

AI & CHATBOT

Introdução à Estatística com
Python

Prof. André Tritiack

FIAP
GRADUAÇÃO

Motivação

- Quando trabalhamos com inteligência artificial, principalmente algoritmos de Aprendizado de Máquina, precisamos trabalhar com grande volume de dados;
- Uma ferramenta básica para se trabalhar com muitos dados é a **estatística**;
- A estatística é um ramo da matemática que usa probabilidade para **modelar** (criar fórmulas matemáticas para descrever) eventos e observações;
- Na IA existem três aplicações principais da estatística:
 - Como pré-processamento dos dados;
 - Como métrica de desempenho;
 - Interno as técnicas/algoritmos;

Estatística I

Medidas de tendência central

Mediana

Mediana é o valor que separa a exata metade dos dados quando estes estão ordenados. Por exemplo:

$$X = \{1, 3, 3, 6, 7, 8, 9\}$$

$$\text{Mediana de } X = 6$$

$$X = \{3, 2, 1, 5, 4\}$$

$$\text{Mediana de } X = 3$$

$$X = \{3, 5, 7, 9\}$$

$$\text{Mediana de } X = \frac{5+7}{2} = 6$$

$$X = \{3, 1, 5, -2, 3, 3, 1, 20, -2, -2, -2\}$$

$$\text{Mediana de } X = 1$$

Moda

Moda é valor que mais se repete em um conjunto de dados. Atenção, pode ter mais de uma moda nos dados e os dados podem não ter moda. Vejamos os exemplos:

$$X = \{1, 1, 1, 1, 1\}$$

Moda de $X = 1$

$$X = \{1, 2, 3, 4, 5\}$$

X não tem moda

$$X = \{3, 3, 5, 7, 1, 1\}$$

Moda de $X = 1$ e 3

$$X = \{3, 3, 5, -2, 3, 1, 1, 20, -2, -2, -2\}$$

Moda de $X = -2$

Média

- Média é uma **medida de tendência central**. De maneira intuitiva, a média é um valor que nos explica como vários dados observacionais de um mesmo atributo se comportam como um todo. Uma boa pergunta é: se tenho vários valores sobre o mesmo atributo, qual é o **valor esperado** (valor médio) dessa atributo?
- Existem várias formas de calcular a média, cada uma recebendo um nome específico. Temos a **média aritmética**, a **média geométrica**, a **média harmônica** e a **média ponderada**. Arquitas de Tarento já havia dado o nome para a média aritmética, geométrica e harmônica em 400 a.C. há 2423 anos!
- Para cada problema, um tipo de média é mais indicado. Em geral, a média aritmética e a média ponderada são as mais usadas!

Média Aritmética

A média aritmética é determinada pela soma das observações dividida pelo número total de observações:

Valor médio de x

Primeira observação de x

Segunda observação de x

Notação compacta Σ (letra grega que lê-se sigma) e aqui significa somatório

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i$$

Número de observações

Índice do somatório, vai da observação 1 até N

Média Geométrica

A média geométrica é determinada pelo produto das observações elevado ao inverso do número de observações:

Valor médio de x

Primeira observação de x

Segunda observação de x

Notação compacta π (letra grega que lê-se pi) e aqui significa produto

$$\bar{x} = (x_1 x_2 x_3 \dots x_N)^{\frac{1}{N}} = \left(\prod_{i=1}^N x_i \right)^{\frac{1}{N}}$$

Número de observações

Índice do produto, vai da observação 1 até N

Média Harmônica

A média harmônica é determinada pela soma dos inversos das observações multiplicada pelo número total de observações:

Valor médio de x

Número de observações

$$\bar{x} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_N}} = \frac{N}{\sum_{i=1}^N \frac{1}{x_i}}$$

Primeira observação de x

Segunda observação de x

Índice do somatório, vai da observação 1 até N

Pense na notação compacta como um loop for que está indo 0 até N-1 e onde x_i é o vetor x na posição i, $x[i]$

Média Ponderada

A média ponderada é uma média aritmética na qual as observações x_i são multiplicadas por um coeficiente (peso) w_i :

Valor médio de x

Primeira observação de x

Segunda observação de x

Peso da primeira observação

Peso da segunda observação

Peso da n -ésima observação

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_N x_N}{w_1 + w_2 + w_3 + \cdots + w_N} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

Exercícios

Não usar pacotes prontos!



1) Faça funções em Python em células do Jupyter Notebook para calcular a média:

- a. Aritmética
- b. Geométrica
- c. Harmônica
- d. Verifique que: $\bar{x}_{aritmética} > \bar{x}_{geométrica} > \bar{x}_{harmônica}$

Use como entrada $x = [39, 38, 27, 22, 20, 17, 10, 10, 10, 10, 7, 7, 7, 7, 6]$

Dica:

Para fazer
a operação
de potência

```
1 a = 5
2 b = 2
3 x = a**b
4 print(x)
25
```

2) Faça uma função em Python em uma célula do Jupyter Notebook para calcular a média ponderada. Ela deve receber dois vetores (listas) como entrada. Use o mesmo x do exercício anterior com os pesos $w = [113, 88, 58, 65, 71, 46, 36, 33, 37, 40, 24, 21, 20, 15, 20]$

Exercícios



3) Faça funções em Python em células do Jupyter Notebook para calcular a moda e a mediana:

Para testar, use como entrada $x = [39, 38, 27, 22, 20, 17, 10, 10, 10, 10, 7, 7, 7, 7, 6]$

Dicas:

para ordenar um vetor em python podemos usar o método `.sort()`

```
1 y = [3, 3, 5, -2, 3, 1, 1, 20, -2, -2, -2]
2 y.sort()
3 print(y)
```

`[-2, -2, -2, -2, 1, 1, 3, 3, 3, 5, 20]`

para contar o número de ocorrências de um valor em um vetor em python podemos usar o método `.count(valor)`

```
[-2, -2, -2, -2, 1, 1, 3, 3, 3, 5, 20]
```

```
1 y.count(3)
```

`3`

Estatística II

Medidas de dispersão

Variância da População

A variância é uma medida de quanto os dados estão distribuídos em torno do valor esperado (média). Existem duas formas de calcular a variância: em **relação a população** de dados e em **relação a uma amostra** estatística dos dados (subconjunto).

Notação compacta
 Σ (letra grega que lê-se sigma) e aqui significa somatório

Variância da população da variável x

Primeira observação de x

Média de x denotada pela letra grega μ (mi)

Segunda observação de x

Número total de elementos da população

Índice do somatório, vai do primeiro elemento até o último elemento N da população

$$\sigma_x^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \dots + (x_N - \mu)^2}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Variância da Amostra

A variância da amostra é bem semelhante a variância da população, diferindo apenas no fato de divisão (variância não viciada):

The diagram illustrates the formula for sample variance, s_x^2 , with several annotations pointing to its components:

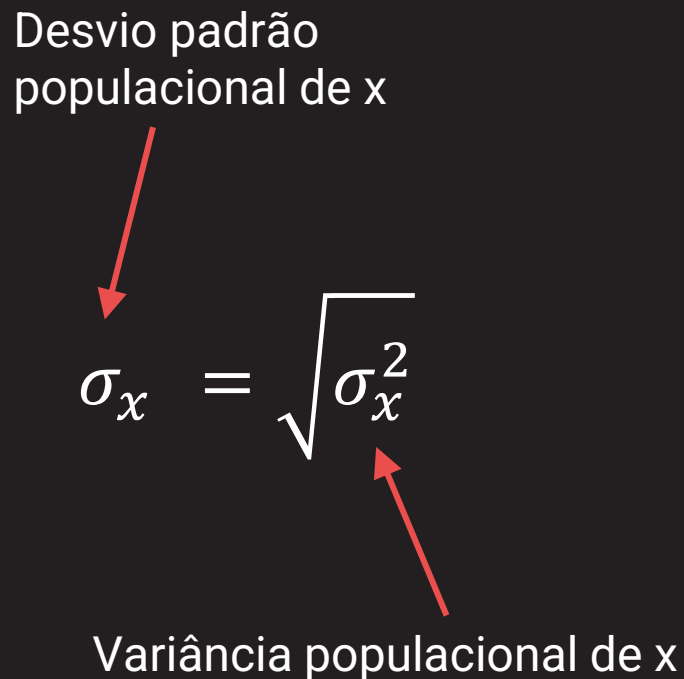
- Variância amostral da variável x**: Points to the symbol s_x^2 .
- Primeira observação de x**: Points to x_1 in the first term of the sum.
- Média de x**: Points to \bar{x} in the first term of the sum.
- Segunda observação de x**: Points to x_2 in the second term of the sum.
- Número de amostras**: Points to $n - 1$ in the denominator.
- Notação compacta**: Points to the summation symbol \sum in the compact form.
- Índice do somatório, vai da observação 1 até n**: Points to the index i in the compact form.

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Desvio Padrão Populacional e Amostral

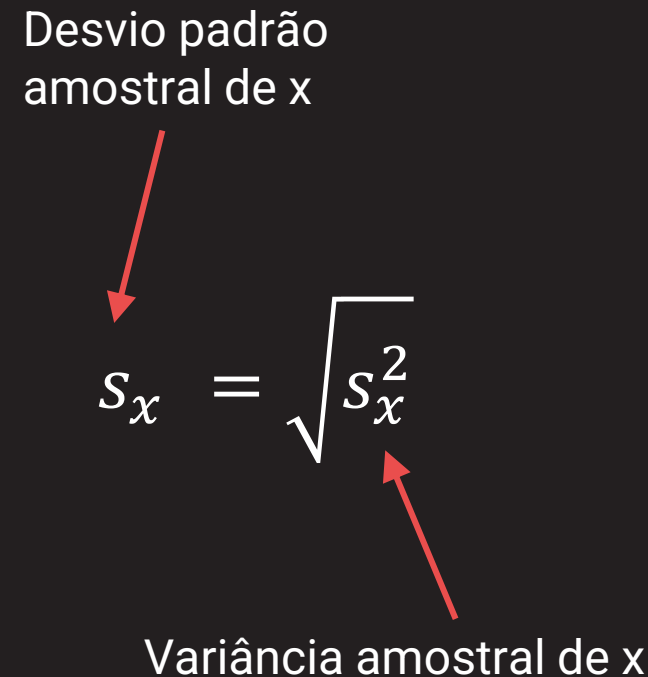
O desvio padrão é uma medida de dispersão cuja unidade é igual a unidade da variável aleatória medida (i.e. de x). Ele é calculado pela raiz quadrada da variância e é denotado pela letra grega σ ou pela letra latina s :

Desvio padrão
populacional de x


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

Variância populacional de x

Desvio padrão
amostral de x


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

Variância amostral de x

Populacional vs Amostral

Variance in Python

```
import numpy as np
vec = [1, 2, 3, 4, 5, 6, 7]
np.var(vec)
```

4.0

Variance in R

```
library(stats)
vec <- c(1, 2, 3, 4, 5, 6, 7)
stats::var(vec)
```

[1] 4.666667

R utilizes Bessel's correction when calculating variance, which changes the formula from returning **population variance** to **sample variance**

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

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N - 1}$$

✅ Using Bessel's correction gives an unbiased estimator as demonstrated in this example (*sd of 2 implies a variance of 4*)

```
map_dbl(1:100000, ~ {
  x <- rnorm(n = 5, mean = 0, sd = 2)
  sum((x - mean(x))^2) / length(x)
}) |> mean()
```

[1] 3.190721

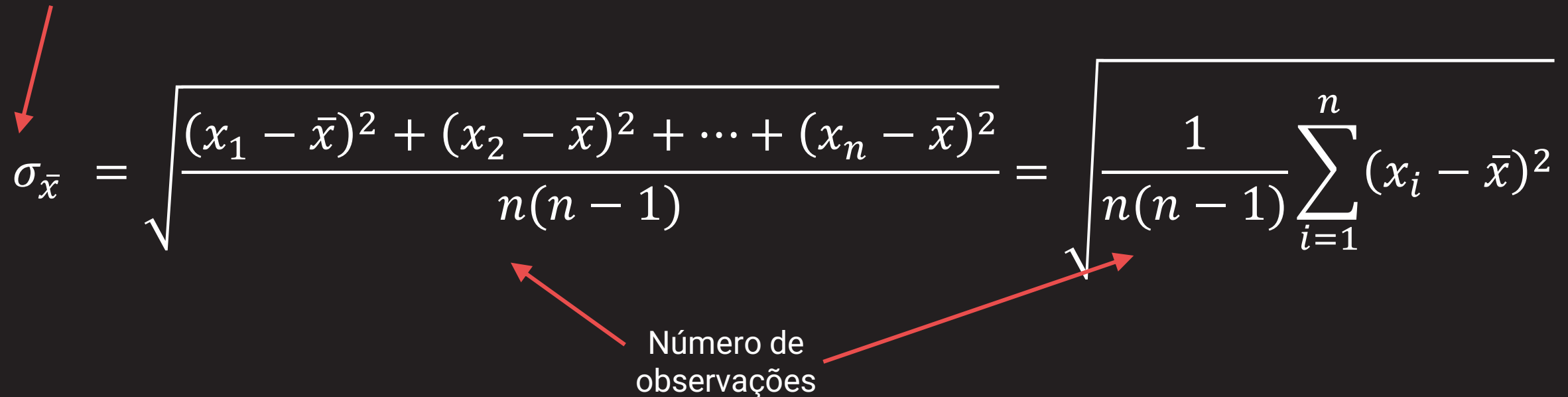
```
map_dbl(1:100000, ~ {
  x <- rnorm(n = 5, mean = 0, sd = 2)
  sum((x - mean(x))^2) / (length(x) - 1)
}) |> mean()
```

[1] 3.993807

Desvio Padrão (amostral) da Média

O desvio padrão da média determina qual é a incerteza (erro) associada ao valor da média. No inglês é chamada de **standard error of the mean** (sem). Matematicamente, podemos calcular o desvio padrão da média como:

Desvio padrão da média de x



The diagram illustrates the formula for the standard error of the mean ($\sigma_{\bar{x}}$). A red arrow points from the text 'Desvio padrão da média de x' to the symbol $\sigma_{\bar{x}}$. Another red arrow points from the text 'Número de observações' to the term $n(n-1)$ in the denominator of the first formula. A third red arrow points from the same text to the square root symbol in the second formula.

$$\sigma_{\bar{x}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n(n-1)}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Exercícios

Não usar pacotes prontos!



4) Faça funções em Python em células do Jupyter Notebook para calcular a:

- a. Variância amostral
- b. Variância populacional
- c. Desvio padrão amostral
- d. Desvio padrão populacional
- e. Incerteza da média

Cada função deve receber apenas o vetor/lista de dados numéricos.

Use como entrada $x = [39, 38, 27, 22, 20, 17, 10, 10, 10, 10, 7, 7, 7, 7, 6]$

Dica: Você pode usar as funções dos exercícios anteriores

Bibliotecas

Usando funções prontas

Bibliotecas

- No Python, assim como em outras linguagem de programação, há diversas bibliotecas (pacotes) que podem ser adicionados ao seu código e que já implementam funções prontas;
- Bibliotecas são nada mais do que programas (scripts) com código para rodar uma ou mais funções. Normalmente elas são feitas no paradigma de programação orientada a objetos (POO);
- Além de facilitarem o fato de não termos que implementar a função que queremos, as bibliotecas **podem** ter vantagens adicionais de **suporte constante** (por membros da comunidade) e **melhor performance**.
- Cuidado, sempre escolha bibliotecas com comunidades ativas, com boa documentação e que se preocupem com a performance (caso sua aplicação precise).

Numpy

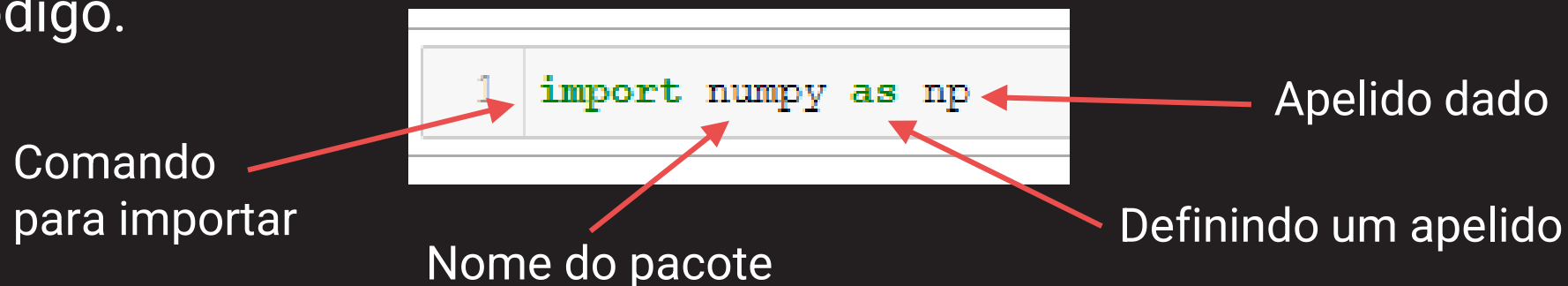


O Numpy é uma biblioteca para trabalhar com vetores e matrizes em Python, implementando uma série de funções matemáticas prontas para Geometria Analítica, Álgebra Linear, Cálculo Numérico e Estatística.

Vejamos a documentação:

<https://numpy.org/doc/stable/reference/index.html>

Uma vez instalada, para usar, devemos primeiro importar a biblioteca no nosso código.



Exercícios

Usar pacotes prontos!



- 5) Refaça os exercícios anteriores usando as funções prontas de estatística do Numpy. Você pode ver como usá-las aqui:

<https://numpy.org/doc/stable/reference/routines.statistics.html>

- 6) Alguma função você não encontrou pronta no Numpy? Procure na internet outra biblioteca que já tenha a função implementada. Instale (se for necessário) e teste para as mesmas entradas.
- 7) Faça uma comparação de desempenho das funções testadas usando um vetor de 100 mil valores aleatórios entre 0 e 99. Use o timeit:

```
1 %timeit -n 10 np.mean(y)

13.6 ms ± 1.25 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

Próximos Passos

O que veremos na próxima aula

Nas próxima aulas...

- Visualização de dados;
- Introdução à Ciência de dados;



Copyright © 2023

Slides do Prof. Henrique Ferreira - FIAP

Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem o consentimento formal, por escrito, do Professor (autor).