

# Tópicos Especiais em Computação I

## Análise de Dados

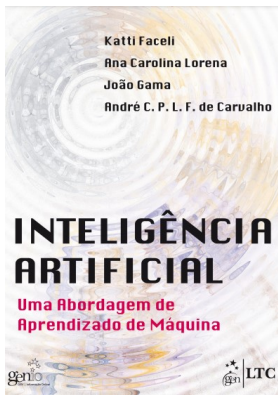
Patrícia Lucas

Bacharelado em Sistemas de Informação  
IFNMG - Campus Salinas

Salinas  
Março 2021

# Referência

## Análise dos dados



## Capítulo 2: Análise de dados.

Inteligência Artificial: Uma abordagem de aprendizado de máquina. Katti Faceli...[et al.]. - Rio de Janeiro: LTC, 2011.

# Conjuntos de dados

## Introdução

Apesar do crescente número de bases de dados disponíveis, na maioria das vezes não é possível utilizar algoritmos de aprendizado de máquina diretamente sobre esses dados.

A análise dos dados e as técnicas de pré-processamento são frequentemente utilizadas para tornar os conjuntos de dados mais adequados para o uso desses algoritmos.

# Análise dos dados

## Análise dos dados

A análise das características presentes em um conjunto de dados permite a descoberta de padrões e tendências que podem fornecer informações valiosas que ajudem a compreender o processo que gerou os dados.

Muitas dessas características podem ser obtidas por meio da aplicação de fórmulas estatísticas simples ou podem ser observadas por meio do uso de técnicas de visualização.

# Exploração de dados

## Análise dos dados

Uma grande quantidade de informações úteis pode ser extraída de um conjunto de dados por meio de sua análise ou exploração.

Informações obtidas durante a exploração podem ajudar na escolha da técnica mais apropriada de pré-processamento e também do aprendizado.

A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados como, por exemplo:

- Frequência.
- Localização ou tendências central (ex: média).
- Dispersão ou espalhamento (ex: desvio padrão).
- Distribuição ou formato.

# Exploração de dados

## Dados univariados

**Dados univariados:** quando o objeto possui apenas um atributo.

*Medidas de localidade:* definem pontos de referência nos dados.

Ex: moda, média, mediana e os quartis.

- Média: dado um conjunto de  $n$  valores numéricos  $x = x_1, x_2, \dots, x_n$ , o valor médio desse conjunto é dado pela Equação 1:

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

# Exploração de dados

## Dados univariados

**OBS: a média é sensível a *outliers*!**

*Outliers* são valores muito diferentes dos demais valores observados para um mesmo atributo. Vamos verificar isso...

Uma turma de 7 alunos possui as seguintes idades:

18, 19, 20, 20, 23, 24, 24.

A média de idades dessa turma é: 21, 1.

Supondo que um aluno de 80 anos entre para essa turma, a média passaria a ser 28, 5.

Dessa forma, percebemos que a média é um bom indicador do meio de um conjunto apenas se os valores estão distribuídos simetricamente.

# Exploração de dados

## Dados univariados

O problema com os *outliers* é minimizado com o uso da mediana.

- Mediana: dado um conjunto de  $n$  valores numéricos ordenado  $x = x_1, x_2, \dots, x_n$ , a mediana desse conjunto é dada pela Equação 2:

$$\text{mediana}(x) = \begin{cases} \frac{1}{2}(x_r + x_{r+1}) & \text{se } n \text{ for par} \\ x_{r+1} & \text{se } n \text{ for ímpar} \end{cases} \quad (2)$$

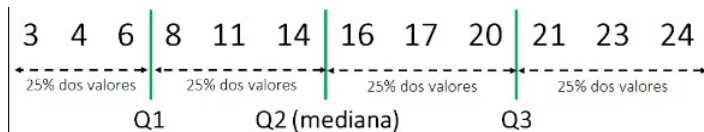
Voltando ao exemplo da turma, a mediana dos conjunto de idades seria 21,5, que é um valor bem mais realista.



# Exploração de dados

## Dados univariados

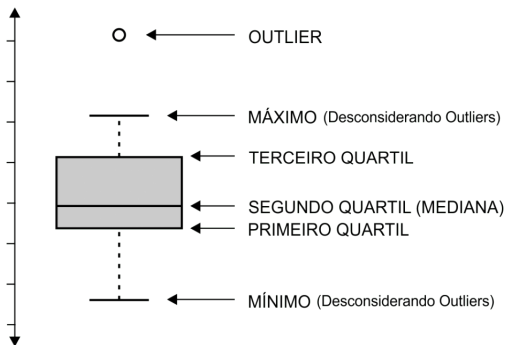
- Moda: é o valor encontrado com maior frequência para um atributo.
- Quartis: dividem os valores do conjunto ordenado da seguinte forma:



# Exploração de dados

## Dados univariados

Uma técnica para visualizar os quartis, mediana e *outliers* é o *box-plots*.



# Exploração de dados

## Dados univariados

*Medidas de espalhamento:* medem a dispersão de um conjunto de valores, permitindo observar se esses estão amplamente espalhados ou relativamente concentrados em torno de um valor.

A medidas mais comuns são:

- Intervalo.
- Variância.
- Desvio padrão.

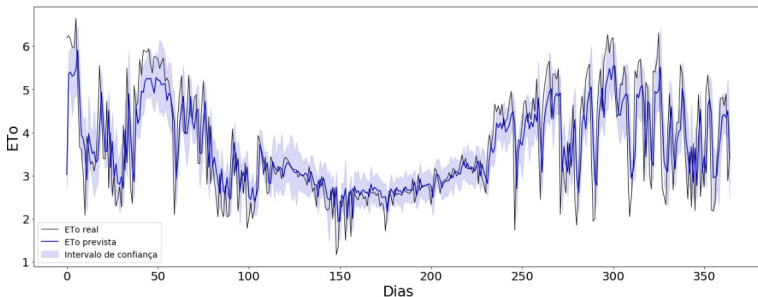
# Exploração de dados

## Dados univariados

**Intervalo:** mostra o espalhamento máximo entre os valores de um conjunto.

Sejam  $x = x_1, x_2, \dots, x_n$  os valores do atributo para  $n$  objetos. O intervalo desse conjunto é medido pela Equação 3:

$$\text{intervalo}(x) = \max_{i=1, \dots, n} (x_i) - \min_{i=1, \dots, n} (x_i) \quad (3)$$



# Exploração de dados

## Dados univariados

**Variância:** avalia o espalhamento de valores em um conjunto.

Sejam  $x = x_1, x_2, \dots, x_n$  os valores do atributo para  $n$  objetos. A variância desse conjunto é medida pela Equação 4:

$$\text{variância}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4)$$

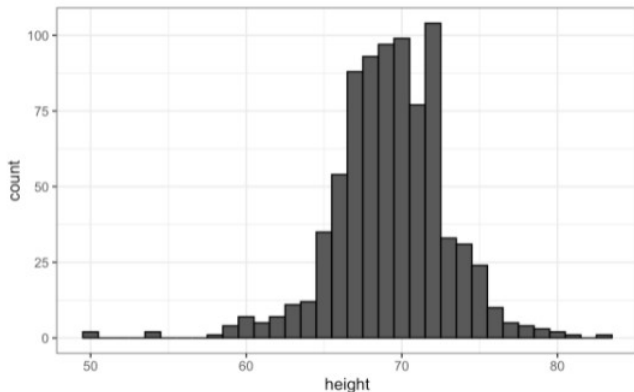
Onde:  $\bar{x}$  é a média dos valores de  $x$ .

Outra medida de espalhamento é o desvio padrão, calculado a partir da raiz quadrada da variância.

# Exploração de dados

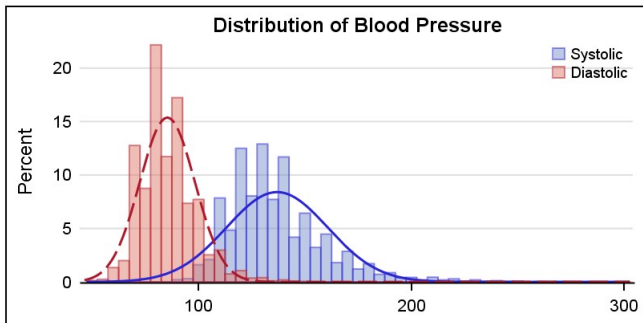
## Dados univariados

Para visualizar o desvio padrão e a distribuição dos dados, podemos representá-los em um **histograma**. Cada barra do histograma possui uma altura proporcional ao número de elementos com aquele valor no conjunto de dados.



# Exploração de dados

## Dados univariados



# Exploração de dados

## Dados multivariados

**Dados multivariados:** quando o objeto possui mais de um atributo.

Nesse caso as medidas de localidade e de espalhamento podem ser obtidas calculando a medida para cada atributo separadamente.

Dados multivariados ainda permitem análises da relação entre dois ou mais atributos através da medida de *correlação*.



# Exploração de dados

## Dados multivariados

*Medida de correlação:* apresenta uma indicação da força da relação linear entre dois atributos.

A *matriz de correlação* apresenta a correlação entre cada possível par de atributos de um conjunto de dados, onde de cada elemento tem seu valor definido pela Equação 5:

$$\text{correlação}(x^i, x^j) = \frac{\text{covariância}(x^i, x^j)}{s_i s_j} \quad (5)$$

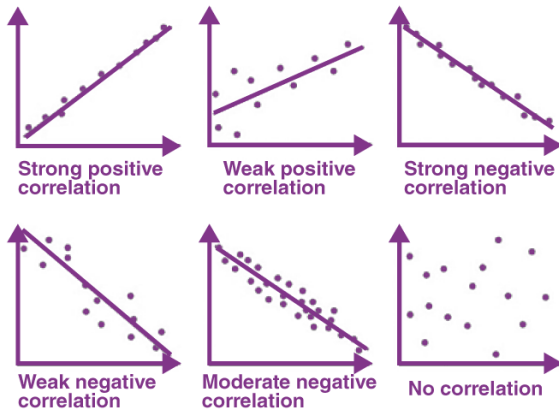
Onde:  $x^i$  é o i-ésimo atributo e  $s_i$  é o desvio padrão dos valores desse atributo.

- A correlação( $x^i, x^i$ ) = 1.
- Correlação positiva: o aumento do valor  $x^i$  é acompanhado pelo aumento de  $x^j$  e vice-versa.
- Correlação negativa: a redução do valor  $x^i$  é acompanhado pelo aumento de  $x^j$  e vice-versa.

# Exploração de dados

## Dados multivariados

Para visualizar a relação entre diferentes atributos, podemos usar um *scatter plot*.



# Exploração de dados

## Correlação X causalidade



Fonte da imagem: <https://medium.com/analytics-vidhya/correlation-causation-977f71bb1e36>