# Princípios básicos da Análise Exploratória de Dados UFCD: 10807

CARGA HORÁRIA: 25 horas

ALEXANDRA CAMPOS VIDAL DE SOUZA
FORMADORA

# Aula 4

- **Variáveis estatísticas e tipos de dados**
- **Métodos de análise de dados para uma variável**
  - Medidas de tendência central
  - Medidas de dispersão
  - Distribuições de frequências e histogramas
  - Medidas de localização
  - Outros indicadores

# Variáveis estatísticas e tipos de dados

# Tipos de Variáveis de Dados

Um dos elementos mais críticos da estatística e da análise de dados é a habilidade de escolher a técnica certa para cada tarefa a ser realizada. Carpinteiros e mecânicos sabem a importância de se ter a ferramenta certa na hora certa e os problemas que podem ocorrer caso a ferramenta errada seja utilizada. Também sabem que usar a ferramenta certa aumenta suas chances de conseguir o resultado desejado logo na primeira tentativa, segundo a abordagem do “seja inteligente e economize esforço”

# O que são Variáveis?

Variáveis são as características ou atributos medidos ou registrados para cada observação.

Elas descrevem as propriedades, qualidades ou quantidades que variam entre as observações

Variáveis

Observações

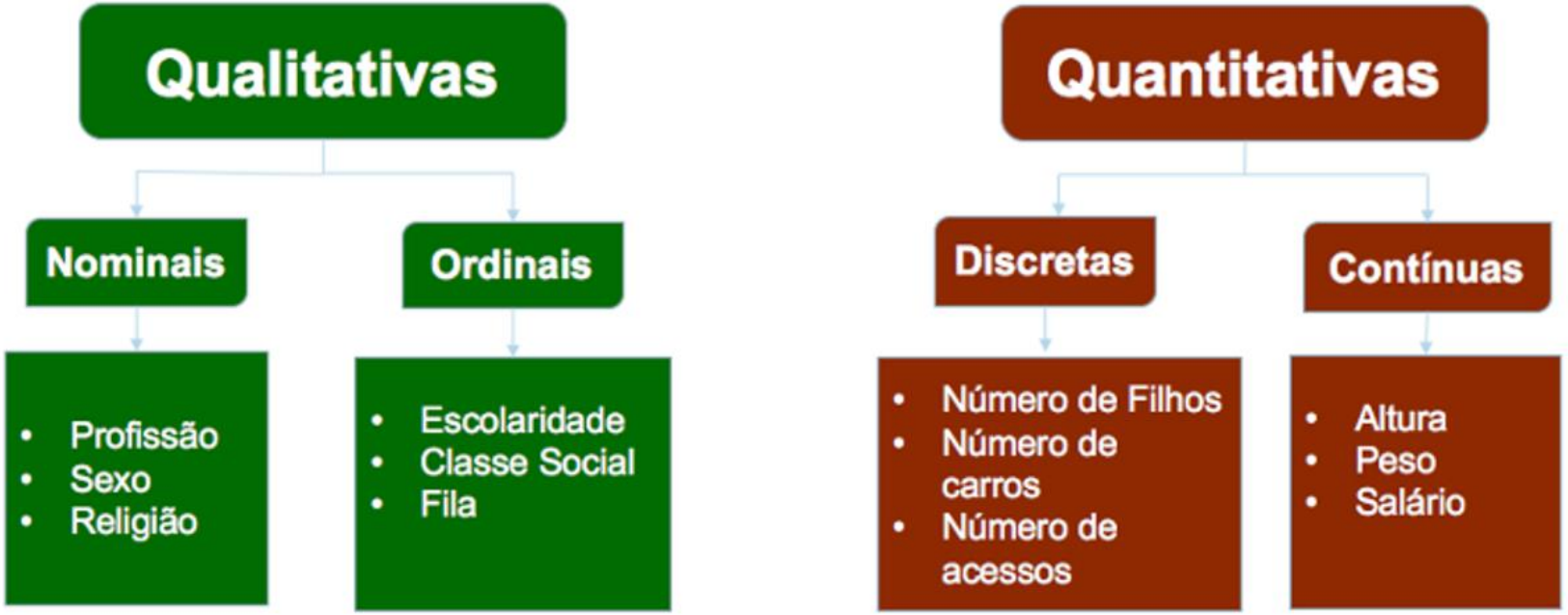
| | Idade | Sexo | Peso | Cor dos olhos |
|---|---|---|---|---|
| Indivíduo 1 | 42 | M | 59 | Verde |
| Indivíduo 2 | 34 | M | 54 | Castanho |
| Indivíduo 3 | 56 | F | 89 | Azul |
| Indivíduo 4 | 41 | M | 76 | Castanho |
| Indivíduo 5 | 23 | F | 65 | Castanho |

Como exemplo a altura e o peso seriam variáveis no conjunto de dados dos estudantes. Cada estudante (observação) teria um valor específico para a altura e outro para o peso (variáveis).

# Os dados podem conter variáveis

- **Qualitativas –** utilizam termos descritivos para descrever algo de interesse. Ex: cor dos olhos, estado civil, religião, gênero, grau de escolaridade, classe social, tipo sanguíneo, cor da pele, etc...
- **Quantitativas** – representadas por valores numéricos que podem ser contados ou medidos. Ex: número de crianças em uma sala de aula, peso do corpo humano, idade, número de filhos, etc...

Dentro desta classificação, podemos ter variáveis:


Qualitativas
Nominais
Profissão
Sexo
Religião
Ordinais
Escolaridade
Classe Social
Fila
Quantitativas
Discretas
Número de Filhos
Número de carros
Número de acessos
Contínuas
Altura
Peso
Salário

- **Qualitativas nominais**(não há uma ordem natural), como, por exemplo, o sexo de uma pessoa.
- **Qualitativas ordinais**(possuem uma ordem natural), como, por exemplo, o índice de aprovação de um político: péssimo, ruim, regular, bom ou ótimo
- **Quantitativas discretas**(os possíveis valores são contáveis), como o número de alunos em uma sala ou o número de acessos
- **Quantitativas contínuas**(podem ser observados quaisquer valores dentro de um intervalo), como a altura ou peso de uma pessoa.

Mas, atenção!!!!

- Um dado classificado como "idade" pode ser quantitativo.
- Ex:. 11, 15, 18, 25, 42 anos.
- Entretanto, se esse dado for informado por "faixa etária" ele é qualitativo (ordinal).
- Ex: 0 –5 anos, 6 –12 anos, 13 –18 anos, 19 –28anos

- A classificação dos dados é muito importante, pois uma vez classificado da forma correta, facilitará a escolha do melhor teste estatístico a ser utilizado na analise de dados.
- Essas informações são muito importantes tambem para tratamento de dados.

# Como Escolher a visualização Estatístico Baseado nos Tipos de Variáveis e Dados

- **Qualitativas:** Usar gráficos de barras, tabelas de frequência e proporções.
- **Quantitativas:** Aplicar medidas de tendência central (média, mediana) e dispersão (desvio padrão, variância), além de gráficos como histogramas.

# Principais objetivos

- **Resumo dos dados:** Identificar informações essenciais, como média, mediana e moda.
- **Compreensão da distribuição:** Observar como os dados estão espalhados ou concentrados.
- **Identificação de padrões ou anomalias**: Encontrar tendências ou outliers.
- **Preparação para análises mais complexas:** Auxiliar no pré-processamento de dados para análises multivariadas ou modelagens.

# Medidas de tendência central ou Posição (Média, Mediana e Moda)

- Os **métodos de análise de dados para uma variável** são ferramentas e técnicas utilizadas para explorar e descrever os dados de **uma única variável,** ou seja, uma característica ou atributo isolado de um conjunto de dados. Essa análise é também conhecida como **análise univariada** e é essencial para entender os padrões básicos, características e distribuição dos dados.

# Medidas de tendência central

# Principais Métodos e Técnicas

## Medidas de Posição

(Média, Mediana e Moda)

As medidas de posição, também conhecidas como medidas de tendência central, são valores que descrevem o centro ou a posição central de um conjunto de dados. As três medidas de posição mais comuns são a **média, mediana e a moda**. Cada uma delas oferece uma maneira diferente de resumir e representar a distribuição dos dados.

# Média

A média é a soma de todos os valores de um conjunto de dados dividida pelo número total de valores. É uma das medidas de tendência central mais comuns e frequentemente usada para representar o valor "típico" de um conjunto de dados. A média pode ser afetada por valores extremos (outliers) e pode não ser a melhor representação do centro dos dados em tais casos.

Média=n∑xi

xi: Cada valor dos dados.

n: Número total de valores.

Exemplo: Para os dados 2,4,6,82,4,6,8: $\text{Média} = \frac{2+4+6+8}{4} = 5$

# Considerações sobre as Medidas de Posição

## A Média

- **A média** é o número que fica equidistante de todos os números da distribuição…
- **A média** é o centro de massa da distribuição dos dados…
- **A média** é o primeiro chute, na verdade a estimativa mais imediata…
- **A média** é o modelo mais simples para estimar, generalizar um dado…
- **A média** só depende de uma soma e de uma divisão…
- **A média** é a grande "mãe", ela abraça todos os valores, mesmo os mais distantes…

# Mediana

A mediana é o valor que separa um conjunto de dados ordenado em duas metades iguais. Se o número total de valores no conjunto de dados é ímpar, a mediana é o valor do meio. Se o número total de valores é par, a mediana é a média dos dois valores centrais. A mediana é menos sensível a valores extremos e pode ser uma medida mais representativa do centro dos dados quando a distribuição é assimétrica ou contém outliers(valores extremos).

# Mediana

Se **n** (número de valores) for ímpar: A mediana é o valor no meio. Se **n** for par: A mediana é a média dos dois valores centrais.

Passos:

- Ordene os dados.
- Encontre o valor central.

Exemplo:

Dados ímpares: 1,3,5,7,9 → Mediana = 5.

Dados pares: 2,4,6,8 → Mediana = 4+6/2=5.

# Considerações sobre as Medidas de Posição

## A Médiana

- **A mediana** significa “aquele número que está no meio”…
- **A mediana** é bastante intuitiva…
- **A mediana** depende de uma ordenação prévia dos dados…
- **A mediana** NÃO é sensível a valores extremos…

# Moda

A moda **é o valor que ocorre com maior frequência** em um conjunto de dados. Um conjunto de dados pode ter nenhuma moda, uma moda (unimodal) ou várias modas (multimodal). A moda pode ser usada para dados numéricos ou categóricos e é uma medida útil da tendência central, especialmente quando a média e a mediana não são aplicáveis ou não fornecem uma representação adequada do centro dos dados.

A moda é o valor que aparece com maior frequência.

# Moda

Pode haver uma moda **(unimodal),** mais de uma moda **(multimodal)** ou nenhuma moda.

Exemplo:

Dados: 1,2,2,3,41,2,2,3,4 → Moda = 2. (unimodal)

Dados: 1,2,2,3,3,41,2,2,3,4,3 → Moda = 2 e 3. (multimodal)

Dados: 1,2,3,41,4 → Sem moda.

# Considerações sobre as Medidas de Posição

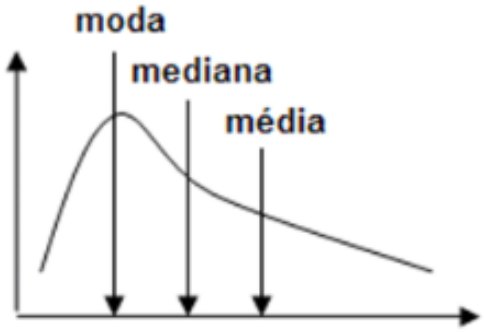
## A Moda

- **A moda** é o chute que damos para algo categórico…
- **A moda** é tão simples que nem tem fórmula…
- **A moda** não é necessariamente única…

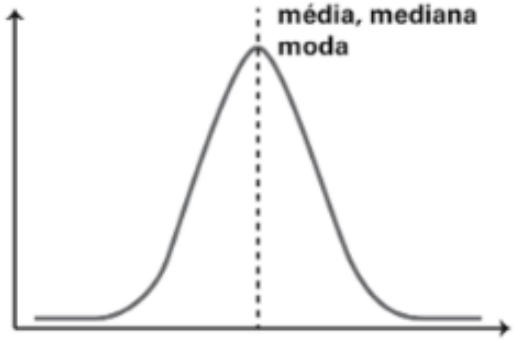
- Pode ser feito um algoritmo para achar os numeros que mais aparecem numconjunto de dados

# Comparar Media e Mediana

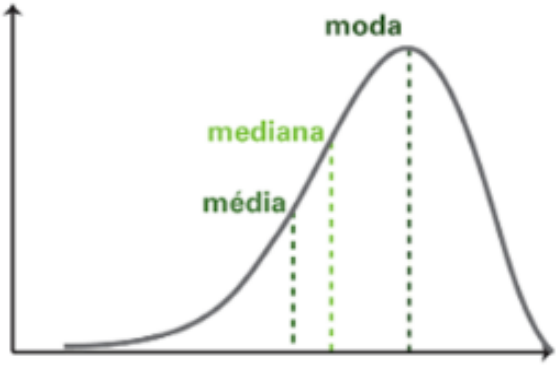
Posição de acordo com o comportamento dos dados.




Distribuição assimétrica direita




Distribuição simétrica




Distribuição assimétrica esquerda

# Para que usar a Media ou mediana

- Qual a vantagem de se usar tanto a média quanto a mediana , se elas tem o mesmo objetivo?

A vantagem está na diferença que temos entre elas(ou seja, sensibilidade a valores extremos)

# E quanto a Valores Extremos?

- **O chamado outliers**, são dados, observações que destoam das demais informações, e que aparece quando comparamos a média e mediana.

# Obrigado!