Estatísticas Descritiva

Prof. Sandro Jerônimo de Almeida

Medidas de Espalhamento

Intervalo

- Diferença entre o máximo e o mínimo:

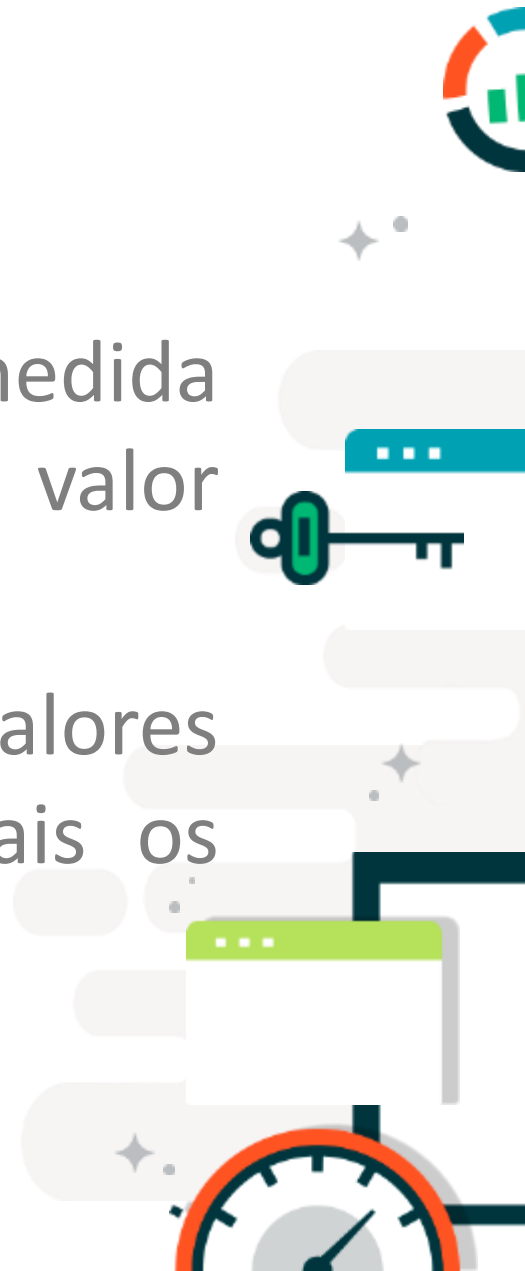
$$intervalo(x) = max(x) - min(x)$$

- Exemplo: 17, 18, 6, 29, 38, 21, 22, 40
- O intervalo do conjunto é: 40 – 6 = 34

Medidas de Espalhamento

Variância

- Dado um conjunto de dados, a variância é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está do valor central (médio).
- Quanto menor é a variância, mais próximos os valores estão da média; mas quanto maior ela é, mais os valores estão distantes da média.

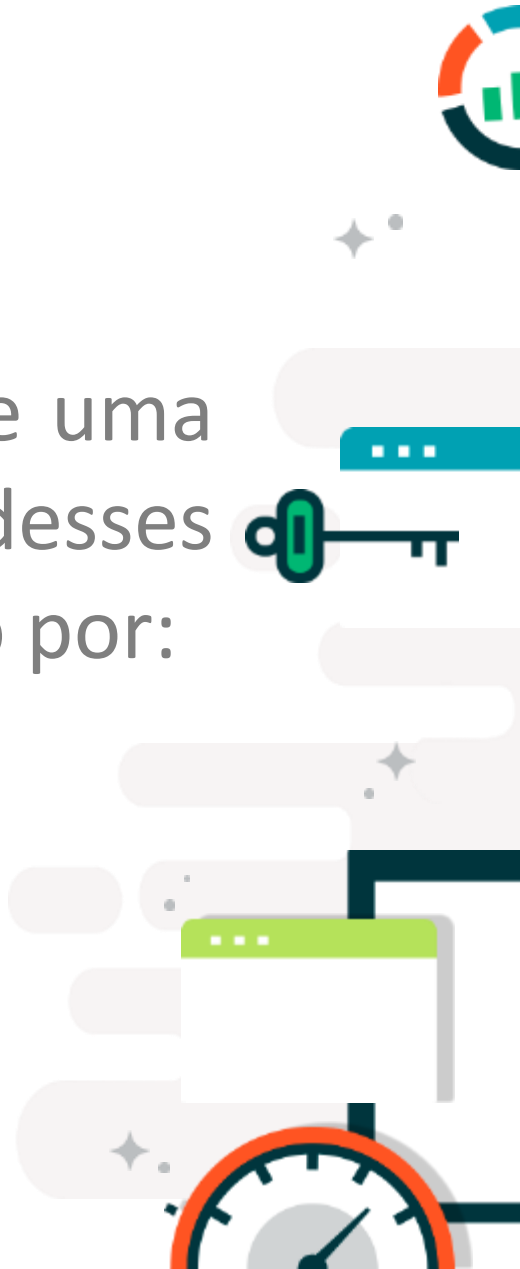


Medidas de Espalhamento

Variância

- Considere que x_1, x_2, ..., x_n são os n elementos de uma amostra e que X é a média aritmética desses elementos. O cálculo da variância amostral é dado por:

- $$Var = \frac{(x_1 - X)^2 + (x_2 - X)^2 + (x_3 - X)^2 + \ldots + (x_n - X)^2}{n - 1}$$



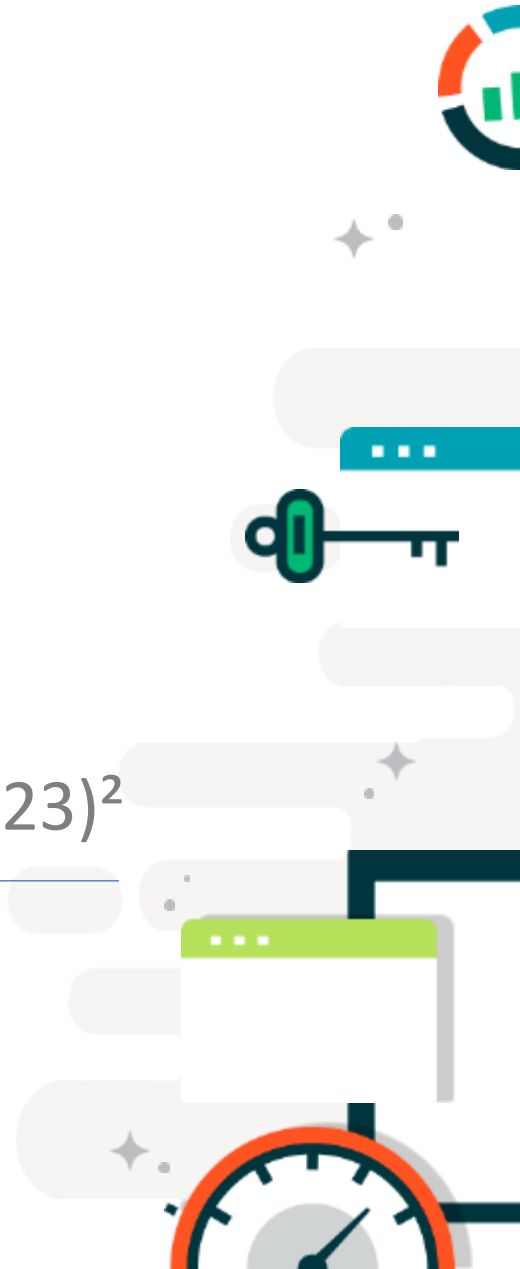
Medidas de Espalhamento

Variância

- Exemplo de idades: {25, 28, 20, 19, 23}
- Média das idades: 115 / 5 = 23 anos

- $\text{Var} = \dfrac{(25-23)^2 + (28-23)^2 + (20-23)^2 + (19-23)^2 + (23-23)^2}{(5-1)}$

$= (4 + 25 + 9 + 16 + 0) / 4 = \mathbf{13,5}$

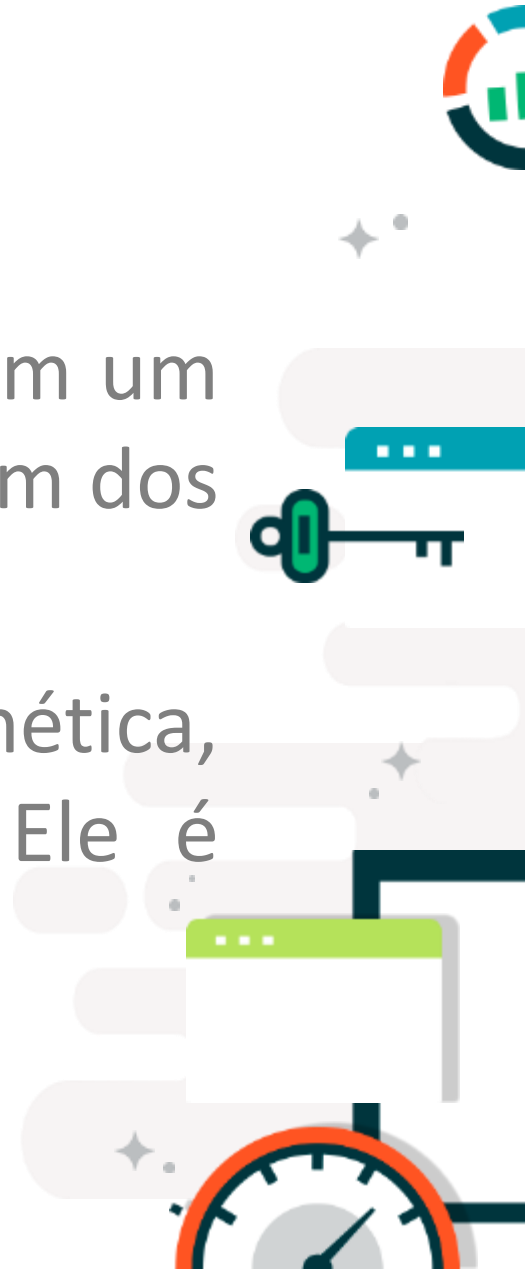


Medidas de Espalhamento

Desvio Padrão

- O desvio padrão é capaz de identificar o “erro” em um conjunto de dados, caso quiséssemos substituir um dos valores coletados pela média aritmética.
- O desvio padrão aparece junto à média aritmética, informando o quão “confiável” é esse valor. Ele é apresentado da seguinte forma:

  média aritmética (x) ± desvio padrão (dp)



Medidas de Espalhamento

Variância (σ2)

- $variância(x) = \sigma^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Desvio padrão (σ): raiz quadrada da variância

$$desvio\ padrao(x) = \sigma \quad (x) = \sqrt{\sigma^2(x)}$$

- Ambos também são sensíveis a outliers

Medidas de Espalhamento

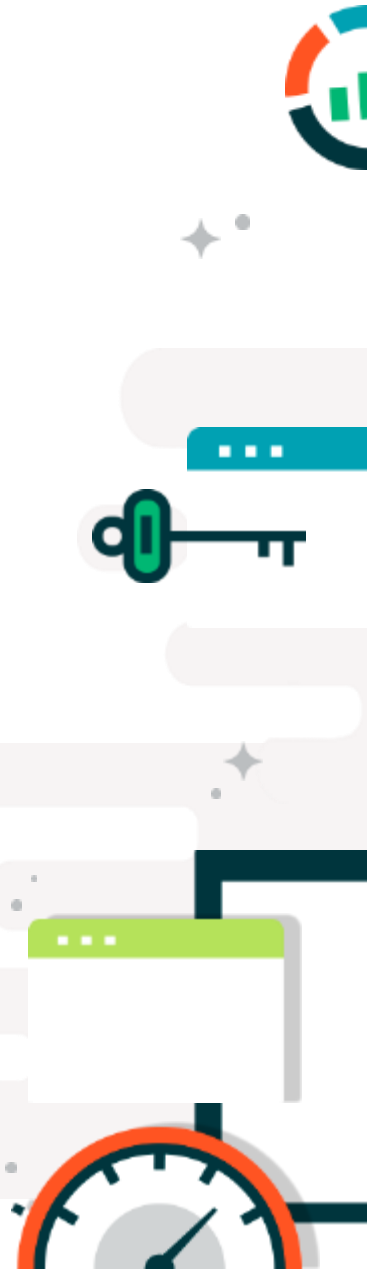
Desvio Padrão

- Exemplo de idades: {25, 28, 20, 19, 23}
- Média das idades: 115 / 5 = 23 anos

- Var = $\dfrac{(25-23)^2 + (28-23)^2 + (20-23)^2 + (19-23)^2 + (23-23)^2}{(5-1)}$

  = (4 + 25 + 9 + 16 + 0) / 4 = **13,5**

- **Desvio Padrão =** $\sqrt{Var} = \sqrt{13,5} = 3,67$ ⟶ Interpretação: 23 ± 3,67



Medidas de Espalhamento

Algumas outras medidas

Desvio médio absoluto (AAD – *absolute average deviation*):

$$AAD(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Desvio mediano absoluto (MAD – *median absolute deviation*):

$$MAD(x) = mediana(\{|x_1 - \bar{x}|, \cdots, |x_n - \bar{x}|\})$$

Intervalo interquertil (IQR – *interquartil interval*):

$$IQR(x) = P_{75\%} - P_{25\%}$$

Base de Dados IRIS

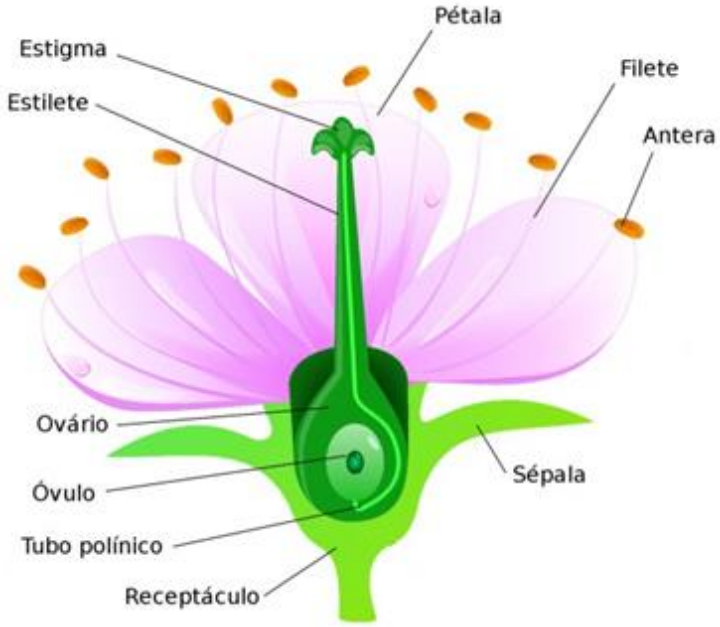
Dados sobre de espécies de lírios

- Iris é uma base de dados disponível no *Machine Learning Repository* da UC Irvine.
- http://archive.ics.uci.edu/ml/

Base de Dados IRIS

Estatística Descritiva

	Comprimento da sépala (cm)	Largura da sépala (cm)	Comprimento da pétala (cm)	Largura da pétala (cm)
count	150,00000	150,00000	150,00000	150,00000
mean	5,84333	3,05400	3,75867	1,19867
std	0,82807	0,43359	1,76442	0,76316
min	4,30000	2,00000	1,00000	0,10000
25%	5,10000	2,80000	1,60000	0,30000
50%	5,80000	3,00000	4,35000	1,30000
75%	6,40000	3,30000	5,10000	1,80000
max	7,90000	4,40000	6,90000	2,50000




PUC Minas
Virtual