

# Medidas Separatrizes: Conceito e Aplicação

---

As **medidas separatrizes** são estatísticas que dividem um conjunto ordenado de dados em partes iguais. Essas medidas ajudam a analisar a **distribuição dos dados**, facilitando a interpretação de valores e permitindo comparações entre diferentes grupos ou percentuais da amostra.

Elas são amplamente utilizadas em estatística descritiva e aplicadas em áreas como economia, educação, análise de desempenho e ciência de dados.

---

## 1. Tipos de Medidas Separatrizes

As principais medidas separatrizes são:

1. **Quartis (Qi)** → Dividem os dados em **4 partes** de 25%.
2. **Decis (Di)** → Dividem os dados em **10 partes** de 10%.
3. **Percentis (Pi)** → Dividem os dados em **100 partes** de 1%.

Cada uma dessas medidas determina um valor que separa os dados em frações específicas, permitindo análises como:

- Comparar desempenhos individuais dentro de um grupo.
- Identificar valores extremos (outliers).
- Avaliar a distribuição de um fenômeno (salários, notas, tempos de resposta etc.).

### Fórmula Geral

Para qualquer medida separatriz, o valor  $P_k$  (onde  $k$  é o percentil, decil ou quartil desejado) pode ser obtido pela fórmula:

$$P_k = \left( \frac{k}{m} \right) \times (n+1)$$

onde:

-  $k$  = número da medida separatriz desejada (exemplo: 1º quartil, 3º decil, 90º percentil).

-  $m$  = número total de divisões (4 para quartis, 10 para decis, 100 para percentis).

-  $n$  = total de elementos no conjunto de dados ordenado.

Se o índice encontrado não for um número inteiro, deve-se **interpolar** entre os valores adjacentes.

---

## 2. Quartis (Qi)

Os quartis dividem um conjunto de dados ordenado em **quatro partes iguais**, com **três valores de referência**:

- **Q<sub>1</sub> (Primeiro Quartil, 25%)** → 25% dos dados estão abaixo desse valor.
- **Q<sub>2</sub> (Mediana, 50%)** → 50% dos dados estão abaixo desse valor.
- **Q<sub>3</sub> (Terceiro Quartil, 75%)** → 75% dos dados estão abaixo desse valor.

## Exemplo

Conjunto de dados ordenado:

\$2, 5, 7, 10, 14, 18, 20, 22, 25, 30\$

- $Q_1$  (25° percentil)  $\rightarrow \$ (7+10)/2 = 8.5\$$
- $Q_2$  (mediana, 50° percentil)  $\rightarrow \$ (14+18)/2 = 16\$$
- $Q_3$  (75° percentil)  $\rightarrow \$ (22+25)/2 = 23.5\$$

A amplitude interquartil (**IQR**) é  $Q_3 - Q_1$ , usada para identificar **outliers**.

---

## 3. Decis (Di)

Os **decis** dividem os dados em **dez partes iguais**. São representados por  $D_1$  a  $D_9$ , onde cada um corresponde a **10% da distribuição**.

Por exemplo:

- $D_1$  (10%)  $\rightarrow$  10% dos dados estão abaixo desse valor.
- $D_5$  (50%)  $\rightarrow$  Equivalente à mediana.
- $D_9$  (90%)  $\rightarrow$  90% dos dados estão abaixo desse valor.

## Exemplo

Queremos encontrar  $D_4$  (**40% dos dados abaixo**) no mesmo conjunto:

$$D_4 = \frac{4}{10} \times (10+1) = 4.4$$

Interpolação entre o **4° e 5° elemento** (10 e 14):

$$D_4 = 10 + 0.4 \times (14 - 10) = 11.6$$

---

## 4. Percentis (Pi)

Os **percentis** dividem os dados em **100 partes iguais** e são muito usados em testes de desempenho, como ENEM e SAT.

- $P_1$  (1%)  $\rightarrow$  1% dos valores estão abaixo.
- $P_{25}$  (25%)  $\rightarrow$  Igual ao  $Q_1$ .
- $P_{50}$  (50%)  $\rightarrow$  Igual à **mediana**.
- $P_{90}$  (90%)  $\rightarrow$  90% dos valores estão abaixo.

## Exemplo

Para calcular  $P_{90}$  (**90° percentil**):

$$P_{90} = \frac{90}{100} \times (10+1) = 9.9$$

Interpolação entre **9° e 10° elemento** (25 e 30):

---

$$P_{\{90\}} = 25 + 0.9 \times (30 - 25) = 29.5$$

## 5. Comparação Geral

Medida Separatriz	Número de Divisões	Interpretação
Quartis (Qi)	4 partes (25%)	$Q_1, Q_2, Q_3$ mostram a distribuição em 4 seções.
Decis (Di)	10 partes (10%)	$D_1, D_2, \dots, D_9$ separam os dados em 10%.
Percentis (Pi)	100 partes (1%)	$P_k$ indica a posição em 1% de intervalo.

## 6. Aplicações Práticas

### 1 Educação (Notas e Classificações)

- No ENEM, estar no  $P_{90}$  significa que o aluno teve desempenho superior a 90% dos candidatos.

### 2 Finanças e Economia

- Análise de **distribuição de renda**: Quartis ajudam a entender desigualdade salarial.
- Bancos usam percentis para avaliar **riscos de crédito**.

### 3 Saúde e Ciência

- Medicina**: Percentis são usados em crescimento infantil (ex.: um bebê no  $P_{75}$  é maior que 75% dos bebês da mesma idade).
- Pesquisas científicas**: Outliers são detectados analisando a **amplitude interquartil (IQR)**.

### 4 Big Data e Inteligência Artificial

- Usado para identificar padrões em grandes volumes de dados (ex.: detecção de fraudes, segmentação de clientes).

Aqui está um exemplo de como calcular **quartis, decis e percentis** no Excel usando funções integradas.

### Passo a Passo no Excel

1 **Crie um conjunto de dados** em uma coluna (exemplo: A1:A10):

A (Dados)
2
5

## A (Dados)

7

10

14

18

20

22

25

30

## 1 Quartis no Excel

Para calcular os quartis, use a função **QUARTIL.INC**:

- **Q1 (25%)** → **=QUARTIL.INC(A1:A10,1)**
- **Q2 (Mediana, 50%)** → **=QUARTIL.INC(A1:A10,2)**
- **Q3 (75%)** → **=QUARTIL.INC(A1:A10,3)**

## 2 Decis no Excel

Para calcular um **Decil**, usamos a função **PERCENTIL.INC**:

- **D4 (40%)** → **=PERCENTIL.INC(A1:A10, 0.4)**

## 3 Percentis no Excel

Para calcular um percentil, usamos também a função **PERCENTIL.INC**:

- **P90 (90%)** → **=PERCENTIL.INC(A1:A10, 0.9)**

## Explicação das Funções

- **QUARTIL.INC(intervalo, N)** → Retorna o quartil N de um conjunto de dados.
- **PERCENTIL.INC(intervalo, k)** → Retorna o valor do percentil k (exemplo: **0.9** para 90%).

## Saída esperada no Excel

Se você aplicar essas fórmulas corretamente, os resultados serão semelhantes a:

```
Q1 = 8,5
Q2 (Mediana) = 16,0
Q3 = 23,5
D4 = 11,6
P90 = 29,5
```

Aqui está um exemplo em Python para calcular **quartis, decis e percentis** de um conjunto de dados usando a biblioteca **numpy**.

### Exemplo: Quartis, Decis e Percentis

```
import numpy as np

# Conjunto de dados ordenado
dados = np.array([2, 5, 7, 10, 14, 18, 20, 22, 25, 30])

# 1 Quartis (Q1, Q2, Q3)
Q1 = np.percentile(dados, 25) # Primeiro quartil (25%)
Q2 = np.percentile(dados, 50) # Mediana (50%)
Q3 = np.percentile(dados, 75) # Terceiro quartil (75%)

print(f"Q1 (25%): {Q1}")
print(f"Q2 (Mediana, 50%): {Q2}")
print(f"Q3 (75%): {Q3}")

# 2 Decis (exemplo: D4 - 40%)
D4 = np.percentile(dados, 40) # 40% dos dados abaixo desse valor
print(f"D4 (40%): {D4}")

# 3 Percentis (exemplo: P90 - 90%)
P90 = np.percentile(dados, 90) # Percentil 90
print(f"P90 (90%): {P90}")
```

### Explicação do Código

- O conjunto de dados está **ordenado**.
- **np.percentile(dados, x)** retorna o valor do percentil **x**.
- Como os **quartis** correspondem aos percentis **25%, 50% e 75%**, eles são calculados diretamente.
- O **4º decil** equivale a **40%**, então usamos **np.percentile(dados, 40)**.
- O **percentil 90** é obtido com **np.percentile(dados, 90)**.

---

### Saída esperada

Q1 (25%): 8.5  
Q2 (Mediana, 50%): 16.0  
Q3 (75%): 23.5  
D4 (40%): 11.6  
P90 (90%): 29.5

## 7. Conclusão

As **medidas separatrizes** são fundamentais para entender e interpretar a distribuição dos dados. Elas permitem **comparações objetivas** e facilitam a tomada de decisões em diversas áreas.

- **Quartis** ajudam a visualizar a dispersão geral.
- **Decis** fornecem divisões mais detalhadas.
- **Percentis** permitem uma análise minuciosa, útil em estatísticas educacionais e financeiras.

Essas medidas são amplamente usadas em **ciência de dados, economia, educação e análise de risco**, tornando-se ferramentas essenciais para transformar números em insights acionáveis.

## Medidas de dispersão

---

### Medidas de Dispersão em Estatística

As medidas de dispersão são estatísticas descritivas que quantificam o grau de variação ou dispersão de um conjunto de dados em relação à sua tendência central. Enquanto medidas de tendência central, como média, mediana e moda, fornecem um valor representativo dos dados, as medidas de dispersão indicam o quanto os valores individuais diferem desse ponto central. A análise dessas medidas é essencial para compreender a variabilidade dos dados e para comparar diferentes distribuições estatísticas.

## Principais Medidas de Dispersão

### 1. Amplitude

A **amplitude** é uma das medidas de dispersão mais simples em estatística, sendo utilizada para quantificar a diferença entre o maior e o menor valor de um conjunto de dados. Apesar de sua simplicidade, a amplitude desempenha um papel fundamental na análise exploratória dos dados, permitindo uma **avaliação inicial da variabilidade** dentro de uma distribuição.

---

#### 1. Definição e Cálculo da Amplitude

A amplitude pode ser definida matematicamente da seguinte forma:

$$\text{Amplitude} = X_{\max} - X_{\min}$$

onde:

-\$X\_{\max}\$ representa o maior valor do conjunto de dados.

$-X_{\min}$  representa o menor valor do conjunto de dados.

A amplitude expressa a **distância total** coberta pelos valores dentro de uma distribuição. Se os dados forem muito dispersos, a amplitude será grande; se forem concentrados, a amplitude será pequena.

Por exemplo, considere o seguinte conjunto de dados:

#### Exemplo 1:

$\{5, 8, 12, 14, 17, 21, 24\}$

- O maior valor ( $X_{\max}$ ) é **24**.
- O menor valor ( $X_{\min}$ ) é **5**.

Logo, a amplitude será:

$$\text{Amplitude} = 24 - 5 = 19$$

---

## 2. Características da Amplitude

A amplitude é uma **medida de dispersão absoluta**, ou seja, seu valor depende da escala dos dados. Algumas de suas principais características incluem:

1. **Simplicidade:** Fácil de calcular e interpretar.
2. **Sensibilidade a Outliers:** A presença de valores extremos pode distorcer a amplitude, tornando-a uma medida instável.
3. **Não considera a distribuição interna dos dados:** Dois conjuntos de dados podem ter a mesma amplitude, mas distribuições completamente diferentes.

#### Exemplo Comparativo

Considere os dois conjuntos de dados abaixo:

- **Conjunto A** 😞  $\{10, 12, 14, 16, 18\}$
- **Conjunto B** 😞  $\{2, 5, 10, 20, 30\}$

Para ambos os conjuntos:

$$\text{Amplitude} = X_{\max} - X_{\min}$$

- Para o **Conjunto A**:  $18 - 10 = 8$
- Para o **Conjunto B**:  $30 - 2 = 28$

Embora o Conjunto B tenha uma amplitude maior, sua variabilidade real pode ser melhor representada por medidas mais robustas, como a variância ou o desvio padrão.

---

## 3. Tipos de Amplitude

A amplitude pode ser calculada de diferentes formas, dependendo do contexto da análise:

### 3.1. Amplitude Total (Simples)

É a diferença entre o maior e o menor valor do conjunto de dados, conforme definido anteriormente.

### 3.2. Amplitude Interquartil (IQR - Interquartile Range)

O **intervalo interquartil (IQR)** é uma medida de dispersão mais robusta, pois ignora os extremos da distribuição e foca na variação dos valores centrais. Ele é calculado como:

$$IQR = Q_3 - Q_1$$

onde:

-  $Q_1$  (primeiro quartil) representa o valor abaixo do qual 25% dos dados estão localizados.

-  $Q_3$  (terceiro quartil) representa o valor abaixo do qual 75% dos dados estão localizados.

### 3.3. Amplitude Relativa

A amplitude relativa expressa a dispersão dos dados em relação ao valor médio da amostra:

$$\text{Amplitude Relativa} = \frac{X_{\max} - X_{\min}}{\bar{X}}$$

onde  $\bar{X}$  é a média aritmética dos valores do conjunto.

---

## 4. Aplicações da Amplitude

### 4.1. Controle de Qualidade

A amplitude é usada em **cartas de controle** para monitorar a variação dos processos industriais. Pequenas amplitudes indicam processos estáveis, enquanto grandes amplitudes podem sinalizar problemas na produção.

### 4.2. Estudos Ambientais e Meteorológicos

Na meteorologia, a amplitude térmica diária é calculada como a diferença entre a temperatura máxima e mínima registrada em um dia.

### 4.3. Análise Financeira

Em finanças, a amplitude dos preços de um ativo ao longo do tempo pode indicar volatilidade e riscos.

### 4.4. Educação e Avaliação

Na análise de desempenho acadêmico, a amplitude das notas pode indicar discrepâncias na dificuldade de provas ou na preparação dos alunos.

## 5. Limitações da Amplitude

Apesar de sua utilidade, a amplitude apresenta algumas limitações:

- **Extrema sensibilidade a valores atípicos:** Pequenos desvios extremos podem alterar drasticamente a amplitude.



- **Não reflete a distribuição dos dados:** Dois conjuntos podem ter a mesma amplitude, mas diferentes concentrações de valores.
- **Pouco informativa para conjuntos grandes:** À medida que o número de observações aumenta, a amplitude tende a crescer naturalmente, tornando-a uma medida menos eficaz para grandes amostras.

Por isso, muitas análises complementam a amplitude com medidas mais robustas, como o desvio padrão e a variância.

---

## 6. Cálculo da Amplitude no Python e no Excel

### 6.1. Em Python

Podemos calcular a amplitude usando a biblioteca `numpy`:

```
import numpy as np

dados = [5, 8, 12, 14, 17, 21, 24]

amplitude = np.ptp(dados) # ptp retorna a diferença entre max e min

print(f"Amplitude: {amplitude}")
```

### 6.2. Em Excel

Para calcular a amplitude no Excel:

1. Insira os dados em uma coluna (por exemplo, de A1 a A7).
2. Use a fórmula:

```
=MAX(A1:A7) - MIN(A1:A7)
```

Isso retornará a amplitude dos dados inseridos.

## 7. Conclusão

A amplitude é uma medida estatística **simples, porém fundamental** para uma análise preliminar da dispersão dos dados. Embora sua aplicação seja útil em diversos contextos, sua **sensibilidade a outliers e falta de detalhamento** exigem o uso complementar de medidas mais robustas, como a variância, o desvio padrão e o intervalo interquartil.

Ao interpretar a amplitude, é essencial considerar **a natureza dos dados, o tamanho da amostra e a presença de valores extremos**, garantindo uma análise estatística mais precisa e confiável.

---

## 8. Referências

- BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 8. ed. São Paulo: Saraiva, 2017.
  - TRIOLA, M. F. *Introdução à Estatística*. 13. ed. Pearson, 2020.
  - MONTGOMERY, D. C.; RUNGER, G. C. *Applied Statistics and Probability for Engineers*. 7th ed. John Wiley & Sons, 2018.
  - TUKEY, J. W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.
- 

## 2. Desvio Médio

O **Desvio Médio** é uma medida de dispersão que indica **o afastamento médio dos valores de um conjunto de dados em relação à sua média aritmética**. Ele expressa, em termos absolutos, o quanto os dados se espalham em torno da média, sendo útil para entender a variabilidade dos dados sem considerar a direção das diferenças.

### 1. Conceito e Definição

O desvio médio é calculado como a média das diferenças absolutas entre cada valor do conjunto de dados e a média aritmética dos valores. Sua fórmula geral é dada por:

#### 1.1. Para uma População

$$DM = \frac{\sum |X_i - \mu|}{N}$$

onde:

- \$DM\$ = Desvio médio
- \$X\_i\$ = Cada valor do conjunto de dados
- \$\mu\$ = Média populacional
- \$N\$ = Número total de elementos da população

#### 1.2. Para uma Amostra

$$DM = \frac{\sum |X_i - \bar{X}|}{n}$$

onde:

- \$\bar{X}\$ = Média amostral
- \$n\$ = Número de elementos da amostra

O uso da **média aritmética** como referência para medir as diferenças permite entender como os dados se comportam em relação ao valor central.

---

## 2. Importância do Desvio Médio

O desvio médio é amplamente utilizado porque:

- **Fornecer uma medida intuitiva de dispersão:** Ao trabalhar com valores absolutos, evita cancelamentos que podem ocorrer ao somar diferenças brutas.

- **É útil para dados simétricos:** Quando a distribuição dos dados é aproximadamente simétrica, o desvio médio é uma boa representação da dispersão.
- **Pode ser mais fácil de interpretar:** Em algumas situações, o desvio médio pode ser mais intuitivo que a variância e o desvio padrão, pois não envolve elevação ao quadrado.

Segundo **Triola (2020)**, o desvio médio é uma alternativa útil ao desvio padrão em alguns contextos, pois fornece uma medida de dispersão **sem amplificar grandes desvios individuais**, como ocorre na variância.

### 3. Diferença entre Desvio Médio, Variância e Desvio Padrão

Medida	Vantagens	Desvantagens
<b>Desvio Médio</b>	Fácil de interpretar e menos sensível a valores extremos	Não é amplamente utilizado em estatísticas inferenciais
<b>Variância</b>	Leva em conta o quadrado dos desvios, sendo útil para cálculos estatísticos	Pode ser difícil de interpretar, pois está em unidades quadradas
<b>Desvio Padrão</b>	Expressa dispersão na mesma unidade dos dados e é amplamente utilizado	É mais sensível a valores extremos do que o desvio médio

O **desvio médio** é menos utilizado que a **variância** e o **desvio padrão**, principalmente porque não tem tantas aplicações em estatísticas inferenciais e não aparece em fórmulas de distribuições estatísticas.

No entanto, em **análises exploratórias** ou quando se deseja uma medida intuitiva de dispersão, ele pode ser útil.

### 4. Exemplo Prático de Cálculo

#### 4.1. Cálculo Manual

Considere o seguinte conjunto de dados:

$\{5, 10, 15, 20, 25\}$

##### Passo 1: Calcular a Média

$$\bar{X} = \frac{5 + 10 + 15 + 20 + 25}{5} = 15$$

##### Passo 2: Calcular os Desvios Absolutos

$X_i$	$ X_i - \bar{X} $
5	$ 5 - 15  = 10$
10	$ 10 - 15  = 5$
15	$ 15 - 15  = 0$

$X_i$	$ X_i - \bar{X} $
20	$ 20 - 15  = 5$
25	$ 25 - 15  = 10$

### Passo 3: Calcular o Desvio Médio

$$DM = \frac{10 + 5 + 0 + 5 + 10}{5} = \frac{30}{5} = 6$$

Assim, o **desvio médio é 6**.

## 4.2. Cálculo em Python

Podemos calcular o desvio médio usando **NumPy** e **pandas**:

```
import numpy as np
import pandas as pd

dados = np.array([5, 10, 15, 20, 25])

media = np.mean(dados) # Calcula a média
desvio_medio = np.mean(np.abs(dados - media)) # Calcula o desvio médio

print(f"Média: {media}")
print(f"Desvio Médio: {desvio_medio}")
```

**Saída esperada:**

```
Média: 15.0
Desvio Médio: 6.0
```

## 4.3. Cálculo no Excel

1. Insira os valores na **coluna A** (exemplo: **A1:A5** → {5, 10, 15, 20, 25}).
2. Calcule a média com a fórmula:

```
=MÉDIA(A1:A5)
```

3. Em outra coluna, calcule os desvios absolutos:

```
=ABS(A1 - MÉDIA(A1:A5))
```

4. Para obter o desvio médio, use:

```
=MÉDIA(B1:B5)
```

Isso retornará o valor **6.0**.

---

## 5. Aplicações do Desvio Médio

O desvio médio é utilizado em diversas áreas, incluindo:

- **Economia e Finanças:** Para medir a volatilidade de preços e retornos de ativos financeiros.
- **Controle de Qualidade:** Para avaliar a variação de medidas em processos industriais.
- **Ciências Sociais:** Para analisar distribuições de renda e desigualdade econômica.
- **Engenharia:** Para avaliar variações em medições e testes de produtos.

De acordo com **Montgomery e Runger (2018)**, o desvio médio pode ser usado como uma **medida robusta de variabilidade**, especialmente quando se deseja evitar a influência de valores extremos, tornando-o adequado para alguns tipos de controle estatístico de processos.

---

## 6. Considerações Finais

O **desvio médio** é uma medida de dispersão **simples, intuitiva e fácil de interpretar**, embora não seja tão amplamente usada quanto o desvio padrão. Ele fornece uma visão clara da variabilidade dos dados em torno da média, sendo útil para análises exploratórias e aplicações práticas em diversas áreas.

Embora seja menos utilizado em estatísticas inferenciais, seu papel em análises descritivas e aplicações industriais o torna uma ferramenta valiosa para compreender a dispersão dos dados.

---

## 3. Variância

A **variância** é uma medida estatística que expressa a dispersão dos dados em relação à média. Em outras palavras, ela quantifica **o quão distantes os valores estão da média** de um conjunto de dados. Essa medida é fundamental para entender a estabilidade e a confiabilidade de um conjunto de informações, sendo amplamente utilizada em estatística descritiva, inferência estatística, machine learning e diversas áreas aplicadas.

---

### 1. Conceito de Variância

A variância mede a média dos **quadrados das diferenças** entre cada valor e a média aritmética do conjunto de dados. Ela é representada pelas seguintes fórmulas:

- **Para uma população inteira** (variância populacional):

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

- **Para uma amostra** (variância amostral):

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

onde:

- $X_i$  representa cada valor do conjunto de dados,
- $\mu$  é a média populacional,
- $\bar{X}$  é a média amostral,
- $N$  é o número total de elementos da população,
- $n$  é o número total de elementos da amostra,
- $\sigma^2$  é a variância populacional,
- $s^2$  é a variância amostral.

A principal diferença entre **variância populacional** e **variância amostral** é o denominador. No caso amostral, subtrai-se **1** do número de elementos (graus de liberdade) para corrigir o viés da estimativa, garantindo que a variância amostral seja uma estimativa não tendenciosa da variância populacional.

## 1.1. Interpretação Intuitiva

A variância **mede o espalhamento** dos dados. Quando a variância é **baixa**, os valores do conjunto de dados estão próximos da média. Quando a variância é **alta**, os dados apresentam grande dispersão em torno da média.

Por exemplo, considere dois conjuntos de dados com a mesma média:

- **Conjunto A:** 10, 11, 10, 9, 10
- **Conjunto B:** 5, 15, 10, 2, 18

Embora ambos tenham média **10**, o **Conjunto B tem uma variância maior** porque seus valores estão mais espalhados.

## 2. Importância da Variância

A variância é fundamental para diversas análises estatísticas, pois:

- **Ajuda a quantificar a dispersão dos dados**, permitindo comparar diferentes distribuições.
- **É usada no cálculo do desvio padrão**, que é a raiz quadrada da variância e mais intuitivo para interpretação.
- **Tem aplicações diretas em modelagem estatística**, como regressão, análise de variância (ANOVA) e testes de hipóteses.
- **É utilizada na teoria da probabilidade**, ajudando a determinar a incerteza e a estabilidade de um conjunto de dados.
- **É essencial em finanças e economia**, onde mede o risco e a volatilidade de ativos financeiros.

Como afirma **Montgomery e Runger (2018)**, a variância é uma das medidas estatísticas mais importantes, pois permite entender o comportamento de um conjunto de dados e sua previsibilidade.

### 3. Cálculo da Variância: Exemplo Passo a Passo

Vamos calcular a variância para o seguinte conjunto de dados:

$$X = \{4, 8, 6, 5, 3\}$$

#### Passo 1: Calcular a Média

$$\bar{X} = \frac{4 + 8 + 6 + 5 + 3}{5} = \frac{26}{5} = 5.2$$

#### Passo 2: Calcular as Diferenças em Relação à Média

$X_i$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
4	$4 - 5.2 = -1.2$	$(-1.2)^2 = 1.44$
8	$8 - 5.2 = 2.8$	$(2.8)^2 = 7.84$
6	$6 - 5.2 = 0.8$	$(0.8)^2 = 0.64$
5	$5 - 5.2 = -0.2$	$(-0.2)^2 = 0.04$
3	$3 - 5.2 = -2.2$	$(-2.2)^2 = 4.84$

#### Passo 3: Calcular a Variância

$$s^2 = \frac{1.44 + 7.84 + 0.64 + 0.04 + 4.84}{5 - 1}$$

$$s^2 = \frac{14.8}{4} = 3.7$$

Portanto, a variância amostral é **3.7**.

---

### 4. Cálculo da Variância em Python

Aqui está um exemplo prático usando **numpy**:

```
import numpy as np

dados = [4, 8, 6, 5, 3]

# Variância populacional
variancia_populacional = np.var(dados)

# Variância amostral (ddof=1 para corrigir viés)
variancia_amostral = np.var(dados, ddof=1)

print(f"Variância populacional: {variancia_populacional:.2f}")
print(f"Variância amostral: {variancia_amostral:.2f}")
```

Saída:

Variância populacional: 2.96  
Variância amostral: 3.70

---

## 5. Relação entre Variância e Desvio Padrão

O **desvio padrão** é a **raiz quadrada da variância**, tornando a medida mais interpretável, pois mantém a mesma unidade dos dados.

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

Se a variância amostral for **3.7**, então o desvio padrão será:

$$s = \sqrt{3.7} \approx 1.92$$

O desvio padrão facilita a análise de dispersão porque está na mesma unidade dos dados, enquanto a variância é expressa na unidade **ao quadrado**.

---

## 6. Aplicações Práticas da Variância

A variância é utilizada em diversas áreas:

- **Finanças:** Para medir a volatilidade de ativos financeiros.
- **Engenharia:** Para avaliar a variabilidade em processos de produção.
- **Ciências sociais:** Para analisar diferenças de desempenho entre grupos.
- **Machine Learning:** Para otimizar algoritmos e reduzir overfitting.

Segundo **Bussab e Morettin (2017)**, a variância é uma medida essencial para entender a estabilidade de um conjunto de dados e prever seu comportamento futuro.

---

A variância é uma das medidas mais importantes da estatística, pois permite quantificar a **dispersão dos dados** e avaliar sua estabilidade. Seu cálculo, embora simples, tem **amplas aplicações práticas**, desde finanças até inteligência artificial.

Entender a variância ajuda a tomar **decisões mais embasadas** e aprimorar a análise de dados, tornando-se uma ferramenta indispensável na estatística moderna.

## 4. Desvio Padrão

O **desvio padrão** é uma das medidas de dispersão mais importantes da estatística, pois indica o grau de variação ou dispersão dos dados em relação à média. Essa métrica é amplamente utilizada em diversas áreas, como ciência de dados, economia, engenharia, ciências sociais e análise de riscos, pois fornece uma visão quantitativa da estabilidade e previsibilidade de um conjunto de dados.

---

## 1. Definição do Desvio Padrão



O **desvio padrão** (representado por  $\sigma$  para populações e  $s$  para amostras) mede **o quanto os valores de um conjunto de dados se afastam da média**. Ele é definido como a raiz quadrada da variância, permitindo que a dispersão seja expressa na mesma unidade dos dados originais.

A fórmula do desvio padrão **populacional** ( $\sigma$ ) é:

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Onde:

- $X_i$  representa cada valor do conjunto de dados,
- $\mu$  é a média populacional,
- $N$  é o número total de elementos na população.

Para amostras, o **desvio padrão amostral** ( $s$ ) é calculado com um pequeno ajuste, substituindo  $N$  por  $n - 1$  (graus de liberdade):

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

Esse ajuste é necessário para corrigir a subestimação da variabilidade da população quando usamos uma amostra.

---

## 2. Interpretação do Desvio Padrão

O valor do desvio padrão indica a dispersão dos dados:

- **Desvio padrão pequeno:** os dados estão próximos da média, indicando pouca variação.
- **Desvio padrão grande:** os dados estão muito espalhados, sugerindo maior variabilidade.

Por exemplo, considere duas turmas de alunos com médias de notas iguais a 7,0:

- **Turma A:** Notas: [6,8,7,7,7] → **Desvio padrão pequeno**
- **Turma B:** Notas: [4,10,3,9,9] → **Desvio padrão grande**

Embora ambas tenham média 7, a Turma B tem uma dispersão muito maior, o que indica **uma maior variabilidade no desempenho dos alunos**.

---

## 3. Relação com a Distribuição Normal e a Regra Empírica (68-95-99,7)

O desvio padrão é essencial para compreender distribuições de dados, especialmente a **distribuição normal** (ou Gaussiana). A **Regra Empírica** afirma que, em uma distribuição normal:

- **68%** dos valores estão dentro de **1 desvio padrão** da média ( $\mu \pm \sigma$ ).
- **95%** dos valores estão dentro de **2 desvios padrão** ( $\mu \pm 2\sigma$ ).
- **99,7%** dos valores estão dentro de **3 desvios padrão** ( $\mu \pm 3\sigma$ ).

Essa regra permite prever a dispersão dos dados e identificar valores atípicos (**outliers**) quando um dado se encontra além de 3 desvios padrão da média.

---

## 4. Comparação com Outras Medidas de Dispersão

O desvio padrão **é mais robusto que a amplitude** (que só considera valores extremos), mas **é sensível a outliers**, pois eleva ao quadrado as diferenças em relação à média. Alternativas incluem:

- **Intervalo Interquartil (IQR)**: mede a dispersão sem ser afetado por outliers.
- **Coeficiente de Variação (CV)**: expressa o desvio padrão como uma porcentagem da média, permitindo comparações entre conjuntos de dados com unidades diferentes.

---

## 5. Aplicação Prática com Exemplo em Python

Aqui está um exemplo de cálculo do desvio padrão em **Python**, usando a biblioteca **numpy**:

```
import numpy as np

# Dados
dados = [10, 12, 23, 23, 16, 23, 21, 16, 18, 19]

# Cálculo do desvio padrão populacional
desvio_padrao_populacional = np.std(dados)

# Cálculo do desvio padrão amostral
desvio_padrao_amostral = np.std(dados, ddof=1)

print(f"Desvio Padrão Populacional: {desvio_padrao_populacional:.2f}")
print(f"Desvio Padrão Amostral: {desvio_padrao_amostral:.2f}")
```

Saída esperada:

```
Desvio Padrão Populacional: 4.87
Desvio Padrão Amostral: 5.13
```

### 5.1 Exemplo python com altura

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

# Parâmetros da distribuição de altura (em metros)
mu = 1.70 # média de altura
sigma = 0.10 # desvio padrão

# Gerar os dados da distribuição normal
x = np.linspace(mu - 4*sigma, mu + 4*sigma, 1000)
y = stats.norm.pdf(x, mu, sigma)

# Mostrar uma amostra dos dados
print("Amostra de valores de altura (m) e densidade de probabilidade:")
```

```

for xi, yi in zip(x[::100], y[::100]):
    print(f"Altura = {xi:.2f} m, Densidade = {yi:.5f}")

# Criar o gráfico
plt.figure(figsize=(10, 5))
plt.plot(x, y, label="Distribuição Normal de Alturas", color="black")

# Regiões da regra empírica
for i, alpha in zip(range(1, 4), [0.3, 0.2, 0.1]):
    plt.fill_between(x, y, where=(mu - i*sigma <= x) & (x <= mu +
i*sigma),
                    color="blue", alpha=alpha,
                    label=f"{68 if i == 1 else 95 if i == 2 else 99.7}%
dentro de {i}σ")

# Linhas de média e desvios padrão
plt.axvline(mu, color='red', linestyle='dashed',
            label=f'Média (1.70 m)')
plt.axvline(mu, color='red', linestyle='dashed', label=f'Desvio Padrão
({sigma:.2f} m)')
plt.axvline(mu - sigma, color='green', linestyle='dashed', label='1σ
(±0.10 m)')
plt.axvline(mu + sigma, color='green', linestyle='dashed')
plt.axvline(mu - 2*sigma, color='blue', linestyle='dashed', label='2σ
(±0.20 m)')
plt.axvline(mu + 2*sigma, color='blue', linestyle='dashed')
plt.axvline(mu - 3*sigma, color='gray', linestyle='dashed', label='3σ
(±0.30 m)')
plt.axvline(mu + 3*sigma, color='gray', linestyle='dashed')

# Configurações do gráfico
plt.title("Distribuição Normal de Alturas - Regra Empírica (68-95-
99.7)")
plt.xlabel("Altura (metros)")
plt.ylabel("Densidade de Probabilidade")
plt.legend()
plt.grid(True)

# Salvar dados da distribuição em CSV e Excel
dados = pd.DataFrame({'Valor': x, 'Densidade': y})

# CSV
dados.to_csv('dados_distribuicao.csv', index=False)

# Excel
dados.to_excel('dados_distribuicao.xlsx', index=False)

# >>> SALVA A IMAGEM COMO PNG <<<
plt.savefig("distribuicao_normal_altura.png", dpi=300)

plt.show()

```

---

## Explicação do Código: Distribuição Normal e Regra Empírica

O código em Python acima tem como objetivo **visualizar a Regra Empírica (68-95-99.7)** aplicada a uma **distribuição normal** e também **exportar os dados gerados para análise posterior** (em CSV, Excel e imagem PNG).

### Etapas do Código

#### Importação das bibliotecas

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pandas as pd
```

Essas bibliotecas permitem:

- Criar vetores e fazer cálculos (**numpy**)
- Plotar gráficos (**matplotlib**)
- Trabalhar com distribuições estatísticas, como a normal (**scipy.stats**)
- Exportar dados em formatos como **.csv** e **.xlsx** (**pandas**)

---

#### Definição dos parâmetros

```
mu = 100      # média
sigma = 15    # desvio padrão
```

Define os parâmetros da distribuição normal: a média ( $\mu = 100$ ) e o desvio padrão ( $\sigma = 15$ ). Esses valores podem representar, por exemplo, **pontuações de testes padronizados**.

---

#### Geração dos dados da curva normal

```
x = np.linspace(mu - 4*sigma, mu + 4*sigma, 1000)
y = stats.norm.pdf(x, mu, sigma)
```

- Cria um vetor **x** com 1000 pontos entre  $-4\sigma$  e  $+4\sigma$  ao redor da média.
- Calcula a **função densidade de probabilidade** (PDF) da distribuição normal para cada ponto de **x**.

## Construção do gráfico

```
plt.plot(x, y, label="Distribuição Normal", color="black")
```

- Plota a curva da distribuição normal.

---

## Aplicação da Regra Empírica

```
for i, alpha in zip(range(1, 4), [0.3, 0.2, 0.1]):  
    ...
```

- Preenche as áreas sob a curva dentro de  $1\sigma$ ,  $2\sigma$  e  $3\sigma$  da média.
- Essas áreas correspondem, aproximadamente, a:
  - 68% dos dados em  $\pm 1\sigma$
  - 95% dos dados em  $\pm 2\sigma$
  - 99.7% dos dados em  $\pm 3\sigma$

---

## Linhas verticais de referência

```
plt.axvline(mu, ...)
```

- Adiciona linhas tracejadas na média e nos desvios padrão ( $\pm 1\sigma$ ,  $\pm 2\sigma$ ,  $\pm 3\sigma$ ) para facilitar a leitura da curva.

---

## Finalização e salvamento do gráfico

```
plt.savefig("distribuicao_normal.png")
```

- Mostra o gráfico e salva a imagem em formato **.png**.

---

## Exportação dos dados para análise

```
dados = pd.DataFrame({'Valor': x, 'Densidade': y})  
dados.to_csv('dados_distribuicao.csv', index=False)  
dados.to_excel('dados_distribuicao.xlsx', index=False)
```

- Cria uma tabela (**DataFrame**) com os valores de **x** (pontuação) e **y** (densidade).
  - Salva os dados como:
    - **.csv**: compatível com editores de texto e Excel.
    - **.xlsx**: arquivo do Excel nativo.
- 

O que é a Regra Empírica?

A **Regra Empírica** afirma que em uma distribuição normal:

- Cerca de **68%** dos dados estão dentro de **1 desvio padrão** da média.
- Cerca de **95%** dentro de **2 desvios padrões**.
- Cerca de **99,7%** dentro de **3 desvios padrões**.

Ela é baseada nas propriedades matemáticas da distribuição normal e **ajuda a entender rapidamente a dispersão dos dados em torno da média**.

---

## 6. Aplicação em Excel

No Excel, o desvio padrão pode ser calculado com as funções:

- **POPULAÇÃO**: **=DESPAD.P(A1:A10)**
- **AMOSTRAL**: **=DESPAD(A1:A10)**

Essas funções ajudam a calcular rapidamente a dispersão de um conjunto de dados em planilhas.

---

## 7. Aplicações do Desvio Padrão no Mundo Real

O desvio padrão tem aplicações práticas em diversas áreas, como:

- **Finanças**: Mede o risco de investimentos; um ativo com maior desvio padrão tem retornos mais voláteis.
  - **Controle de Qualidade**: Empresas usam o desvio padrão para verificar a consistência da produção.
  - **Medicina**: Avalia a variabilidade em testes clínicos, como a resposta de pacientes a um novo tratamento.
  - **Ciência de Dados**: Ajuda na detecção de outliers e no entendimento da dispersão de variáveis.
- 

## 8. Conclusão

O desvio padrão é uma medida estatística essencial que indica **o grau de variação dos dados** em relação à média. Ele desempenha um papel fundamental em análises estatísticas e previsão de eventos, ajudando na **tomada de decisões** informadas em diversas áreas.

Apesar de ser amplamente utilizado, é importante **combiná-lo com outras medidas de dispersão** para obter uma visão mais completa dos dados.

---

## 5. Coeficiente de Variação (CV)

O **Coeficiente de Variação (CV)** é uma medida de dispersão relativa que expressa o grau de variabilidade de um conjunto de dados em relação à média, sendo, portanto, uma forma de comparar a dispersão de diferentes distribuições, especialmente quando essas distribuições têm unidades ou magnitudes diferentes. Ele é particularmente útil em contextos onde desejamos comparar a variabilidade entre duas ou mais séries de dados que podem ter escalas ou unidades diferentes, mas que se refere a uma mesma característica ou fenômeno.

### 1. Definição e Fórmula

O **Coeficiente de Variação** é calculado pela razão entre o **desvio padrão** e a **média aritmética**, multiplicado por 100 para expressá-lo como uma porcentagem:

$$CV = \left( \frac{\sigma}{\mu} \right) \times 100$$

onde:

-\$\sigma\$ é o **desvio padrão** da distribuição dos dados.

-\$\mu\$ é a **média aritmética** dos dados.

### Interpretação do Coeficiente de Variação

O Coeficiente de Variação é uma medida **adimensional**, ou seja, não possui unidade, o que facilita comparações entre conjuntos de dados de unidades diferentes. O **CV** expressa a quantidade de variação relativa, em relação à média, de um conjunto de dados. Quanto maior o valor do **CV**, maior será a variabilidade em relação à média.

Se o **CV** for baixo, significa que os dados estão mais concentrados em torno da média, ou seja, a **variabilidade é pequena**. Se o **CV** for alto, significa que os dados estão mais dispersos em relação à média, com **maior variabilidade**.

### 2. Vantagens e Aplicações do Coeficiente de Variação

#### 2.1. Comparação de Dados com Diferentes Unidades ou Escalas

Uma das principais vantagens do Coeficiente de Variação é a sua capacidade de **comparar dados com diferentes unidades ou escalas**. Por exemplo, ao comparar o risco de investimentos em diferentes mercados financeiros, um **CV alto** indica maior risco (variabilidade maior) em relação à média de retorno.

#### Exemplo prático:

- Em um mercado de ações, um ativo com um retorno médio de 10% e desvio padrão de 3% tem um CV de 30%. Já outro ativo, com um retorno médio de 30% e desvio padrão de 15%, terá um CV de 50%. O Coeficiente de Variação revela que o segundo ativo, apesar de ter um retorno médio maior, tem uma maior **variabilidade** em torno da média.

#### 2.2. Avaliação da Incerteza

O Coeficiente de Variação é amplamente utilizado na **avaliação de risco e incerteza**, especialmente em **modelos financeiros** e **gestão de investimentos**. Ao comparar o risco relativo de diferentes investimentos ou variáveis, o **CV** ajuda a entender não apenas o valor médio, mas também a **consistência ou previsibilidade** de um ativo ou fenômeno.

**Exemplo:** Se dois investimentos têm o mesmo retorno médio, mas o **CV** de um deles for mais alto, o investimento com maior CV terá mais **incerteza** associada ao seu desempenho futuro.

### 2.3. Medida de Dispersão Normalizada

Como o **CV** é uma medida relativa, ele serve como uma **medida de dispersão normalizada**. Isso é especialmente útil quando estamos lidando com séries de dados com magnitudes diferentes. Em vez de confiar apenas na magnitude absoluta dos desvios padrão ou amplitude, o **CV** oferece uma forma de medir a dispersão proporcionalmente ao valor médio.

**Exemplo:** Ao comparar duas fábricas, uma que produz 100 unidades de um produto por mês com uma variação de 10 unidades (desvio padrão de 10) e outra que produz 1.000 unidades com uma variação de 100 unidades, o desvio padrão absoluto não diz muito sobre a comparação da variabilidade relativa. Mas ao calcular o **CV**, podemos concluir qual fábrica tem maior dispersão relativa, dado que o **CV** de ambas pode ser comparado diretamente.

## 3. Limitações do Coeficiente de Variação

Apesar das vantagens, o Coeficiente de Variação tem algumas **limitações** que devem ser observadas:

### 3.1. Sensibilidade a Valores Negativos

O **Coeficiente de Variação** só pode ser calculado quando a **média** dos dados for positiva. Caso a média seja zero ou negativa, o **CV** perde seu significado, pois a fórmula de cálculo envolve uma divisão pela média. Isso pode ocorrer, por exemplo, em distribuições com **valores negativos** ou quando há uma **média muito baixa**. Em tais casos, o **CV** não fornece uma medida válida de dispersão.

### 3.2. Extrapolação Limitada em Distribuições Não Simétricas

Embora o **CV** seja útil para distribuições simétricas ou moderadamente assimétricas, ele pode não ser tão eficaz em distribuições extremamente assimétricas. O **CV** é uma **medida de dispersão proporcional**, mas pode ser distorcido por **extremos ou outliers** em distribuições com caudas longas, como distribuições **exponenciais** ou **log-normais**.

## 4. Exemplo Prático do Coeficiente de Variação

Aqui está um exemplo prático de como calcular o Coeficiente de Variação em Python para duas séries de dados:

```
import numpy as np

# Dados de exemplo 1 (poderiam ser, por exemplo, os retornos de um
investimento)
```



```

dados1 = [12, 15, 18, 14, 16, 17, 15]

# Dados de exemplo 2 (outro conjunto de dados para comparação)
dados2 = [100, 150, 120, 180, 140, 160, 130]

# Função para calcular o Coeficiente de Variação
def coeficiente_variacao(dados):
    media = np.mean(dados)
    desvio_padrao = np.std(dados, ddof=1)
    cv = (desvio_padrao / media) * 100
    return cv

# Calculando o Coeficiente de Variação para ambos os conjuntos de dados
cv1 = coeficiente_variacao(dados1)
cv2 = coeficiente_variacao(dados2)

print(f"Coeficiente de Variação para dados1: {cv1:.2f}%")
print(f"Coeficiente de Variação para dados2: {cv2:.2f}%")

```

Neste exemplo, você calcularia o **CV** para dois conjuntos de dados e poderia comparar sua **variabilidade relativa**. Se o **CV** de um conjunto for maior, isso indica maior **dispersão relativa** em comparação ao outro conjunto, independentemente das suas magnitudes absolutas.

O **Coeficiente de Variação** é uma das medidas de dispersão mais úteis para comparar a **variabilidade relativa** de diferentes conjuntos de dados, especialmente quando esses conjuntos têm unidades ou magnitudes diferentes. Sua capacidade de fornecer uma **mensuração normalizada da dispersão** o torna uma ferramenta poderosa em estatística e análise de dados, particularmente em áreas como **finanças, gestão de risco, controle de qualidade e ciências sociais**.

No entanto, é importante lembrar suas limitações, como a **sensibilidade a dados negativos** e sua adequação apenas para distribuições não altamente assimétricas. O Coeficiente de Variação deve ser utilizado com cautela, complementado por outras medidas de dispersão e análise de dados quando necessário.

## Importância das Medidas de Dispersão

As medidas de dispersão são fundamentais para diversas aplicações estatísticas:

- **Comparação de variabilidade** entre diferentes distribuições.
- **Identificação de outliers** e padrões em conjuntos de dados.
- **Base para inferência estatística**, como intervalos de confiança e testes de hipóteses.

Medidas de dispersão complementam as medidas de tendência central ao fornecer uma visão detalhada sobre a variabilidade dos dados. O desvio padrão e a variância são amplamente utilizados devido à sua aplicabilidade em modelos estatísticos e inferência, enquanto o coeficiente de variação é útil para comparações entre diferentes contextos.

## Diferença entre Amplitude, Desvio Médio, Variância, Desvio Padrão e Coeficiente de Variação

Todos esses conceitos estatísticos medem a **dispersão dos dados**, ou seja, o quão espalhados os valores estão em relação à média. Cada um tem um propósito específico.

---

## \$ 1. Amplitude – "A Diferença Entre o Maior e o Menor Valor"

A **amplitude** é a forma mais simples de medir a dispersão. Ela **considera apenas os extremos** e ignora os valores intermediários.

### Fórmula da Amplitude:

\$

$$\text{Amplitude} = X_{\text{máx}} - X_{\text{mín}}$$

\$

### ✓ Quando usar?

- Quando precisamos de uma **medida rápida e fácil** da dispersão.
- Pode ser **enganosa** se houver outliers, pois considera apenas dois valores.

### Exemplo Prático:

Se os tempos de entrega de pizza forem **25, 30, 28, 22 e 35 minutos**, a amplitude será:

\$

$$\text{Amplitude} = 35 - 22 = 13$$

\$

Ou seja, a maior diferença entre os tempos foi de **13 minutos**.

---

## 2. Desvio Médio – "Média das Diferenças Absolutas"

O **desvio médio** calcula a **média das diferenças absolutas** em relação à média.

### Fórmula do Desvio Médio:

\$

$$DM = \frac{\sum |X_i - \mu|}{n}$$

\$

### ✓ Quando usar?

- Quando queremos uma medida de dispersão **intuitiva e fácil de interpretar**.
- **Menos sensível a outliers** do que a variância e o desvio padrão.

### Exemplo Prático:

Se a média do tempo de entrega for **28 minutos**, e as diferenças absolutas forem **3, 2, 0, 6 e 7**, o desvio médio será:

\$

$$\frac{3+2+0+6+7}{5} = 3.6$$

\$

---

### 3. Variância $(\sigma^2)$ – "Média das Diferenças Elevadas ao Quadrado"

A **variância** mede a dispersão dos dados **elevando ao quadrado** as diferenças entre cada ponto e a média.

**Fórmula da Variância:**

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$$

✓ **Quando usar?**

- Quando queremos uma **medida mais rigorosa da dispersão**.
- Usada em cálculos estatísticos como **regressão e machine learning**.

**Exemplo Prático:**

Se a média do tempo de entrega for **28 minutos**, e as diferenças ao quadrado forem **9, 4, 0, 36 e 49**, a variância será:

$$\frac{9+4+0+36+49}{5} = 19.6$$

---

### 4. Desvio Padrão $(\sigma)$ – "Raiz Quadrada da Variância"

O **desvio padrão** é simplesmente a **raiz quadrada da variância**, mantendo a mesma unidade dos dados.

**Fórmula do Desvio Padrão:**

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{n}}$$

✓ **Quando usar?**

- Quando queremos uma **medida de dispersão intuitiva** na **mesma unidade dos dados**.
- Muito usado em **estatística descritiva e inferencial**.

**Exemplo Prático:**

Se a variância dos tempos de entrega for **19.6**, então o desvio padrão será:

$$\sqrt{19.6} \approx 4.43$$

Ou seja, em média, os tempos de entrega variam **4.43 minutos** da média.

---

### 5. Coeficiente de Variação (CV) – "Dispersão Relativa"

O **coeficiente de variação** mede a **dispersão em relação à média**. Diferente dos outros métodos, ele é **expresso em porcentagem**, permitindo comparar dispersões de diferentes conjuntos de dados.

## Fórmula do Coeficiente de Variação:

\$

$$CV = \left( \frac{\sigma}{\mu} \right) \times 100\%$$

\$

### ✓ Quando usar?

- Quando queremos **comparar a variabilidade de dois conjuntos de dados com unidades diferentes**.
- Útil para avaliar **consistência em medidas financeiras, industriais e científicas**.

### Exemplo Prático:

Se a média do tempo de entrega for **28 minutos** e o desvio padrão for **4.43 minutos**, então:

\$

$$CV = \left( \frac{4.43}{28} \right) \times 100\% \approx 15.8\%$$

\$

Isso significa que **a variação dos tempos de entrega representa cerca de 15.8% da média**.

---

## Resumo

Medida	O que significa?	Fórmula
<b>Amplitude</b>	Diferença entre o maior e o menor valor	$X_{\text{máx}} - X_{\text{mín}}$
<b>Desvio Médio</b>	Média das diferenças absolutas em relação à média	$\frac{\sum$
<b>Variância</b>	Média das diferenças quadradas em relação à média	$\frac{\sum (X_i - \mu)^2}{n}$
<b>Desvio Padrão</b>	Raiz quadrada da variância, mantém a unidade original dos dados	$\sqrt{\frac{\sum (X_i - \mu)^2}{n}}$
<b>Coeficiente de Variação</b>	Dispersão em relação à média, expresso em %	$\left( \frac{\sigma}{\mu} \right) \times 100\%$

### Dica prática:

- **Amplitude:** Boa para uma análise inicial, mas não confiável.
- **Desvio Médio:** Mais intuitivo e fácil de interpretar.
- **Variância:** Mais precisa, mas difícil de entender.
- **Desvio Padrão:** Melhor medida geral de dispersão.
- **Coeficiente de Variação:** Melhor para **comparar dados de naturezas diferentes**.

## Exemplo Prático em Python

Aqui está um exemplo de como calcular medidas de dispersão em Python usando a biblioteca **numpy**:

```
import numpy as np

dados = [10, 12, 23, 23, 16, 23, 21, 16, 18, 19]

amplitude = np.ptp(dados) # Diferença entre máximo e mínimo
variancia = np.var(dados, ddof=1) # Variância amostral
desvio_padrao = np.std(dados, ddof=1) # Desvio padrão amostral
cv = (desvio_padrao / np.mean(dados)) * 100 # Coeficiente de variação
iqr = np.percentile(dados, 75) - np.percentile(dados, 25) # Intervalo interquartil

print(f"Amplitude: {amplitude}")
print(f"Variância: {variancia:.2f}")
print(f"Desvio Padrão: {desvio_padrao:.2f}")
print(f"Coeficiente de Variação: {cv:.2f}%")
print(f"Intervalo Interquartil (IQR): {iqr}")
```

Exemplo de interpretação. Dado o contexto:

- Média ( $\mu$ ) = 5
- Variância ( $\sigma^2$ ) = 12
- Valor observado (x) = 4

Queremos **interpretar o valor 4** dentro desse conjunto.

### Etapa 1: Entendendo o que é a variância

A **variância** mede **o quão espalhados** estão os dados em relação à média.

- Se a variância é **baixa**, os dados estão **concentrados perto da média**.
- Se a variância é **alta**, os dados estão **mais espalhados**.

No seu caso, a variância é 12. Isso indica um espalhamento **razoável** (nem muito pequeno, nem gigantesco).

### Etapa 2: Interpretar o valor 4 em relação à média

A média é 5. O valor 4 está **abaixo da média**:

$$\begin{aligned} & \$ \\ x - \mu &= 4 - 5 = -1 \\ & \$ \end{aligned}$$

Ou seja, esse valor está **1 unidade abaixo da média**.

### Etapa 3: Transformar isso em desvio padrão

Para entender o quanto esse "1" representa, a gente precisa converter em **desvio padrão**, pois a variância sozinha é difícil de interpretar.

\$

$$\sigma = \sqrt{12} \approx 3.46$$

\$

---

#### Etapa 4: Calcular o Z-score

Vamos ver quantos desvios padrão o valor 4 está afastado da média:

\$

$$Z = \frac{x - \mu}{\sigma} = \frac{4 - 5}{3.46} \approx -0.29$$

\$

---

#### Interpretação:

- Um Z-score de **-0.29** significa que o valor 4 está **0,29 desvios padrão abaixo da média**.
- Como isso está **próximo de zero**, podemos dizer que:
  - **É um valor comum, nada extremo.**
  - **Está dentro da variação esperada** dos dados.
  - Não é considerado outlier, nem um valor incomum.

---

#### Resumo Didático:

Se a variância é 12 e a média é 5, um valor 4 está **levemente abaixo da média**, mas **totalmente dentro do esperado**, pois a dispersão dos dados é grande ( $\sigma \approx 3.46$ ), então essa diferença é pequena em comparação com a "espalhabilidade" do conjunto.

---

#### Exemplo: Temperatura corporal em uma clínica

Suponha que uma clínica médica coletou as temperaturas corporais (em °C) de 10 pacientes:

[36.7, 36.9, 37.0, 36.8, 36.5, 36.6, 36.9, 37.1, 36.8, 39.0]

Repare que todas as temperaturas estão próximas de 37, **menos uma: 39.0°C**.

---

#### Passo 1: Calcular a média e o desvio padrão

Vamos calcular:

- **Média ( $\mu$ ):** soma de todos os valores ÷ número de valores
- **Desvio padrão ( $\sigma$ ):** raiz da variância

```
import numpy as np

dados = [36.7, 36.9, 37.0, 36.8, 36.5, 36.6, 36.9, 37.1, 36.8, 39.0]
media = np.mean(dados)
desvio_padrao = np.std(dados)

print(f"Média: {media:.2f}")
print(f"Desvio padrão: {desvio_padrao:.2f}")
```

### Resultado:

```
Média: 37.03
Desvio padrão: 0.66
```

---

### Passo 2: Calcular o Z-score para o valor 39.0

Agora aplicamos a fórmula do Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{39.0 - 37.03}{0.66} \approx 2.98$$

---

### Interpretação:

- O valor **39.0°C** tem um **Z-score de aproximadamente 2.98**
- Isso significa que ele está **quase 3 desvios padrão acima da média**
- Como regra geral:
  - Valores com  $|Z| > 2$  são **potencialmente extremos**
  - Valores com  $|Z| > 3$  são **prováveis outliers**

### Conclusão:

O valor **39.0°C** é um **outlier**, pois está  **muito distante da média** comparado aos demais.

---

### Visualização (opcional)

Você pode usar **matplotlib** para plotar os dados e destacar o outlier:

```
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 4))
plt.plot(dados, marker='o', linestyle='-', label="Temperaturas")
plt.axhline(media, color='green', linestyle='--', label="Média")
```

```
plt.axhline(media + 2*desvio_padrao, color='red', linestyle='--',
label="+2σ")
plt.axhline(media - 2*desvio_padrao, color='red', linestyle='--',
label="-2σ")
plt.title("Temperaturas com possível outlier")
plt.legend()
plt.grid()
plt.show()
```

## Identificação de dispersão

### Z-Score: Entendendo o Cálculo do Desvio Padrão com Z-Score

O **Z-score** (ou **pontuação z**) é uma medida estatística que descreve a posição de um valor em relação à média de um conjunto de dados. Ele indica quantos **desvios padrões** um valor está afastado da média. O Z-score é frequentemente usado para identificar valores extremos ou outliers, especialmente em distribuições normais.

#### Fórmula do Z-Score

A fórmula básica do Z-score é:

$$Z = \frac{X - \mu}{\sigma}$$

Onde:

- **X**: O valor individual que estamos analisando.
- **\$\mu\$** ( $\mu$ ): A média dos dados.
- **\$\sigma\$** ( $\sigma$ ): O desvio padrão dos dados.

#### Explicando os Componentes:

1. **X**: Este é o valor específico para o qual queremos calcular o Z-score. Pode ser, por exemplo, a nota de um aluno em uma prova ou a altura de uma pessoa em um estudo de crescimento.
2. **\$\mu\$**: A **média** de todos os valores no conjunto de dados. Ela é calculada somando todos os valores e dividindo pela quantidade de elementos:

$$\mu = \frac{1}{n} \sum_{i=1}^n X_i$$

Onde  $n$  é o número total de dados e  $X_i$  são os valores individuais.

3. **\$\sigma\$**: O **desvio padrão** indica a dispersão dos dados em relação à média. Ele é calculado pela fórmula:



\$

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

\$

### Interpretação do Z-score

- **Z = 0**: O valor \$ X ) está exatamente na média.
- **Z > 0**: O valor \$ X ) está acima da média.
- **Z < 0**: O valor \$ X ) está abaixo da média.
- **Z > 3 ou Z < -3**: O valor \$ X ) é considerado um **outlier**, pois está mais de 3 desvios padrões da média, o que é uma diferença significativa.

---

### Exemplo Prático de Cálculo do Z-score

Vamos calcular o Z-score de um valor usando um conjunto de dados simples. Suponha que temos as notas de 5 alunos em uma prova:

70, 75, 80, 85, 90

Queremos calcular o Z-score para o aluno que obteve a nota **85**.

#### Passo 1: Calcular a Média (\$\mu\$)

A média das notas é:

\$

$$\mu = \frac{70 + 75 + 80 + 85 + 90}{5} = \frac{400}{5} = 80$$

\$

#### Passo 2: Calcular o Desvio Padrão (\$\sigma\$)

Agora, vamos calcular o desvio padrão das notas. A fórmula é:

\$

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

\$

Substituindo os valores:

\$

$$\sigma = \sqrt{\frac{1}{5} \left( (70 - 80)^2 + (75 - 80)^2 + (80 - 80)^2 + (85 - 80)^2 + (90 - 80)^2 \right)}$$

\$

\$

$$\sigma = \sqrt{\frac{1}{5} \left( 100 + 25 + 0 + 25 + 100 \right)} = \sqrt{\frac{250}{5}} = \sqrt{50} \approx$$

7.07

\$

### Passo 3: Calcular o Z-score

Agora, podemos calcular o Z-score para a nota **85**:

\$

$$Z = \frac{X - \mu}{\sigma} = \frac{85 - 80}{7.07} = \frac{5}{7.07} \approx 0.71$$

\$

O Z-score da nota **85** é **0.71**. Isso significa que a nota do aluno está **0.71 desvios padrões acima da média**.

---

### Como Usar o Z-score para Encontrar Outliers

Uma das utilidades mais comuns do Z-score é **identificar outliers**. Em uma distribuição normal (ou quase normal), valores com Z-scores maiores que 3 ou menores que -3 são considerados outliers. Isso ocorre porque, em uma distribuição normal padrão:

- **68%** dos dados estarão dentro de **1 desvio padrão** da média (Z entre -1 e 1).
- **95%** dos dados estarão dentro de **2 desvios padrões** da média (Z entre -2 e 2).
- **99.7%** dos dados estarão dentro de **3 desvios padrões** da média (Z entre -3 e 3).

Portanto, qualquer valor com um Z-score superior a **3** ou inferior a **-3** está consideravelmente afastado da média e pode ser classificado como um outlier.

---

### Vantagens do Z-score

- **Facilidade de interpretação:** O Z-score é intuitivo, pois quantifica o quão distante um valor está da média em termos de desvios padrões.
- **Universalidade:** Pode ser aplicado a qualquer distribuição de dados, desde que os dados não sejam extremamente assimétricos.

### Desvantagens do Z-score

- **Sensibilidade a distribuições não normais:** O Z-score pode ser menos útil em distribuições assimétricas ou com caudas longas, onde os dados não seguem uma distribuição normal.
  - **Assume normalidade:** A interpretação do Z-score assume que os dados se aproximam de uma distribuição normal. Para distribuições muito diferentes da normal, outras técnicas podem ser mais apropriadas para detectar outliers.
- 

### Conclusão

O Z-score é uma maneira poderosa de medir a posição de um valor dentro de um conjunto de dados, especialmente para identificar outliers. Ele utiliza a média e o desvio padrão para determinar quantos desvios padrões um valor está afastado da média, ajudando a identificar valores extremos que podem

distorcer análises estatísticas. Com esse entendimento, é possível avaliar de forma mais rigorosa a consistência e a confiabilidade dos dados em diferentes cenários.

# identificação de outliers

---

A **identificação de outliers** é um passo essencial na análise de dados, pois esses valores atípicos podem distorcer medidas estatísticas e comprometer a qualidade dos modelos preditivos. Um *outlier* é um valor que se distancia significativamente da maioria dos dados, podendo ser resultado de erros de medição, entrada de dados ou, em alguns casos, indicar uma descoberta importante.

## Principais Métodos de Identificação de Outliers

### 1. Método do Intervalo Interquartil (IQR)

Um dos métodos mais utilizados consiste em calcular o intervalo interquartil (IQR), definido como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). Valores que estão abaixo de  $Q1 - 1.5 * IQR$  ou acima de  $Q3 + 1.5 * IQR$  são considerados outliers (Tukey, 1977).

### 2. Z-score (ou escore padronizado)

O Z-score mede quantos desvios-padrão um dado está distante da média. Valores com Z-score superior a 3 ou inferior a -3 são geralmente classificados como outliers (Barnett & Lewis, 1994).

### 3. Métodos Baseados em Modelos

Algoritmos como *Isolation Forest*, *Local Outlier Factor (LOF)* e *DBSCAN* são utilizados em contextos mais complexos e de alta dimensionalidade (Breunig et al., 2000).

## Por Que Detectar Outliers?

- **Melhora a qualidade dos dados**
- **Aumenta a robustez dos modelos de aprendizado de máquina**
- **Evita conclusões estatísticas enganosas**

## Citações Importantes

- "Outlier detection is an essential step in data preprocessing and has applications in fraud detection, fault diagnosis, and system health monitoring." – *Chandola et al., 2009*
- "Outliers, or extreme observations, may carry valuable information about the process under study, or they may be simply due to errors." – *Barnett & Lewis, 1994*

Se quiser, posso te ajudar com um exemplo prático em Python ou um gráfico ilustrativo. Deseja seguir por esse caminho?

## Referências

- BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 8. ed. São Paulo: Saraiva, 2017.
- MONTGOMERY, D. C.; RUNGER, G. C. *Applied Statistics and Probability for Engineers*. 7th ed. John Wiley & Sons, 2018.
- TRIOLA, M. F. *Introdução à Estatística*. 13. ed. Pearson, 2020.
- TUKEY, J. W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

