Tópico 2: Estatística Descritiva com R

MRS

1. Estatística Descritiva com R

- A estatística descritiva é o conjunto de técnicas que tem como objetivo **resumir, organizar e apresentar** os dados de forma clara e informativa.
- É o primeiro passo em qualquer análise de dados.
- Utiliza medidas de tendência central e medidas de dispersão.
- A comunicação de resultados é feita através de gráficos.

2. Medidas de Tendência Central

Essas medidas indicam um valor central em torno do qual os dados se distribuem.

- **Média**: soma dos valores dividida pelo número de observações.
- **Mediana**: valor central dos dados ordenados.
- **Moda**: valor mais frequente (requer o pacote `modeest` em R).

Exemplo

```
dados = c(12, 15, 14, 10, 18, 14, 17, 12, 12)

mean(dados)      # média
```

```
[1] 13.77778
```

```
median(dados)    # mediana
```

```
[1] 14
```

```r
# moda
# install.packages("modeest")
library(modeest)
mfv(dados)       # moda (most frequent value)
```

```
[1] 12
```

Interpretação

```r
# Dados
dados = c(12, 15, 14, 10, 18, 14, 17, 12, 12)

# Medidas
media = mean(dados)
mediana = median(dados)

# Instalar e carregar pacote para moda
if (!require(modeest)) install.packages("modeest", quiet = TRUE)
library(modeest)
moda = mfv(dados)

# Tabela com interpretações
tabela = data.frame(
  Medida = c("Média", "Mediana", "Moda"),
  Valor = c(
    round(media, 2),
    mediana,
    paste(moda, collapse = ", ")
  ),
  Interpretação = c(
    paste("É o valor médio dos dados. Se todos os dados fossem iguais, teriam valor", round(
    paste("É o valor central dos dados ordenados. 50% dos valores estão abaixo e 50% acima d
    paste("É o(s) valor(es) que mais se repetem no conjunto. Aqui:", paste(moda, collapse = "
  )
)

# Mostrar tabela formatada
knitr::kable(tabela, caption = "Medidas de Tendência Central", align = "l")
```

Table 1: Medidas de Tendência Central

Medida	Valor	Interpretação
Média	13.78	É o valor médio dos dados. Se todos os dados fossem iguais, teriam valor 13.78 .
Mediana	14	É o valor central dos dados ordenados. 50% dos valores estão abaixo e 50% acima de 14 .
Moda	12	É o(s) valor(es) que mais se repetem no conjunto. Aqui: 12 .

Exemplo interpretado

```r
# Dados
dados <- c(12, 15, 14, 10, 18, 14, 17, 12, 12)

# Medidas
media <- mean(dados)
mediana <- median(dados)

# Instalar e carregar pacote para moda
if (!require(modeest)) install.packages("modeest", quiet = TRUE)
library(modeest)
moda <- mfv(dados)
```

```
Table: Medidas de Tendência Central

|Medida  |Valor |Interpretação                                                                                |
|:-------|:-----|:--------------------------------------------------------------------------------------------|
|Média   |13.78 |É o valor médio dos dados. Se todos os dados fossem iguais, teriam valor 13.78 .             |
|Mediana |14    |É o valor central dos dados ordenados. 50% dos valores estão abaixo e 50% acima de 14 .      |
|Moda    |12    |É o(s) valor(es) que mais se repetem no conjunto. Aqui: 12 .                                 |
```

Conclusão da análise

- A média de 13.78 representa o valor médio do conjunto. Mostra a tendência geral dos dados.
- A mediana de 14 indica que 50% dos valores estão abaixo e 50% acima desse ponto, o que reforça uma distribuição equilibrada.
- A moda de 12 revela que esse é o valor mais frequente no conjunto, aparecendo 3 vezes.
- Como média, mediana e moda estão próximas, os dados estão relativamente bem distribuídos, sem grandes assimetrias.
- Há uma leve assimetria à esquerda, pois a moda $<$ média $<$ mediana, mas não é suficiente para indicar uma distorção significativa.

- No geral, os dados apresentam uma distribuição estável e centralizada, com poucas repetições extremas.

3. Medidas de Dispersão

Indicam o grau de variação dos dados em torno da média ou entre si. Ajudam a entender se os valores estão concentrados ou espalhados.

- **Desvio padrão (sd)**:
  - Medida de dispersão mais comum.
  - Mede o quão distantes, em média, os dados estão da média.
  - Quanto maior, mais dispersos os dados.
- **Variância (var)**:
  - Desvio padrão ao quadrado.
  - Mede igualmente a dispersão, mas em unidades ao quadrado.
  - É útil para cálculos estatísticos, mas menos intuitiva.
- **Amplitude (diff(range())):**:
  - Diferença entre o maior e o menor valor.
  - Simples, mas sensível a outliers.
- **Intervalo interquartil (IQR)**:
  - Diferença entre o 3º e 1º quartil.
  - Mostra a dispersão dos 50% centrais dos dados.
  - É resistente a outliers.
- **Limites (range):**
  - Limites do intervalo de dados.

```r
dados = c(12, 15, 14, 10, 18, 14, 17, 12, 12)

sd(dados)           # desvio padrão
```

```
[1] 2.587362
```

```r
var(dados)           # variância
```

```
[1] 6.694444
```

```r
diff(range(dados))  # amplitude
```

```
[1] 8
```

```r
IQR(dados)           # intervalo interquartil
```

```
[1] 3
```

```r
range(dados)         # menor e maior valor
```

```
[1] 10 18
```

Interpretação

```r
# Dados
dados = c(12, 15, 14, 10, 18, 14, 17, 12, 12)

# Medidas
dp = round(sd(dados), 2)
variancia = round(var(dados), 2)
amplitude = diff(range(dados))
iqr = IQR(dados)
valores = range(dados)

dp
```

```
[1] 2.59
```

```r
variancia
```

```
[1] 6.69
```

```r
amplitude
```

```
[1] 8
```

```r
iqr
```

```
[1] 3
```

```r
valores
```

```
[1] 10 18
```

```r
# Tabela com interpretações
tabela = data.frame(
  Medida = c(
    "Desvio padrão",
    "Variância",
    "Amplitude",
    "Intervalo interquartil (IQR)",
    "Range (menor, maior)"
  ),
  Valor = c(
    dp,
    variancia,
    amplitude,
    iqr,
    paste(valores[1], valores[2], sep = ", ")
  ),
  interpretacao = c(
    paste("Em média, os valores estão a", dp, "unidades da média."),
    "É o desvio padrão ao quadrado. Serve para cálculos estatísticos.",
    paste("A diferença entre o maior (", valores[2], ") e o menor (", valores[1], ") valor."
    paste("Os 50% centrais dos dados estão distribuídos dentro de um intervalo de", iqr, ".")
    "Indica os limites do intervalo de dados."
  )
)

# Mostrar tabela formatada
knitr::kable(tabela, caption = "Medidas de dispersão", align = "l")
```

Table 2: Medidas de dispersão

Medida	Valor	interpretacao
Desvio padrão	2.59	Em média, os valores estão a 2.59 unidades da média.
Variância	6.69	É o desvio padrão ao quadrado. Serve para cálculos estatísticos.
Amplitude	8	A diferença entre o maior (18) e o menor (10) valor.
Intervalo interquartil (IQR)	3	Os 50% centrais dos dados estão distribuídos dentro de um intervalo de 3 .
Range (menor, maior)	10, 18	Indica os limites do intervalo de dados.

Exemplo interpretado

```
# Dados
dados <- c(12, 15, 14, 10, 18, 14, 17, 12, 12)

# Medidas
dp <- round(sd(dados), 2)
variancia <- round(var(dados), 2)
amplitude <- diff(valores)
iqr <- IQR(dados)
valores <- range(dados)
```

```
Table: Medidas de dispersão

|Medida                      |Valor  |Interpretação                                                                |
|:---------------------------|:------|:----------------------------------------------------------------------------|
|Desvio padrão               |2.59   |Em média, os valores estão a 2.59 unidades da média.                         |
|Variância                   |6.69   |É o desvio padrão ao quadrado. Serve para cálculos estatísticos.             |
|Amplitude                   |8      |A diferença entre o maior (18) e o menor (10) valor.                         |
|Intervalo interquartil (IQR) |3      |Os 50% centrais dos dados estão distribuídos dentro de um intervalo de 3 .  |
|Range (menor, maior)        |10, 18 |Indica os limites do intervalo de dados.                                     |
```

Conclusão da análise

- O desvio padrão de 2.59 indica que a variação média dos dados em torno da média é moderada.
- A variância de 6.69 complementa essa noção, usada mais em fórmulas estatísticas.
- A amplitude de 8 mostra que o conjunto não é muito amplo, mas é sensível a extremos.
- O IQR de 3 revela que a metade central dos dados está bem agrupada, ou seja, não há grande dispersão no “miolo” dos dados.
- Os valores 10 e 18 são os extremos, mas como o IQR ainda é pequeno, os dados não parecem ter outliers significativos.

4. Resumos Estatístico

Funções para gerar rapidamente resumos estatísticos:

- summary(): mínimo, Q1, mediana (Q2), média, Q3, máximo.
- table(): mostra quantas vezes cada categoria aparece.
- prop.table(): calcula a proporção relativa de cada categoria (em % se multiplicar por 100).

Sumário

Para dados numéricos.

```
dados = c(12, 15, 14, 10, 18, 14, 17, 12, 12)
summary(dados)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   12.00   14.00   13.78   15.00   18.00
```

Tabelas de frequência

Para dados categóricos (qualidades, categorias ou rótulos; nominais: cor; ordinais: classificação) ou discretos (números inteiros: nº de formandos).

```
categorias = c("A", "B", "A", "C", "B", "A")
table(categorias)                # frequência absoluta
```

```
categorias
A B C
3 2 1
```

```
prop.table(table(categorias))    # frequência relativa
```

```
categorias
        A         B         C
0.5000000 0.3333333 0.1666667
```

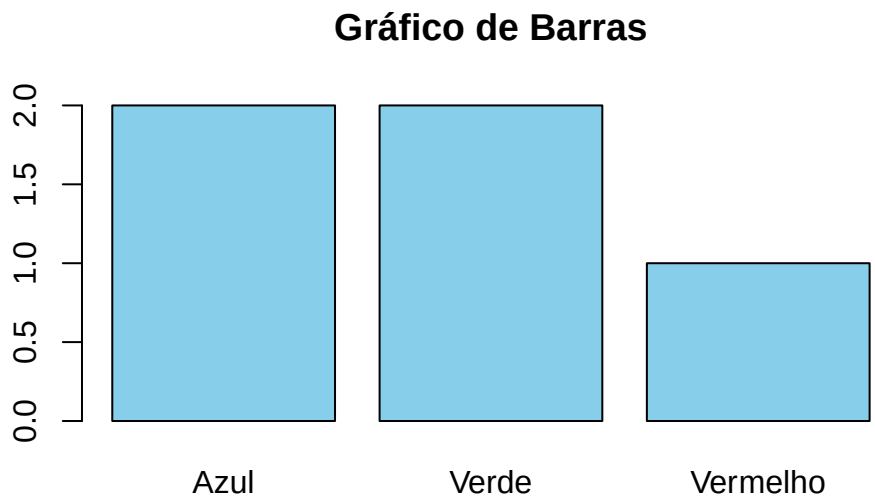
5. Gráficos

Gráficos são ferramentas essenciais para visualização e comunicação dos dados.

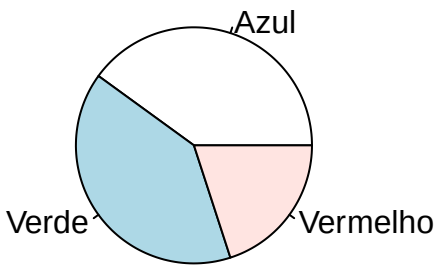
- **Gráfico de barras**
  - Para dados categóricos ou discretos.
  - Mostra frequências de categorias.
  - Mostra a frequência ou contagem de cada categoria.
  - Cada barra representa uma categoria.
- **Gráfico circular**
  - Para representar proporções de categorias (dados categóricos).
  - Mostra partes de um todo em forma de fatias.
  - Melhor para poucos grupos.
- **Gráfico de linhas**
  - Representa evolução ou tendência ao longo do tempo.
  - Conecta pontos com linhas.
  - Ideal para séries temporais ou dados ordenados
- **Histograma**
  - Para dados numéricos contínuos.
  - Mostra a distribuição de frequências por intervalos (bins).
  - Indica a forma da distribuição (simétrica, assimétrica, etc.).
- **Boxplot**
  - Resume a distribuição dos dados numéricos.
  - Mostra mediana, quartis e outliers.
- **Gráfico de dispersão**
  - Para relação entre duas variáveis numéricas.
  - Útil para detectar correlação.

Exemplo

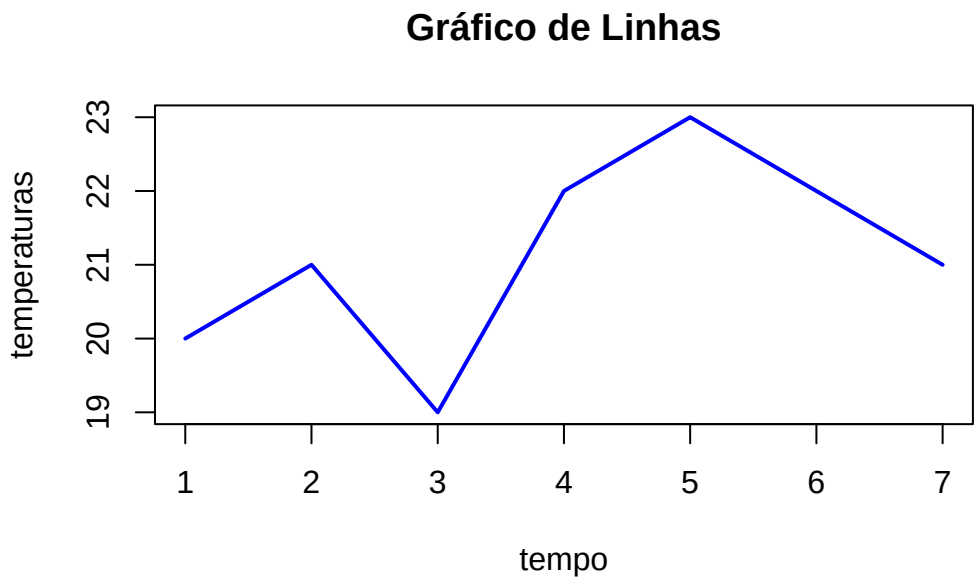
```r
# Gráfico de Barras
dados_categoricos = c("Azul", "Verde", "Azul", "Vermelho", "Verde")
barplot(table(dados_categoricos), main = "Gráfico de Barras", col = "skyblue")
```



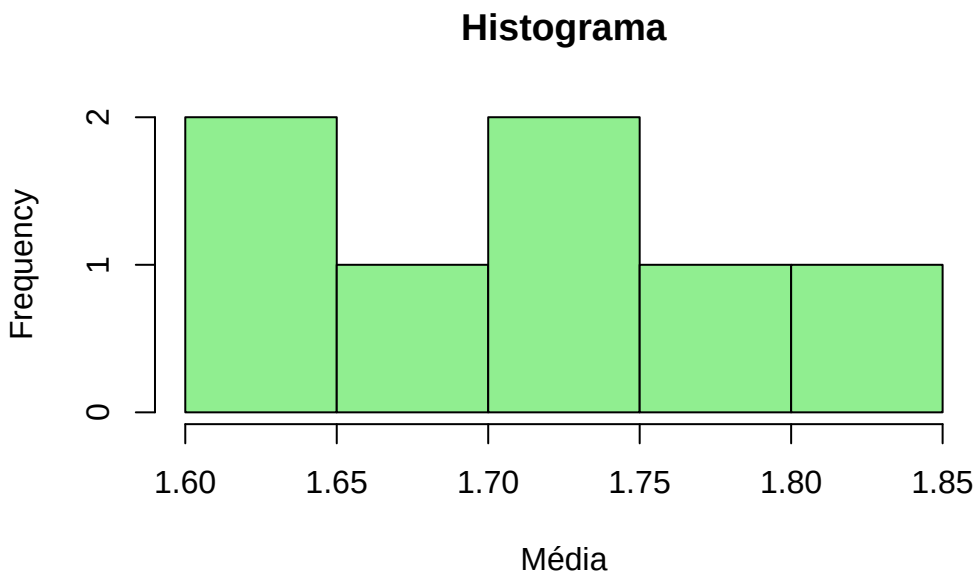

```r
# Gráfico Circular
pie(table(dados_categoricos), main = "Gráfico Circular")
```



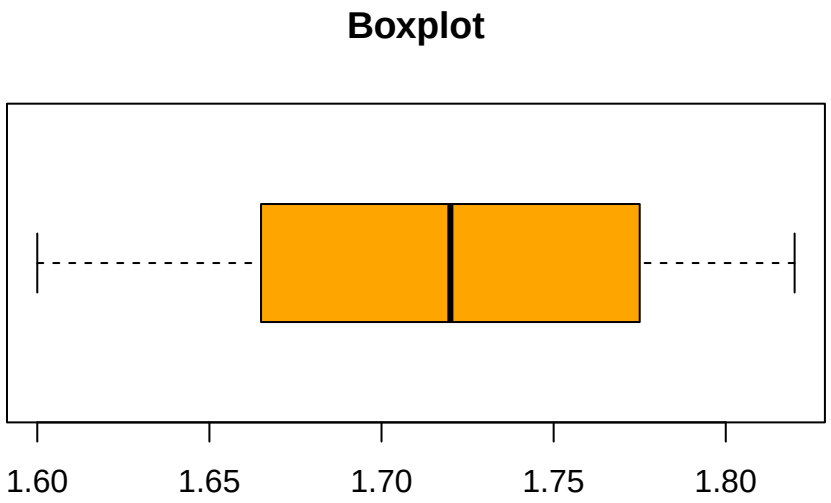
```r
# Gráfico de Linhas
tempo = 1:7
temperaturas = c(20, 21, 19, 22, 23, 22, 21)
plot(tempo, temperaturas, type = "l", main = "Gráfico de Linhas", col = "blue", lwd = 2)
```



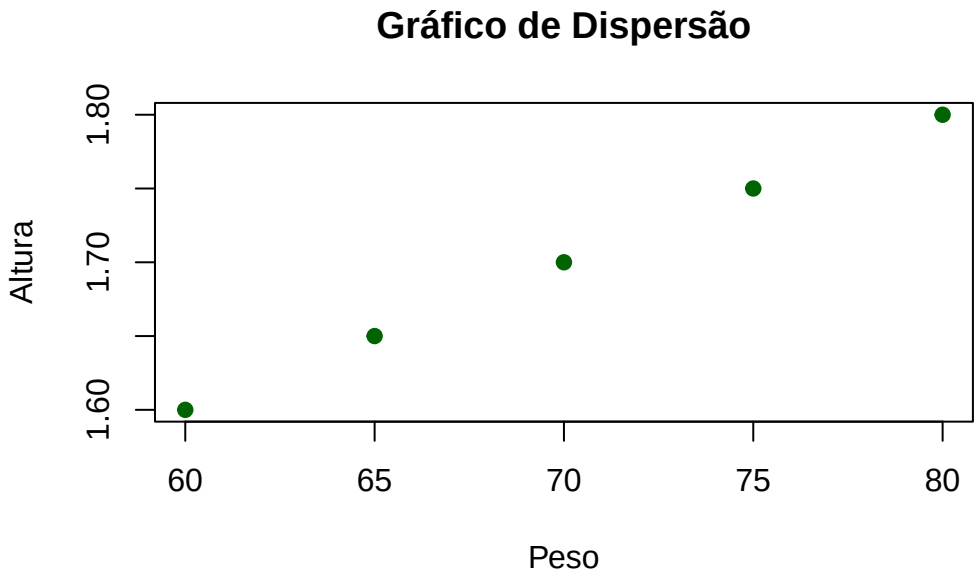

```r
# Histograma
dados_continuos = c(1.72, 1.65, 1.80, 1.75, 1.68, 1.82, 1.60)
hist(dados_continuos, main = "Histograma", xlab = "Média", col = "lightgreen")
```




```
# Boxplot
boxplot(dados_continuos, main = "Boxplot", col = "orange", horizontal = TRUE)
```



```r
# Gráfico de dispersão
peso = c(60, 65, 70, 75, 80)
altura = c(1.60, 1.65, 1.70, 1.75, 1.80)
plot(peso, altura, main = "Gráfico de Dispersão", xlab = "Peso", ylab = "Altura", pch = 19, c
```



Exercício

Utilize o vetor abaixo para calcular e visualizar estatísticas descritivas:

```r
notas = c(5.5, 7.0, 8.5, 6.0, 9.0, 7.5, 6.5, 8.0)
```

1. Calcule a média, mediana, desvio padrão, amplitude e IQR.
2. Verifique a moda (utilize `modeest`).
3. Gere um histograma e um boxplot.

Solução

```r
notas = c(5.5, 7.0, 8.5, 6.0, 9.0, 7.5, 6.5, 8.0)

mean(notas)
```

```
[1] 7.25
```

```
median(notas)
```

```
[1] 7.25
```

```
sd(notas)
```

```
[1] 1.224745
```

```
diff(range(notas))
```

```
[1] 3.5
```

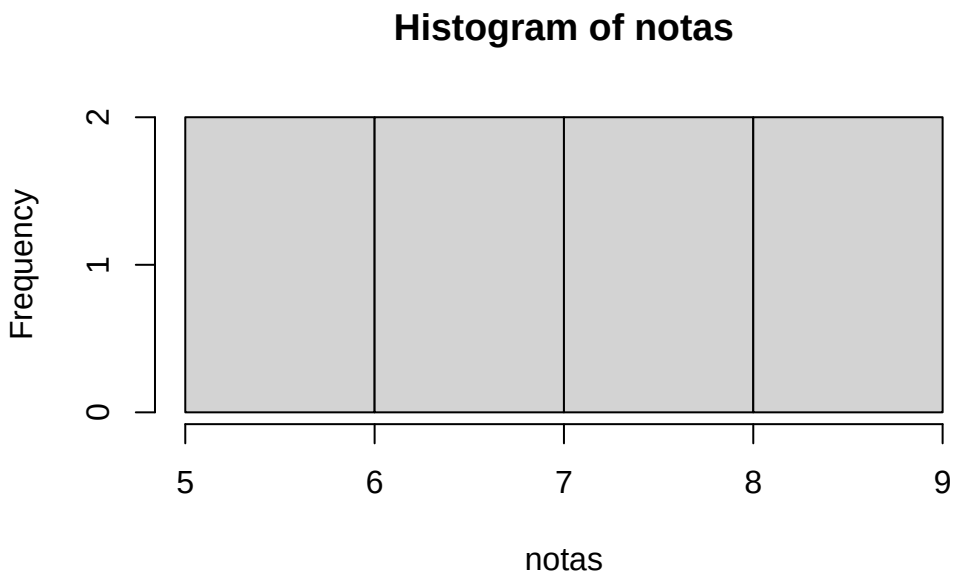
```
IQR(notas)
```

```
[1] 1.75
```

```
# Moda
library(modeest)
mfv(notas)
```

```
[1] 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
```

```
# Gráficos
hist(notas)
```




```
boxplot(notas)
```