importantes

Suponha que os pacientes com uma dada **doença fatal** possuem a seguinte **expectativa de vida**:



Uma nova droga, muito **cara** e que apresenta efeitos colaterais, foi aplicada em todos os pacientes que, agora, possuem a seguinte **expectativa de vida**:



A droga vale a pena?

Analisando (erroneamente) apenas a **mediana**, não!

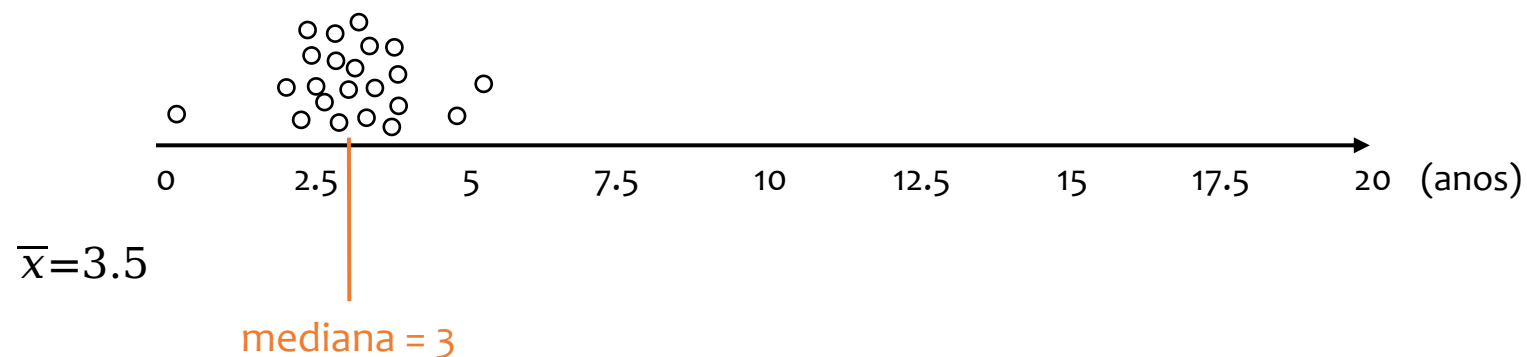
Nossos “outliers” **são vidas** e devem ser **relevantes** em nossa decisão.

Pela melhora considerável de muitos pacientes, a **droga vale a pena!**

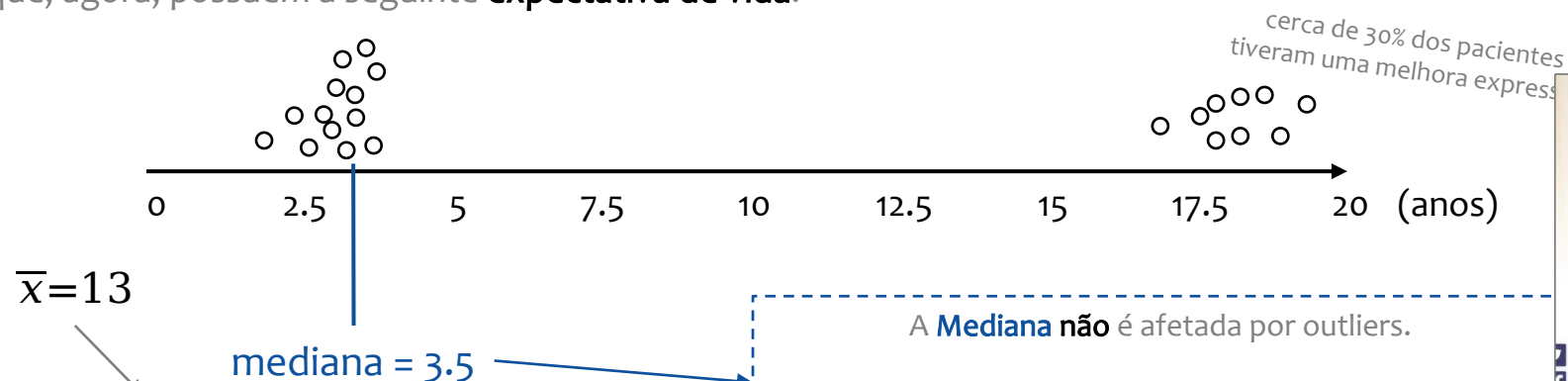
A **Média** é afetada pela dispersão dos dados (outliers)

Outliers podem ser importantes

Suponha que os pacientes com uma dada **doença fatal** possuem a seguinte **expectativa de vida**:



Uma nova droga, muito **cara** e que apresenta efeitos colaterais, foi aplicada em todos os pacientes que, agora, possuem a seguinte **expectativa de vida**:



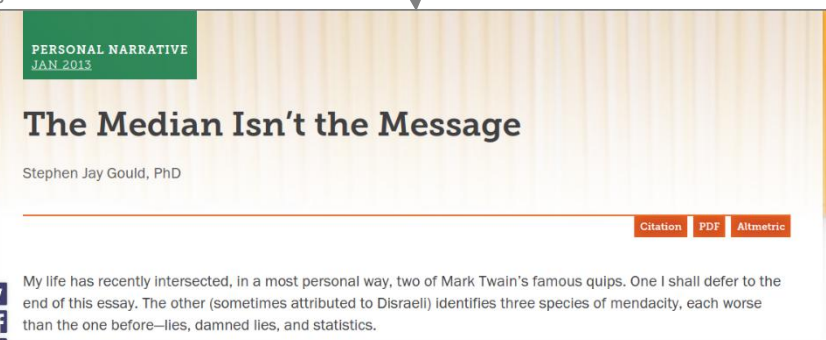
A **Média** é afetada pela dispersão dos dados (outliers)

A droga vale a pena?

Analisando (erroneamente) apenas a **mediana**, não!

Nossos “outliers” **são vidas** e devem ser **relevantes** em nossa decisão.

Pela melhora considerável de muitos pacientes, a **droga vale a pena!**



Quando usar a Média e a Mediana?

Quando usar a Média e a Mediana?

Depende se os **outliers** **distorcem** o que está sendo descrito ou, em vez disso, são uma **parte importante** da análise (mensagem).

Não há nenhuma regra para usar apenas uma delas: **análise estatísticas** costumam apresentar as **2 medidas**.

Quando apenas uma delas é usada:

- Pode ser por questões de brevidade/simplicidade; ou
- Alguém pode estar tentando persuadí-lo com a estatística.



Estimativas de Variabilidade

Qual é o Peso Médio?



Cenário 1: Vôo internacional

$$\overline{x_1} = 70\text{kg}$$



Cenário 2: Maratona

$$\overline{x_2} = 70\text{kg}$$

Qual é o Peso Médio?



Cenário 1: Vôo internacional



Cenário 2: Maratona

$$\overline{x}_1 = 70\text{kg}$$

Os pesos dos dois grupos têm (aproximadamente)
o mesmo centro/meio.

$$\overline{x}_2 = 70\text{kg}$$

Isso significa que tanto os **passageiros** do
vôo internacional quanto os
atletas da maratona
possuem pesos similares, correto?

Qual é o Peso Médio?



Cenário 1: Vôo internacional



Cenário 2: Maratona

$$\overline{x}_1 = 70\text{kg}$$

Os pesos dos dois grupos têm (aproximadamente)
o mesmo centro/meio.

$$\overline{x}_2 = 70\text{kg}$$

Isso significa que tanto os **passageiros** do
vôo internacional quanto os
atletas da maratona
possuem pesos similares, correto?

NÃO!

Por quê?

Qual é o Peso Médio?



Cenário 1: Vôo internacional



Cenário 2: Maratona

$$\overline{x}_1 = 70\text{kg}$$

Os pesos dos dois grupos têm (aproximadamente)
o mesmo centro/meio.

$$\overline{x}_2 = 70\text{kg}$$

Isso significa que tanto os **passageiros** do
vôo internacional quanto os
atletas da maratona
possuem pesos similares, correto?

NÃO!

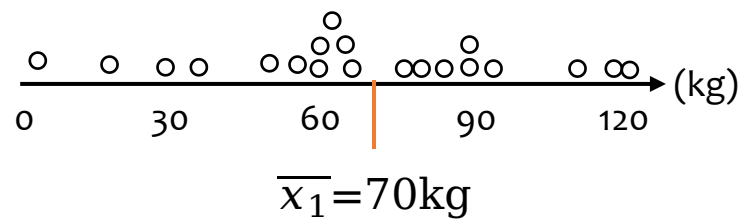
Por quê?

Os pesos dos **passageiros** são **mais espalhados/dispersos de sua média** do que os pesos dos **atletas**.

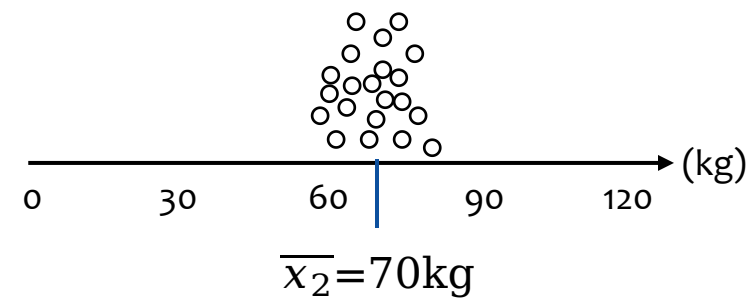
Há mais variabilidade nos pesos dos passageiros.

Dispersão dos Pesos

Cenário 1: Vôo internacional

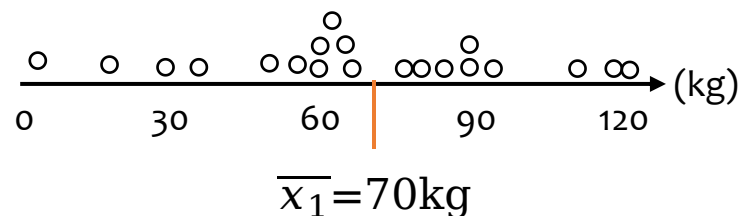


Cenário 2: Maratona

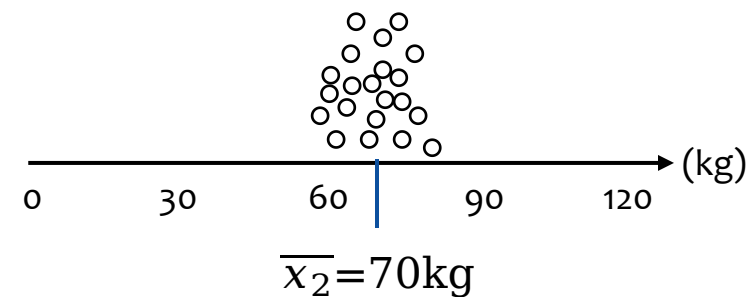


Dispersão dos Pesos

Cenário 1: Vôo internacional



Cenário 2: Maratona



Precisamos de maneiras para medir
essa **variabilidade/dispersão**

- Percentis e Quartis
- Mean absolute deviation
- Variance
- Standard Deviation

Percentis e Quartis

Estimativas de dispersão baseadas em **dados ordenados (ranqueados)**.

Percentis e Quartis

Estimativas de dispersão baseadas em **dados ordenados (ranqueados)**.

Percentis

O **k-ésimo percentil P_k** é o valor x_k em que **pelo menos k%** dos dados **são menores que x_k** ;

P. ex: O valor do P_{95} (95º percentil) indica que há 95% dos dados inferiores ao seu valor.

Percentis e Quartis

Estimativas de dispersão baseadas em **dados ordenados (ranqueados)**.

Percentis

O **k-ésimo percentil** P_k é o valor x_k em que **pelo menos k%** dos dados **são menores que x_k** ;

P. ex: O valor do P_{95} (95º percentil) indica que há 95% dos dados inferiores ao seu valor.

Quartis

Dividem a distribuição em quatro partes iguais de 25%:

- O 1º quartil Q_1 (ou P_{25}) separa os **25%** de dados inferiores;
- O 2º quartil Q_2 (ou P_{50}) separa os **50%** de dados inferiores --> **mediana**;
- O 3º quartil Q_3 (ou P_{75}) separa os **75%** de dados inferiores

Percentis e Quartis

Estimativas de dispersão baseadas em **dados ordenados (ranqueados)**.

Percentis

O **k-ésimo percentil P_k** é o valor x_k em que **pelo menos k%** dos dados **são menores que x_k** ;

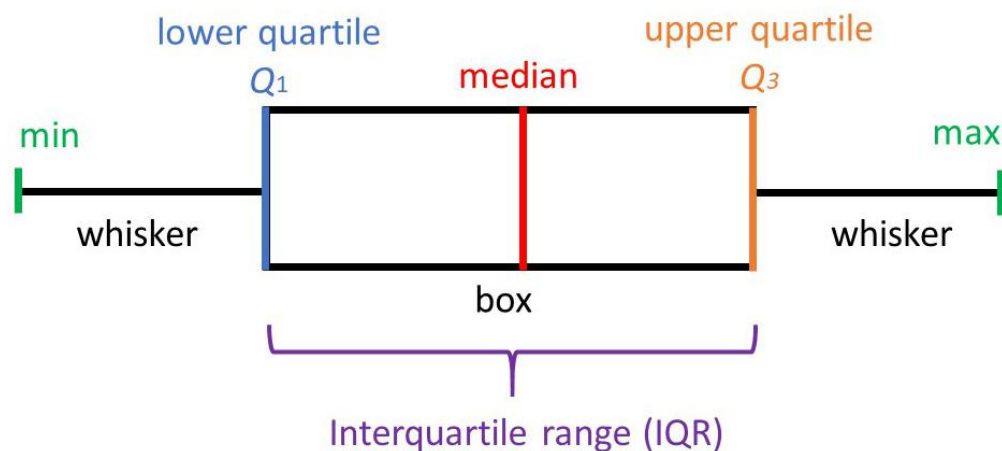
P. ex: O valor do P_{95} (95º percentil) indica que há 95% dos dados inferiores ao seu valor.

Quartis

Dividem a distribuição em quatro partes iguais de 25%:

- O 1º quartil Q_1 (ou P_{25}) separa os **25%** de dados inferiores;
- O 2º quartil Q_2 (ou P_{50}) separa os **50%** de dados inferiores --> **mediana**;
- O 3º quartil Q_3 (ou P_{75}) separa os **75%** de dados inferiores

Boxplot



Percentis e Quartis

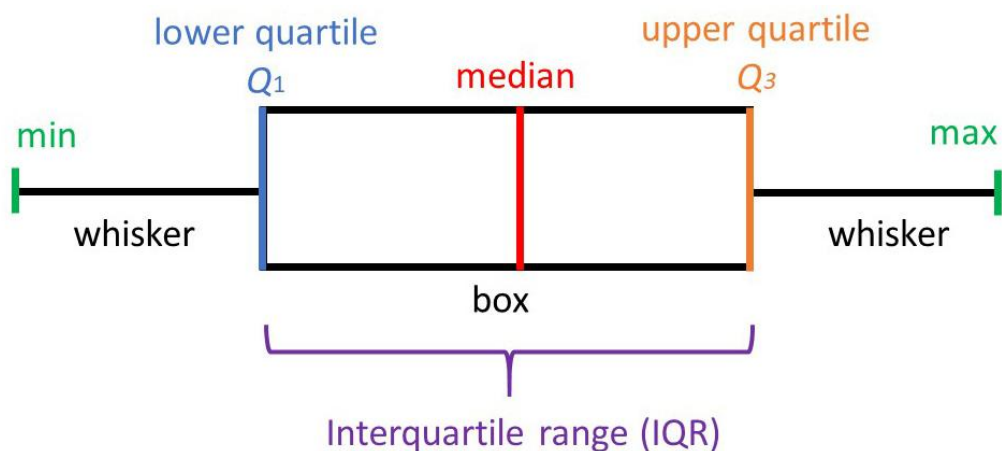
Estimativas de dispersão baseadas em **dados ordenados (ranqueados)**.

Percentis

O **k-ésimo percentil P_k** é o valor x_k em que **pelo menos k%** dos dados **são menores que x_k** ;

P. ex: O valor do P_{95} (95º percentil) indica que há 95% dos dados inferiores ao seu valor.

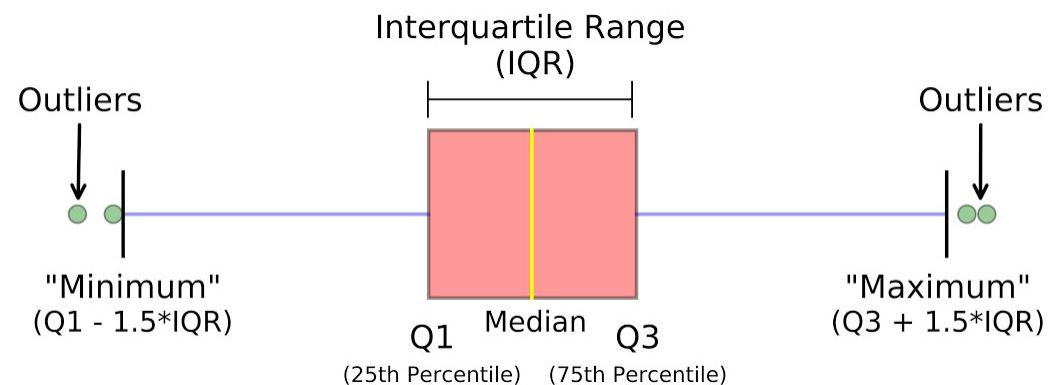
Boxplot



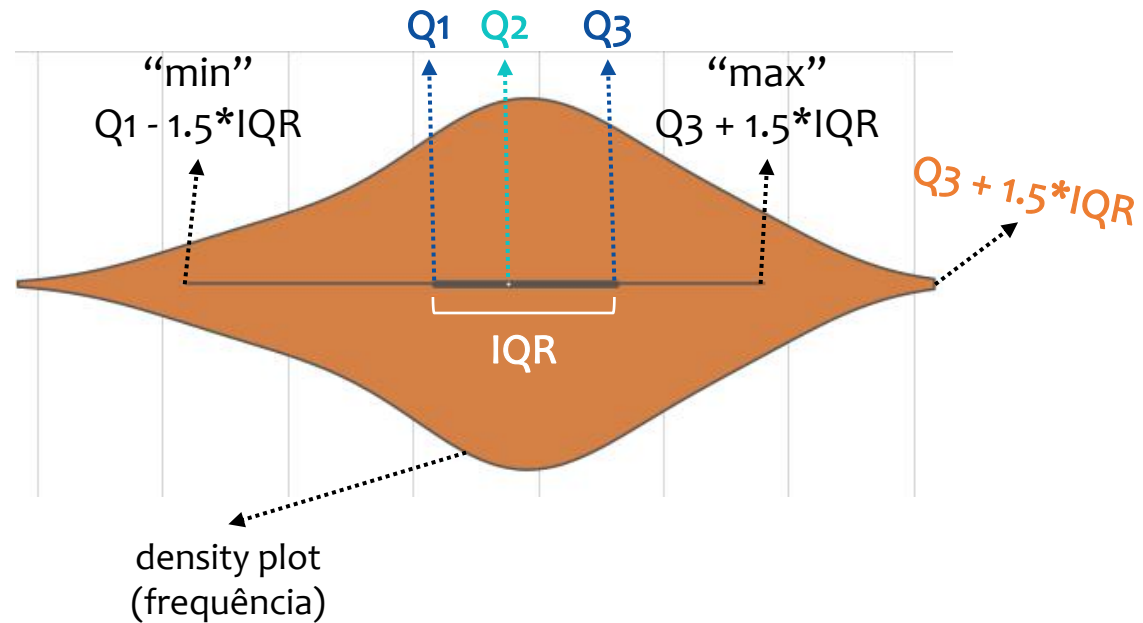
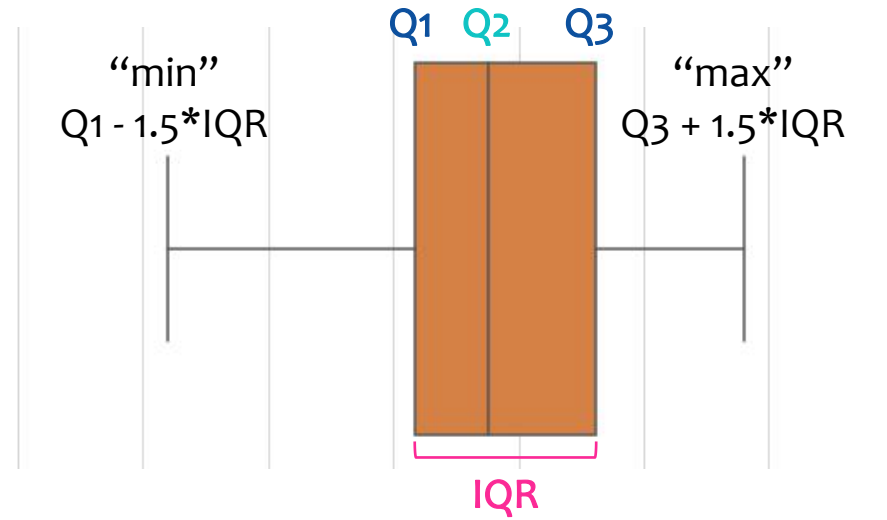
Quartis

Dividem a distribuição em quatro partes iguais de 25%:

- O 1º quartil Q_1 (ou P_{25}) separa os 25% de dados inferiores;
- O 2º quartil Q_2 (ou P_{50}) separa os 50% de dados inferiores --> **mediana**;
- O 3º quartil Q_3 (ou P_{75}) separa os 75% de dados inferiores

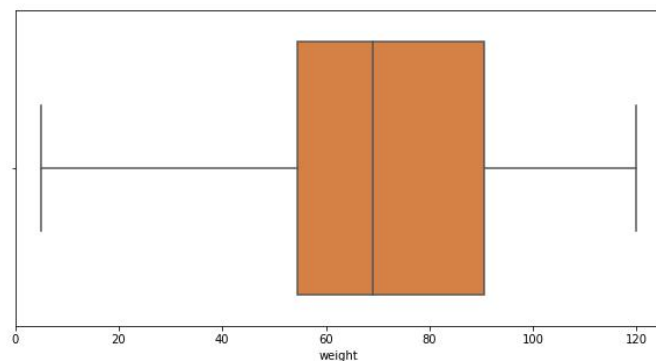
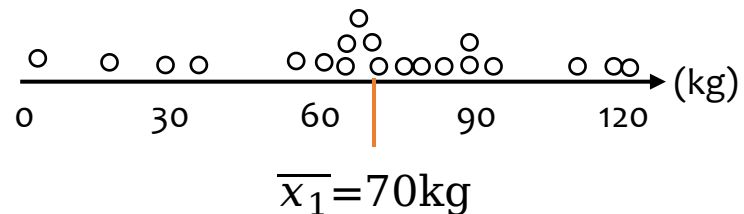


Violin Plot

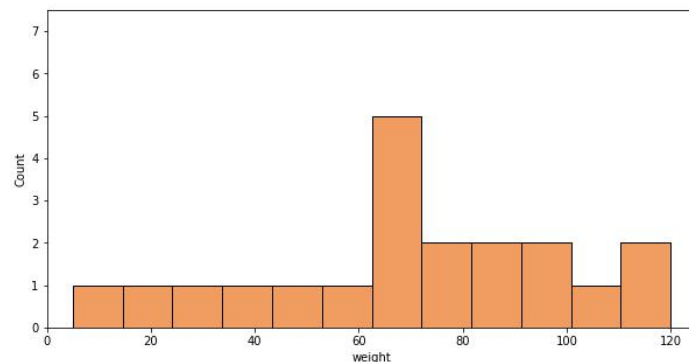


Dispersão dos Pesos

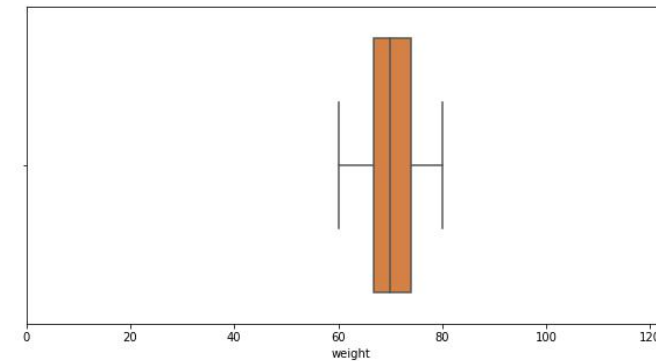
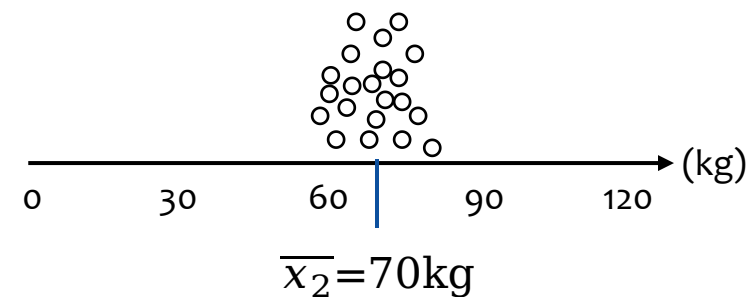
Cenário 1: Vôo internacional



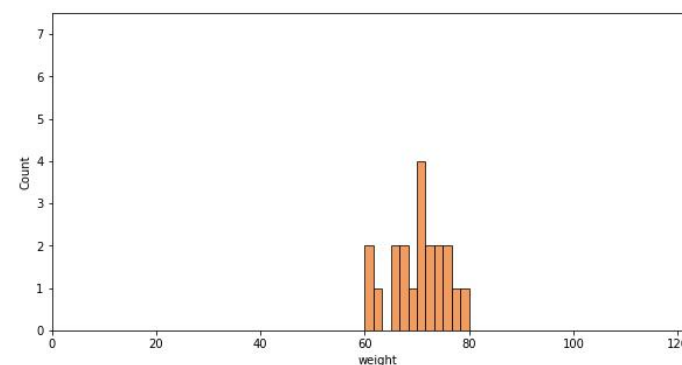
Q1: 54.5 kg
Q2: 69 kg (**mediana**)
Q3: 90.5



Cenário 2: Maratona

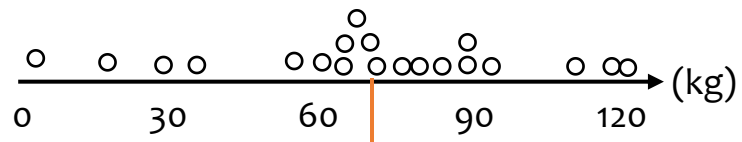


Q1: 66.75 kg
Q2: 70 kg (**mediana**)
Q3: 74.5

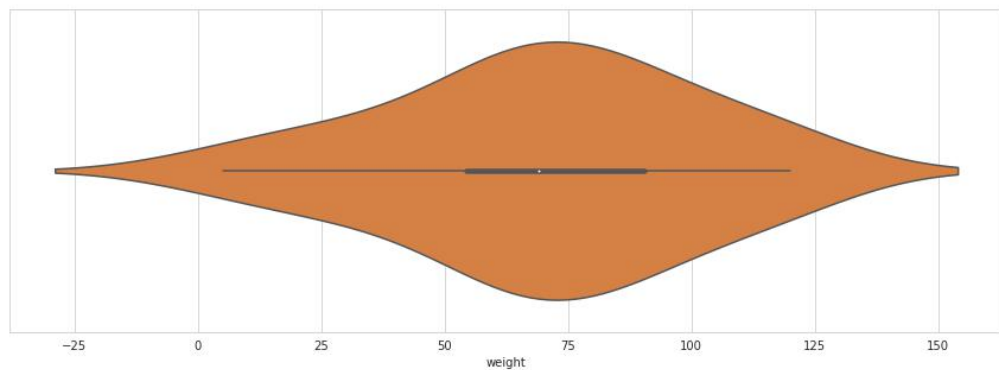


Dispersão dos Pesos

Cenário 1: Voo internacional

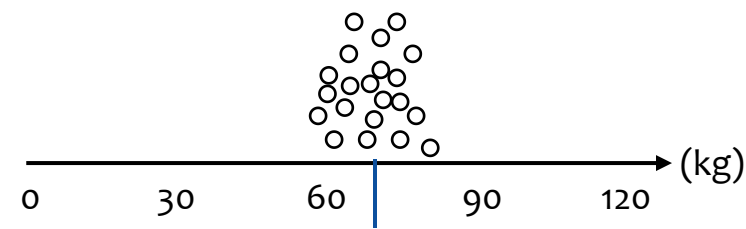


$$\overline{x}_1 = 70 \text{ kg}$$

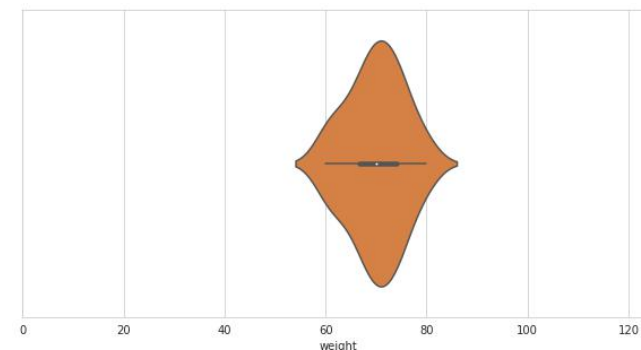


Q1: 54.5 kg
Q2: 69 kg (**mediana**)
Q3: 90.5

Cenário 2: Maratona



$$\overline{x}_2 = 70 \text{ kg}$$



Q1: 66.75 kg
Q2: 70 kg (**mediana**)
Q3: 74.5

Medidas de Dispersão

Mean Absolute Deviation (MAD)

$$MAE_x = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Variance

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

Standard Deviation

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

x_i : iésimo valor/registo

\bar{x} : peso do iésimo valor/registo

n : número de valores/registros

Medidas de Dispersão

Mean Absolute Deviation (MAD)

$$MAE_x = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Variance

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

Standard Deviation

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

Mais fáceis de interpretar por estarem na
mesma escala dos dados originais.

x_i : iésimo valor/registo
 \bar{x} : peso do iésimo valor/registo
 n : número de valores/registos

Medidas de Dispersão

Mean Absolute Deviation (MAD)

$$MAE_x = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Variance

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

Standard Deviation

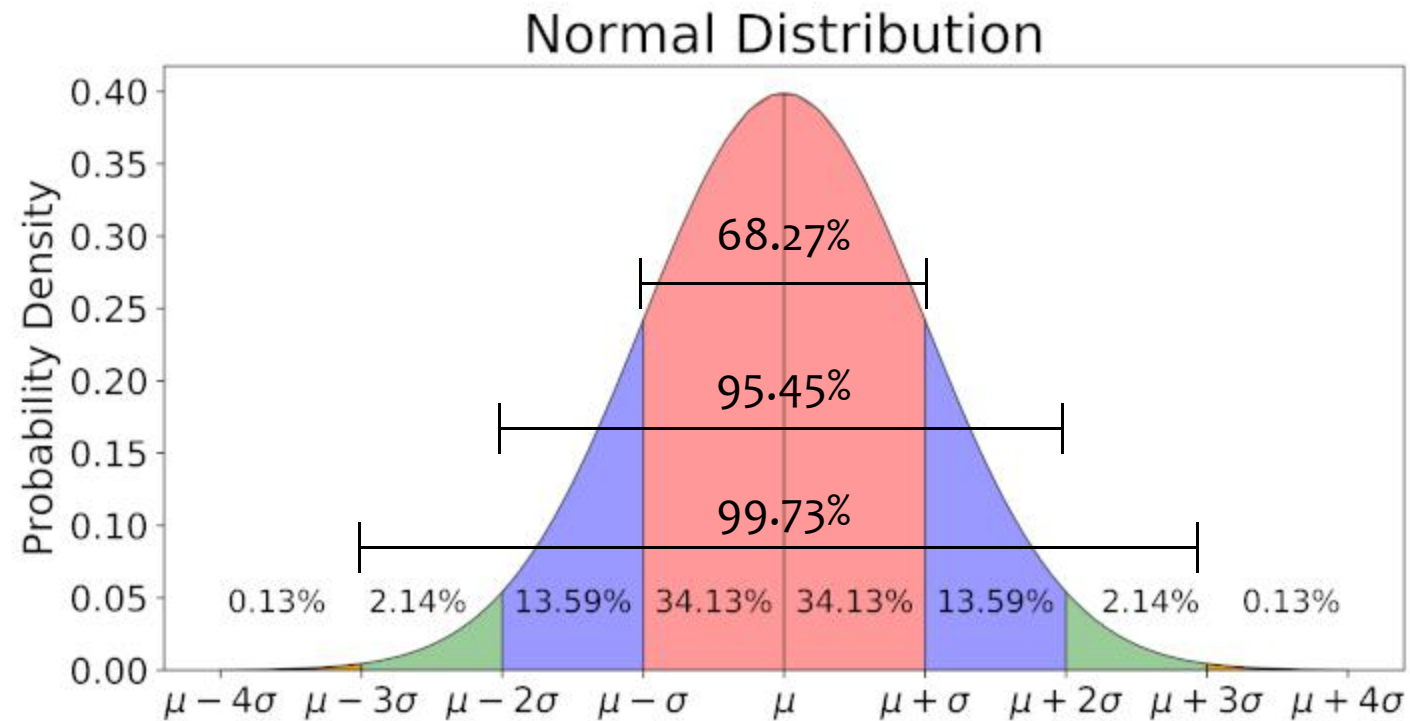
$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Mais fáceis de interpretar por estarem na **mesma escala** dos dados originais.

É **preferível** do que o MAD pois, matematicamente, é muito **mais conveniente** trabalhar com **valores ao quadrado** do que **valores absolutos**, especialmente, para modelos estatísticos.

x_i : iésimo valor/registo
 \bar{x} : peso do iésimo valor/registo
 n : número de valores/registros

Distribuição Normal (e a Regra 68-95-99.7)



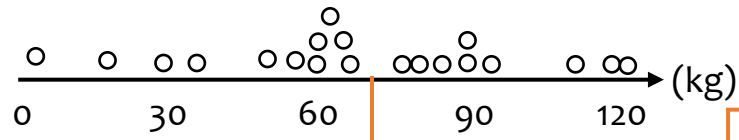
μ : média populacional

σ : desvio padrão

Dispersão dos Pesos

Cenário 1: Vôo internacional

Cenário 2: Maratona

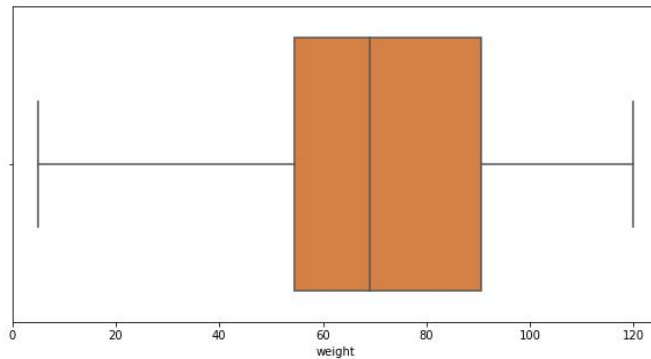


$\bar{x}_1 = 70\text{kg}$

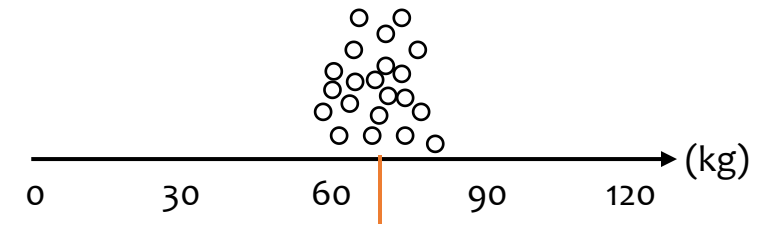
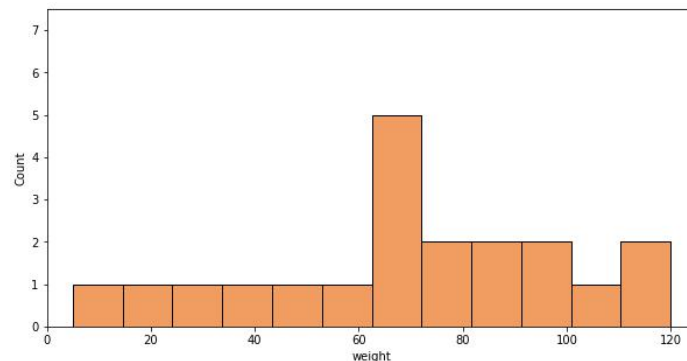
$$MAD_{x_1} = 23.7\text{kg}$$

$$\sigma_{x_1}^2 = 958\text{kg}$$

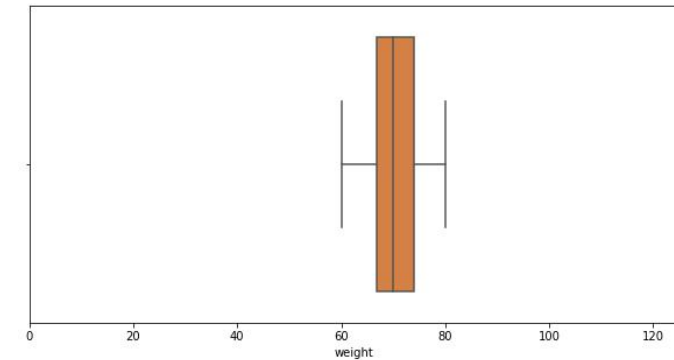
$$\sigma_{x_1} = 30.95\text{kg}$$



Q1: 54.5 kg
Q2: 69 kg (**mediana**)
Q3: 90.5



$\bar{x}_2 = 70\text{kg}$

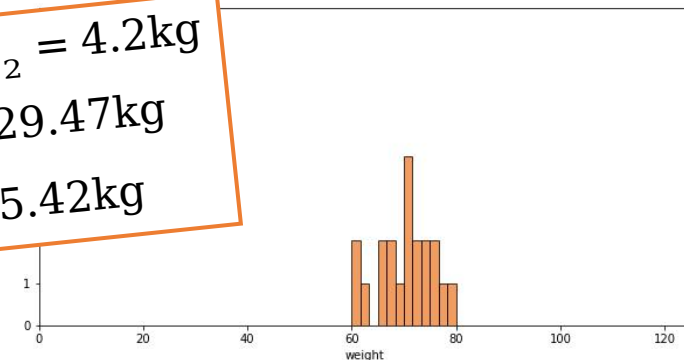


Q1: 66.75 kg
Q2: 70 kg (**mediana**)
Q3: 74.5

$$MAD_{x_2} = 4.2\text{kg}$$

$$\sigma_{x_2}^2 = 29.47\text{kg}$$

$$\sigma_{x_2} = 5.42\text{kg}$$



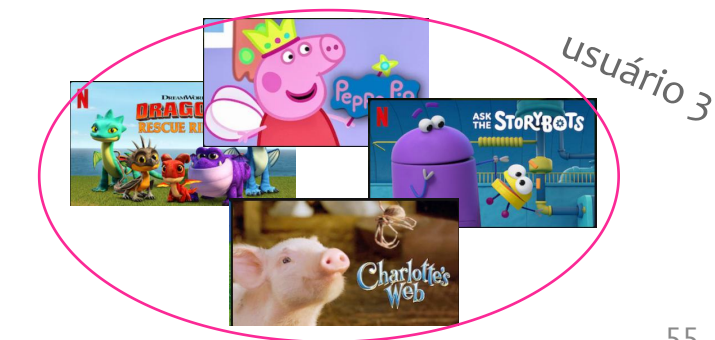
Correlação

Como a Netflix sabe quais filmes eu gosto?



Um sistema de recomendação simples

Filmes/Séries
que **eu** gostei / assisti muito



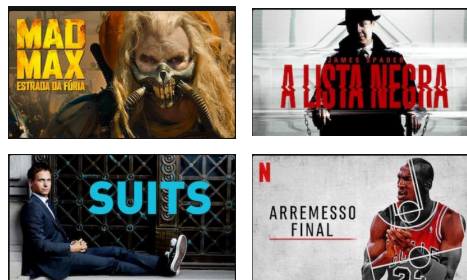
PS: Obviamente, o sistema de recomendações da Netflix é muito mais complexo e robusto do que esse =)

Um sistema de recomendação simples

Filmes/Séries
que **eu** gostei / assisti muito



recomendações



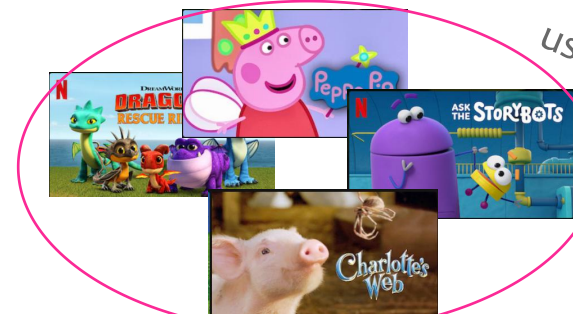
usuário 1



usuário 2



usuário 3



PS: Obviamente, o sistema de recomendações da Netflix é muito mais complexo e robusto do que esse =)

Um sistema de recomendação simples



PS: Obviamente, o sistema de recomendações da Netflix é muito mais complexo e robusto do que esse =)

Coeficiente de Correlação

Mede a **relação** (**associação linear**) entre duas variáveis dentro de uma **mesma escala métrica**.

Variáveis X e Y possuem **correlação positiva** se os **valores das variáveis** movem-se juntos:

- Ao aumentar os valores de X, os valores de Y também aumentam.

Variáveis X e Y possuem **correlação negativa** se os **valores das variáveis** movem-se em direções opostas:

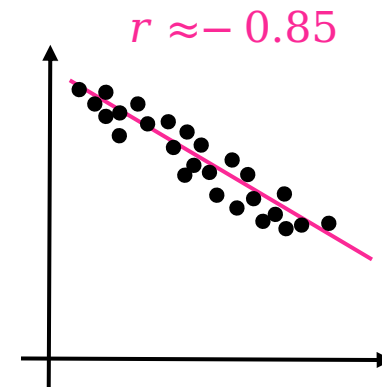
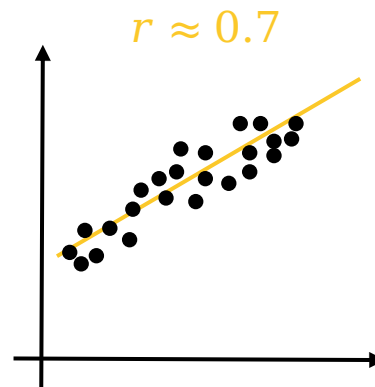
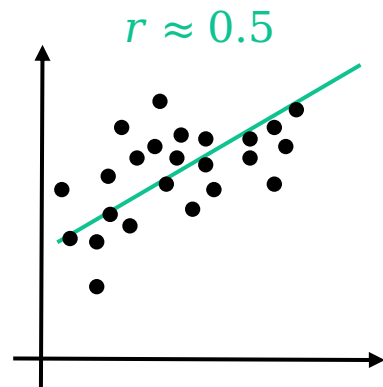
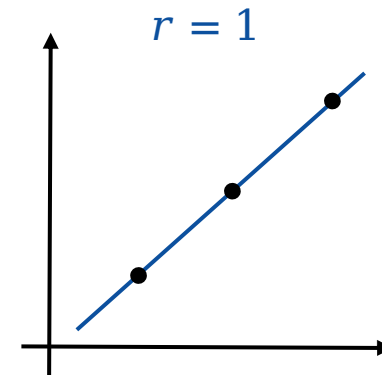
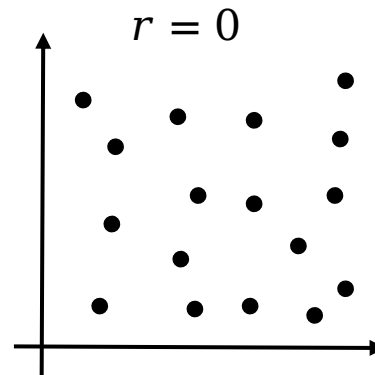
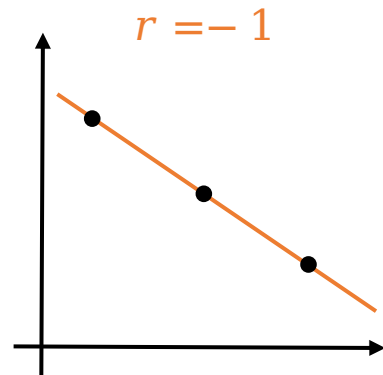
- Ao aumentar os valores de X, os valores de Y diminuem.

O **coeficiente de correlação** pode variar de **-1** (associação negativa perfeita) e **+1** (associação positiva perfeita).

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

Coeficiente de Correlação

scatterplots

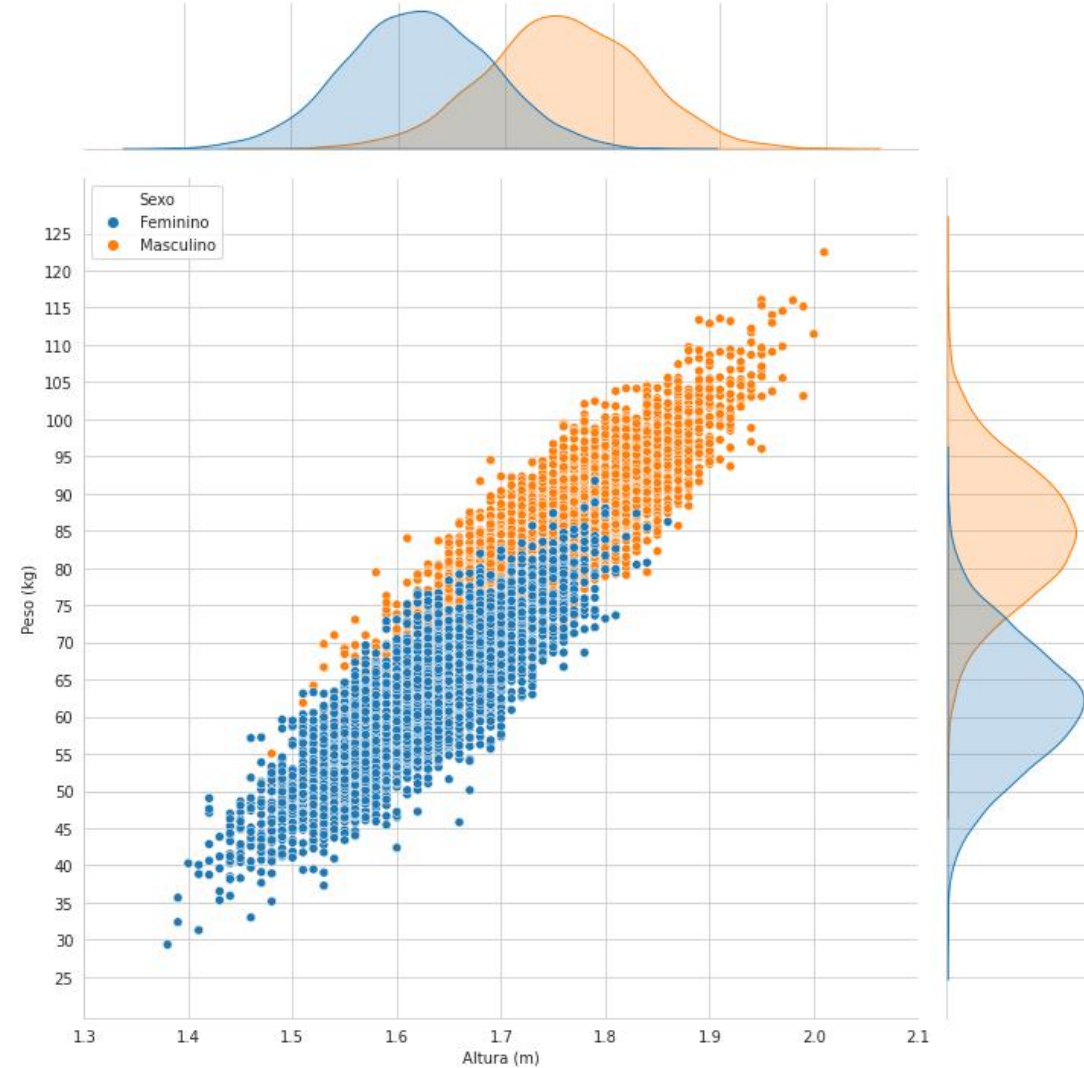


Ex: Peso x Altura

| | | m | kg |
|------|-----------|--------|--------|
| | Sexo | Altura | Peso |
| 0 | Masculino | 1.88 | 109.72 |
| 1 | Masculino | 1.75 | 73.62 |
| 2 | Masculino | 1.88 | 96.50 |
| 3 | Masculino | 1.82 | 99.81 |
| 4 | Masculino | 1.77 | 93.60 |
| ... | ... | ... | ... |
| 9995 | Feminino | 1.68 | 62.04 |
| 9996 | Feminino | 1.70 | 77.50 |
| 9997 | Feminino | 1.62 | 58.28 |
| 9998 | Feminino | 1.75 | 74.32 |
| 9999 | Feminino | 1.57 | 51.55 |

10000 rows x 3 columns

depois do
pré-processamento



Dataset: Weight-Height dataset: <https://www.kaggle.com/mustafaali96/weight-height>

Mão na Massa!

Faça uma Análise Exploratória no conjunto de dados “Gas Prices in Brazil”.

<https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil>

D1EAD – Análise Estatística para Ciência de Dados 2021.1



Análise Exploratória de Dados

Prof. Ricardo Sovat

sovat@ifsp.edu.br

Prof. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br

