# Sumarização de dados

## Média, mediana, moda, variância, desvio padrão, distribuição de frequência

André Gustavo dos Santos

Departamento de Informática
Universidade Federal de Viçosa

INF222 - 2022/2

UFV

## – Fonte do material

O conteúdo a seguir foi preparado e gentilmente cedido para uso nesta disciplina pela profª Elizabeth Wanner, do PPGMMC, CEFET-MG.

UFV

# Tópicos da aula

1 Estatística: ciência dos dados

2 Medidas de Centro

3 Medidas de Variação

4 Distribuição de frequência

UFV

# Estatística

- A estatística surgiu por demanda do Estado para coletar dados com fins tributários
- Contudo, foi a partir do século XX que a Estatística desenvolveu-se como uma área específica do conhecimento: Estatística Inferencial
- Resumidamente: estatística é a ciência dos dados
- Trata-se de uma ciência MEIO, que objetiva instrumentalizar a coleta, classificação, sumarização, organização, análise e interpretação de dados

UFV

# Probabilidade X Estatística

**Probabilidade**
Dada a piscina de bolas, qual é a probabilidade de se obter uma certa combinação de cores?

**Estatítica**
Dadas as cores de poucas bolas, o que eu conheço sobre a piscina de bolas?

**Inferência Estatística**: usando amostras para obter conclusões sobre populações

# População, Amostra e Observação

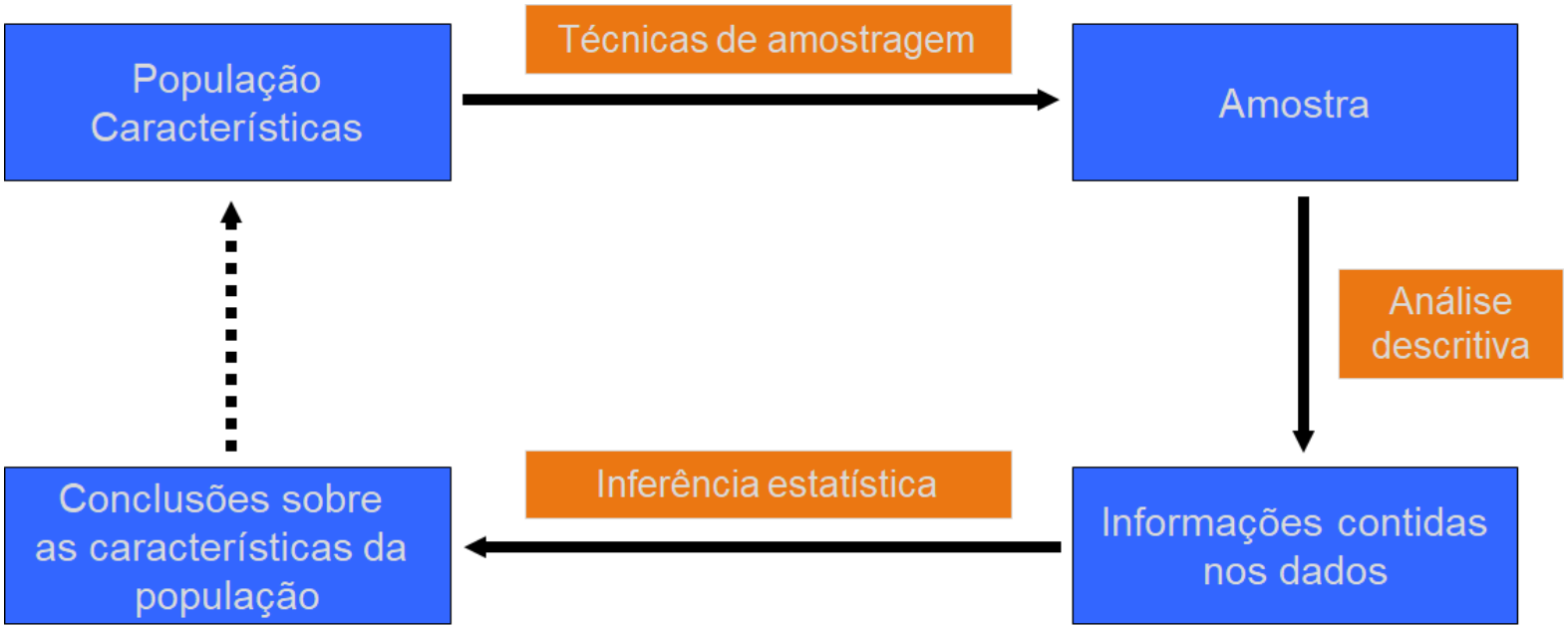
**População:** a totalidade (universo) de itens, objetos ou pessoas que possuem uma determinada característica em comum

**Amostra:** uma parte representativa da população

**Observação:** é um único elemento de uma amostra, um ponto individualmente coletado

UFV

# Estatística



# Motivação

**Case 1:** Idade das Melhores Atrizes e dos Melhores Atores listadas em ordem desde a primeira cerimônia de premiação do Oscar em 1928.

**PERGUNTAS:**

- Os prêmios na Academia envolvem discriminação com base na idade?
- Há diferenças sérias e importantes entre as idades das Melhores Atrizes e as idades dos Melhores Atores?
- Existe uma tendência das Melhores Atrizes serem mais jovens que os Melhores Atores?

UFV

# Idade das Melhores Atrizes e dos Melhores Atores

**Melhores atrizes**

| | | | | | |
|---|---|---|---|---|---|
| 22 | 37 | 28 | 63 | 32 | 26 |
| 31 | 27 | 27 | 28 | 30 | 26 |
| 29 | 24 | 38 | 25 | 29 | 41 |
| 30 | 35 | 35 | 33 | 29 | 38 |
| 54 | 24 | 25 | 46 | 41 | 28 |
| 40 | 39 | 29 | 27 | 31 | 38 |
| 29 | 25 | 35 | 60 | 43 | 35 |
| 34 | 34 | 17 | 37 | 42 | 41 |
| 36 | 32 | 41 | 33 | 31 | 74 |
| 33 | 50 | 38 | 61 | 21 | 41 |
| 26 | 80 | 42 | 19 | 33 | 35 |
| 45 | 49 | 39 | 34 | 26 | 25 |
| 33 | 35 | 35 | 28 | | |

**Melhores atores**

| | | | | | |
|---|---|---|---|---|---|
| 44 | 41 | 62 | 52 | 41 | 34 |
| 34 | 52 | 41 | 37 | 38 | 34 |
| 32 | 40 | 43 | 56 | 41 | 39 |
| 49 | 57 | 41 | 38 | 42 | 52 |
| 51 | 35 | 30 | 39 | 41 | 44 |
| 49 | 35 | 47 | 31 | 47 | 37 |
| 57 | 42 | 45 | 42 | 44 | 62 |
| 43 | 42 | 48 | 49 | 56 | 38 |
| 60 | 30 | 40 | 42 | 36 | 76 |
| 39 | 53 | 45 | 36 | 62 | 43 |
| 51 | 32 | 42 | 54 | 52 | 37 |
| 38 | 32 | 45 | 60 | 46 | 40 |
| 36 | 47 | 29 | 43 | | |

UFV

# Entendendo os dados

**OBJETIVOS:**

- Melhor compreensão dos dados com o propósito da apresentação
- Organizar e resumir um conjunto de dados em tabelas
- Organizar e resumir um conjunto de dados em gráficos

UFV

# Características dos dados

**Algumas definições:**

- **Dados brutos:** dados como foram oferecidos ou recolhidos
- **Centro:** valor representativo ou médio do conjunto
- **Variação:** quanto os valores dos dados variam entre eles
- **Distribuição:** natureza ou forma da distribuição dos dados
- ***Outliers:*** valores que se localizam muito longe da grande maioria dos outros valores amostrais
- **Tempo:** características dos dados que mudam com o tempo

UFV

# Medidas de Centro

**Definição**: um número que representa o valor central de um conjunto de dados

- média
- mediana
- moda
- ponto médio

# Média Aritmética

- É a medida de centro encontrada pela adição dos valores e divisão do total pelo número de valores

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{x_1 + \ldots + x_n}{n}$$

em que $\{x_1,\ x_2,\ x_3,\ \cdots,\ x_n\}$ é uma amostra de dados de tamanho $n$

- Vantagem: fácil de entender e de calcular
- Desvantagem: sensível a qualquer valor; um valor excepcional pode afetar drasticamente a média
- Notação para a média populacional: $\mu$

UFV

# Mediana

- É a medida de centro que é o valor do meio quando os dados originais estão arranjados em ordem crescente de magnitude
- Cerca de metade dos valores no conjunto de dados está abaixo da mediana e metade está acima dela

- Procedimento:
  1. Ordene os valores
  2. Se o número de valores for ímpar, a mediana será o número localizado no meio exato da lista ordenada
  3. Se o número de valores for par, a mediana será encontrada pelo cálculo da média aritmética dos dois números do meio
- Vantagem: insensível a valores extremos

UFV

# Exemplo

Considere o conjunto de dados amostrais:

$$\{5.40,\ \ 1.10,\ \ 0.42,\ \ 0.73,\ \ 0.48,\ \ 1.10\}$$

**Média:**

$$\bar{x} = \frac{5.40 + 1.10 + 0.42 + 0.73 + 0.48 + 1.10}{6} = 1.538$$

**Mediana:**

1. Ordene os valores: 0.42, 0.48, 0.73, 1.10, 1.10, 5.40
2. Como temos 6 valores (número par), a mediana é encontrada pelo cálculo da média dos dois valores do meio (o $3^o$ e o $4^o$):

$$\tilde{x} = \frac{0.73 + 1.10}{2} = 0.915$$

**Obs**: Valores bem diferentes para a média e mediana: efeito do 5.40.

UFV

# Moda

**Definição:** É o valor que ocorre mais frequentemente

- Quando dois valores ocorrem com a mesma maior frequência, cada um é uma moda e o conjunto é dito bimodal
- Quando mais de dois valores de dados ocorrem com a mesma maior frequência, cada um é uma moda e o conjunto é dito multimodal
- Quando nenhum valor se repete, dizemos que não há moda

**Exemplo:**

1. $\{5.40,\ 1.10,\ 0.42,\ 0.73,\ 0.48,\ 1.10\}$: O número 1.10 é a moda pois aparece duas vezes (maior frequência)
2. $\{27,\ 27,\ 27,\ 55,\ 55,\ 55,\ 88,\ 88\}$: Os números 27 e 55 são, ambos, modas e, portanto, o conjunto é dito bimodal

UFV

# Ponto Médio

**Definição:** É a medida de centro que é exatamente o valor a meio caminho entre o maior valor e o menor valor no conjunto de dados

$$\text{Ponto Médio} = \frac{\text{valor máximo} + \text{valor mínimo}}{2}$$

- Raramente usado porém fácil de calcular
- Sensível a casos extremos

**Exemplo:**

$$\{5.40,\ \ 1.10,\ \ 0.42,\ \ 0.73,\ \ 0.48,\ \ 1.10\}$$

$$\text{Ponto Médio} = \frac{5.40 + 0.42}{2} = 2.91$$

UFV

# Idade das Melhores Atrizes e dos Melhores Atores

| | Melhores Atrizes | Melhores Atores[1] |
|---|---|---|
| Média | 35.4 | ? |
| Mediana | 33.5 | ? |
| Moda | 35 | ? |
| Ponto Médio | 48.5 | ? |

[1] Exercício para a próxima aula

# Medidas de Variação

Considere três conjuntos de dados:

$$A = \{6,\ \ 6,\ \ 6\}$$

$$B = \{4,\ \ 7,\ \ 7\}$$

$$C = \{1,\ \ 3,\ \ 14\}$$

Observe que a média é a mesma: $\bar{A} = \bar{B} = \bar{C} = 6.0$.

No entanto, as amostras são muito diferentes nas quantidades que os dados variam.

**Objetivo:** Medir tais variações de maneira sistemática.

# Medidas de Variação

- Amplitude
- Variância
- Desvio Padrão
- Coeficiente de Variação

UFV

# Amplitude

**Definição:** é a diferença entre o maior valor e o menor valor

$$A = \{6,\ \ 6,\ \ 6\} \rightarrow \text{Amplitude}_A = 6 - 6 = 0$$

$$B = \{4,\ \ 7,\ \ 7\} \rightarrow \text{Amplitude}_B = 7 - 4 = 3$$

$$C = \{1,\ \ 3,\ \ 14\} \rightarrow \text{Amplitude}_C = 14 - 1 = 13$$

UFV

# Variância

**Definição:** é a média da variação quadrática dos valores em torno da média. Um desvio médio quadrático dos valores em relação à média.

**Variância populacional:**

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

**Variância amostral:**

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

UFV

# Variância

**Propriedades:**

- valor sempre não negativo;
- valores maiores indicam maior variação dos dados;
- pode crescer muito com a inclusão de um ou mais dados que estão muito afastados dos demais;
- a variância de uma amostra é levemente menor que a variância da população: se os dados tiverem sido tirados de uma amostra, é menos provável que inclua dados extremos, portanto estão mais agrupados que os dados da população;
- na variância amostral, o valor $n - 1$ representa o número de graus de liberdade, considerando os desvios $x_i - \bar{x}$;
- isto é importante quando a amostra for pequena; para $n$ grande, existe pouca diferença entre as fórmulas.

UFV

# Desvio Padrão

**Definição:** é raiz quadrada da média da variação quadrática dos valores em torno da média:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \text{ ou } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

**Propriedades:**

- mesmas propriedades da variância;
- porém, com a mesma unidade dos dados originais.

UFV

# Desvio Padrão

Considerando toda a população:

$$A = \{6,\ 6,\ 6\} \rightarrow \sigma_A = \sqrt{\frac{(6-6)^2 + (6-6)^2 + (6-6)^2}{3}} = 0$$

$$B = \{4,\ 4,\ 7\} \rightarrow \sigma_B = \sqrt{\frac{(4-6)^2 + (7-6)^2 + (7-6)^2}{3}} = 1.15$$

$$C = \{1,\ 3,\ 14\} \rightarrow \sigma_C = \sqrt{\frac{(1-6)^2 + (3-6)^2 + (14-6)^2}{3}} = 5.71$$

# Coeficiente de Variação

**Definição:** O coeficiente de variação para um conjunto de dados amostrais não negativos, expresso como um percentual, descreve o desvio padrão relativo à média.

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

Compara a variação para valores originados de diferentes populações.

# Idade das Melhores Atrizes e dos Melhores Atores

| | Melhores Atrizes | Melhores Atores |
|---|---|---|
| Média | 35.4 | ? |
| Mediana | 33.5 | ? |
| Moda | 35 | ? |
| Ponto Médio | 48.5 | ? |
| Amplitude | 63 | ? |
| Desvio Padrão | 11.36 | ? |
| Variância | 129.10 | ? |
| Coef. de Variação | 32.08 % | ? |

UFV

# Dados em tabelas

- Uma tabela ou série estatística é o conjunto de dados agrupados segundo algum critério específico.
- A forma mais simples é aquela em que se organiza os dados por **categorias ou classes**, de acordo com o que se quer estudar.
- Esta tabela se chama **distribuição de frequência** dos dados: listar os valores dos dados individualmente ou por grupos de intervalo, juntamente com suas frequências correspondentes: absoluta (em valores inteiros) e relativa (em valores percentuais), bem como na versão acumulada de cada uma.

UFV

# Classificação dos dados

- As classes ou categorias devem ser abrangentes, escolhidas de forma que toda resposta possa ser incluída.
  Sugestão: classe “outros” ou “100 ou mais”.
- As categorias devem ser distintas e não ser ambíguas.
  Sugestão: “de 0 até 20” e “maior que 20 até 40”.
- As categorias não devem ser muito numerosas, para não ficar muito difícil de se analisar.
  Sugestão: algo entre 5 ou 10 classes.

UFV

# Construção da tabela

- Seja claro sobre o que você deseja que a tabela mostre.
- Liste primeiro a variável de interesse.
- Uma tabela pode resumir mais de uma variável ao mesmo tempo.
- Duas ou três tabelas simples são melhores que uma grande e complicada.
- Toda linha e coluna deve receber um título claro.
- Se possível, inclua os totais de cada linha e coluna.
- A fonte de informação deve ser informada abaixo da tabela.
- Inclua qualquer outra informação útil em notas de rodapé.

UFV

# Distribuição de Frequência

**Estudo das Idades das Melhores Atrizes**

| Idade das atrizes | Frequência absoluta |
|---|---|
| $\leq 30$ | 28 |
| 31-40 | 30 |
| 41-50 | 12 |
| 51-60 | 2 |
| 61-70 | 2 |
| $\geq 71$ | 2 |

**Conceitos:**

- limites inferiores e superiores das classes
- amplitude das classes
- pontos médios das classes
- fronteiras de classes

# Construindo a distribuição de frequência

**Passo 1:** Escolher o número de classe desejado: $n = 6$

**Passo 2:** Calcular a amplitude de classe:

$$A = \frac{max - min}{n} = \frac{80 - 21}{6} = 9.833 \approx 10$$

**Passo 3:** Definir os limites superiores e inferiores:
$1^a$ Classe 21 - 30 ; $2^a$ Classe 31 - 40; $3^a$ Classe 41 - 50 ; $4^a$ Classe 51 - 60; $5^a$ Classe 61-70; $6^a$ Classe 71 - 80

**Passo 4:** Calcular frequência absoluta *F*: fazer a contagem de elementos em cada classe.

**Passo 5:** Calcular frequência relativa *Fr* e acumulada *Fac*, além da relativa acumulada *Frac*.

UFV

# Frequência relativa e acumulada

| Idade das atrizes | F | Fr | Fac | Frac |
|---|---|---|---|---|
| $\leq 30$ | 28 | 37 % | 28 | 37 % |
| 31 - 40 | 30 | 39 % | 58 | 76 % |
| 41-50 | 12 | 15 % | 70 | 91 % |
| 51-60 | 2 | 3 % | 72 | 94 % |
| 61-70 | 2 | 3 % | 74 | 97 % |
| $\geq 71$ | 2 | 3 % | 76 | 100 % |
| TOTAL ($\Sigma$) | 76 | 100 % | – | – |

# Atividade 1 *(consulte enunciado completo e forma de entrega no PVANet Moodle)*

- Calcular as medidas de centro e de variação para os dados dos atores.
- Construir a distribuição de frequência para os dados dos atores.
- Procurar responder às perguntas colocadas:
  - Os prêmios na Academia envolvem discriminação com base na idade?
  - Há diferenças sérias e importantes entre as idades das Melhores Atrizes e as idades dos Melhores Atores?
  - Existe uma tendência das Melhores Atrizes serem mais jovens que os Melhores Atores?

UFV