AI & CHATBOT

Aula 14 – Introdução à Estatística com Python

Prof. Henrique Ferreira

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}
Mediana de X = 6

X = {3, 5, 7, 9}
Mediana de X = $\frac{5+7}{2} = 6$

X = {3, 2, 1, 5, 4}
Mediana de X=3

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

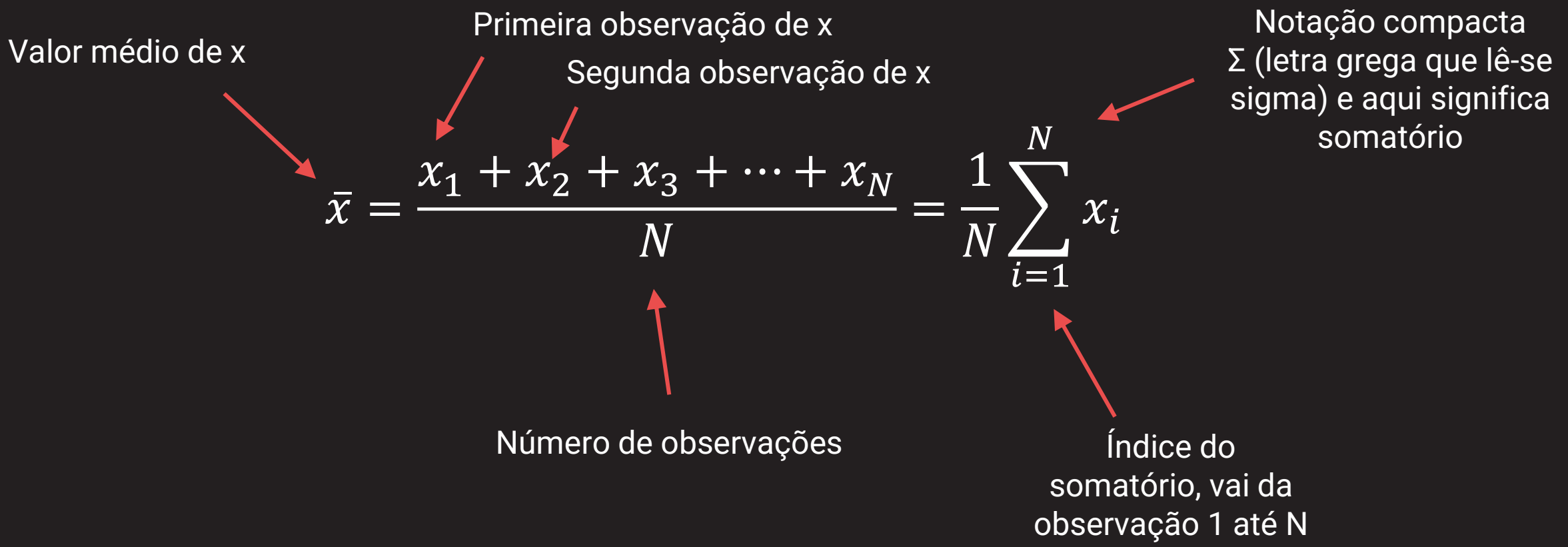
- Média é uma **medida de tendência central**. De maneira intuitiva, a média é um valor que nos explica como vários dados observacionais de um mesmo atributo se comportam como um todo. Uma boa pergunta é: se tenho vários valores sobre o mesmo atributo, qual é o **valor esperado** (valor médio) dessa 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 2421 anos!

- Para cada problema, um tipo de média é mais indicado. Em geral, a média aritmética e a média ponderada são as mais usadas!

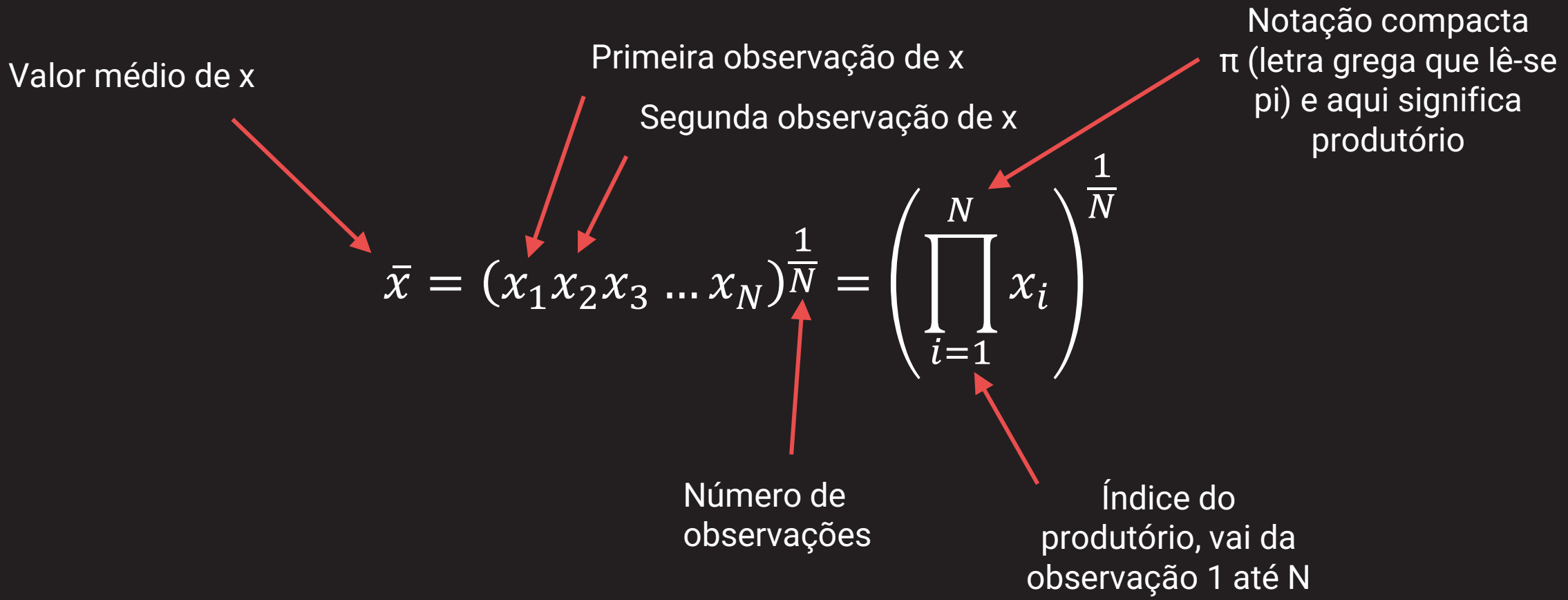
Média Aritmética

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



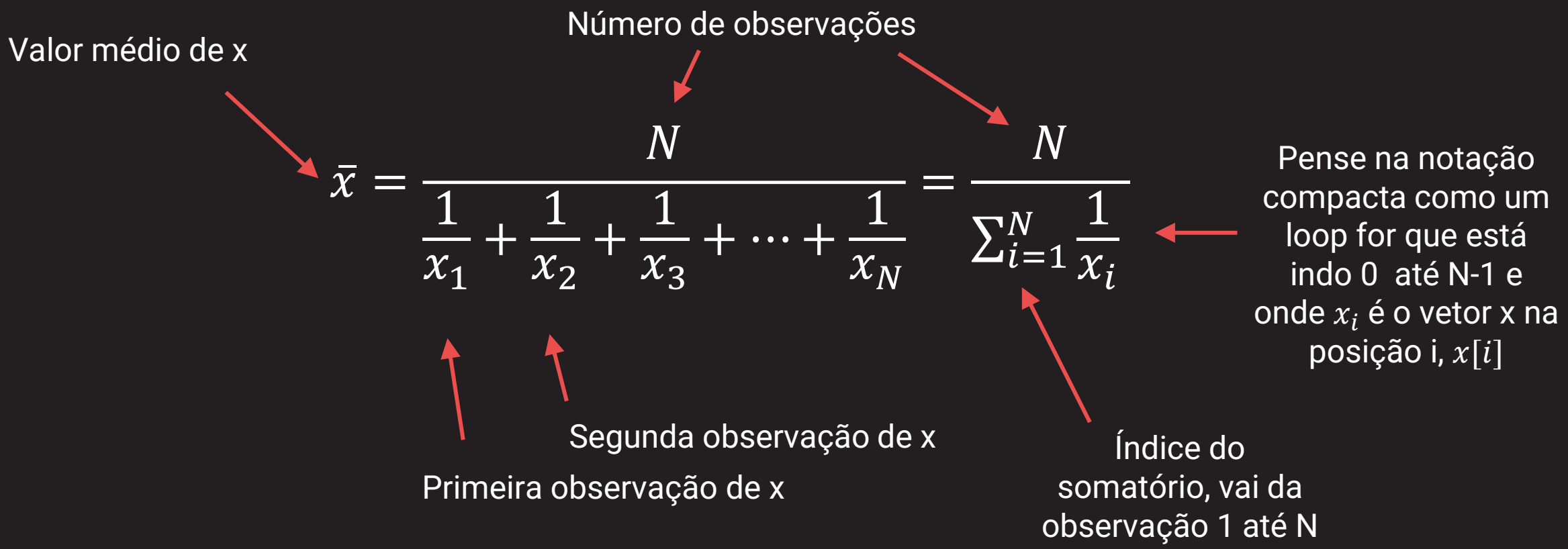
Média Geométrica

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



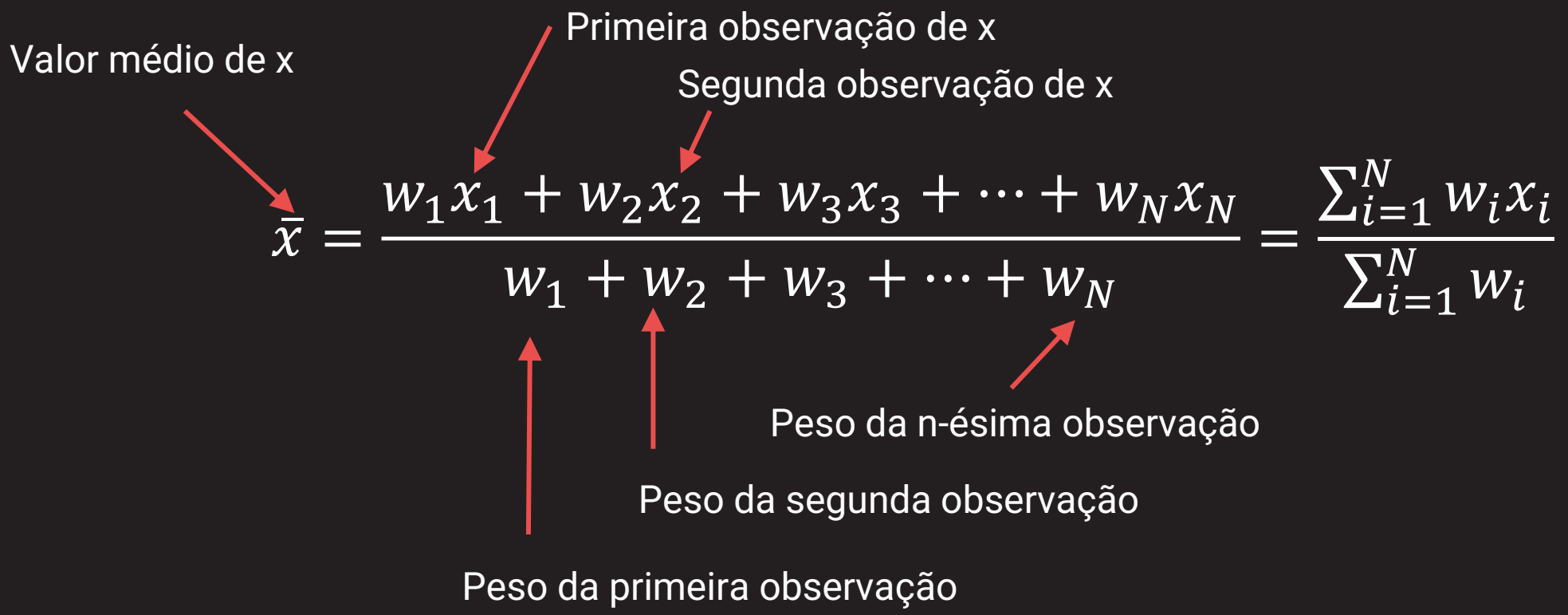
Média Harmônica

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



Média Ponderada

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

Variância da população da variável x

Primeira observação de x

Média de x denotada pela letra grega μ (mi)

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

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

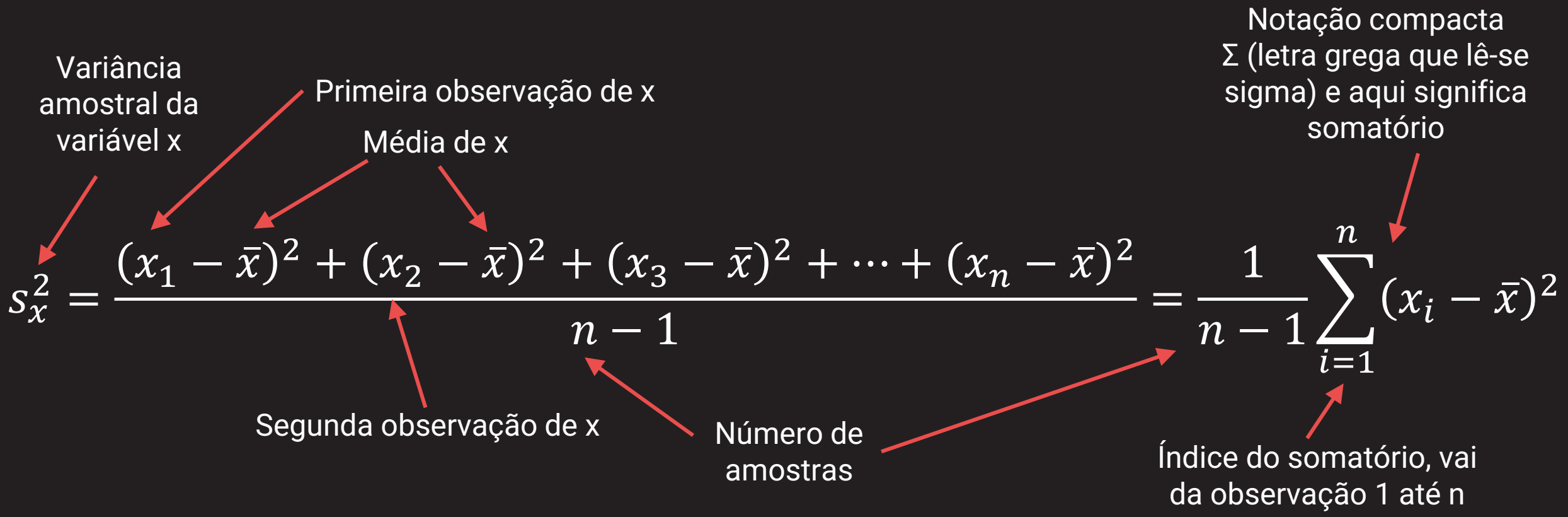
Segunda observação de x

Número total de elementos da população

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

Variância da Amostra

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



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**

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

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

✅ 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

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

Número de observações

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

Estatística III

Medidas de relação

Covariância

A covariância mede a variabilidade conjunta de duas variáveis aleatórias X e Y.

Covariância entre X e Y

Média da variável X

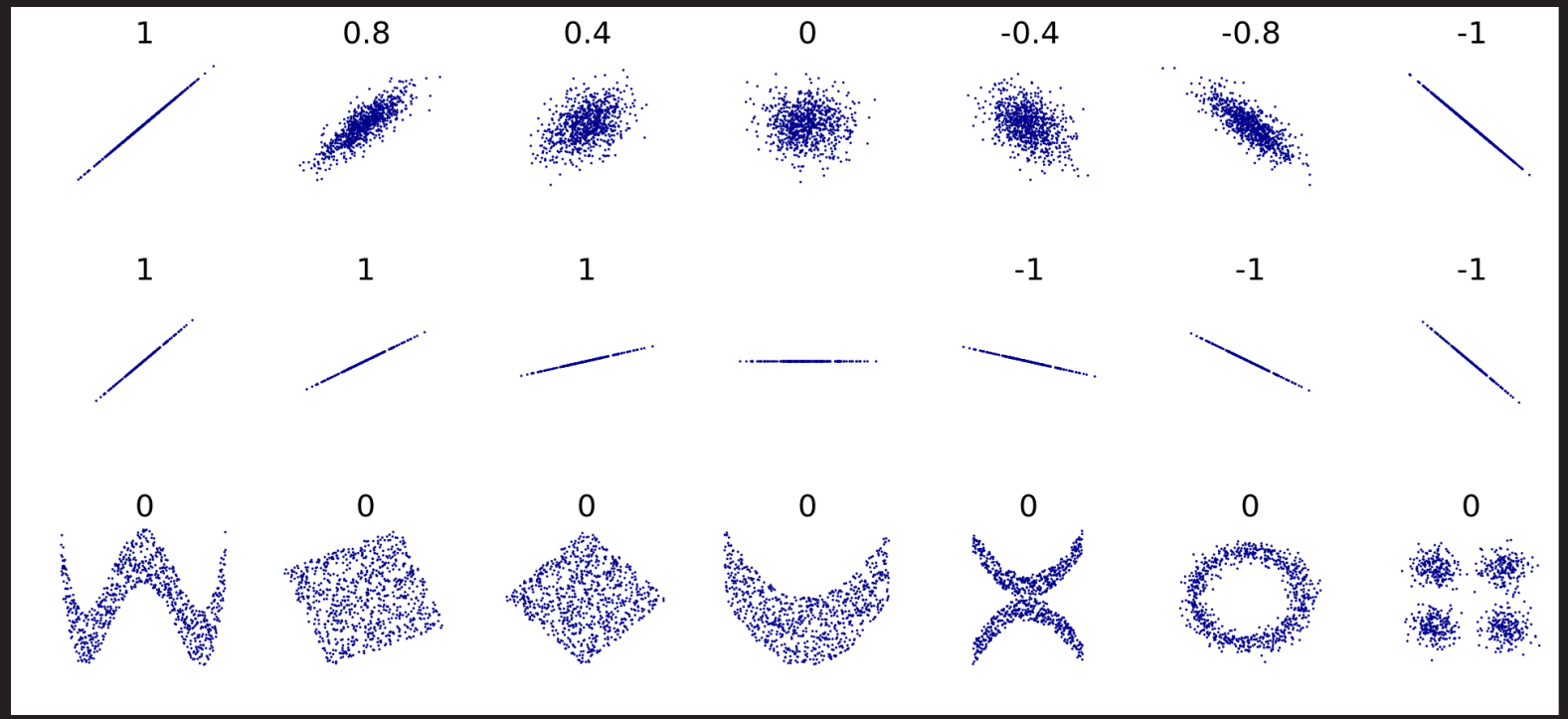
$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]$$

Número de observações

Média da variável Y

Correlação

A correlação mede a relação mútua entre duas variáveis aleatórias. Existem várias formas de calcular correlação, sendo a mais tradicional pelo Coeficiente de Correlação de Pearson, denotado por ρ:



Correlação entre X e Y

Covariância entre X e Y

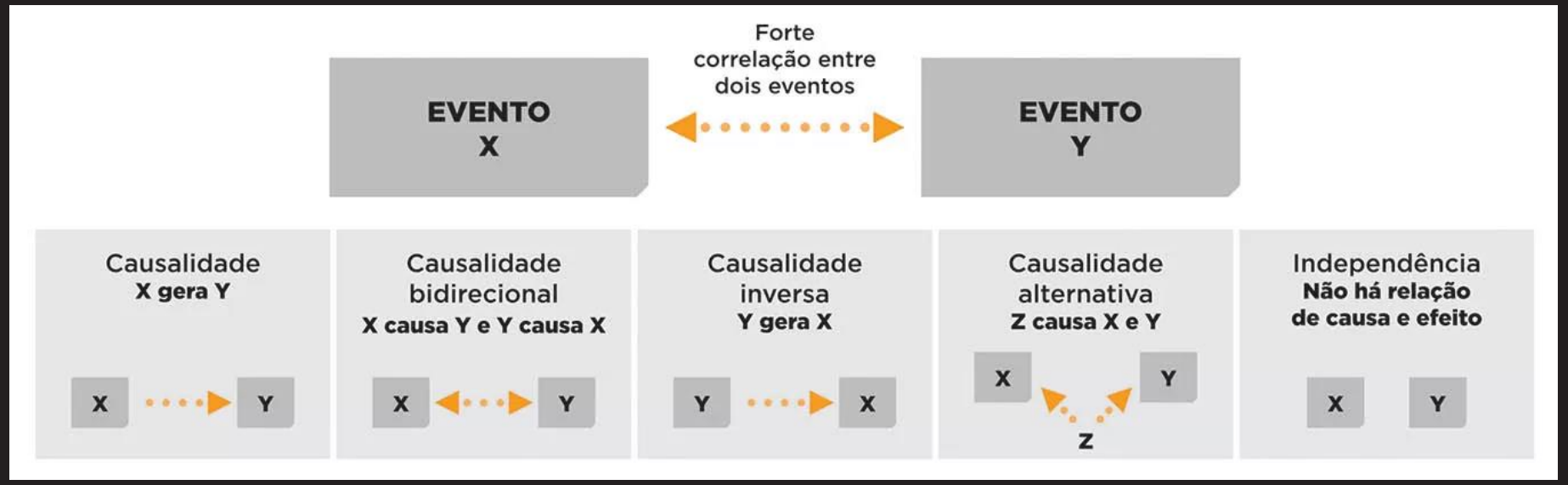
$$\rho = CORR(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

Variância de X

Variância de Y

Correlação vs Causalidade

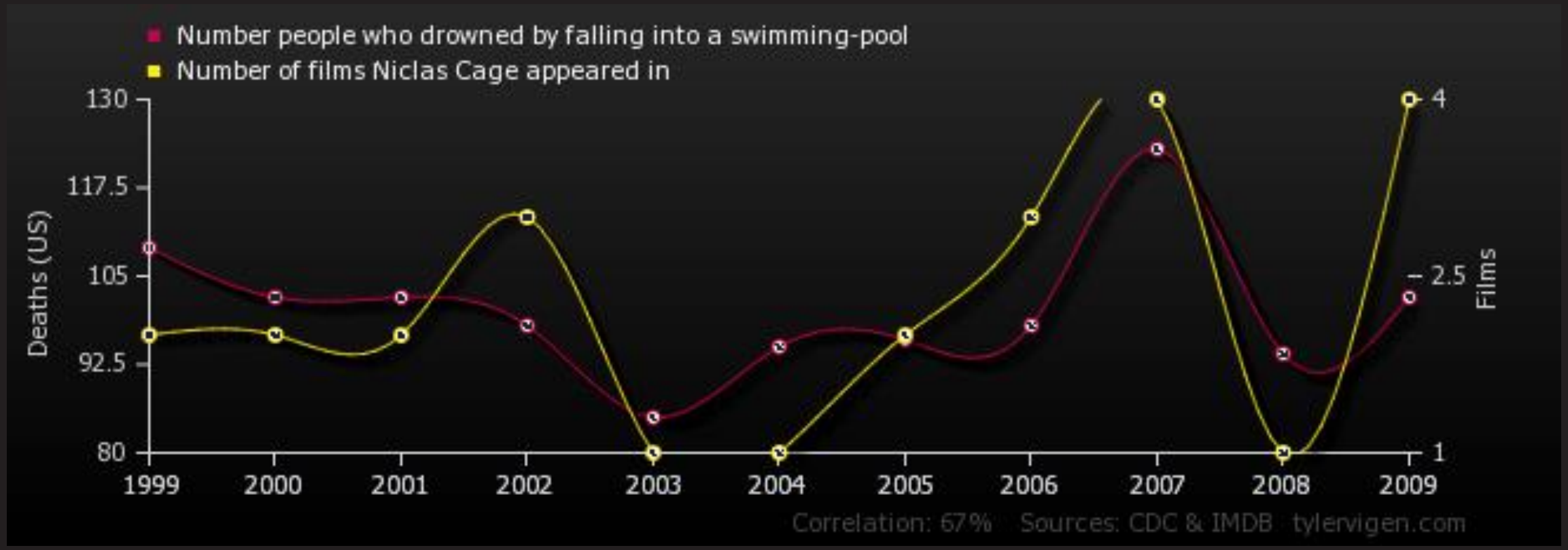
- Correlação é a relação entre duas ou mais variáveis;
- Causalidade é o conceito de precedência de um evento que causa outro;
- Duas coisas podem estar relacionadas mas não serem, necessariamente, uma a causa da outra;



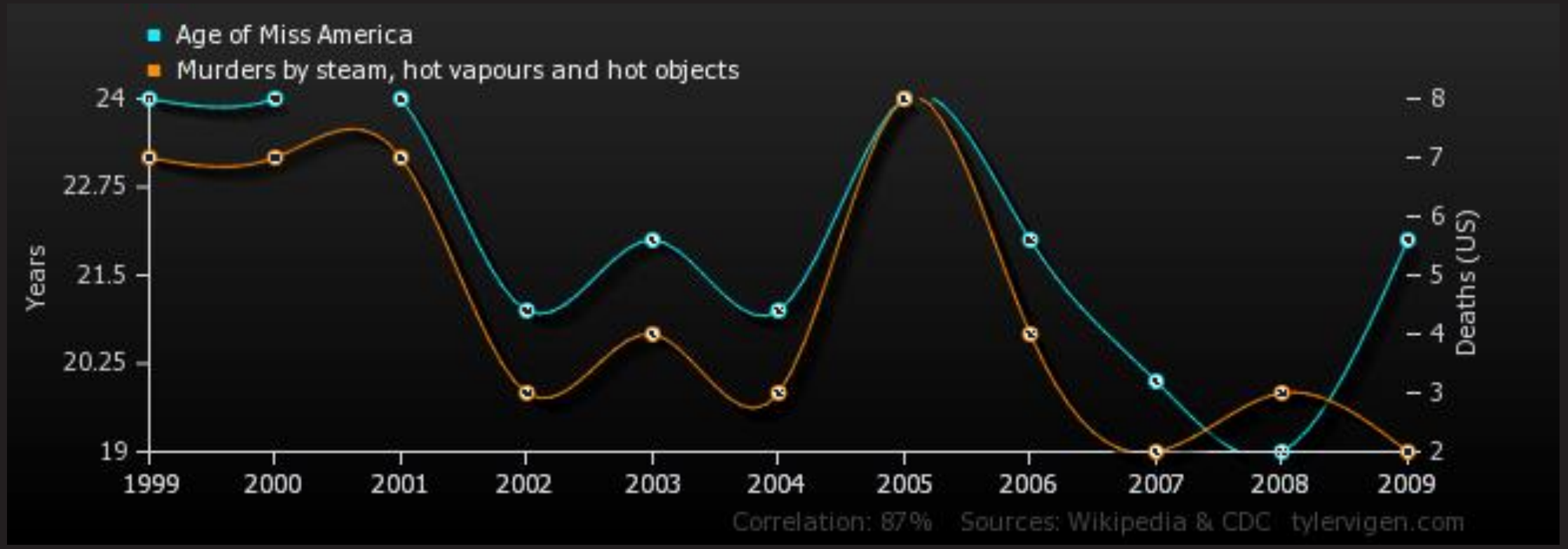
Em 1659, no antigo vilarejo de Carcassonne, todos os habitantes comeram pães. Hoje estão todos mortos. Logo, pão mata!

Correlação vs Causalidade

https://www.wnycstudios.org/podcasts/otm/articles/spurious-correlations



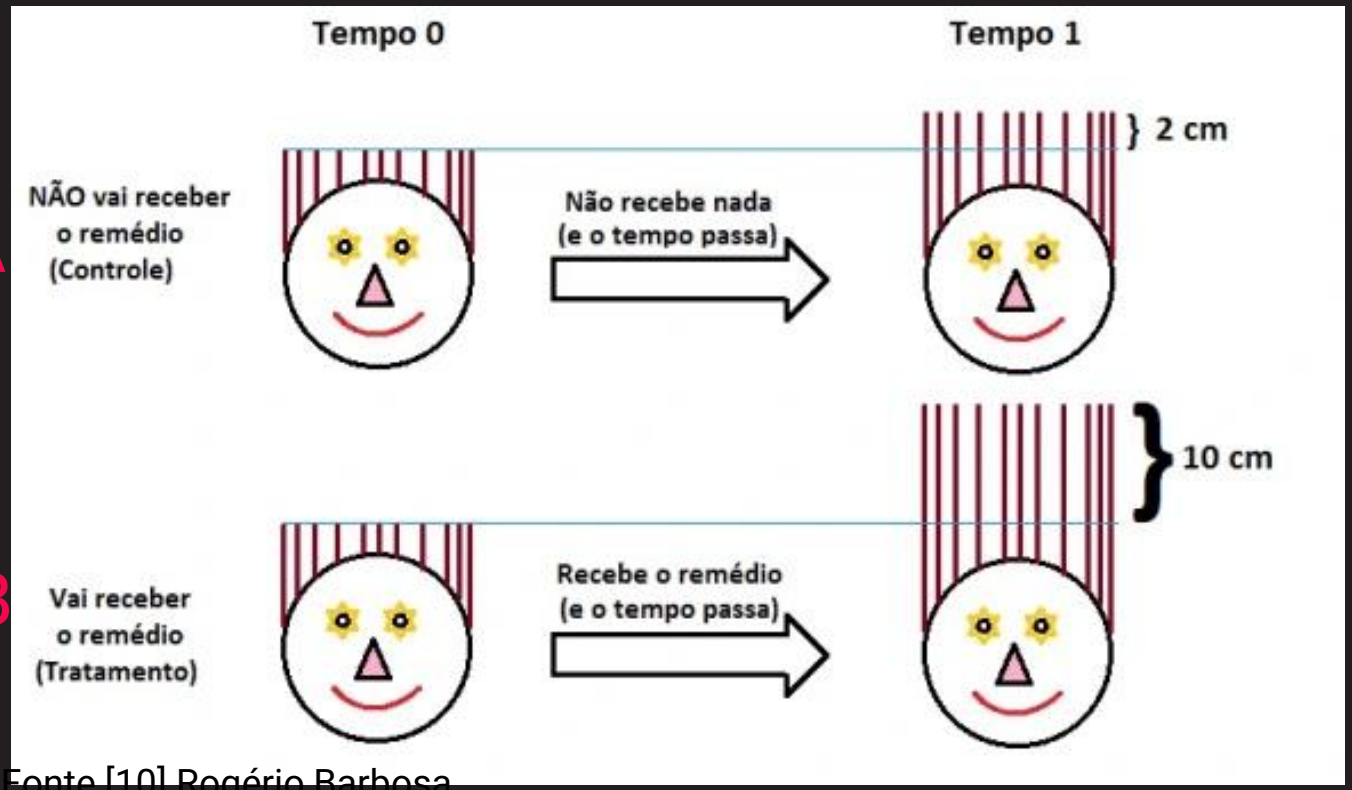
Correlação vs Causalidade



Correlação vs Causalidade

- Imaginemos um remédio para fazer crescer cabelo e um experimento para ver se o remédio realmente funciona;
- Pegamos dois indivíduos A e B e medimos o comprimento do cabelo em dois momentos: Tempo 0 antes de dar o remédio e Tempo 1 um mês depois de terem tomado o remédio;

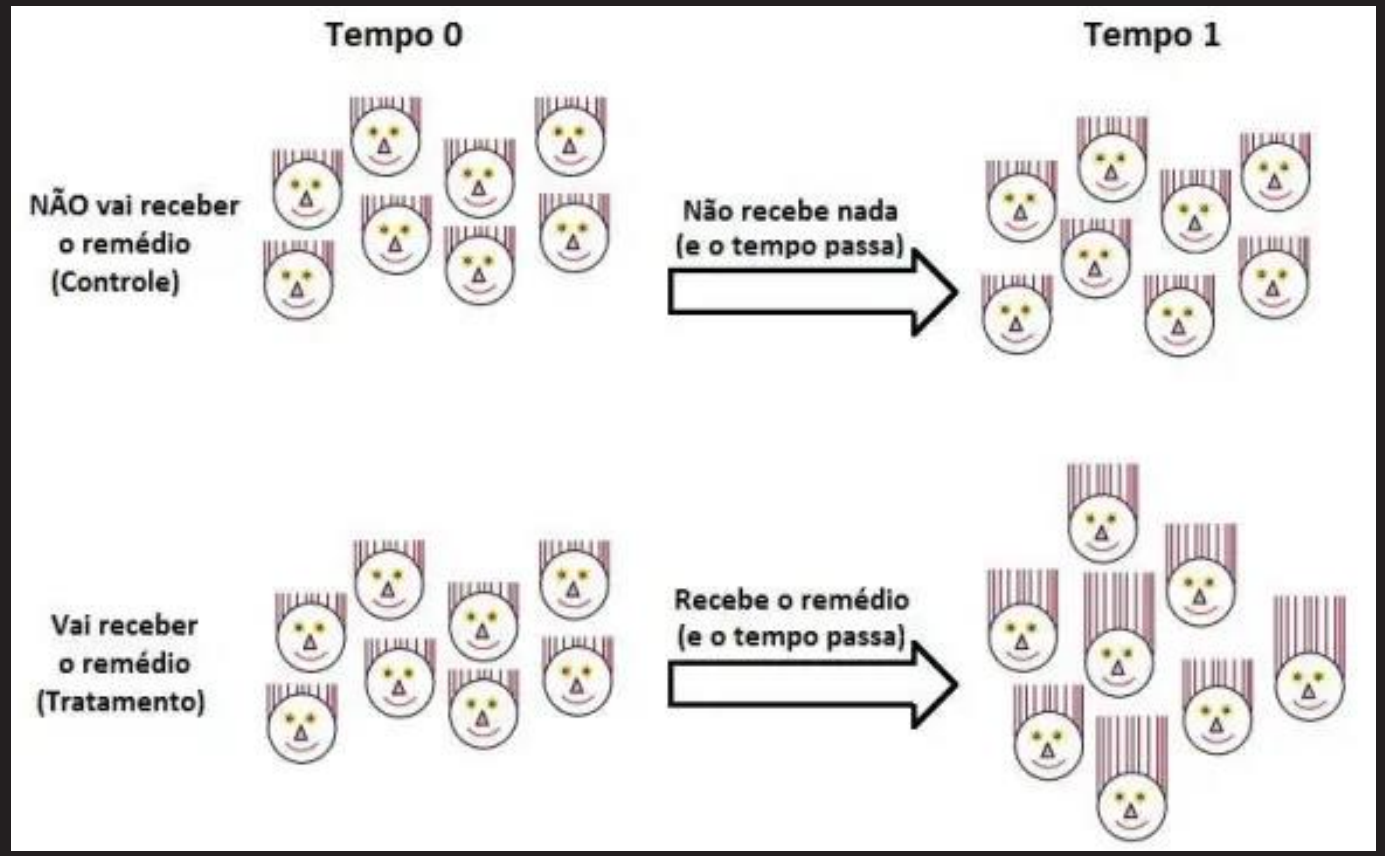
O remédio fez o cabelo crescer?



Fonte [10] Rogério Barbosa

Correlação vs Causalidade

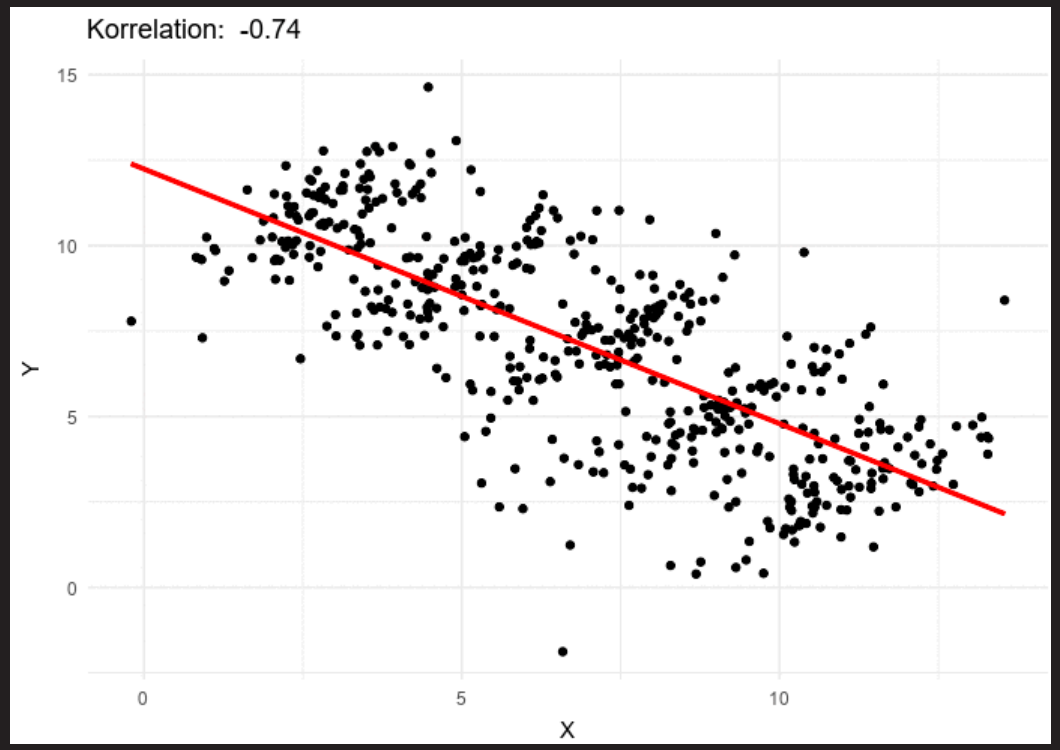
- Idealmente gostaríamos de fazer experimentos perfeitos mas as pessoas, e as coisas em geral, carregam muita variabilidade intrínseca que não permite isolar todas as variáveis;
- Por isso precisamos de estatística para ter uma ideia mais clara da causalidade (efeito causal médio).

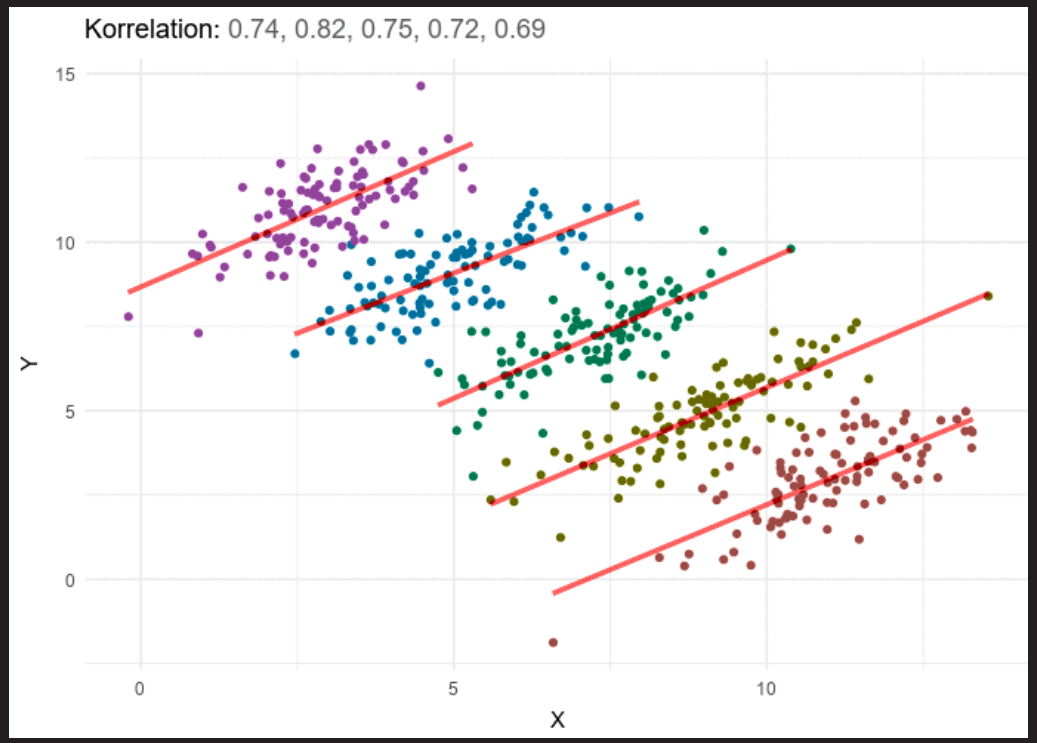


https://sociaisemetodos.wordpress.com/2013/11/01/se-correlacao-nao-e-causalidade-o-que-e-parte-2-viagem-no-tempo-contrafactuais-e-experimentos-cientificos/

Paradoxo de Yule-Simpson

O Paradoxo de Udny Yule (1903) e Edward Simpson (1951) é um fenômeno no qual uma tendência observada em um grupo de dados desaparece ou se reverte quando observada em grupos combinados:





Paradoxo de Yule-Simpson

Exemplo dos Hospitais (Ronaldo Pratti, MD 2021):

Absoluto

Hospital		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

Relativo

Hospital		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**3%**	97%	100%
	Hospital B	**2%**	98%	100%

Pergunta: em qual hospital se tratar?

Paradoxo de Yule-Simpson

- Aparentemente o hospital A tem uma mortalidade 50% maior que o hospital B;
- Pergunta: E se o hospital A receber casos mais graves que o hospital B?
- Nesse caso, devemos levar em consideração a **variável oculta** severidade da doença;

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	63	2037	2100
	Hospital B	16	784	800
	Total	79	2821	2900

Accounting for the lurking variable: **"severity of illness"**

Patients severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	57	1443	1500
	Hospital B	8	192	200
	Total	65	1635	1700

Patients *not* severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	6	594	600
	Hospital B	8	592	600
	Total	14	1186	1200

Paradoxo de Yule-Simpson

Patients severly ill

		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**3.8%**	96.2%	100%
	Hospital B	**4.0%**	96.0%	100%

Patients *not* severly ill

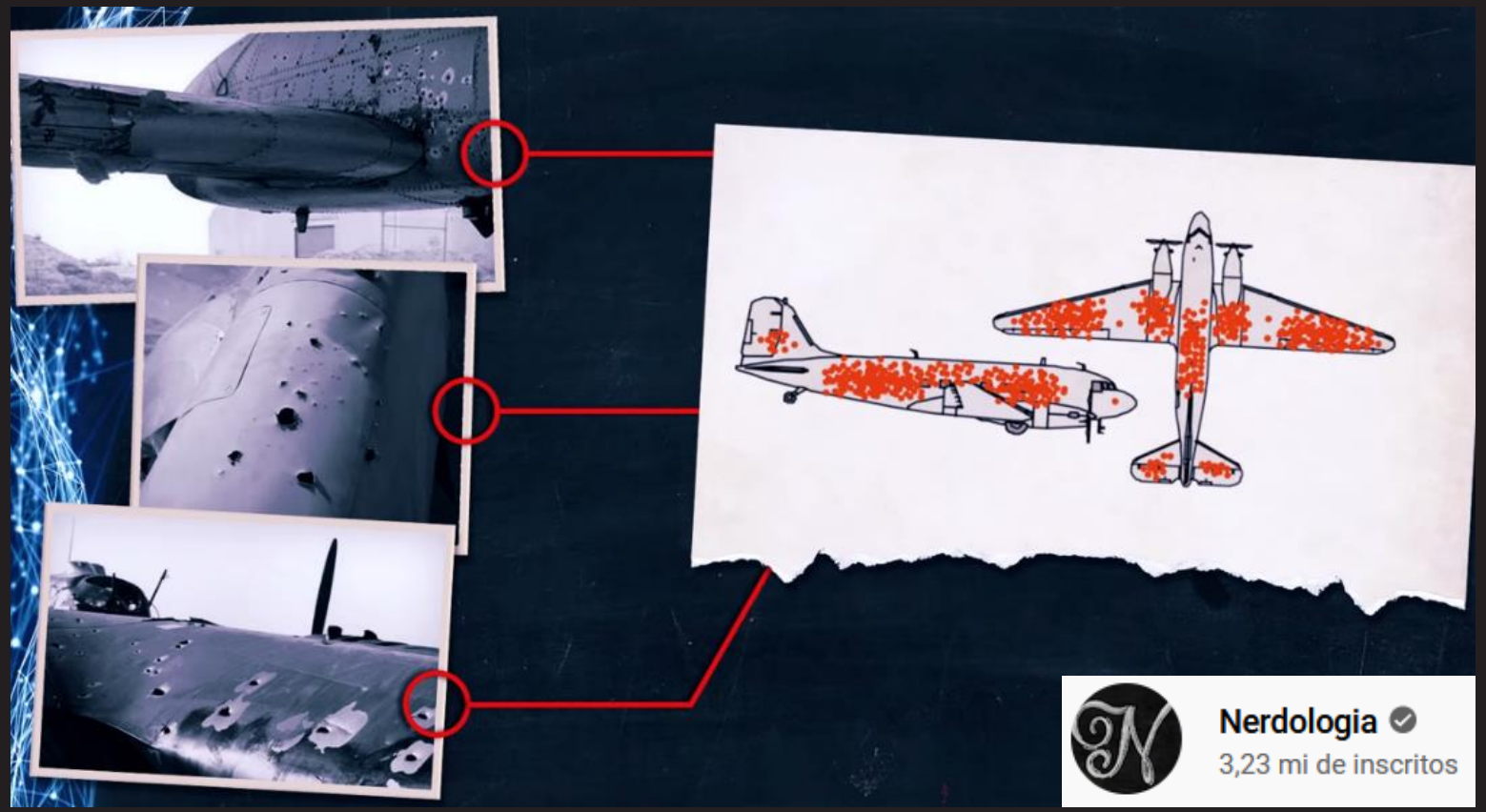
		Patient's Status		
		Died	Survived	**Total**
Hospital	Hospital A	**1.0%**	99.0%	100%
	Hospital B	**1.3%**	98.7%	100%

- O uso da variável oculta pode levar a uma alteração no sentido da associação!
- Este é o Paradoxo de Yule-Simpson! Tirar duas conclusões opostas com o mesmo conjunto de dados!
- Para resolver o paradoxo de Simpson precisamos realizar experimentos controlados e levar em consideração a causalidade;

Pensando como um Cientista de Dados

Durante a Segunda Guerra Mundial, Abraham Wald (1902-1950), liderando o Statistical Research Group na Universidade de Columbia, aplicou estatística para tratar problemas de guerra.

Onde reforçar a blindagem dos aviões?



https://www.youtube.com/watch?v=ykSILAQQu6o&ab_channel=Nerdologia

Exercícios

5) Faça uma função em Python para calcular a covariância entre duas listas X e Y. Utilize para testar a função as listas X=[1, 2, 3, 4, 5] e Y=[10, 20, 30, 40, 50].

6) Faça uma função em Python para calcular a correlação de Pearson entre duas listas X e Y. Utilize o mesmo X e Y do exercício 5 para testar.

7) A correlação de Pearson mede linearidade direta. Utilizando os dados do exercício 6, calcule a correlação para Z=[10,-20,30,-40,50]. Os valores de CORR(X,Y) e CORR(X,Z) são iguais ou diferentes? O que você pode concluir disso?

Bibliotecas

Usando funções prontas

Bibliotecas

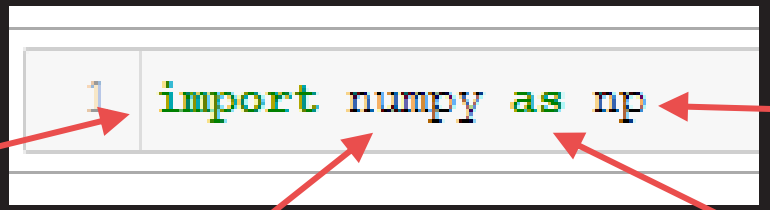
- No Python, assim como em outras linguagem de programação, há diversas bibliotecas (pacotes) que podem ser adicionados ao seu código e que já implementam funções prontas;
- Bibliotecas são nada mais que do que programas (scripts) com código para rodar uma ou mais funções. Normalmente elas são feitas 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!

8) 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

9) 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.

10) 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)
```

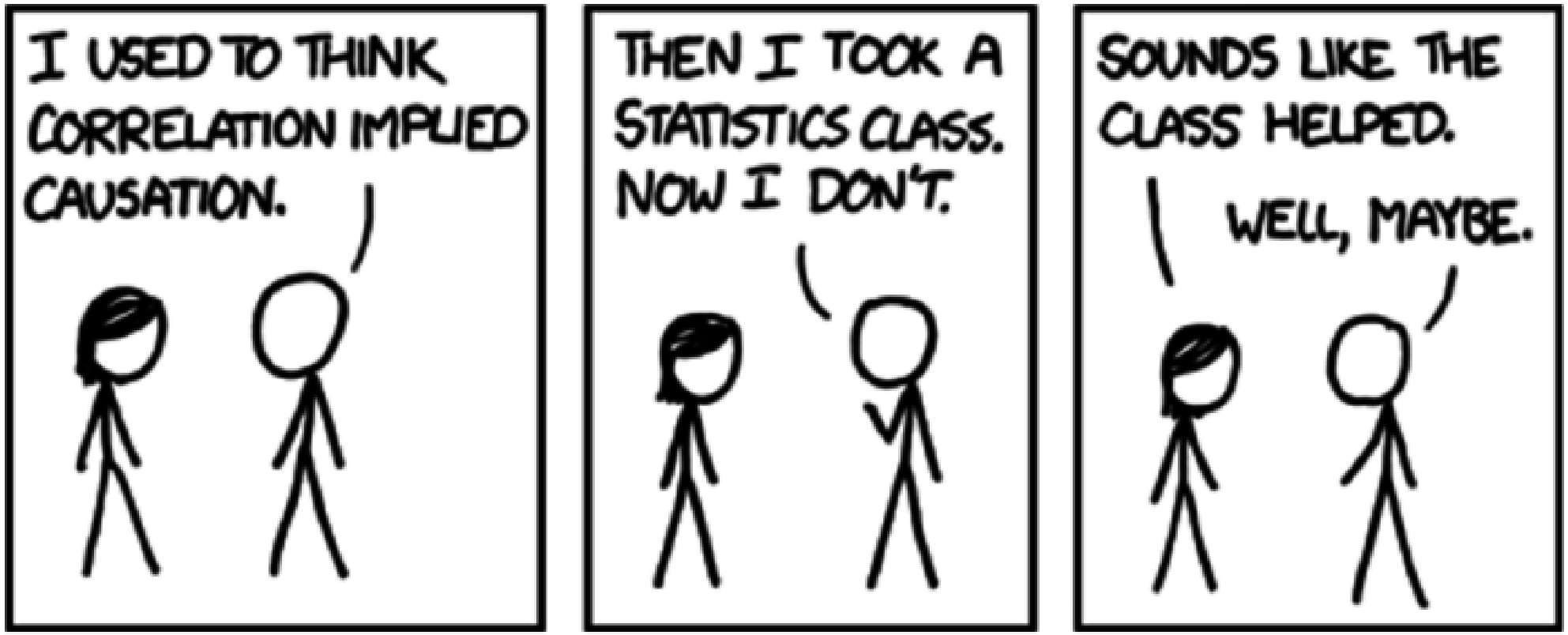
```
13.6 ms ± 1.25 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

Próximos Passos

O que veremos na próxima aula

Nas próxima aulas...

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


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.