Introdução à Estatística Descritiva

Descrever e Resumir Dados Utilizando Medidas de Tendência Central e Dispersão

Márcio Nicolau

2025-09-03

Table of contents

List of Figures

Introdução e Objetivos

Após explorarmos os fundamentos da análise combinatória e da probabilidade, estamos prontos para mergulhar no mundo da **Estatística Descritiva**. Esta área da estatística tem como objetivo principal organizar, resumir e apresentar dados de forma que suas características essenciais possam ser facilmente compreendidas. Em vez de lidar com a incerteza de eventos futuros, a estatística descritiva nos ajuda a entender o que já aconteceu, transformando grandes volumes de dados brutos em informações concisas e significativas.

Para qualquer cientista de dados ou analista, a estatística descritiva é a primeira linha de defesa ao abordar um novo conjunto de dados. Ela nos permite identificar padrões, detectar anomalias e formular perguntas importantes para análises mais aprofundadas. Nesta aula, focaremos nas medidas numéricas que nos permitem descrever o "centro" e a "dispersão" de um conjunto de dados.

Objetivos de Aprendizagem

Ao final desta aula, você será capaz de:

- Distinguir entre diferentes tipos de dados e sua relevância para as medidas descritivas.
- Calcular e interpretar as principais medidas de tendência central: média, mediana e moda.
- Calcular e interpretar as principais medidas de dispersão: amplitude, variância e desvio padrão.
- Identificar qual medida de tendência central e dispersão é mais apropriada para diferentes tipos de dados e distribuições.
- Utilizar Python e R para calcular e visualizar essas medidas descritivas.

Tipos de Dados

Antes de mergulharmos nas medidas, é crucial relembrar e entender os tipos de dados, pois a escolha da medida descritiva mais adequada depende diretamente da natureza da variável que está sendo analisada.

Dados Qualitativos (ou Categóricos)

Representam categorias ou qualidades e não podem ser medidos numericamente de forma significativa.

- **Nominais:** Não há ordem intrínseca entre as categorias.
  - **Exemplos:** Cor dos olhos (azul, castanho, verde), sexo (masculino, feminino), estado civil (solteiro, casado, divorciado).
- **Ordinais:** Há uma ordem intrínseca, mas as diferenças entre as categorias não são quantificáveis ou consistentes.

  - **Exemplos:** Grau de escolaridade (fundamental, médio, superior), nível de satisfação (muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito).

Dados Quantitativos (ou Numéricos)

Representam quantidades e são medidos em uma escala numérica.

- **Discretos:** Resultam de contagens e assumem valores inteiros, geralmente com lacunas entre eles.
  - **Exemplos:** Número de filhos, número de carros em uma casa, número de defeitos em um produto.
- **Contínuos:** Resultam de medições e podem assumir qualquer valor dentro de um intervalo.
  - **Exemplos:** Altura, peso, tempo, temperatura, renda.

Medidas de Tendência Central

As medidas de tendência central nos fornecem um valor que busca representar o "centro" de um conjunto de dados, um ponto em torno do qual os dados se agrupam. (Bussab; Morettin, 2017, p. 19–24)

Média Aritmética ($\bar{x}$ ou μ)

A **média aritmética** é a soma de todos os valores de um conjunto de dados dividida pelo número de valores. É a medida de tendência central mais comum.

- **Fórmula (Amostra):** $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **Fórmula (População):** $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Onde:

- x_i representa cada valor individual no conjunto de dados.
- n é o número de observações na amostra.
- N é o número de observações na população.

Características:

- É sensível a valores extremos (outliers).
- A soma dos desvios de cada valor em relação à média é sempre zero.
- É o "centro de gravidade" dos dados.

> Quando usar
>
> Para dados quantitativos sem outliers significativos e com distribuição aproximadamente simétrica.

Mediana ($\tilde{x}$ ou Md)

A **mediana** é o valor do meio em um conjunto de dados ordenado. Divide os dados em duas metades iguais, com 50% dos valores abaixo dela e 50% acima.

Como calcular:

1. Ordene os dados em ordem crescente.
2. Se o número de observações (n) for ímpar, a mediana é o valor central.
   - Posição da mediana: $\frac{n+1}{2}$.
3. Se o número de observações (n) for par, a mediana é a média dos dois valores centrais.
   - Posições da mediana: $\frac{n}{2}$ e $\frac{n}{2} + 1$.

Características:

- É robusta a valores extremos (outliers), ou seja, não é significativamente afetada por eles.

> Quando usar
>
> Para dados quantitativos com outliers, distribuições assimétricas (enviesadas), ou quando você precisa do ponto médio exato. Também pode ser usada para dados ordinais.

Moda (Mo)

A **moda** é o valor ou valores que aparecem com maior frequência em um conjunto de dados.

Características:

- Pode não existir (todos os valores aparecem com a mesma frequência).
- Pode haver uma moda (unimodal), duas modas (bimodal) ou múltiplas modas (multimodal).
- É a única medida de tendência central que pode ser usada com dados nominais.

> Quando usar
>
> Para qualquer tipo de dado, mas é especialmente útil para dados nominais e para identificar picos em distribuições.

Exemplo de Cálculo Manual

Considere o conjunto de dados: $\{10, 12, 12, 15, 18, 20, 22\}$

- **Média:** $(10 + 12 + 12 + 15 + 18 + 20 + 22)/7 = 109/7 \approx 15.57$
- **Mediana:** Os dados já estão ordenados. $n = 7$ (ímpar). Posição: $(7 + 1)/2 = 4$. O 4º valor é 15. Mediana = 15.
- **Moda:** O valor 12 aparece duas vezes, que é a maior frequência. Moda = 12.

Código para Medidas de Tendência Central

Python

```
import numpy as np
import pandas as pd
from scipy import stats
```

```python
data = [10, 12, 12, 15, 18, 20, 22]

# Usando NumPy
mean_np = np.mean(data)
median_np = np.median(data)
# NumPy não tem uma moda direta, mas scipy.stats tem
mode_np_scipy = stats.mode(data, keepdims=False) # keepdims=False para compatibilidade futura

print(f"Dados: {data}")
print(f"Média (NumPy): {mean_np:.2f}")
print(f"Mediana (NumPy): {median_np:.2f}")
print(f"Moda (SciPy): {mode_np_scipy:.2f}")

# Usando Pandas Series
s_data = pd.Series(data)
mean_pd = s_data.mean()
median_pd = s_data.median()
mode_pd = s_data.mode() # Retorna uma série, pode ter múltiplas modas

print("\n--- Usando Pandas ---")
print(f"Média (Pandas): {mean_pd:.2f}")
print(f"Mediana (Pandas): {median_pd:.2f}")
print(f"Moda (Pandas): {list(mode_pd.values)}") # Converte para lista se houver múltiplos
```

R

```r
data <- c(10, 12, 12, 15, 18, 20, 22)

# Média
mean_r <- mean(data)

# Mediana
median_r <- median(data)

# Moda (R base não tem uma função direta para moda, precisamos criar uma ou usar um pacote)
# Função simples para calcular a moda
get_mode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
mode_r <- get_mode(data)

# Se houver múltiplas modas, a função acima retornará apenas a primeira.
# Para retornar todas as modas, é um pouco mais complexo:
```

```r
get_all_modes <- function(v) {
  t <- table(v)
  max_freq <- max(t)
  modes <- as.numeric(names(t[t == max_freq]))
  return(modes)
}
all_modes_r <- get_all_modes(data)


cat(sprintf("Dados: %s\n", paste(data, collapse = ", ")))
cat(sprintf("Média (R): %.2f\n", mean_r))
cat(sprintf("Mediana (R): %.2f\n", median_r))
cat(sprintf("Moda (R, simples): %.2f\n", mode_r))
cat(sprintf("Moda (R, todas as modas): %s\n", paste(all_modes_r, collapse = ", ")))
```

Comparação e Distribuição

A escolha da medida de tendência central depende da forma da distribuição dos dados.

Medidas de Dispersão

As medidas de dispersão nos informam sobre a variabilidade ou o espalhamento dos dados em torno da medida de tendência central. Um conjunto de dados com pouca dispersão é mais homogêneo, enquanto um com muita dispersão é mais heterogêneo. (Bussab; Morettin, 2017, p. 25–32)

Amplitude (Range)

A **amplitude** é a diferença entre o valor máximo e o valor mínimo em um conjunto de dados.

- **Fórmula:** Amplitude = Valor Máximo - Valor Mínimo

Características:

- É a medida de dispersão mais simples.
- Extremamente sensível a outliers, pois utiliza apenas os dois valores extremos.
- Não considera a distribuição dos dados entre o mínimo e o máximo.

Variância (s^2 ou σ^2)

A **variância** mede a dispersão média dos dados em torno da média, calculando a média dos quadrados dos desvios de cada observação em relação à média.

- **Fórmula (Amostra):** $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - (O denominador $n-1$ é usado para amostras para fornecer uma estimativa não viciada da variância populacional.)
- **Fórmula (População):** $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

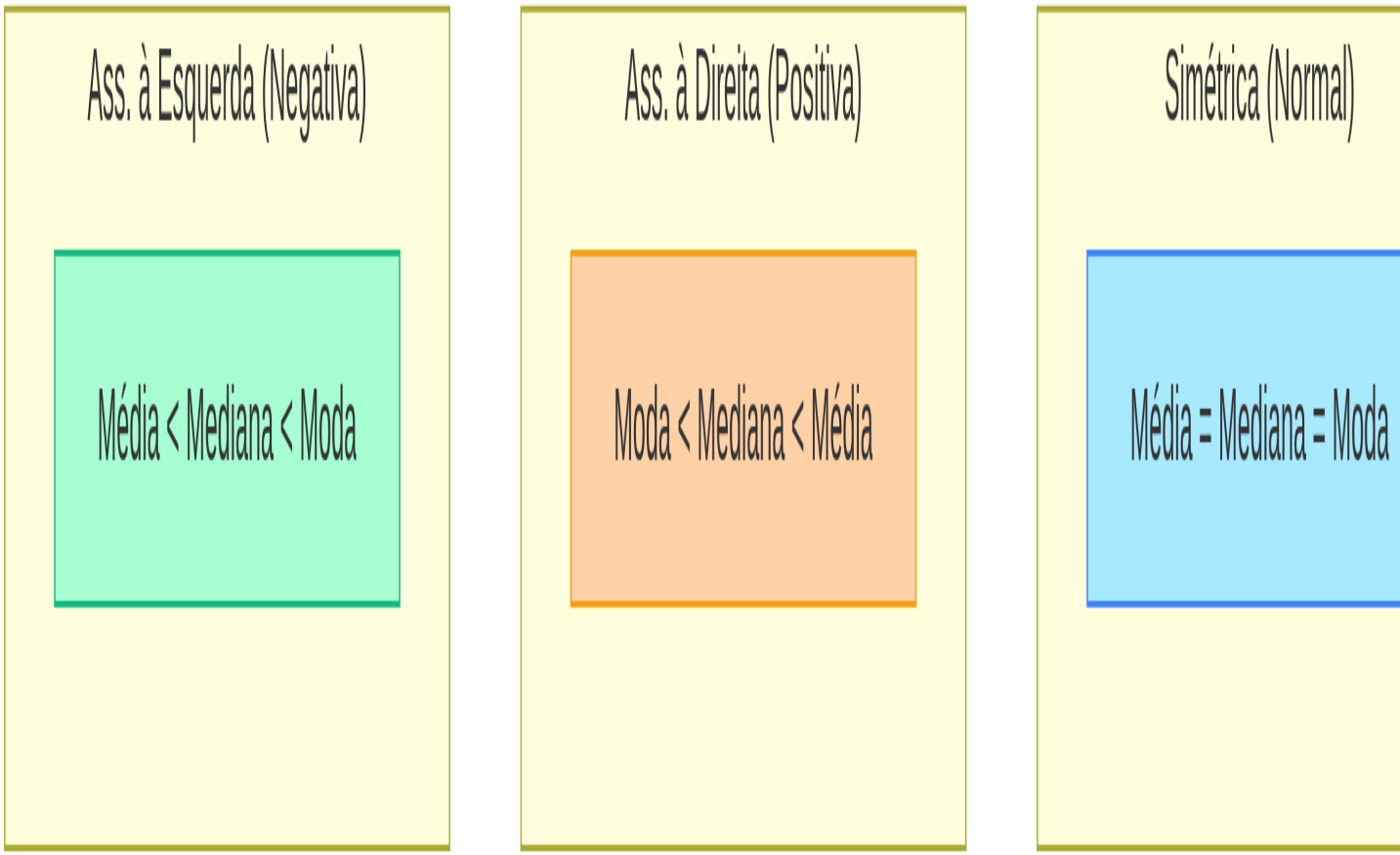

Figure 1: Diagrama ilustrando a posição das medidas de tendência central em diferentes distribuições. **Ass.: Assimetria**.

Características:

- Sua unidade de medida é o quadrado da unidade original dos dados, o que dificulta a interpretação direta.
- É fundamental para o cálculo do desvio padrão.

Desvio Padrão (s ou σ)

O **desvio padrão** é a raiz quadrada da variância. É a medida de dispersão mais utilizada, pois retorna a dispersão à unidade de medida original dos dados, tornando-a mais interpretável.

- **Fórmula (Amostra):** $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- **Fórmula (População):** $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

Características:

- Indica a dispersão típica dos valores em relação à média.
- Um desvio padrão baixo indica que os pontos de dados tendem a estar próximos da média; um desvio padrão alto indica que os pontos de dados estão espalhados por uma ampla gama de valores.
- Também é sensível a outliers.

Código para Medidas de Dispersão

Python

```python
import numpy as np
import pandas as pd
from scipy import stats

data = [10, 12, 12, 15, 18, 20, 22]

# Usando NumPy
range_np = np.max(data) - np.min(data)
variance_np = np.var(data, ddof=1) # ddof=1 para variância amostral (n-1)
std_dev_np = np.std(data, ddof=1) # ddof=1 para desvio padrão amostral (n-1)
mean_np = np.mean(data) # Recalculando a média para o CV

print(f"Dados: {data}")
print(f"Amplitude (NumPy): {range_np:.2f}")
print(f"Variância Amostral (NumPy): {variance_np:.2f}")
print(f"Desvio Padrão Amostral (NumPy): {std_dev_np:.2f}")

# Usando Pandas Series
s_data = pd.Series(data)
range_pd = s_data.max() - s_data.min()
variance_pd = s_data.var() # Por padrão, Pandas usa ddof=1 para var()
std_dev_pd = s_data.std() # Por padrão, Pandas usa ddof=1 para std()
```

```python
mean_pd = s_data.mean()

print("\n--- Usando Pandas ---")
print(f"Amplitude (Pandas): {range_pd:.2f}")
print(f"Variância Amostral (Pandas): {variance_pd:.2f}")
print(f"Desvio Padrão Amostral (Pandas): {std_dev_pd:.2f}")
```

R

```r
data <- c(10, 12, 12, 15, 18, 20, 22)

# Amplitude
range_r <- max(data) - min(data)

# Variância Amostral (R usa n-1 por padrão)
variance_r <- var(data)

# Desvio Padrão Amostral (R usa n-1 por padrão)
std_dev_r <- sd(data)

# Média (para CV)
mean_r <- mean(data)

cat(sprintf("Dados: %s\n", paste(data, collapse = ", ")))
cat(sprintf("Amplitude (R): %.2f\n", range_r))
cat(sprintf("Variância Amostral (R): %.2f\n", variance_r))
cat(sprintf("Desvio Padrão Amostral (R): %.2f\n", std_dev_r))
```

Relação entre Medidas de Tendência Central e Dispersão

As medidas de tendência central nos dizem onde o centro dos dados está, enquanto as medidas de dispersão nos informam o quão "espalhados" os dados estão em torno desse centro. Juntas, elas fornecem uma imagem mais completa da distribuição de um conjunto de dados.

Um conjunto de dados pode ter a mesma média, mas dispersões muito diferentes.

- Exemplo 1: $\{10, 10, 10, 10, 10\}$ ($\bar{x} = 10, s = 0$) - Sem dispersão.
- Exemplo 2: $\{5, 10, 15, 10, 10\}$ ($\bar{x} = 10, s \approx 3.54$) - Alguma dispersão.
- Exemplo 3: $\{1, 10, 19, 10, 10\}$ ($\bar{x} = 10, s \approx 6.93$) - Maior dispersão.

Verificação de Aprendizagem

Resolva os problemas abaixo, calculando as medidas descritivas solicitadas e interpretando os resultados.

1. **Salários de uma Equipe:**

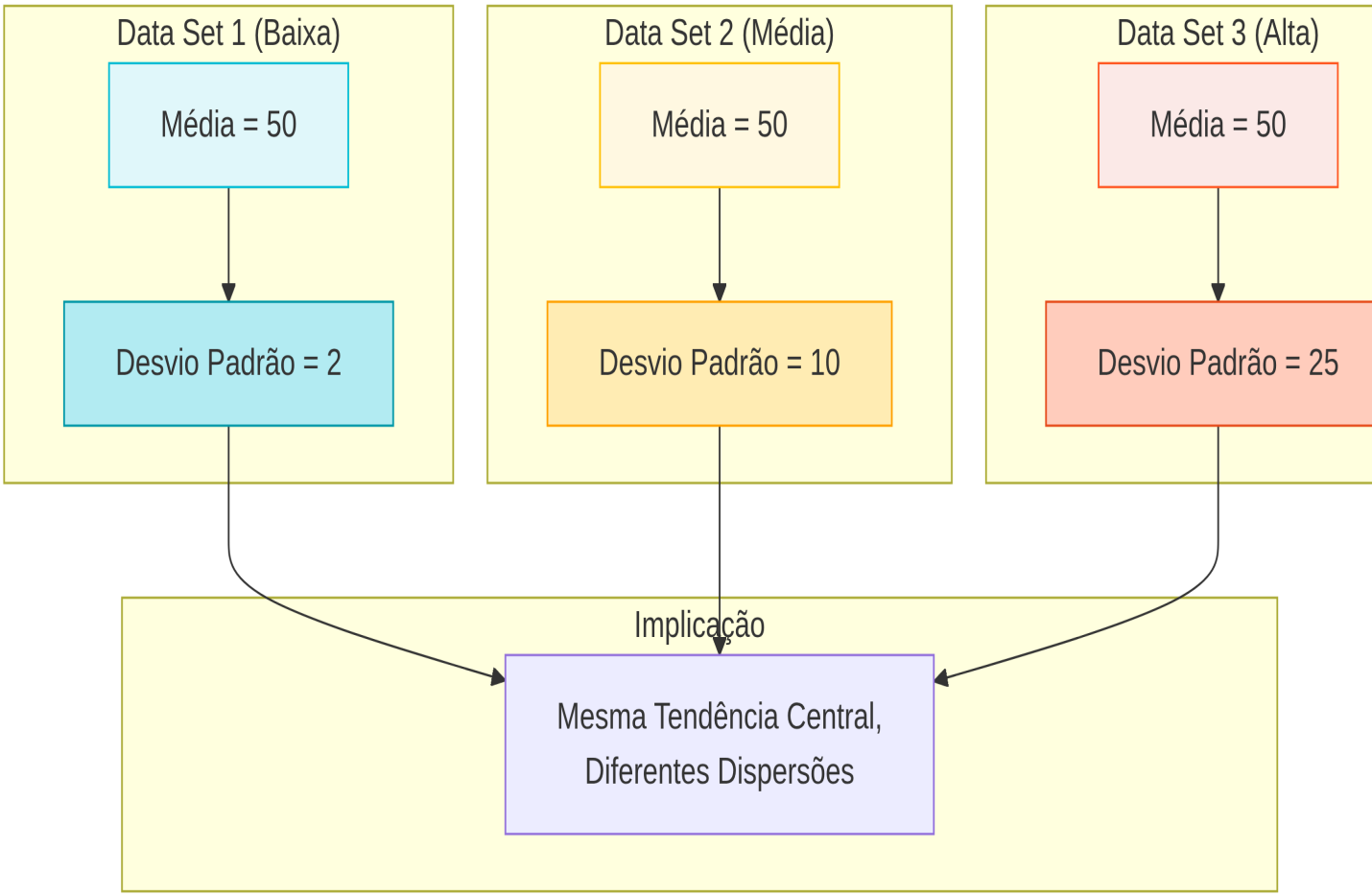

Figure 2: Diagrama ilustrando a relação entre a média e o desvio padrão em distribuições com a mesma média mas dispersões diferentes.

Os salários mensais (em R$) de uma pequena equipe são: `[3000, 3500, 3200, 3800, 4000, 3100, 20000]`.

a) Calcule a média, mediana e moda dos salários.
b) Qual medida de tendência central melhor representa o "salário típico" desta equipe e por quê?
c) Calcule a amplitude, variância amostral e desvio padrão amostral.

2. **Pontuações em um Teste:**

   Um professor aplicou um teste e as pontuações (de 0 a 10) de 10 alunos foram: `[7, 8, 5, 9, 7, 6, 8, 7, 10, 4]`.

   a) Calcule a média, mediana e moda das pontuações.
   b) Calcule a amplitude e o desvio padrão amostral.
   c) Se um aluno tivesse tirado 0 em vez de 4, como isso afetaria a média e a mediana? E o desvio padrão?

3. **Tempo de Atendimento:**

   Os tempos de atendimento (em minutos) em um call center em uma hora foram: `[2.5, 3.1, 2.8, 3.5, 2.9, 3.0, 2.7, 3.2, 2.8, 3.0]`.

   a) Calcule a média e o desvio padrão amostral dos tempos de atendimento.
   b) Se o objetivo é que o tempo médio de atendimento seja no máximo 3.0 minutos com pouca variação, o que os resultados indicam?

4. **Cores Favoritas (Dados Categóricos):**

   Em uma pesquisa, as cores favoritas de 15 pessoas foram: `[Azul, Verde, Azul, Vermelho, Amarelo, Azul, Verde, Azul, Vermelho, Azul, Amarelo, Verde, Azul, Vermelho, Azul]`.

   a) Qual a medida de tendência central mais apropriada para estes dados? Calcule-a.
   b) Faz sentido calcular o desvio padrão ou a média para estes dados? Justifique.

Referências Bibliográficas

BUSSAB, Luiz O. de M.; MORETTIN, Pedro A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.