A similaridade entre os termos **desvio padrão** e **erro padrão** podem gerar um equívoco quanto ao significado, utilização e até mesmo o cálculo dessas duas medidas. Buscando superar qualquer engano, nessa atividade vamos aprender sobre ambos os conceitos.

Para facilitar a explicação, vamos começar com um exemplo. Suponhamos que Valquíria é gerente de uma empresa que vende barras de chocolate. Nessa empresa, existe uma máquina que produz as barras de chocolate com um peso nominal de 100 gramas.

Valquíria quer garantir que, em média, cada pacote contenha a quantidade certa de chocolates, mas está ciente de que pode haver alguma variabilidade no peso de cada barra.

Ela instruiu seus funcionários a coletar 10 barras de chocolate da máquina e medir o peso de cada uma. Após a coleta dos dados, o funcionário calcula a média do peso das barras e descobre que a média da amostra é de, aproximadamente, 100 gramas. No entanto, ele avisa que o peso de algumas barras não era exatamente 100 gramas e mostra a tabela de pesos:

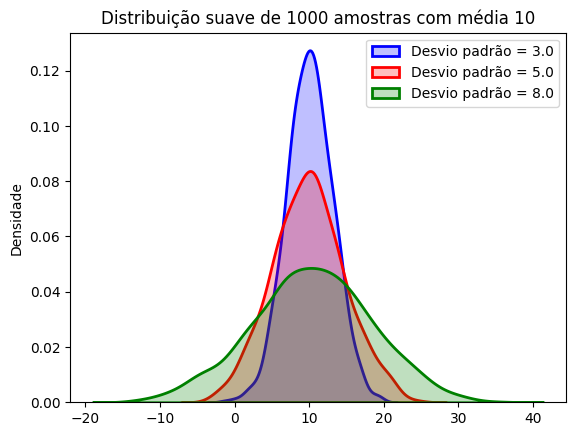
|  | **Peso das barras de chocolate (g)** |
| --- | --- |
| 0 | 99.4197 |
| 1 | 101.821 |
| 2 | 98.5708 |
| 3 | 100.63 |
| 4 | 94.7449 |
| 5 | 99.3241 |
| 6 | 97.6067 |
| 7 | 100.679 |
| 8 | 108.739 |
| 9 | 99.0544 |

Desejando entender a **dispersão dos pesos na amostra em relação a média amostral**, Valquíria recorre ao cálculo do desvio padrão.

## **Desvio padrão**

O **desvio padrão é uma medida de dispersão que quantifica a variabilidade ou espalhamento dos dados em uma única distribuição**. Isso significa que podemos utilizar essa métrica para medir a variabilidade de pesos dentro das amostras, indicando quão uniforme é aquela produção em termos de peso das barras de chocolate.

Uma forma de visualizar como o valor do desvio padrão retrata a variabilidade da amostra é observando o gráfico abaixo.



No gráfico, temos a distribuição de 3 conjuntos amostrais de 1000 amostras, em ambos a média amostral é igual a 10. O conjunto azul apresenta o menor desvio padrão, enquanto o conjunto verde possui o maior desvio padrão. O conjunto vermelho se encontra entre os dois. Observe que o conjunto em azul está muito mais concentrado em torno do valor da média (10) em comparação com os outros conjuntos, sendo o conjunto verde o que menos se centraliza em torno da média.

A partir dessa representação, podemos observar que um desvio padrão menor indicaria que as amostras tendem a ter valores mais convergentes à média, enquanto um desvio padrão maior indicaria uma maior variabilidade nesses valores.

Trazendo para o exemplo das barras de chocolate, um valor baixo de desvio padrão poderia indicar que as barras tendem a ter pesos mais consistentes dentro da amostra, próximos a 100 gramas.

Mas como podemos calcular essa medida?

### **Calculando o desvio padrão matematicamente**

O cálculo matemático do desvio padrão envolve várias etapas. Vamos percorrer essas etapas passo a passo:

#### **Etapa 1: Cálculo da média amostral.**

Já temos a informação que a média amostral é 100.0. No entanto, para ter um valor exato, vamos somar todos os valores e dividir pela quantidade de amostras.

* Soma de todos os valores amostrais: 1000.5896;
* Quantidade de amostras: 10;
* Média: 1000.5896 / 10 = 100.05896.

#### **Etapa 2: Cálculo da diferenças entre cada valor e a média**

Para cada valor nos dados, subtrai-se a média calculada na Etapa 1.

Para exemplificar vamos executar esse passo na primeira linha dos dados. Coletamos 99.4197, subtraímos 100.05896 e anotamos o valor resultante, 99.4197 - 100.05896 = -0.63926.

**Resultado**

|  | **amostras** | **diferença valor e média** |
| --- | --- | --- |
| 0 | 99.4197 | -0.63926 |
| 1 | 101.821 | 1.76204 |
| 2 | 98.5708 | -1.48816 |
| 3 | 100.63 | 0.57104 |
| 4 | 94.7449 | -5.31406 |
| 5 | 99.3241 | -0.73486 |
| 6 | 97.6067 | -2.45226 |
| 7 | 100.679 | 0.62004 |
| 8 | 108.739 | 8.68004 |
| 9 | 99.0544 | -1.00456 |

#### **Etapa 3: Elevar cada diferença ao quadrado**

Todas as diferenças de valores são elevadas ao quadrado. Este passo é crucial porque considera tanto as diferenças positivas quanto as negativas e dá mais peso às diferenças maiores.

**Resultado:**

|  | **amostras** | **diferença valor e média** | **quadrado da diferença** |
| --- | --- | --- | --- |
| 0 | 99.4197 | -0.63926 | 0.408653 |
| 1 | 101.821 | 1.76204 | 3.10478 |
| 2 | 98.5708 | -1.48816 | 2.21462 |
| 3 | 100.63 | 0.57104 | 0.326087 |
| 4 | 94.7449 | -5.31406 | 28.2392 |
| 5 | 99.3241 | -0.73486 | 0.540019 |
| 6 | 97.6067 | -2.45226 | 6.01358 |
| 7 | 100.679 | 0.62004 | 0.38445 |
| 8 | 108.739 | 8.68004 | 75.3431 |
| 9 | 99.0544 | -1.00456 | 1.00914 |

#### **Etapa 4: Cálculo da soma das diferenças ao quadrado**

Somando todos os valores na coluna quadrado da diferença obtemos o valor 117.584

#### **Etapa 5: Divisão da soma pelo número total de observações menos 1:**

Dividindo o valor 117.584 pelo total de observações menos 1, 10 - 1 = 9, obtemos 13.065. Esse resultado é chamado de [variância](https://pt.wikipedia.org/wiki/Vari%C3%A2ncia) amostral.

A principal razão para a subtração de 1 na fórmula da variância da amostra é compensar o fato de que estamos estimando a variabilidade com base em uma amostra e não em toda a população. Essa correção, conhecida como correção de Bessel, reduz o viés amostral na estimativa da variância.

#### **Etapa 6: Cálculo da raiz quadrada da variância para obter o desvio padrão**

O desvio padrão é a raiz quadrada da variância. Isso é feito para voltar à escala original dos dados, uma vez que a variância é medida em unidades ao quadrado.

O resultado aproximado da raiz quadrada de nossa variância é, aproximadamente, 3.614.

### **Calculando o desvio padrão com Python**

Todo esse cálculo pode ser simplificado através da programação em Python. Utilizamos o [método std() da biblioteca Pandas](https://pandas.pydata.org/pandas-docs/version/2.1/reference/api/pandas.DataFrame.std.html) no nosso dataframe com os valores de pesos amostra\_10\_pesos:

amostra\_10\_pesos['Peso das barras de chocolate (g)'].std()

**Saída:**

3.614533349072264

Verificamos que obtemos o mesmo resultado. O método std() do pandas por padrão já considera o cálculo do desvio padrão amostral.

Se tiver interesse em observar a distribuição suave dos pesos ou realizar os cálculos matemáticos apresentados na atividade, você pode utilizar o código abaixo:

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

amostra\_10\_pesos = pd.DataFrame({'Peso das barras de chocolate (g)': [99.4197, 101.821, 98.5708, 100.63, 94.7449, 99.3241, 97.6067, 100.679, 108.739, 99.0544]})

dp = amostra\_10\_pesos['Peso das barras de chocolate (g)'].std()

sns.kdeplot(amostra\_10\_pesos['Peso das barras de chocolate (g)'], linewidth=2, fill= True, label=f'Desvio padrão = {dp:.3f}')

plt.legend()

plt.ylabel('Frequência')

plt.title('Distribuição suave pesos de barra amostrais')

plt.show()

Após o cálculo do desvio padrão, Valquíria conseguiu entender a medida de dispersão da sua amostra. No entanto, logo surgiu uma dúvida: “Quão confiável é a média amostral para refletir a verdadeira média populacional do peso das barras?”

Para tirar essa dúvida, ela recorreu ao cálculo do erro padrão.

## **Erro padrão**

O erro padrão é uma **medida de variação da média amostral** em relação à **média da população**.

No contexto das barras de chocolate, o erro padrão representaria a variabilidade na média do peso das barras entre diferentes pacotes. Com esse valor calculado, Valquíria não precisa pedir para seus funcionários medirem os pesos de todas as barras que a máquina produziu, pois ela poderá utilizar essa medida para, posteriormente, definir um **intervalo de confiança**.

Observação: O conceito de intervalo de confiança será tratado mais adiante no curso.

### **Calculando o erro padrão matematicamente**

Conseguimos calcular matematicamente o erro padrão amostral dividindo o valor do desvio padrão amostral pela raiz quadrada da quantidade de amostras. Assim, teremos:

* Desvio padrão amostral: 3.614;
* Número de amostras: 10;
* Erro padrão amostral: 3.614 / sqrt(10) = 1.143, aproximadamente.

### **Calculando o erro padrão com Python**

Com Python, podemos aplicar o cálculo matemático que aprendemos durante a atividade:

import numpy as np

import pandas as pd

amostra\_10\_pesos = pd.DataFrame({'Peso das barras de chocolate (g)': [99.4197, 101.821, 98.5708, 100.63, 94.7449, 99.3241, 97.6067, 100.679, 108.739, 99.0544]})

# Desvio padrão

dp\_amostral = amostra\_10\_pesos['Peso das barras de chocolate (g)'].std()

# Tamanho da amostra

tamanho\_amostra = len(amostra\_10\_pesos)

# Calculando o Erro Padrão da Amostra

erro\_padrao\_amostral = dp\_amostral / np.sqrt(tamanho\_amostra)

print('Erro padrão:', erro\_padrao\_amostral )

**Saída:**

Erro padrão: 1.1430158061704814

Um modo mais direto para obter esse valor é utilizar a função [stats.sem](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.sem.html) da biblioteca [scipy](https://docs.scipy.org/doc/scipy/index.html):

import pandas as pd

amostra\_10\_pesos = pd.DataFrame({'Peso das barras de chocolate (g)': [99.4197, 101.821, 98.5708, 100.63, 94.7449, 99.3241, 97.6067, 100.679, 108.739, 99.0544]})

## Calculando o Erro Padrão da Amostra

from scipy import stats

erro\_padrao\_amostral = stats.sem(amostra\_10\_pesos['Peso das barras de chocolate (g)'])

print('Erro padrão:', erro\_padrao\_amostral)

**Saída:**

Erro padrão: 1.1430158061704814

Um erro padrão menor indica que as médias amostrais estão mais consistentemente próximas da verdadeira média populacional, enquanto um erro padrão maior sugere maior variabilidade nas médias amostrais.

## **Diferença entre desvio padrão e erro padrão**

Em termos simples, enquanto o desvio padrão nos diz o quão espalhados estão os dados individuais, o erro padrão nos mostra o quão precisa é a média de uma amostra como estimativa da média da população.

Em termos técnicos, o desvio padrão é a medida de variabilidade individual dos dados e o erro padrão é a medida de variabilidade associada à média amostral. Portanto, podemos resumir que desvio padrão é sobre dados individuais, enquanto o erro padrão diz respeito à confiabilidade da média de uma amostra.