# Estatística para Ciência de Dados


Profa. Rebeca Valgueiro

# Quem sou eu?

- Graduada em Engenharia Civil
- MBA em Gestão Empresarial
- Trabalho a mais de 4 anos no mercado de tecnologia atuando em projetos de:
  - Desenvolvimento web
  - Desenvolvimento desktop- windows
  - Data Science
  - IA e visão computacional

https://www.linkedin.com/in/rebecavalgueiro/

# Estatística Descritiva

**Imagine que você tem um conjunto de dados**

- **As alturas de todos os alunos da sua turma**
- **O número de vendas de um produto ao longo do último ano.**

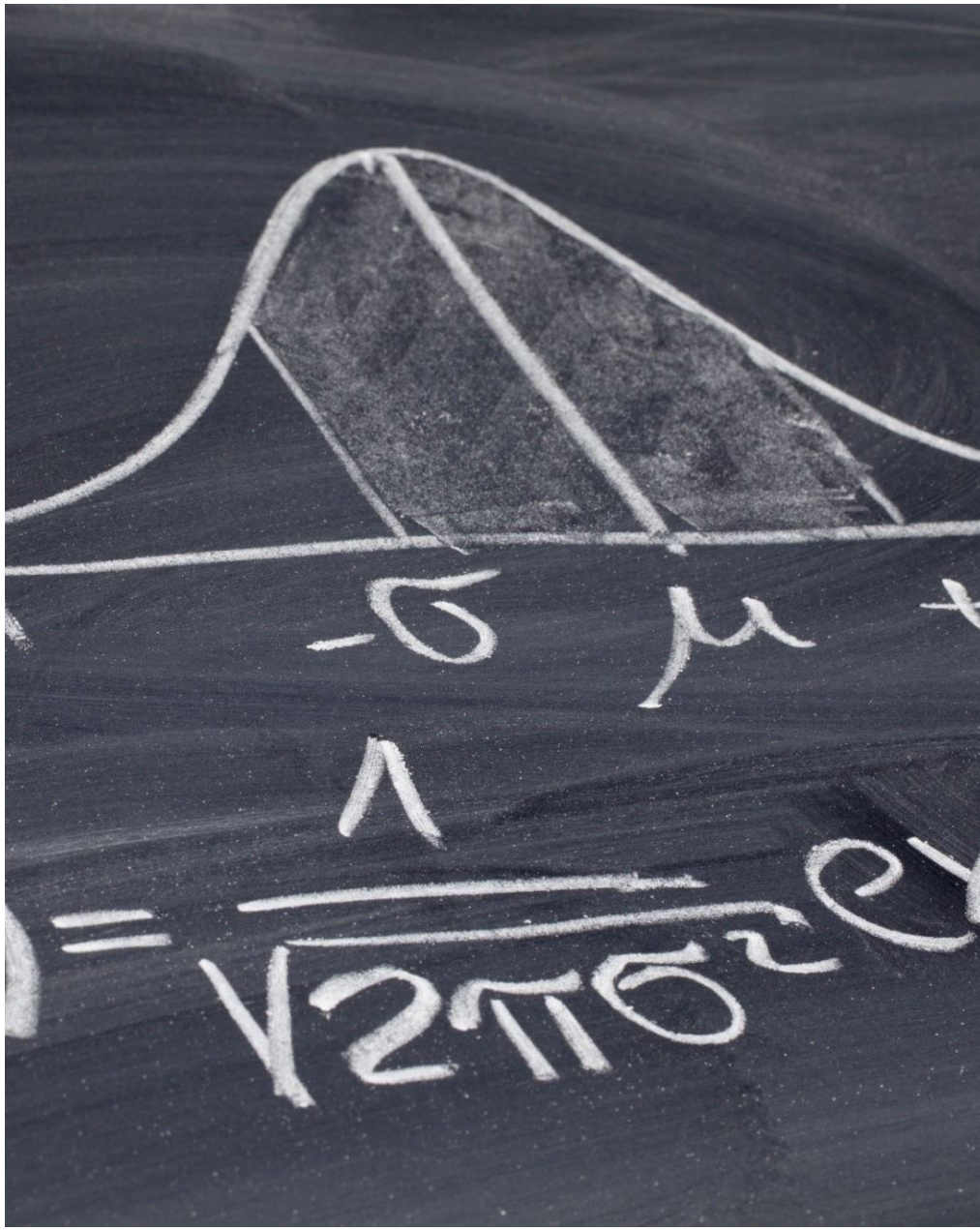
# Estatística Descritiva

**Conjunto de métodos que visam tornar os dados coletados mais fáceis de entender por meio de:**

- ➔ **Organização;**
- ➔ **Simplificação;**
- ➔ **Descrição e**
- ➔ **Apresentação de dados.**

**Usa tabelas, gráficos e medidas que resumem os dados brutos**

- Medidas de Tendência Central
  - Média
  - Mediana
  - Moda
- Medidas de Dispersão
  - Amplitude
  - Variância
  - Desvio Padrão
  - Percentis e Quartis
- Visualizações de Dados
  - Histogramas
  - Gráficos de Barras
  - Gráficos de Pizza
  - Box Plots (Diagramas de Caixa)
  - Tabelas de Frequência



# Medidas de Tendência Central

Servem para identificar um valor "típico" ou central em um conjunto de dados. São como um resumo rápido do nível geral dos seus dados.

# Média Aritmética (ou Média)

É a medida de tendência central mais comum e intuitiva. Ela é calculada somando todos os valores do conjunto de dados e dividindo essa soma pelo número total de valores.

$$\overline{X} = \frac{(X_1 + X_2 + X_3 + \ldots + X_n)}{n}$$

# Média Aritmética (ou Média)

## Vantagens:

- ➔ É fácil de calcular e entender.
- ➔ Utiliza todos os valores do conjunto de dados.
- ➔ É amplamente utilizada em diversas análises estatísticas.

## Desvantagens:

- ➔ É **muito sensível a valores extremos** (outliers). Um único valor muito alto ou muito baixo pode distorcer significativamente a média, fazendo com que ela não represente bem o centro da maioria dos dados.

# Média Aritmética (ou Média)

## Exemplo

Considere as notas de um aluno em 5 provas: 7, 8, 6, 9, 5.

A média seria:

$$\frac{7+8+6+9+5}{5} = \frac{35}{5} = 7$$

Nesse caso, a nota média do aluno é 7.

# Média Aritmética (ou Média)

**Exemplo -** Calcule a média salarial dos funcionários dessa empresa:

| | |
|---|---|
| Funcionário 1 | R$ 2.500 |
| Funcionário 2 | R$ 2.800 |
| Funcionário 3 | R$ 3.000 |
| Funcionário 4 | R$ 3.200 |
| Funcionário 5 | R$ 3.500 |
| SOMA | R$ 15.000 |

**MÉDIA =** $\frac{15000}{5} = R\$3.000$

# Média Aritmética (ou Média)

**Exemplo -** Um novo CEO foi contratado, qual a nova média?

| | |
|---|---|
| Funcionário 1 | R$ 2.500 |
| Funcionário 2 | R$ 2.800 |
| Funcionário 3 | R$ 3.000 |
| Funcionário 4 | R$ 3.200 |
| Funcionário 5 | R$ 3.500 |
| Funcionário 6 | R$ 25.000 |
| SOMA | R$ 40.000 |

**MÉDIA =** $\frac{40000}{6} \approx R\$6.666,67$

# Mediana

A mediana é o valor central de um conjunto de dados que foi ordenado do menor para o maior (ou do maior para o menor).

Ela divide o conjunto de dados em duas metades iguais.

1 - Ordene o seu conjunto de dados.

2 - Se o número de valores (n) for ímpar, a mediana é o valor que está exatamente no meio da lista ordenada.

3 - Se o número de valores (n) for par, a mediana é a média dos dois valores centrais da lista ordenada.

# Mediana

## Vantagens:

- ➔ Não é afetada por valores extremos (outliers).
- ➔ É útil para descrever a tendência central de dados assimétricos (onde a distribuição não é simétrica).

## Desvantagens:

- ➔ Não utiliza todos os valores do conjunto de dados no seu cálculo direto (apenas o(s) valor(es) central(is)).
- ➔ Pode ser mais trabalhosa de calcular para conjuntos de dados muito grandes que precisam ser ordenados.

# Mediana

**Exemplo 1 -** Considere as notas: 5, 6, 7, 8, 9.

MEDIANA = 7

**Exemplo 2 -** Considere as notas: 5, 6, 7, 8

MEDIANA = $\frac{6+7}{2} = 6.5$

# Moda

A moda é o valor (ou valores) que ocorre com maior frequência em um conjunto de dados.

A moda nos diz qual o valor mais comum no nosso conjunto de dados.

➔ Pode haver mais de uma moda (bimodal, trimodal, etc.) ou nenhuma moda (se todos os valores aparecerem a mesma quantidade de vezes).

# Moda

## Vantagens:

- ➔ É fácil de identificar.
- ➔ Pode ser usada para dados categóricos (não numéricos), como a cor de olhos mais comum em um grupo de pessoas.

## Desvantagens:

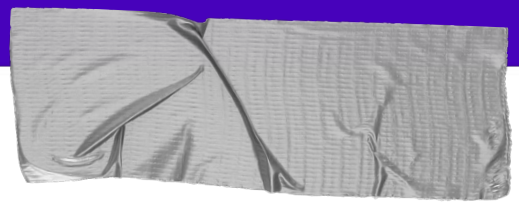
- ➔ Não utiliza todos os valores do conjunto de dados.
- ➔ Pode não ser uma boa representação da tendência central se a frequência mais alta for de um valor muito distante do resto dos dados.

# Moda

**Exemplo 1 (uma moda):** Considere as idades: 20, 22, 25, 22, 28, 22, 30.
A moda é 22, pois aparece 3 vezes, mais do que qualquer outra idade.

**Exemplo 2 (duas modas - bimodal):** Considere as alturas (em cm): 165, 170, 175, 170, 180, 165.
As modas são 165 e 170, pois ambas aparecem 2 vezes.

**Exemplo 3 (sem moda - amodal):** Considere os números: 1, 2, 3, 4, 5.
Não há moda, pois todos os números aparecem apenas uma vez.



# HORA DE PRATICAR!

# **Biblioteca -** Numpy


Pontos Fortes:

- Eficiência para operações numéricas: NumPy é otimizado para cálculos numéricos em arrays multidimensionais.
- Foco em arrays: NumPy é a base para muitas outras bibliotecas de ciência de dados em Python.

Considerações:

- Menos flexível para dados mistos ou rotulados: NumPy arrays são geralmente homogêneos (todos os elementos do mesmo tipo). Se seus dados contiverem diferentes tipos (por exemplo, strings e números) ou se você precisar de rótulos para seus dados, o Pandas oferece estruturas mais adequadas.

# **Biblioteca -** Numpy


Instalação pip install numpy

Importação import numpy as np

# **Biblioteca -** Numpy


| Característica | NumPy | Pandas |
|---|---|---|
| **Estrutura de Dados Principal** | Array N-dimensional | Series , DataFrame |
| **Tipos de Dados** | Homogêneo (um único tipo por array) | Flexível, pode conter múltiplos tipos de dados por coluna (inteiros, floats, strings, booleanos, etc.) |
| **Rótulos/Índices** | Indexação numérica implícita (baseada na posição) | Rótulos para linhas e colunas |
| **Tratamento de Dados Ausentes** | Menos suporte nativo (usa NaN para floats) | Suporte robusto para dados ausentes (NaN) |
| **Operações** | Focado em operações matemáticas e numéricas eficientes em arrays | Ampla gama de operações para manipulação, limpeza, transformação, agrupamento e análise de dados tabulares e séries temporais |
| **Analogia** | Uma caixa de ferramentas com ferramentas matemáticas poderosas para trabalhar com números | Uma planilha avançada ou um banco de dados em memória, com funcionalidades para organizar, limpar e analisar dados |

# **Biblioteca -** Numpy


**np.mean(a):** Calcula a **média aritmética** dos elementos do array a.

```
import numpy as np
data = np.array([1, 2, 3, 4, 5])
media = np.mean(data)  # Resultado: 3.0
```

**np.median(a):** Calcula a **mediana** dos elementos do array a.

```
data = np.array([1, 2, 3, 4, 10])
mediana = np.median(data)  # Resultado: 3.0
```

**Não há uma função direta para a moda na biblioteca base do NumPy:** No entanto, para uma análise mais direta da moda, o Pandas (com sua função pd.Series(a).mode()) ou scipy.stats.mode() são geralmente mais convenientes.

```
lista_ocorrencias, contagem = np.unique(dados, return_counts = True)
moda = lista_ocorrencias[contagem == np.max(contagem)]
print(moda)
```

# **Biblioteca -** Pandas

**`.mean()`:** Calcula a média aritmética dos valores na Series ou nas colunas do DataFrame.

```
data = pd.Series([1, 2, 3, 4, 5])
media = data.mean()  # Resultado: 3.0
```

**`.median()`:** Calcula a mediana dos valores na Series ou nas colunas do DataFrame.

```
data = pd.Series([1, 2, 3, 4, 10])
mediana = data.median()  # Resultado: 3.0
```

**`.mode()`:** Retorna a(s) moda(s) dos valores na Series ou nas colunas do DataFrame. Pode haver múltiplos valores de moda, então o resultado é uma Series.

```
data = pd.Series([1, 2, 2, 3, 3, 3])
moda = data.mode()  # Resultado: Series([3], dtype: int64)
```

# **Exercício 01** - Medidas de Tendência Central

Uma pesquisa coletou o número de passageiros em 10 ônibus diferentes no Terminal Integrado de Passageiros (TIP) em um determinado horário de pico:

**[45, 52, 38, 60, 48, 55, 40, 58, 50, 42]**

a) Calcule a média do número de passageiros.

b) Calcule a mediana do número de passageiros.

```python
import numpy as np

# Dados do número de passageiros
passageiros = np.array([45, 52, 38, 60, 48, 55, 40, 58, 50, 42])

# a) Calcular a média usando NumPy
media_passageiros = np.mean(passageiros)
print("A média do número de passageiros é: ", media_passageiros)

# b) Calcular a mediana usando NumPy
mediana_passageiros = np.median(passageiros)
print("A mediana do número de passageiros é: ",mediana_passageiros)
```

# **Exercício 02** - Medidas de Tendência Central

As temperaturas máximas (em graus Celsius) registradas em 7 dias consecutivos no Recife foram:

**[31.5, 32.0, 31.8, 32.5, 31.5, 33.0, 32.0]**

a) Calcule a média

b) Calcule a mediana

c) Identifique a moda

```python
import numpy as np
import pandas as pd

# Dados das temperaturas máximas
temperaturas_np = np.array([31.5, 32.0, 31.8, 32.5, 31.5, 33.0, 32.0])

# a) Calcular a média da temperatura usando NumPy
media_temperatura = np.mean(temperaturas_np)
print("A média da temperatura é: ", media_temperatura)

# b) Calcular a mediana da temperatura usando NumPy
mediana_temperatura = np.median(temperaturas_np)
print(f"A mediana da temperatura é: ",mediana_temperatura)

# c) Identificar a moda da temperatura usando Pandas
temperaturas_series = pd.Series(temperaturas_np)
moda_temperatura_pandas = temperaturas_series.mode()
print("A moda da temperatura (usando Pandas) é: ", moda_temperatura_pandas.tolist())
```

# **Exercício 03** - Medidas de Tendência Central

Um cientista de dados analisou o tempo de carregamento (em segundos) de um website acessado por 15 usuários em diferentes bairros do Grande Recife:

**[2.1, 2.5, 1.8, 3.0, 2.2, 2.5, 2.8, 2.1, 2.3, 2.5, 3.5, 2.2, 2.7, 2.4, 2.5]**

a) Calcule a média
b) Calcule a mediana
c) Identifique a moda usando NumPy

# **Exercício 03** - Medidas de Tendência Central

```python
import numpy as np

# Dados do tempo de carregamento
tempo_carregamento = np.array([2.1, 2.5, 1.8, 3.0, 2.2, 2.5, 2.8, 2.1, 2.3, 2.5, 3.5, 2.2, 2.7, 2.4, 2.5])

# a) Calcular a média do tempo de carregamento
media_tempo = np.mean(tempo_carregamento)
print("A média do tempo de carregamento é: ", media_tempo)

# b) Calcular a mediana do tempo de carregamento
mediana_tempo = np.median(tempo_carregamento)
print(f"A mediana do tempo de carregamento é: ", mediana_tempo)

# c) Identificar a moda do tempo de carregamento usando NumPy
valores_unicos, contagens = np.unique(tempo_carregamento, return_counts=True)
modas = valores_unicos[contagens == np.max(contagens)]
print("As modas do tempo de carregamento são: ", modas)
```

# **Exercício 04 -** Medidas de Tendência Central

Uma pesquisa de opinião sobre a satisfação dos moradores de Olinda com os serviços públicos utilizou uma escala de 1 (muito insatisfeito) a 5 (muito satisfeito). As respostas de 20 participantes foram:

**[3, 4, 4, 2, 5, 4, 3, 3, 4, 5, 2, 4, 3, 4, 4, 5, 3, 2, 4, 3]**

a) Crie uma Series do Pandas com esses dados.
b) Calcule a média da satisfação usando o método da Series.
c) Calcule a mediana da satisfação usando o método da Series.
d) Identifique a moda da satisfação usando o método da Series.

# **Exercício 04 -** Medidas de Tendência Central

```python
import pandas as pd

# Dados das respostas sobre satisfação
satisfacao = [3, 4, 4, 2, 5, 4, 3, 3, 4, 5, 2, 4, 3, 4, 4, 5, 3, 2, 4, 3]

# a) Crie uma Series do Pandas com esses dados.
series_satisfacao = pd.Series(satisfacao)

# b) Calcule a média da satisfação usando o método da Series.
media_satisfacao = series_satisfacao.mean()
print("A média da satisfação é: ", media_satisfacao)

# c) Calcule a mediana da satisfação usando o método da Series.
mediana_satisfacao = series_satisfacao.median()
print("A mediana da satisfação é:", mediana_satisfacao)

# d) Identifique a moda da satisfação usando o método da Series.
moda_satisfacao = series_satisfacao.mode()
print("A(s) moda(s) da satisfação é(são):",moda_satisfacao.tolist())
```

# Medidas de Dispersão

Servem para nos mostrar o quão "espalhados" ou "variáveis" esses dados estão ao redor desse centro. Elas nos dão uma noção da homogeneidade ou heterogeneidade do conjunto de dados.

# Medidas de Dispersão

**X** = {50, 50, 50, 50, 50} > média = 50
**Y** = {48, 49, 50, 51, 52} > média = 50
**X** = {10, 20, 50, 80, 90} > média = 50

# Amplitude (ou Alcance)

A amplitude é a medida de dispersão mais simples. Ela é calculada como a diferença entre o maior e o menor valor do conjunto de dados.

$$\text{Amplitude} = \text{Valor Máximo} - \text{Valor Mínimo}$$

# Amplitude (ou Alcance)

## Vantagens:

- ➔ É extremamente fácil de calcular e entender.

## Desvantagens:

- ➔ É muito sensível a outliers, pois apenas os valores extremos são considerados. Um único valor muito alto ou muito baixo pode inflacionar a amplitude sem refletir a variabilidade da maioria dos dados.
- ➔ Não leva em consideração a distribuição dos valores entre o máximo e o mínimo.

# Amplitude (ou Alcance)

**Exemplo 1 -** Considere os valores: 5, 6, 7, 8, 9.

A amplitude é $9-5=4$

**Exemplo 2 -** Considere os valores: 5, 6, 7, 8, 9, 15.

A amplitude é $15-5=10$

# Variância

A variância mede o quão longe cada número no conjunto de dados está da média. Uma variância maior indica uma maior dispersão dos dados em torno da média.

É a média dos quadrados das diferenças entre cada valor e a média.

Variância Populacional ($\sigma^2$)

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

Variância Amostral ($s^2$)

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

# Variância

## Variância Populacional ($\sigma^2$)

**População:** 4 estudantes. As idades deles são: 20, 22, 24, 26 anos.

$$\mu = \frac{20 + 22 + 24 + 26}{4} = \frac{92}{4} = 23 \text{ anos}$$

$$\sigma^2 = \frac{(20 - 23)^2 + (22 - 23)^2 + (24 - 23)^2 + (26 - 23)^2}{4}$$

$$\sigma^2 = \frac{(-3)^2 + (-1)^2 + (1)^2 + (3)^2}{4}$$

$$\sigma^2 = \frac{9 + 1 + 1 + 9}{4} = \frac{20}{4} = 5 \text{ anos}^2$$

## Variância Amostral ($s^2$)

**Amostra: 20, 22, 24 anos**

$$\bar{x} = \frac{20 + 22 + 24}{3} = \frac{66}{3} = 22 \text{ anos}$$

$$s^2_{\text{incorreta}} = \frac{(20 - 22)^2 + (22 - 22)^2 + (24 - 22)^2}{3}$$

$$s^2_{\text{incorreta}} = \frac{(-2)^2 + 0^2 + (2)^2}{3} = \frac{4 + 0 + 4}{3} = \frac{8}{3} \approx 2.67 \text{ anos}^2$$

$$s^2_{\text{correta}} = \frac{(-2)^2 + 0^2 + (2)^2}{2} = \frac{4 + 0 + 4}{2} = \frac{8}{2} = 4 \text{ anos}^2$$

# Variância

## Vantagens:

- ➔ Considera todos os valores do conjunto de dados.
- ➔ É fundamental para muitas técnicas estatísticas.

## Desvantagens:

- ➔ A unidade da medida é elevada ao quadrado, dificultando a interpretação em relação aos dados originais.
- ➔ É sensível a outliers, pois as diferenças em relação à média são elevadas ao quadrado, ampliando o efeito de valores extremos.

# Variância

**Exemplo -** Considere as notas: 6, 7, 8

$$s^2 = \frac{(6-7)^2 + (7-7)^2 + (8-7)^2}{3-1} = \frac{(-1)^2 + 0^2 + 1^2}{2} = \frac{1+0+1}{2} = 1$$

# Desvio padrão

É a medida de dispersão mais utilizada e interpretável. Ele é simplesmente a raiz quadrada da variância.

Desvio padrão pequeno → dados estão mais agrupados perto da média

Desvio padrão grande → maior dispersão.

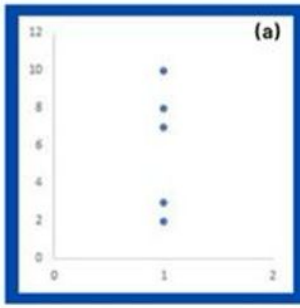
Desvio Padrão Populacional (σ)

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

Desvio Padrão Amostral (s)

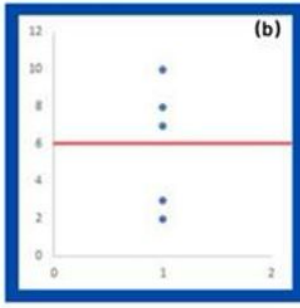
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

# Desvio Padrão



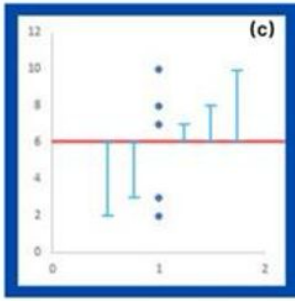

*Vamos pegar o exemplo do seguinte conjunto de dados:*

**2, 3, 7, 8, 10**



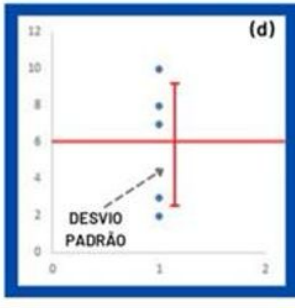

*Primeiro, calculamos a **média** dos dados.*

A **média** deste conjunto é **6**




Em seguida, calculamos o **desvio** de cada ponto da média: subtraindo a média de cada ponto:

**2**-6 = -4, **3**-6 = -3, **7**-6 = 1, **8**-6 = 2, **10**-6 = 4




Finalmente, calculamos o desvio padrão tomando [um tipo de] média desses desvios. Isso nos dá uma ideia de quanto os dados desviam da média, e assim, quanta variação há no conjunto de dados.

# Desvio Padrão

## Vantagens:

- ➔ Está na mesma unidade dos dados originais, facilitando a interpretação.
- ➔ Considera todos os valores do conjunto de dados.
- ➔ É amplamente utilizado em análises estatísticas.

## Desvantagens:

- ➔ Também é sensível a outliers (embora menos que a variância, já que a raiz quadrada "suaviza" o efeito dos valores elevados ao quadrado).

# Intervalo Interquartil (IIQ)

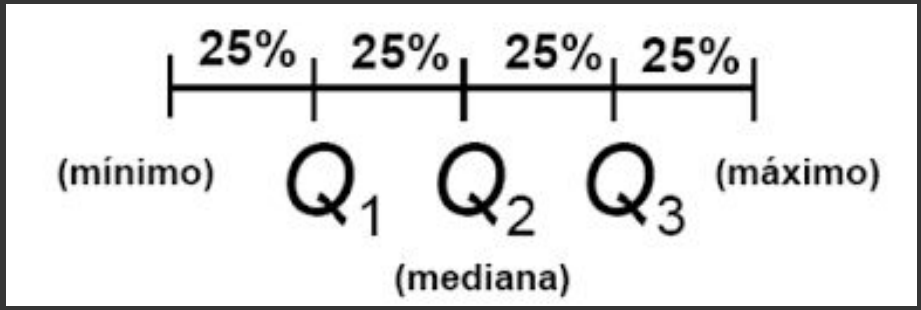
Medida de dispersão que foca na variabilidade da metade central dos seus dados. Diferentemente da amplitude, que é muito sensível a outliers, o IIQ nos dá uma visão mais robusta da dispersão, pois ignora os 25% inferiores e os 25% superiores dos dados.

$$\text{IIQ} = Q3 - Q1$$

# Intervalo Interquartil (IIQ)

**Quartis:**

- Primeiro Quartil (Q1): É o valor abaixo do qual se encontram 25% dos dados quando estes estão ordenados do menor para o maior.
- Segundo Quartil (Q2): É o valor abaixo do qual se encontra 50% dos dados. O Q2 é, na verdade, a mediana do conjunto de dados.
- Terceiro Quartil (Q3): É o valor abaixo do qual se encontra 75% dos dados.



# Intervalo Interquartil (IIQ)

## Vantagens:

- ➔ Não é afetado por valores extremos (outliers). Como ele se concentra nos 50% centrais dos dados, valores muito altos ou muito baixos não influenciam seu cálculo.
- ➔ IIQ é uma boa medida de dispersão para distribuições que não são simétricas (distribuições enviesadas), pois ele descreve a dispersão da parte central dos dados

## Desvantagens:

- ➔ Por focar na metade central, o IIQ não leva em consideração a dispersão dos 25% inferiores e dos 25% superiores dos dados. Portanto, ele não fornece uma imagem completa da variabilidade total do conjunto de dados.
- ➔ Para dados que seguem uma distribuição normal e não possuem outliers significativos, o desvio padrão geralmente é uma medida de dispersão mais eficiente, pois utiliza todos os valores do conjunto de dados.

# Intervalo Interquartil (IIQ)

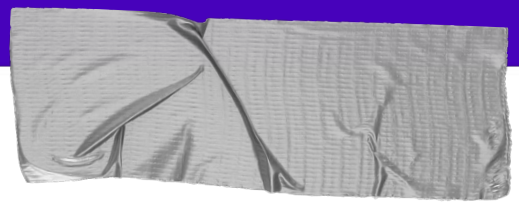
## Exemplo

**Vamos considerar o seguinte conjunto de dados de notas (ordenadas):**

**4, 5, 6, 7, 7, 8, 9, 10, 15**

- ➔ Q1 (25%): O valor na posição (9+1)*1/4 = 2.5 → média entre os valores nas posições 2 e 3 → 5.5
- ➔ Q2 (Mediana, 50%): O valor na posição (9+1)*2/4 =5 → 7
- ➔ Q3 (75%): O valor na posição (9+1)*3/4 = 7.5 → média entre os valores nas posições 7 e 8 → 9.5

$$IIQ = Q3 - Q1 = 9.5 - 5.5 = 4$$



# HORA DE PRATICAR!

# **Biblioteca -** Numpy


**`np.std(a)`:** Calcula o **desvio padrão** dos elementos do array `a`. Por padrão, calcula o desvio padrão populacional. Para calcular o desvio padrão da amostra (com a correção de Bessel), use o argumento `ddof=1`.

```
data = np.array([1, 2, 3, 4, 5])
desvio_padrao_populacional = np.std(data)       # Resultado: 1.414...
desvio_padrao_amostral = np.std(data, ddof=1)  # Resultado: 1.581…
```

**`np.var(a)`:** Calcula a **variância** dos elementos do array `a`. Similar ao `np.std()`, por padrão calcula a variância populacional. Use `ddof=1` para a variância da amostra.

```
data = np.array([1, 2, 3, 4, 5])
variancia_populacional = np.var(data)       # Resultado: 2.0
variancia_amostral = np.var(data, ddof=1)   # Resultado: 2.5
```

**`np.ptp(a)`:** Calcula a **amplitude** (range) dos elementos do array `a` (valor máximo - valor mínimo).

```
data = np.array([1, 2, 3, 4, 10])
amplitude = np.ptp(data)  # Resultado: 9
```

# **Biblioteca -** Numpy


**`np.percentile(a, q)`:** Calcula o **percentil** especificado `q` dos elementos do array `a`. `q` pode ser um único valor (entre 0 e 100).

- **Primeiro Quartil (Q1):** `np.percentile(a, 25)`
- **Segundo Quartil (Q2) - Mediana:** `np.percentile(a, 50)` (equivalente a `np.median(a)`)
- **Terceiro Quartil (Q3):** `np.percentile(a, 75)`

```
data = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
q1 = np.percentile(data, 25)  # Resultado: 3.25
q3 = np.percentile(data, 75)  # Resultado: 7.75
```

**Intervalo Interquartil (IIQ):** Pode ser calculado combinando `np.percentile()`:

```
data = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)
iiq = q3 - q1  # Resultado: 4.5
```

# **Biblioteca -** Pandas

**`.std()`:** Calcula o desvio padrão dos valores na Series ou nas colunas do DataFrame. Por padrão, calcula o desvio padrão da amostra (usa `ddof=1`). Para calcular o desvio padrão populacional, use o argumento `ddof=0`.

```
data = pd.Series([1, 2, 3, 4, 5])
desvio_padrao_amostral = data.std()  # Resultado: 1.5811388300841898
desvio_padrao_populacional = data.std(ddof=0)  # Resultado: 1.4142135623730951
```

**`.var()`:** Calcula a variância dos valores na Series ou nas colunas do DataFrame. Similar ao `.std()`, por padrão calcula a variância da amostra (`ddof=1`). Use `ddof=0` para a variância populacional.

```
data = pd.Series([1, 2, 3, 4, 5])
variancia_amostral = data.var()  # Resultado: 2.5
variancia_populacional = data.var(ddof=0)  # Resultado: 2.0
```

# **Biblioteca** - Pandas

**`.max()`:** Retorna o valor máximo. (para cálculo de amplitude)
**`.min()`:** Retorna o valor mínimo. para cálculo de amplitude)
**Amplitude:** calculado pela subtração do max pelo mínimo

**`.quantile(q=0.5)`:** Calcula o quantil no valor especificado de `q` (entre 0 e 1). O valor padrão de `q` é 0.5 (a mediana).
Primeiro Quartil (Q1): `.quantile(0.25)`
Terceiro Quartil (Q3): `.quantile(0.75)`

```
data = pd.Series([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
q1 = data.quantile(0.25)  # Resultado: 3.25
q3 = data.quantile(0.75)  # Resultado: 7.75
```

**Intervalo Interquartil (IIQ):** Pode ser calculado combinando `.quantile()`:

# **Biblioteca -** Pandas X NumPy

| Medida | Pandas | NumPy |
|---|---|---|
| Média | .mean() | np.mean() |
| Mediana | .median() | np.median() |
| Moda | .mode() (retorna uma Series) | - |
| Desvio Padrão | .std(ddof=1) (amostral, padrão) | np.std(a, ddof=0) (populacional, padrão) |
| | .std(ddof=0) (populacional) | np.std(a, ddof=1) (amostral) |
| Variância | .var(ddof=1) (amostral, padrão) | np.var(a, ddof=0) (populacional, padrão) |
| | .var(ddof=0) (populacional) | np.var(a, ddof=1) (amostral) |
| Amplitude (Range) | .max() - .min() | np.ptp() |
| Percentis/Quartis | .quantile(q) (q entre 0 e 1) | np.percentile(a, q) (q entre 0 e 100) |
| Intervalo Interquartil (IIQ) | .quantile(0.75) - .quantile(0.25) | np.percentile(a, 75) - np.percentile(a, 25) |

# **Exercício 01** - Medidas de Dispersão

Um analista de dados registrou o número de acidentes de trânsito por dia em um cruzamento movimentado da Avenida Agamenon Magalhães durante 10 dias:

**[2, 0, 1, 3, 2, 2, 0, 1, 2, 4]**

a) Calcule a amplitude do número de acidentes

b) Calcule o desvio padrão amostral do número de acidentes

```python
import numpy as np

# Dados do número de acidentes
acidentes = np.array([2, 0, 1, 3, 2, 2, 0, 1, 2, 4])

# a) Calcule a amplitude do número de acidentes usando NumPy.
amplitude_numpy = np.ptp(acidentes)
print("A amplitude do número de acidentes é: ", amplitude_numpy)

# b) Calcule o desvio padrão amostral do número de acidentes usando Pandas.
desvio_padrao = np.std(acidentes)
print("Desvio padrão do número de acidentes é: ", desvio_padrao)
```

# **Exercício 02** - Medidas de Dispersão

As notas de 8 alunos de uma turma de estatística da UFPE na primeira avaliação foram:

**[7.5, 8.0, 6.5, 9.0, 7.0, 8.5, 6.0, 7.5]**

a) Calcule a variância populacional e amostral das notas

b) Calcule o intervalo interquartil (IIQ) das notas.

# **Exercício 02** - Medidas de Dispersão

```python
import numpy as np

# Dados das notas dos alunos
notas = np.array([7.5, 8.0, 6.5, 9.0, 7.0, 8.5, 6.0, 7.5])

# a) Calcule a variância populacional das notas
variancia_populacional = np.var(notas)
print("A variância populacional das notas é:", variancia_populacional)
variancia_amostral = np.var(notas, ddof=1)
print("A variância amostral das notas é:", variancia_amostral)

# b) Calcule o intervalo interquartil (IIQ) das notas.
q1 = np.percentile(notas, 25)
q3 = np.percentile(notas, 75)
iiq = q3 - q1
print("O intervalo interquartil (IIQ) das notas é: ",iiq)
```

# **Exercício 03** - Medidas de Dispersão

Um pesquisador mediu a altura (em metros) de 12 coqueiros em uma praia de Porto de Galinhas:

**[15.2, 16.5, 14.8, 17.0, 15.5, 16.0, 15.8, 16.2, 15.0, 16.8, 17.5, 15.3]**

a) Calcule o desvio padrão populacional das alturas usando NumPy.

b) Calcule o iiq das alturas usando Pandas

# **Exercício 03** - Medidas de Dispersão

```python
import numpy as np
import pandas as pd

# Dados das alturas dos coqueiros
alturas = np.array([15.2, 16.5, 14.8, 17.0, 15.5, 16.0, 15.8, 16.2, 15.0, 16.8, 17.5, 15.3])

# a) Calcule o desvio padrão populacional das alturas usando NumPy.
desvio_padrao_populacional = np.std(alturas)
print(f"O desvio padrão populacional das alturas é:", desvio_padrao_populacional)

# b) Calcule o IIQ das alturas usando Pandas
series_alturas = pd.Series(alturas)
q1_pandas = series_alturas.quantile(0.25)
q3_pandas = series_alturas.quantile(0.75)
iiq_pandas = q3_pandas - q1_pandas
print(f"O Intervalo Interquartil (IIQ) das alturas é: ", iiq_pandas)
```

# **Exercício 04** - Medidas de Dispersão

Os tempos de espera (em minutos) de 15 clientes em uma famosa tapiocaria do Mercado da Boa Vista foram:

**[3, 5, 2, 8, 4, 4, 6, 3, 5, 7, 9, 3, 5, 4, 6]**

a) Crie uma Series do Pandas com esses dados.
b) Calcule a amplitude do tempo de espera usando Pandas
c) Calcule o desvio padrão amostral do tempo de espera usando Pandas
d) Calcule o intervalo interquartil (IIQ) do tempo de espera usando Pandas

# **Exercício 04** - Medidas de Dispersão

```python
import pandas as pd

# Dados dos tempos de espera
tempos_espera = [3, 5, 2, 8, 4, 4, 6, 3, 5, 7, 9, 3, 5, 4, 6]

# a) Crie uma Series com esses dados.
series_espera = pd.Series(tempos_espera)

# b) Calcule a amplitude do tempo de espera
amplitude = series_espera.max() - series_espera.min()
print("A amplitude do tempo de espera é: ", amplitude)

# c) Calcule o desvio padrão amostral do tempo de espera
desvio_padrao_amostral = series_espera.std()
print("O desvio padrão amostral do tempo de espera é: ", desvio_padrao_amostral)

# d) Calcule o intervalo interquartil (IIQ) do tempo de espera
q1 = series_espera.quantile(0.25)
q3 = series_espera.quantile(0.75)
iiq = q3 - q1
print("O Intervalo Interquartil (IIQ) do tempo de espera é: ", iiq)
```

# **Exercício 05** - Desafio

Utilizando o dataset “temperaturas_recife2020.csv” calcule:

a) Média
b) Mediana
c) Moda
d) Amplitude
e) Variância populacional e amostral
f) Desvio padrão populacional e amostral
g) Q1, Q2, Q3 e IQQ