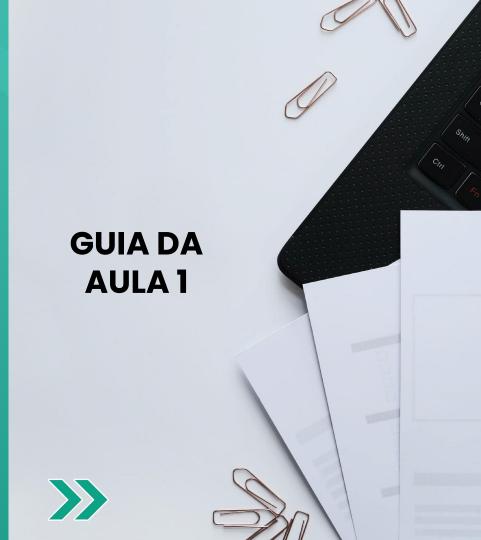


Profissão: Analista de dados





FUNDAMENTOS DE ESTATÍSTICA







Aprenda média e variância



Acompanhe aqui os temas que serão tratados na videoaula

- Introdução
- População e amostra
- Média
- Variância
- Desvio padrão
- Referências







Introdução

Neste módulo, vamos utilizar dados sobre o salário mensal em dólares americanos de jogadores da NBA em 2020. O conjunto de dados está no *link* https://github.com/andre-marcos-perez/ebac-course-utils/blob/main/dataset/wage.csv

que foi inspirado num conjunto de dados do Kaggle, disponível no *link* https://www.kaggle.com/datasets/isaienkov/nba2k20-player-dataset

Download:

```
!wget -q
"https://raw.githubusercontent.com/andre-marcos-perez/ebac-course
```





 Manipulação: vamos ler o arquivo wage.csv com ajuda do Pandas e converter a coluna de interesse do dataframe para um array NumPy.

```
In []:
    import numpy as np
    import pandas as pd
    import seaborn as
    sns

In []:    wage_df =
    pd.read_csv('wage.csv')
    wage_df.head()

In []:    wage_array =
    np.array(wage_df['wage'].astype('int').to_list())
    print(wage_array[0:5])
```









População e amostra

População é um subconjunto composto por todos os elementos de um conjunto. Já a amostra é um subconjunto composto por uma fração dos elementos de um conjunto. O processo de extrair uma amostra de uma população é chamado de **amostragem**. A amostragem é um processo muitas vezes necessário devido a impraticidade do acesso a toda a população (tempo, recursos etc.). Vamos usar como exemplo os dados da cidade de São Paulo que possuía 8.986.687 de eleitores aptos a votar nas eleições municipais de 2020, segundo o tribunal regional eleitoral do estado. Contudo, o Datafolha fez uma pesquisa de intenção de voto com apenas 1.512 (0.017% do total) na cidade de São Paulo com 95% de nível de confiança.

Exemplo:

```
In [ ]: len(wage_array)
```





Média

A média (\$\textbf{x}_m\$) é o valor médio ou média aritmética dos elementos (\$x_i\$) de um conjunto (\$\textbf{x}\$). É definido como a soma dos valores dos elementos dividido pela quantidade dos elementos do conjunto (\$n\$). Quanto maior for o número de elementos de uma amostra, mais próxima a média amostral será da média populacional.

$$\star \int x_i^{n} x_i^{n}$$

Exemple:

```
np.mean(wage_array) # em USD
```





Variância

A variância (\$\sigma^{2}\$) é uma métrica de dispersão representada pelo quadrado do desvio médio dos elementos (\$x_i\$) de um conjunto da sua média (\$\textbf{x}_m\$). É definida como a média da soma dos quadrados da diferença dos valores dos elementos de um conjunto da sua média, corrigidos por um fator amostral.

Nota: Elevar ao quadrado as diferenças evitam que valores negativos impactem a soma.

$$\sigma^{2} = \frac{1}^{n} (x_i-x_m)^{2} {n-1}$$

• Exemplo:





Desvio padrão

O desvio padrão (\$\sigma\$) é uma métrica de dispersão representada pela raiz quadrada da variância. Possui a mesma dimensão da média.

$$\sigma = \sqrt{\sum_{s=0}^{2}}$$

• Exemplo:





Referências

https://agenciabrasil.ebc.com.br/eleicoes-2020/noticia/2020-11/com-33-milhoes-de-eleitores-sp-e-maior-colegio-eleitoral-do-brasil

https://gl.globo.com/sp/sao-paulo/eleicoes/2020/noticia/2020/11/11/pesquisa-datafolha -em-sao-paulo-covas-32percent-boulos-16percent-russomanno-14percent-franca-12p ercent.ghtml

