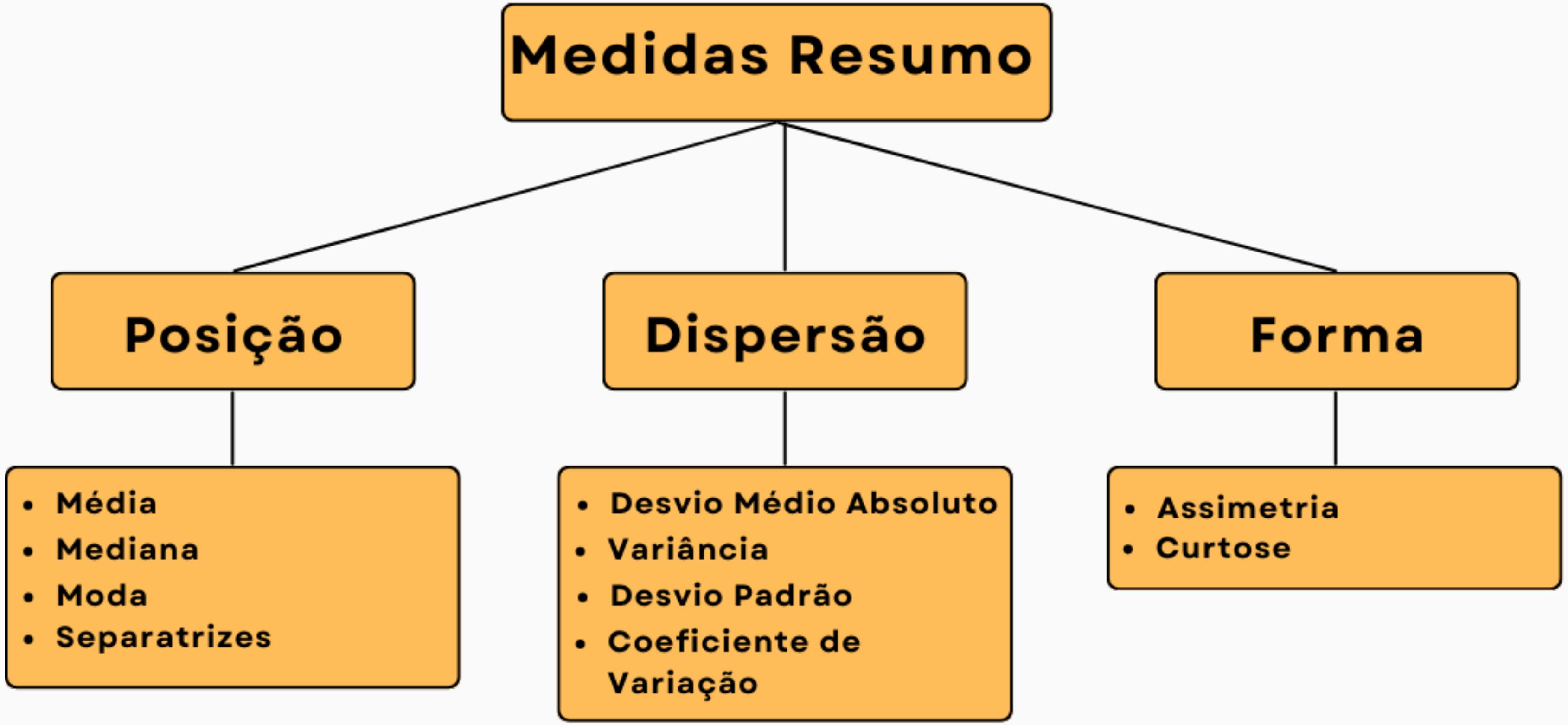
ESTATÍSTICA BÁSICA

Medidas Resumo

Universidade Federal de Santa Maria - UFSM
Prof. Moizés da Silva Melo

Medidas Descritivas



Medidas Descritivas

- As medidas descritivas são utilizadas para **resumir e descrever** um conjunto de dados.
- Elas oferecem informações valiosas sobre a **distribuição dos dados e permitem a análise de características** como centralização, variabilidade e posição dos valores em um conjunto de dados.
- A escolha das medidas apropriadas depende do tipo de dados e dos objetivos da análise estatística.
- As principais medidas descritivas incluem **medidas de posição, medidas de dispersão e medidas de forma**.

Medidas de Posição

Medidas de Posição

Média

- A **média aritmética** é a soma de todos os valores do conjunto dividida pelo número total de observações.

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

- x_i representa cada valor individual no conjunto de dados.
- n é o número total de valores no conjunto de dados.
- $\sum$ denota a soma dos valores.

Medidas de Posição

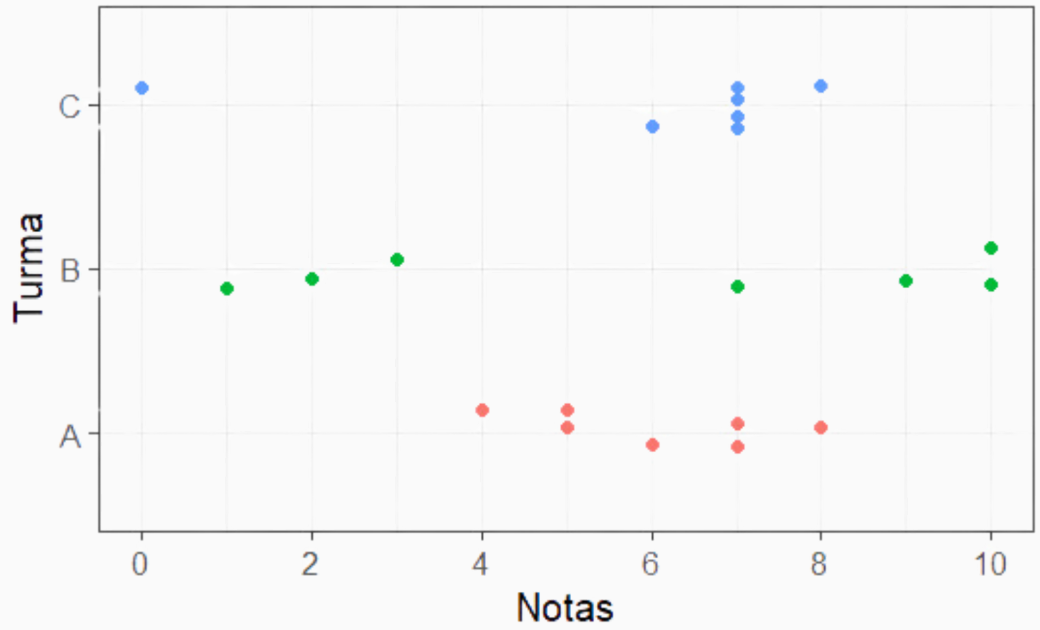
- **Exemplo:** Note que, embora as distribuições das notas das três turmas (A, B, C) sejam distintas, todas apresentam a mesma média (6 pontos).

Turma	Notas dos Alunos	Média da turma
A	**4 5 5 6 7 7 8**	**6**
B	**1 2 3 7 9 10 10**	**6**
C	**0 6 7 7 7 7 8**	**6**

$$\text{Média}_A = \frac{4+5+5+6+7+7+8}{7} = 6$$

$$\text{Média}_B = \frac{0+6+7+7+7+7+8}{7} = 6$$

$$\text{Média}_C = \frac{4+5+5+6+7+7+8}{7} = 6$$



Medidas de Posição

- A média é útil pois permite **resumir** um conjunto de valores e obter uma estimativa central que pode ser **interpretada e comparada facilmente**. No entanto, é importante considerar que a média é **sensível a valores extremos**, também conhecidos como outliers.

- **Exemplo:** Considere uma empresa que está analisando os salários mensais de seus funcionários. A maioria dos funcionários tem salários em torno de R$ 5.000 por mês, mas há um diretor experiente que ganha R$ 100.000. A empresa coletou os salários mensais de cinco funcionários: R$4.000, R$6.000, R$5.500, R$100.000, R$4.500

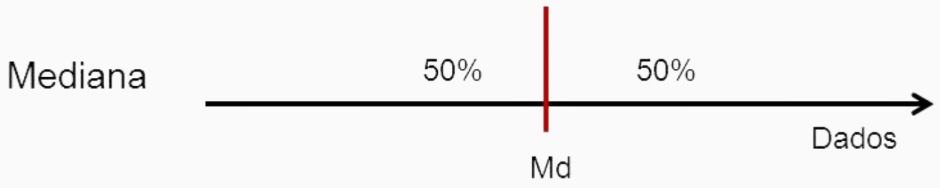
$$\bar{x} = \frac{4.000 + 6.000 + 5.500 + 100.000 + 4.500}{5} = 24.000$$

- Essa média é bastante influenciada pelo salário extremamente alto do diretor. Se olharmos apenas para a média, poderíamos erroneamente concluir que os funcionários em geral ganham em torno de R$24.000, o que não é representativo dos salários da maioria.

Medidas de Posição

Mediana

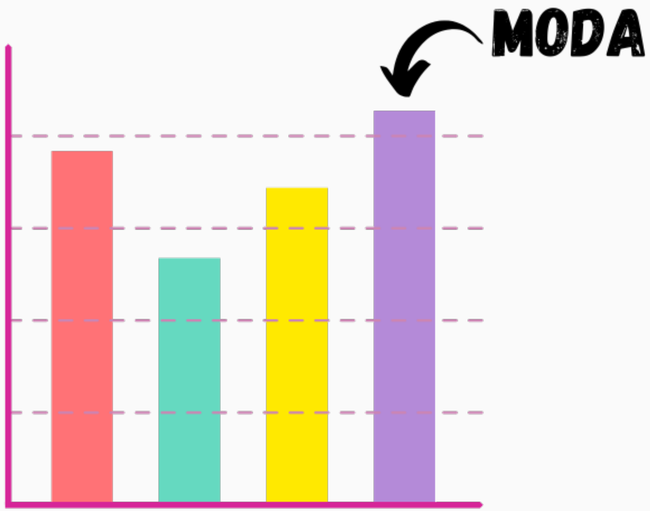
- A **mediana** é o valor que divide o conjunto de dados em duas partes de igual tamanho quando os valores são organizados em ordem.
- O **cálculo** da mediana vai variar ligeiramente dependendo se o número de observações no conjunto de dados é par ou ímpar.
  - Quando o número de observações é ímpar, a mediana é simplesmente o valor do meio após ordenar os dados.
  - Quando o número de observações é par, a mediana é calculada como a média dos dois valores do meio após ordenar os dados.
- A mediana é menos sensível a valores extremos (outliers) do que a média.
- **Exemplo:** Encontre a mediana para os conjuntos de dados a seguir:
  - 17, 12, 20, 15, 10, 23.
  - 17, 12, 20, 15, 10, 23, 21, 24, 98



Medidas de Posição

Moda

- A **moda** é o valor que aparece com maior frequência em um conjunto de dados.



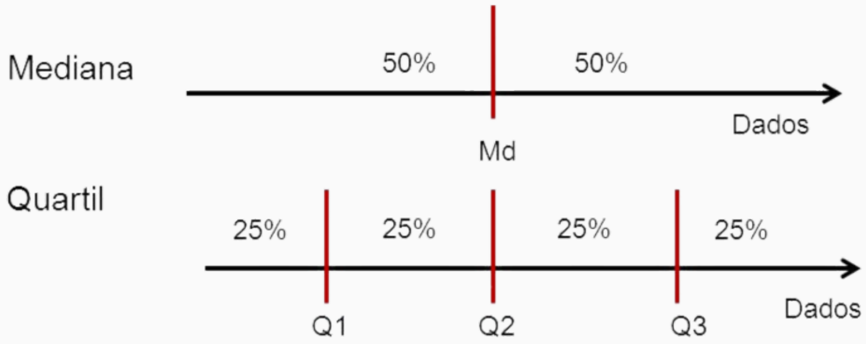
Exemplo

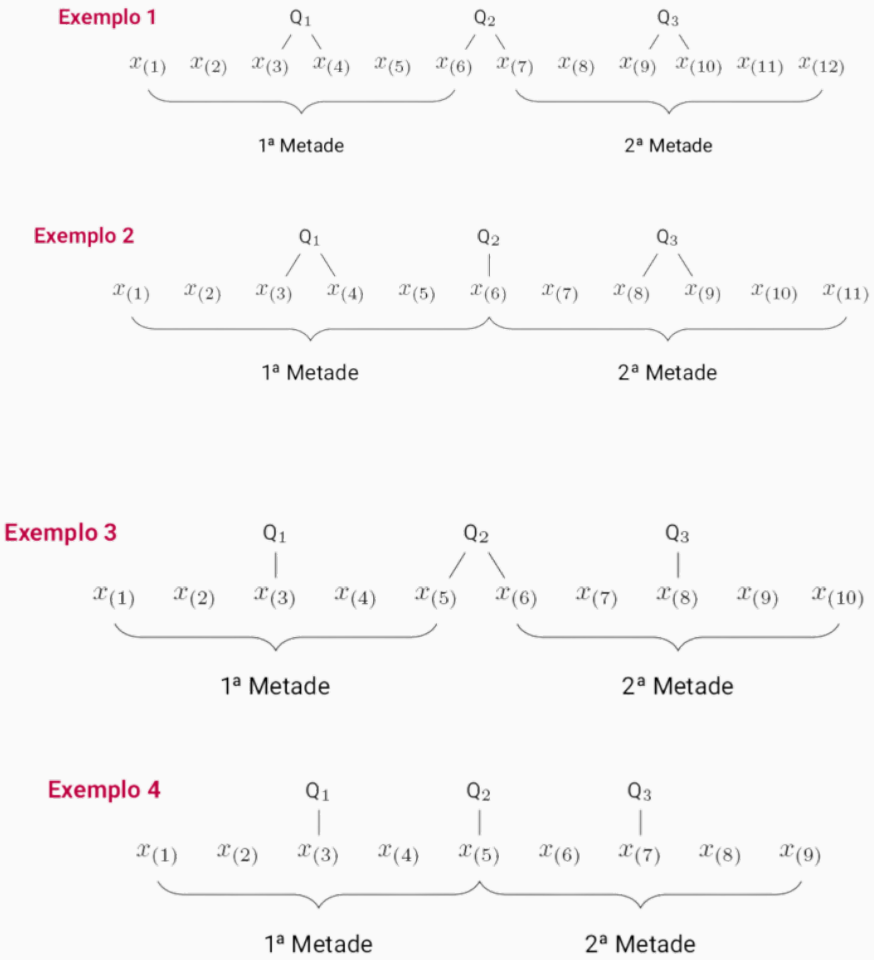
- Dados os conjuntos de observações a seguir, determine a **moda** para cada caso.
    - 4, 3, 4, 5, 3, 7, 2, 3
    - 2, 3, 4, 2, 7, 8, 3
    - 2, 2, 3, 3, 6, 6

Quartis

- Os quartis dividem um conjunto de dados ordenados em quatro partes iguais.

- O segundo quartil é equivalente à mediana.





Medidas de Dispersão

Medidas de Dispersão

Desvio Médio Absoluto

- É a média das diferenças absolutas entre cada valor e a média dos dados.

$$D_m = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$

- n é o número total de observações no conjunto de dados.
- x_i são os valores individuais no conjunto de dados.
- $\overline{x}$ é a média dos valores no conjunto de dados.

Variância

- A variância é uma das **medidas de dispersão mais importantes** e amplamente utilizadas na estatística e na análise de dados. A fórmula da **variância amostral** é a seguinte:

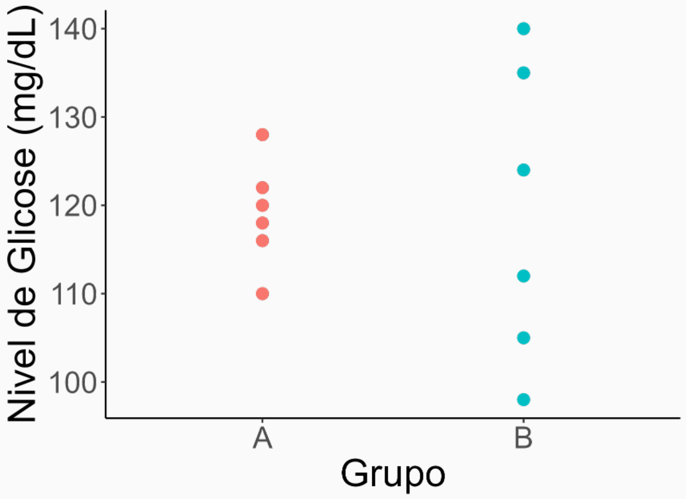
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- n é o número total de observações no conjunto de dados.
- x_i são os valores individuais no conjunto de dados.
- $\overline{x}$ é a média dos valores no conjunto de dados.

Variância

- Os dados apresentados abaixo representam os níveis de glicose no sangue (mg/dL) medidos em dois grupos diferentes de pacientes (Grupo A e Grupo B).

Grupo A	**110**	**120**	**116**	**122**	**128**	**118**
Grupo B	**140**	**105**	**098**	**112**	**135**	**124**



- Embora os dados apresentem comportamentos diferentes, as médias de ambas as amostras são idênticas $\overline{x}_A = \overline{x}_B = \mathbf{119mg/dL}$.

- Ao resumir um conjunto de dados utilizando apenas as medidas de posição, muitas informações relevantes podem se perder. Por isso, é importante medirmos também a dispersão ou variabilidade dos dados.

- Comparando as variâncias, podemos notar que os dados obtidos do grupo B variam mais que os do grupo A.

- $s_A^2 = 36,4$

- $s_B^2 = 281.6$

Medidas de Dispersão

Desvio Padrão

- Às vezes, pode ser mais interessante trabalhar com uma medida de dispersão que esteja na **mesma unidade da variável original**.
- Uma maneira de solucionar o problema é **usar a raiz quadrado da variância**.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{s^2}$$

- No exemplo anterior, temos:

Grupo A

$$s_A = \sqrt{36,4} = 6,03$$

Grupo B

$$s_B = \sqrt{281,6} = 16,78$$

Coeficiente de Variação

- O coeficiente de variação é definido pelo razão entre o desvio padrão e a média:

$$\text{CV} = \frac{s}{\overline{x}} \times 100\%$$

- Essa é uma medida **adimensional**.
- Essa medida é utilizada quando se deseja **comparar a variabilidade** de conjuntos de dados que apresentem diferentes unidades de medição.

Medidas de Dispersão

Coeficiente de Variação (Exemplo)

- Suponha que estamos estudando o índice de massa corporal (IMC) e os níveis de colesterol em um grupo de pacientes. Temos os seguintes dados:
- **IMC:** $\overline{x} = 25,5\text{kg}/\text{m}^2$ e $s = 3,0\text{kg}/\text{m}^2$
- **Nível de colesterol:** $180\text{mg}/\text{dL}$ e $s = 30\text{mg}/\text{dL}$
- Calculando o coeficiente de variação para ambas as variáveis para comparar a variabilidade entre elas:

$$\text{CV} = \frac{3,0\,\text{kg}/\text{m}^2}{25,5\,\text{kg}/\text{m}^2} \times 100\% = 11,76\%$$

$$\text{CV} = \frac{30\,\text{mg}/\text{dL}}{180\,\text{mg}/\text{dL}} \times 100\% = 16,66\%$$