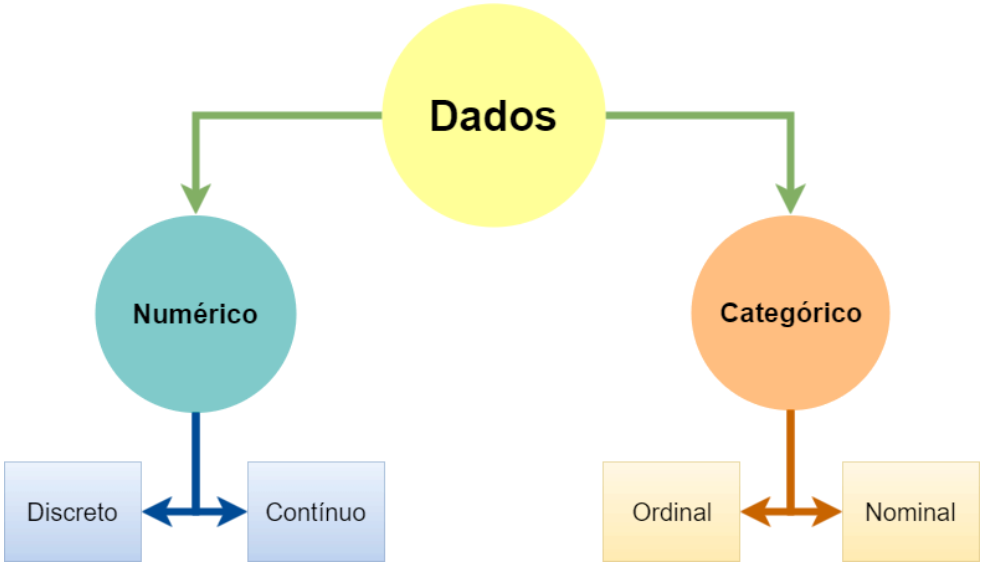
Relatório do bootcamp - 5

Thassiana C. A. Muller

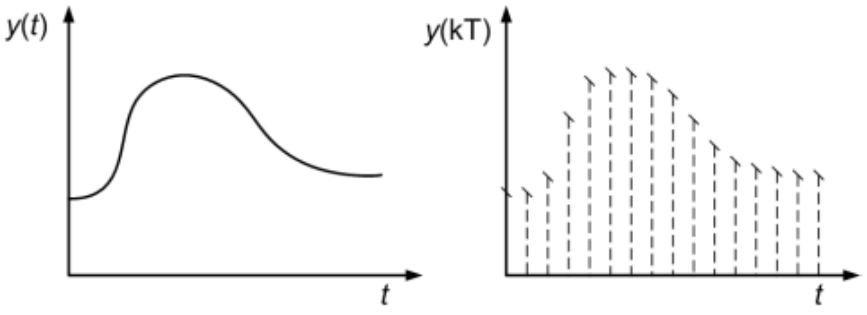
Para realizar boas técnicas de manipulação dos dados é necessário compreender que há diferentes tipos de dados como, por exemplo, temperatura, idade, profissão ou rankings. Assim, a depender do que deseja-se analisar, os dados podem ser do tipo número ou categórico.



Dados Numéricos

Os dados numéricos refletem, em geral, informações de medidas e/ou quantidades e podem ser classificados em duas categorias: discretos e contínuos. Dados discretos representam medidas que não possuem valores intermediários na escala, enquanto dados contínuos podem assumir infinitos valores intermediários.

Os gráficos abaixo, representam respectivamente dados contínuos e dados discretos, no segundo gráfico é possível observar a ausência de informações entre as amostras.



Exemplos de dados numéricos contínuos são:

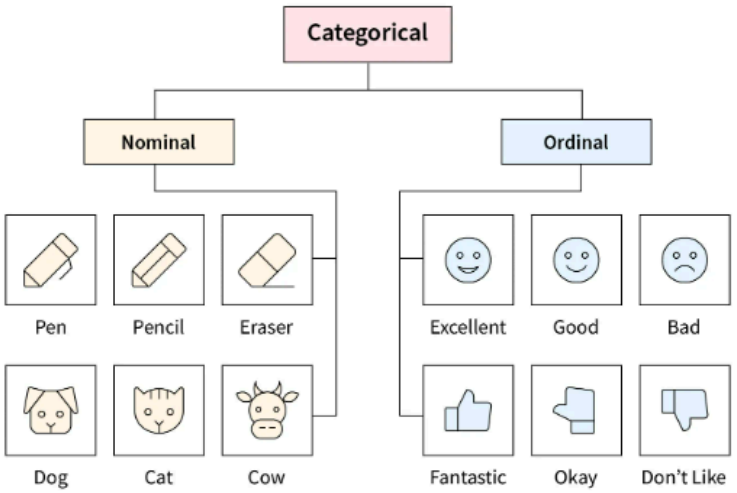
- Quanto tempo leva para um usuário realizar uma compra;
- Volume de chuva;
- Temperatura atual;
- Peso de uma pessoa.

Exemplos de dados numéricos discretos:

- Idade;
- Valor de uma compra;
- Quantidade de itens em estoque;
- Tamanho de roupas.

Dados Categóricos

Os dados categóricos, em geral qualitativos, possuem uma quantidade restrita de possibilidades (categorias) e são utilizados em análises estatísticas para identificar padrões e tendências dentro de grupos distintos. Dados categóricos podem ser ordinais, quando as categorias possuem uma ordem ou hierarquia natural, ou nominais quando as categorias não têm uma ordem específica.



Exemplos de dados categóricos ordinais:

- Classificações de satisfação (como insatisfeito, neutro, satisfeito);
- Níveis de escolaridade;
- Rankings em geral.

Exemplos de dados categóricos nominais:

- Cores;
- Filos de animais;
- Gêneros.

Média, moda e mediana

A média, a moda e a mediana são medidas que retratam diferentes perspectivas sobre um conjunto de dados.

A média aritmética é obtida somando todos os valores e dividindo pelo número total de valores, representando o valor intermediário do conjunto de dados.

$$\overline{x} = \frac{\sum x_i}{n}$$

A moda é o valor que ocorre com mais frequência no conjunto de dados e pode haver mais de uma moda se múltiplos valores ocorrerem com a mesma frequência.

A mediana é o valor central quando os dados são ordenados em ordem crescente, dividindo o conjunto de dados em duas metades iguais.

Exemplo:

```
dados = [10, 15, 20, 25, 25, 30, 35, 40, 40, 40, 45, 50, 55]
```

```
media = np.mean(dados)
moda = stats.mode(dados)[0]
mediana = np.median(dados)

print(f'Média: {media}')
print(f'Moda: {moda}')
print(f'Mediana: {mediana}')
```

```
Média: 33.07692307692308
Moda: 40
Mediana: 35.0
```

Variância e desvio padrão

A variância e o desvio padrão são medidas de dispersão que indicam o grau de espalhamento dos dados em torno da média.

A variância é calculada como a média dos quadrados das diferenças entre cada valor e a média do conjunto de dados, quanto menor é a variância, mais próximos os valores estão da média. No entanto, é necessário levar em consideração se os dados analisados são populacionais ou se são apenas algumas amostras retiradas da população[1].

Exemplo:

Fórmula da variância populacional:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \quad \text{ou} \quad \sigma^2 = \overline{x^2} - (\bar{x})^2$$

Fórmula da variância amostral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \quad \text{ou} \quad s^2 = \left[\overline{x^2} - (\bar{x})^2\right] \times \left(\frac{n}{n-1}\right)$$

μ = valores reais populacionais

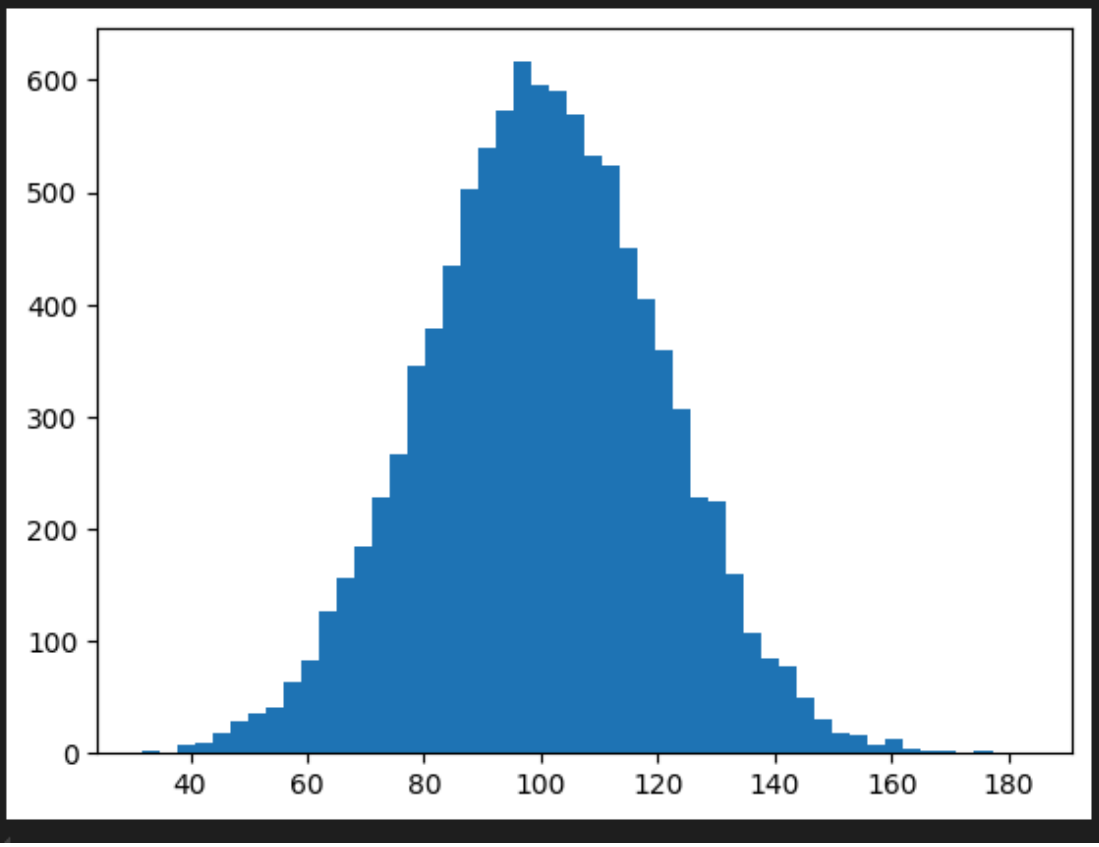
O desvio padrão é a raiz quadrada da variância e fornece uma medida da dispersão com correção de unidade dos dados originais, facilitando a interpretação.

[1]O ajuste (n-1) da fórmula de variância amostral é conhecido como correção de Bessel e é utilizado para obter uma estimativa não viesada da variância a partir de uma amostra.

Exemplo, dado essa distribuição normal:

```
dados = np.random.normal(100.0, 20.0, 10000)
```
✓ 0.0s

```
plt.hist(dados, 50)
plt.show()
```
✓ 0.1s




```
dados.var()
```
✓ 0.0s

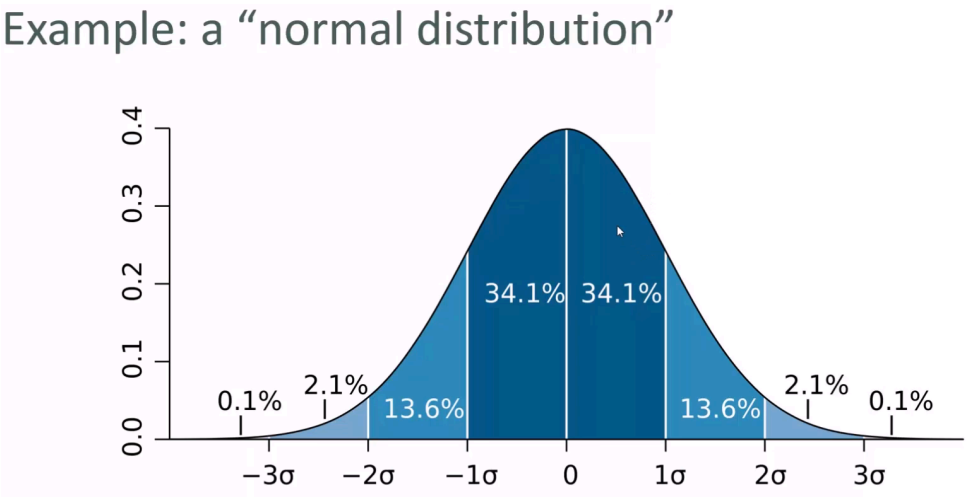
```
406.614442067382
```

```
print(dados.std())
print(np.sqrt(dados.var()))
```
✓ 0.0s

```
20.164683039100368
20.164683039100368
```

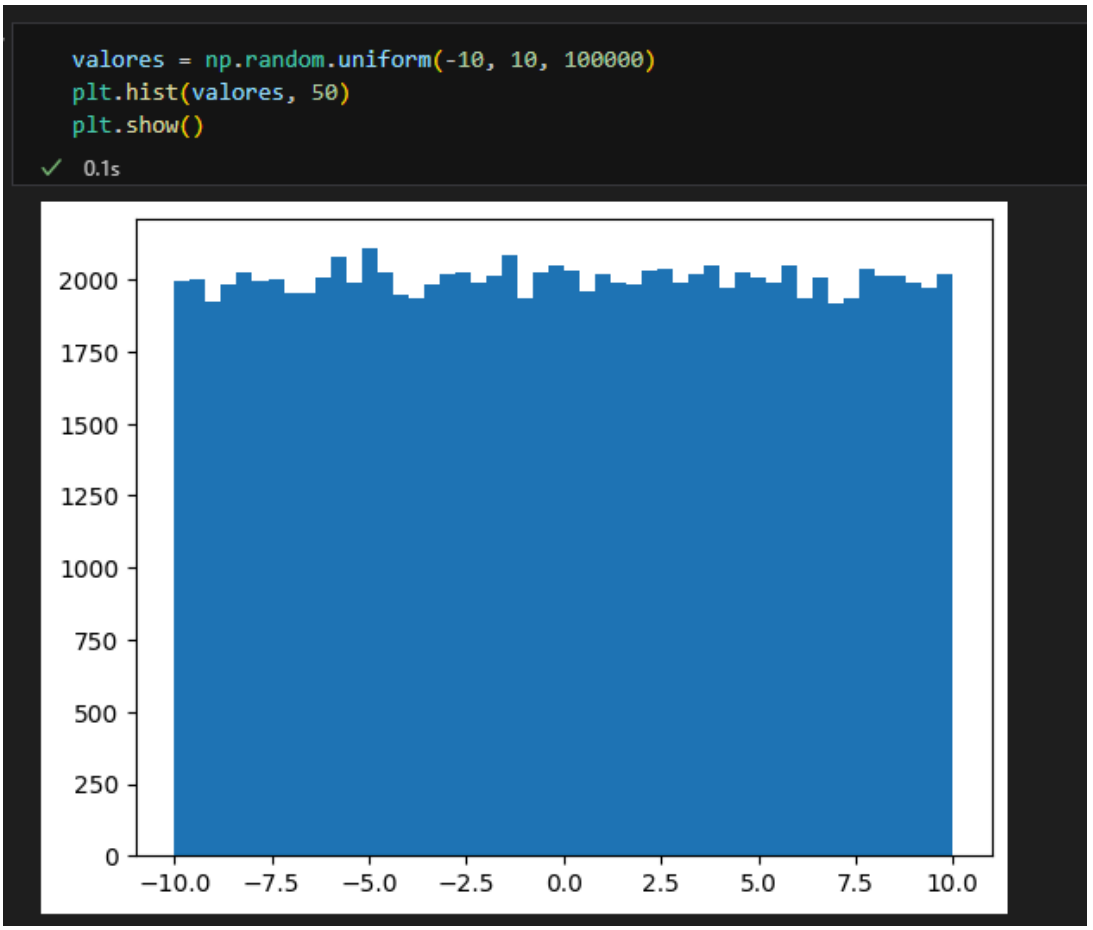
Distribuição normal

A partir dos estudos de desvio padrão (σ), é possível construir um gráfico da distribuição normal (Gaussiana) de um conjunto de dados.



Distribuição uniforme

Como o nome sugere, os dados estão igualmente distribuídos.



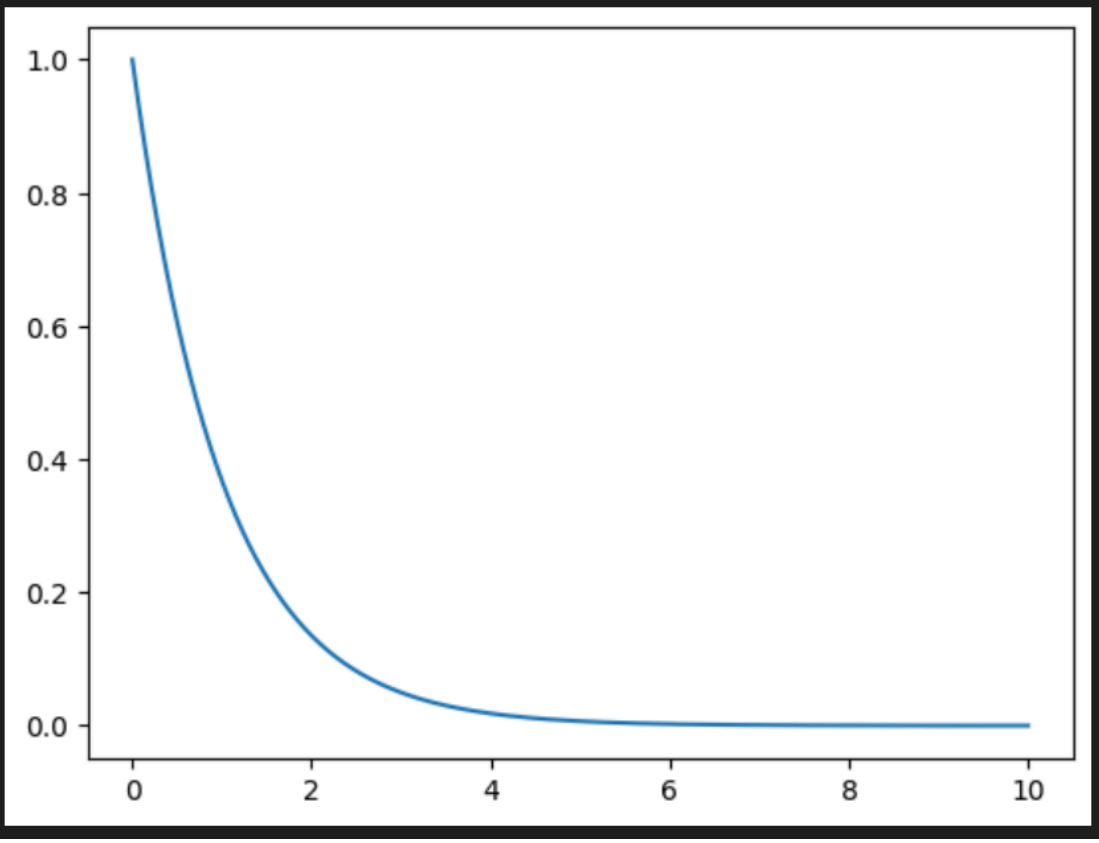
Distribuição exponencial

Evolui exponencialmente

```
dados = np.arange(0, 10, 0.0001)
plt.plot(dados, expon.pdf(dados))
```

✓ 0.1s

```
[<matplotlib.lines.Line2D at 0x174507a9370>]
```



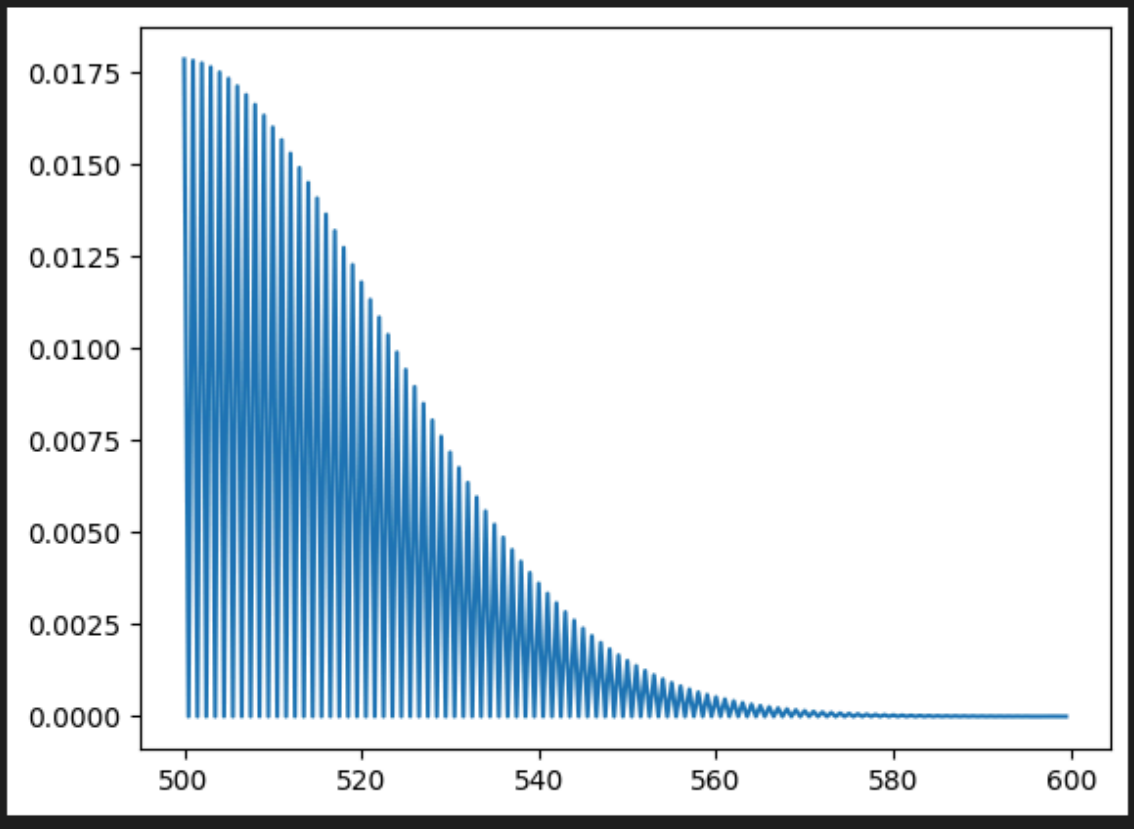
Distribuição Poisson

Diferente da distribuição normal, a distribuição Poisson descreve a probabilidade de um evento ocorrer durante um determinado intervalo de tempo, quando a probabilidade de um evento ocorrer é bem baixa e a quantidade de tentativas é bastante grande. A distribuição de Poisson é frequentemente usada para modelar eventos raros, como o número de chamadas recebidas por um call center em uma hora ou o número de acidentes de carro em um determinado cruzamento por mês.

```
mu = 500
dados = np.arange(500, 600, 0.5)
plt.plot(dados, poisson.pmf(dados, mu))
```

✓ 0.1s

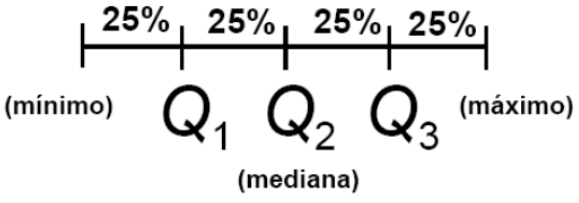
```
[<matplotlib.lines.Line2D at 0x174525152e0>]
```



Percentis

Percentis dividem um conjunto de dados ordenados em 100 partes iguais permitindo identificar a posição relativa de um valor dentro do conjunto de dados. Cada percentil representa a porcentagem de dados que está abaixo desse valor, assim, o 25º percentil, conhecido como primeiro quartil, é o valor abaixo do qual 25% dos dados se encontram, a mediana é definida também como o 50° percentil. Os percentis são amplamente utilizados para entender a distribuição dos dados e identificar valores extremos ou *outliers*.

Exemplo de quartis (Qn)



Covariância e correlação

Covariância e correlação descrevem a relação entre duas variáveis. A covariância indica a direção da relação linear entre variáveis, podendo ser positiva, quando as variáveis aumentam juntas, ou negativa, quando uma variável aumenta enquanto a outra diminui. A correlação, por sua vez, normaliza a covariância, produzindo um valor entre -1 e 1, onde -1 indica uma correlação negativa absoluta, 1 uma correlação positiva absoluta e 0 nenhuma correlação.

$$COV_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{n-1}$$

Onde:

- $\overline{Y}$= é a média aritmética da variável X
- $\overline{X}$ = é a média aritmética da variável Y
- n = número de dados observados
- X_i = dado da série X no instante i
- Y_i = dado da série Y no instante i
- $COV_{X,Y}$ = covariância entre a série X e a série Y

Obs.:
- $(X_i - \overline{X})$ = dispersão de cada um dos dados da série X em relação à média.
- $(Y_i - \overline{Y})$ = dispersão de cada um dos dados da série Y em relação à média.

https://linkconcursos.com.br/covariancia-linear-resultados-significado-e-explicaca/

Probabilidade condicional

A probabilidade condicional é a chance de um evento ocorrer, dado que outro evento já ocorreu. Pode ser calculada:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Por exemplo, sabendo que uma pessoa possui pós-graduação, a probabilidade de ser homem é de 11/20, pois é sabido que apenas 20 pessoas possuem pós-graduação.

Sexo / Escolaridade	Homens	Mulheres	Total
Ensino médio	22	17	39
Ensino superior	15	21	36
Pós-graduação	11	9	20
Total	48	47	95

Teorema de Bayes

O Teorema de Bayes é uma forma de atualizar as probabilidades de hipóteses com base em novas evidências.

$$P(A/B) = \frac{P(B/A)xP(A)}{P(B)}$$

Por exemplo,

```
P_A = 0.01    # Probabilidade de ter a doença
P_B_dado_A = 0.9    # Probabilidade de testar positivo se tiver a doença
P_B_dado_nao_A = 0.05   # Probabilidade de testar positivo se não tiver a doença
P_nao_A = 1 - P_A    # Probabilidade de não ter a doença

P_B = P_B_dado_A * P_A + P_B_dado_nao_A * P_nao_A

P_A_dado_B = (P_B_dado_A * P_A) / P_B

print(f"Probabilidade de ter a doença dado um teste positivo: {P_A_dado_B:.2%}")
```

✓ 0.0s

```
Probabilidade de ter a doença dado um teste positivo: 15.38%
```

- P_A: Probabilidade de uma pessoa ter a doença é 1%.
- P_B_dado_A: Probabilidade de um teste positivo dado que a pessoa tem a doença é 90%.
- P_B_dado_nao_A: Probabilidade de um teste positivo dado que a pessoa não tem a doença é 5%.
- P_nao_A: Probabilidade de uma pessoa não ter a doença é 99%.

Para calcular a probabilidade de um teste Positivo primeiro calcula-se P_B, considerando ambas as possibilidades, ter a doença e não ter a doença, depois usa-se o Teorema de Bayes para calcular P_A_dado_B, que é a probabilidade de uma pessoa ter a doença dado que o teste é positivo.