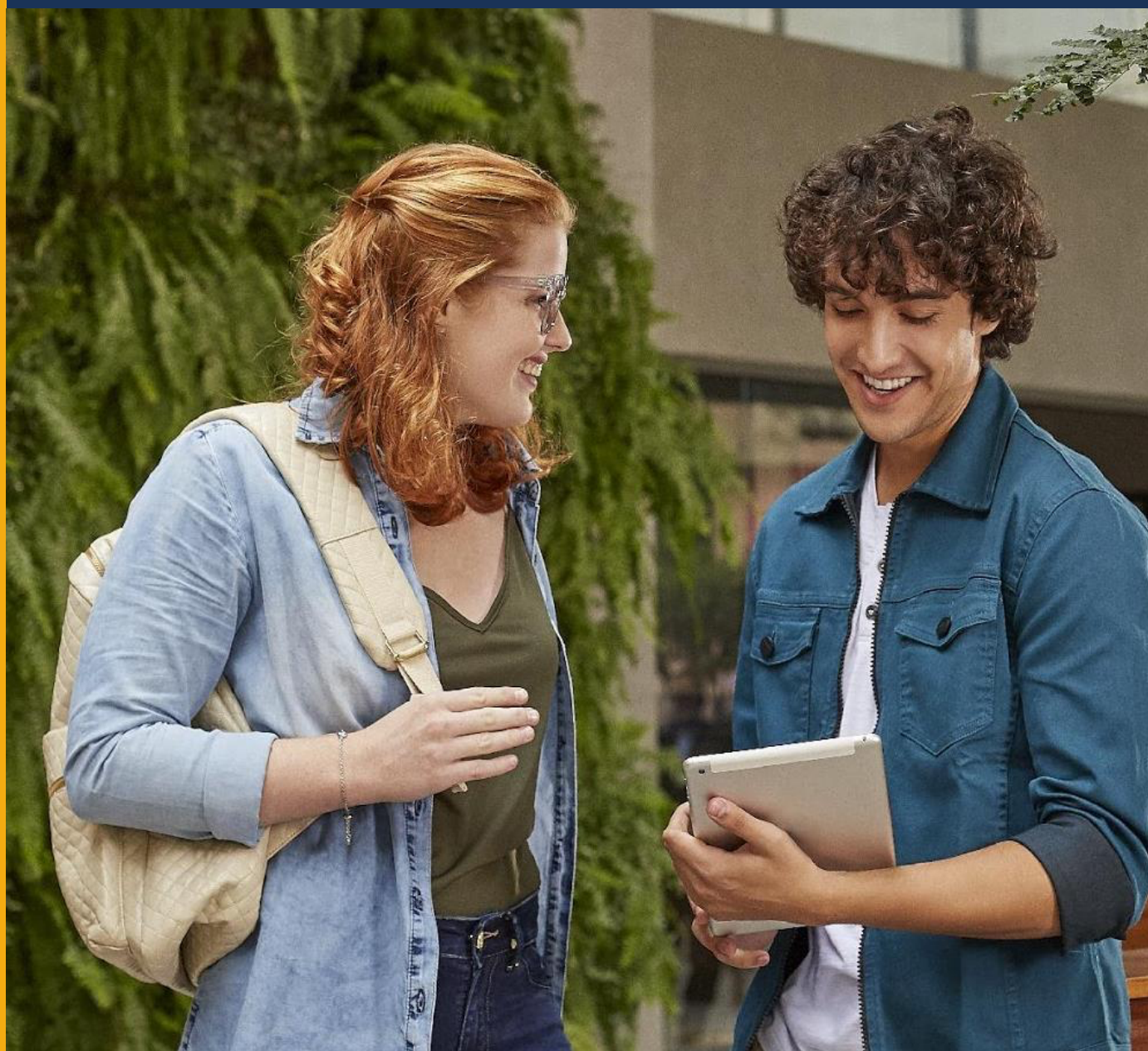
CENTRO UNIVERSITÁRIO IBMEC

CAMPUS BARRA

MBA EM IA, DATA SCIENCE E BIG DATA PARA NEGÓCIOS

ESTATÍSTICA COM R

AULA 2



Ibmec

Sumário

Ibmec

1. Estatística Descritiva

A estatística descritiva é uma área da estatística que se concentra em organizar, resumir e interpretar conjuntos de dados por meio de medidas numéricas e representações gráficas. Seu objetivo é fornecer uma visão geral dos dados, destacando suas principais características sem realizar inferências ou previsões. As ferramentas mais comuns da estatística descritiva incluem medidas de tendência central (como média, mediana e moda), medidas de dispersão (como desvio padrão e variância), medidas de posição relativa (como quartis e percentis) e gráficos (como histogramas e boxplots), que ajudam a visualizar a distribuição e a variabilidade dos dados.

Para ilustrar os conceitos de estatística descritiva utilizaremos o dataset "íris", que já vem junto com o R. O dataset iris é um dos conjuntos de dados mais usados em análises estatísticas e *machine learning*. Ele contém 150 observações de três espécies de flores (Setosa, Versicolor e Virginica), com medições de quatro variáveis numéricas: comprimento da sépala (Sepal.Length), largura da sépala (Sepal.Width), comprimento da pétala (Petal.Length) e largura da pétala (Petal.Width). A seguir, alguns comandos básicos para trabalhar estatística descritiva com esse dataset no R.

```
# Carregar o dataset iris
data(iris)

# Visualizar as primeiras linhas do dataset
head(iris)

# Resumo estatístico básico das variáveis
summary(iris)
```

2. Medidas de Tendência Central

As medidas de tendência central são estatísticas que indicam o ponto ao redor do qual os dados de um conjunto se distribuem. Elas são utilizadas para resumir os dados em um único valor que representa o "centro" ou a "tendência" dos dados. As três principais medidas de tendência central são média, mediana e moda.

2.1. Média

A **média aritmética** é a soma de todos os valores de um conjunto de dados dividida pelo número total de observações. A média é uma medida muito utilizada, mas pode ser influenciada por valores extremos (*outliers*).

Fórmula:

$$\text{Média} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Onde:

- x_i são os valores do conjunto de dados,
- n é o número total de observações.

Exemplo:

Imagine que queremos calcular a média das notas de 5 alunos em uma prova: 7, 8, 9, 6 e 10.

1. Soma das notas: 7+8+9+6+10=40
2. Dividindo pelo número de observações (n=5):

$$\text{Média} = \frac{40}{5} = 8$$

Portanto, a média das notas é **8**.

Exemplo prático em R

```
# Média do comprimento da sépala (Sepal.Length)
mean(iris$Sepal.Length)

# Média de todas as variáveis numéricas
colMeans(iris[, 1:4])
```

2.2. Mediana

A **mediana** é o valor que divide um conjunto de dados ordenados ao meio, de modo que metade dos valores seja menor e a outra metade seja maior que a mediana. A mediana é uma medida robusta, especialmente útil quando há *outliers*, pois não é influenciada por valores extremos.

Como calcular:

- Se o número de observações for ímpar, a mediana é o valor central.
- Se o número de observações for par, a mediana é a média dos dois valores centrais.

Exemplo 1 (Número Ímpar de Observações):

Considere o conjunto de dados com as idades de 7 pessoas: 32, 34, 26, 28, 30, 22, 24.
Ordenando os valores:
22, 24, 26, 28, 30, 32, 34
A mediana é o valor central, neste caso, **28**.

Exemplo 2 (Número Par de Observações):

Agora considere o conjunto de dados com as idades de 6 pessoas: 32, 24, 26, 28, 30, 22.
Ordenando os valores:
22, 24, 26, 28, 30, 32
A mediana será a média dos dois valores centrais: 26 e 28.

$$\text{Mediana} = \frac{26 + 28}{2} = 27$$

Portanto, a mediana das idades é **27**.

Exemplo prático em R

```
# Mediana do comprimento da sépala
median(iris$Sepal.Length)

# Mediana de todas as variáveis numéricas
apply(iris[, 1:4], 2, median)
```

Interpretação

A mediana oferece uma medida mais robusta para dados assimétricos ou com *outliers*. Comparando a média e a mediana, podemos avaliar se os dados estão distribuídos de maneira simétrica ou assimétrica.

2.3. Moda

A **moda** é o valor que ocorre com mais frequência em um conjunto de dados. Enquanto a média e a mediana sempre têm um único valor, a moda pode ter mais de um valor (se houver múltiplos valores com a mesma frequência) ou pode não existir, se todos os valores ocorrerem apenas uma vez.

Exemplo 1 (Moda com um Valor):

Considere o conjunto de dados: 3, 5, 5, 7, 9. Aqui, o valor que aparece com mais frequência é 5, portanto, a moda é 5.

Exemplo 2 (Moda com Dois Valores - Bimodal)

Considere o conjunto de dados: 2, 4, 4, 6, 6, 8. Neste caso, os valores 4 e 6 aparecem duas vezes, e ambos são a moda. Portanto, este conjunto é bimodal, com modas 4 e 6.

Exemplo 3 (Sem Moda)

Considere o conjunto de dados: 1, 2, 3, 4, 5. Todos os valores aparecem apenas uma vez. Então, não há moda.

Exemplo prático em R

O R não tem uma função nativa para calcular a moda. No entanto, podemos implementar uma função que calcule a moda.

```
# Função para calcular a moda
mode_func <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Moda do comprimento da sépala
mode_func(iris$Sepal.Length)
```

3. Medidas de Dispersão

As **medidas de dispersão** são estatísticas que indicam o grau de variabilidade ou espalhamento dos dados em um conjunto em torno de uma medida central, como a média. Elas ajudam a entender o quão concentrados ou dispersos os dados estão. As principais medidas de dispersão incluem a **amplitude** (diferença entre o maior e o menor valor), a **variância** (média dos quadrados das diferenças em relação à média), o **desvio padrão** (raiz quadrada da variância, indicando a dispersão em termos das unidades originais) e o **coeficiente de variação** (desvio padrão como uma porcentagem da média). Essas medidas são fundamentais para complementar as análises de tendência central, fornecendo uma visão mais completa dos dados.

3.1. Amplitude

A **amplitude** é a medida mais simples de dispersão. Ela representa a diferença entre o maior e o menor valor em um conjunto de dados.

Fórmula:

$$\text{Amplitude} = \text{Maior Valor} - \text{Menor Valor}$$

Exemplo:

Considere o conjunto de notas de 5 alunos: 6, 7, 8, 9, 10.

- Maior nota: 10
- Menor nota: 6
- Amplitude = 10 - 6 = 4

Interpretação:

A amplitude indica que a variação das notas é de 4 pontos. No entanto, a amplitude não leva em conta a distribuição dos outros valores, apenas os extremos.

Exemplo prático em R

```
# Amplitude do comprimento da sépala
range(iris$Sepal.Length)
diff(range(iris$Sepal.Length))
```

Interpretação:

A amplitude nos mostra a diferença entre os valores extremos das medições, indicando a extensão dos dados.

3.2. Variância

A **variância** mede o grau de dispersão dos dados em relação à média. Ela calcula a média dos quadrados das diferenças entre cada valor e a média do conjunto de dados. Por utilizar o quadrado das diferenças, a variância dá mais peso aos valores que estão mais afastados da média, o que ajuda a identificar a variabilidade dos dados.

Fórmula:

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

Onde:

- x_i são os valores individuais,
- $\bar{x}$ é a média do conjunto,
- n é o número de observações.

Exemplo:

Considere o conjunto de notas: 6, 7, 8, 9, 10. A média das notas é 8.

1. Diferenças em relação à média:
    - (6 - 8) = -2
    - (7 - 8) = -1
    - (8 - 8) = 0

- (9 - 8) = 1
- (10 - 8) = 2

2. Quadrado das diferenças:
   - (-2)^2 = 4
   - (-1)^2 = 1
   - (0)^2 = 0
   - (1)^2 = 1
   - (2)^2 = 4
3. Somando os quadrados das diferenças:
   - 4 + 1 + 0 + 1 + 4 = 10
4. Dividindo pelo número de observações (5):

$$\text{Variância} = \frac{10}{5} = 2$$

Interpretação:

A variância de 2 indica que, em média, as notas variam 2 unidades quadradas em relação à média. Como os valores estão ao quadrado, a variância não está nas mesmas unidades dos dados originais (neste caso, notas).

Exemplo prático no R

```
# Variância do comprimento da sépala
var(iris$Sepal.Length)

# Variância de todas as variáveis numéricas
apply(iris[, 1:4], 2, var)
```

Interpretação:

A variância e o desvio padrão medem a dispersão dos dados em torno da média. Quanto maiores esses valores, mais espalhados os dados estão.

3.3. Desvio Padrão

Essas medidas ajudam a entender a variabilidade dos dados. O **desvio padrão** é a raiz quadrada da variância e é uma das medidas de dispersão mais utilizadas. Ao extrair a raiz quadrada, ele retorna os valores para as mesmas unidades dos dados originais, tornando sua interpretação mais intuitiva.

Fórmula:

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

Ou seja, o desvio padrão é a raiz quadrada da variância.

Exemplo:

No exemplo anterior, a variância foi 2, então o desvio padrão é:

$$\text{Desvio Padrão} = \sqrt{2} \approx 1,41$$

Interpretação:

O desvio padrão de 1,41 significa que as notas variam, em média, cerca de 1,41 unidades em relação à média de 8. Ele é uma medida mais intuitiva que a variância, pois está na mesma unidade que os dados originais (neste caso, notas).

Exemplo prático no R

```
# Desvio padrão do comprimento da sépala
sd(iris$Sepal.Length)

# Desvio padrão de todas as variáveis numéricas
apply(iris[, 1:4], 2, sd)
```

4. Medidas de Posição Relativa

As **medidas de posição relativa** são estatísticas que indicam a localização ou posição de um valor dentro de um conjunto de dados, em relação aos outros valores. Elas permitem identificar onde um dado específico se situa na distribuição e são úteis para comparações entre diferentes conjuntos de dados. As principais medidas de posição relativa incluem os **quartis** (que dividem os dados em quatro partes iguais), os **percentis** (que dividem os dados em 100 partes) e os **decis** (que dividem os dados em 10 partes). Essas medidas são frequentemente utilizadas para identificar valores extremos (*outliers*) e para avaliar a distribuição de dados em relação a uma média ou mediana.

4.1. Quartis e Percentis

Os quartis e percentis são usados para dividir os dados em porções. O segundo quartil (Q2) é a mediana.

```
# Quartis do comprimento da sépala
quantile(iris$Sepal.Length)

# Percentil 90 do comprimento da sépala
quantile(iris$Sepal.Length, probs = 0.9)
```

Interpretação:

Os quartis são pontos de divisão que segmentam os dados em 25%, 50% (mediana), 75% e 100%. Esses valores ajudam a entender a distribuição e concentração dos dados.

4.2. Boxplot

Um boxplot é uma representação gráfica que exibe os quartis e possíveis *outliers*.

```
# Boxplot para o comprimento da sépala por espécie
boxplot(Sepal.Length ~ Species, data = iris,
        main = "Boxplot do Comprimento da Sépala por Espécie",
        xlab = "Espécie", ylab = "Comprimento da Sépala")
```

Interpretação:

O boxplot é útil para visualizar a mediana, os quartis e identificar possíveis *outliers*. No exemplo, podemos comparar o comprimento da sépala entre as três espécies de flores.

5. Visualização Gráfica dos Dados

A **visualização gráfica dos dados** é a representação visual de informações e dados por meio de gráficos, permitindo uma compreensão rápida e intuitiva das distribuições, tendências e padrões presentes em um conjunto de dados. Gráficos como histogramas, boxplots, gráficos de barras, gráficos de dispersão e gráficos de linhas são comumente utilizados para facilitar a interpretação dos dados. Essa abordagem transforma números e estatísticas em imagens compreensíveis, auxiliando na análise e tomada de decisões, ao mesmo tempo que destaca relações, variações e possíveis outliers de forma mais acessível do que apenas com números.

5.1. Histograma

O histograma mostra a distribuição de frequências dos dados.

```
# Histograma do comprimento da sépala
hist(iris$Sepal.Length,
     main = "Histograma do Comprimento da Sépala",
     xlab = "Comprimento da Sépala",
     col = "lightblue",
     breaks = 10)
```

Interpretação:

O histograma permite visualizar a distribuição dos dados e identificar a forma geral (simétrica, assimétrica, etc.).

5.2. Gráfico de Dispersão

Um gráfico de dispersão mostra a relação entre duas variáveis numéricas.

```
# Gráfico de dispersão entre comprimento da sépala e comprimento da pétala
plot(iris$Sepal.Length, iris$Petal.Length,
     main = "Relação entre Comprimento da Sépala e Pétala",
     xlab = "Comprimento da Sépala",
     ylab = "Comprimento da Pétala",
     col = iris$Species, pch = 19)
```

Interpretação:

O gráfico de dispersão permite identificar padrões e correlações entre duas variáveis. No exemplo, podemos ver como o comprimento da sépala e o comprimento da pétala se relacionam para diferentes espécies.