MAT02018 - Estatística Descritiva

Medidas de variabilidade

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Estatísticas de ordem

Estatísticas de ordem

- A **média** e o **desvio padrão** são bastante conhecidos e muito usados para descrever um conjunto de dados.
  - No entanto, não são as únicas.
- Também podem ser adotadas as medidas de posição relativa, mais conhecidas como **estatísticas de ordem**.
  - Uma destas medidas é a **mediana**.
- Falaremos também dos **quartis**, **decis** e **percentis**.
  - Os quartis serão utilizados também para calcularmos a **amplitude entre quartis**, que pode ser utilizada como uma medida de dispersão.

Quartis

- **Quartis** são pontos que dividem um **conjunto de dados ordenado** em **quatro partes** iguais (quatro quartos).
- Denotamos os quartis por Q_1, Q_2 e Q_3:
  - Q_1 é o **primeiro quartil**. É o valor que tem 25% dos dados iguais ou menores do que ele.
  - Q_2 é o **segundo quartil**. É o valor que tem 50% dos dados iguais ou menores do que ele[1].
  - Q_3 é o **terceiro quartil**. É o valor que tem 75% dos dados iguais ou menores do que ele.

[1] Note que o segundo quartil é também a mediana de um conjunto de dados.

Quartis

Uma forma de encontrar as posições de Q_1 e Q_3[2] na **série ordenada de dados**, como foi visto para o caso da mediana (Q_2).

- posição de Q_1: $\frac{n+1}{4}$
- posição de Q_3: $\frac{3(n+1)}{4}$

No caso do conjunto de dados conter um número par de elementos, Q_1 e Q_3 são obtidos tomando-se a média dos valores vizinhos as suas posições.

[2] Já aprendemos a encontrar a posição de Q_2, a mediana.

Separatrizes

- Da mesma forma que obtemos valores que dividem o conjunto de dados em quatro, podemos generalizar esta ideia para mais grupos.
- **Decis** dividem o conjunto de dados em dez partes iguais.
  - Para obter os decis, organizamos os dados em ordem crescente e depois dividimos o conjunto em dez subconjuntos com igual número de elementos.
  - Os decis são os valores que separam esses subconjuntos.
- **Percentis** dividem o conjunto de dados em cem partes iguais.
  - Para obter os percentis, organizamos os dados em ordem crescente e depois dividimos o conjunto em cem subconjuntos com igual número de elementos.
  - Os percentis são os valores que separam esses subconjuntos.

Separatrizes

Comentários

- Os quartis, decis e percentis são conhecidos como as **separatrizes**.
  - Se divimos o conjunto de dados ordenado em três partes iguais, temos os **tercis**.
  - Se divimos o conjunto de dados ordenado em cinco partes iguais, temos os **quintis**.
- O cálculo dos decis e percentis é mais adequado quando o tamanho da amostra (ou população) é grande.

Amplitude entre quartis

Amplitude entre quartis

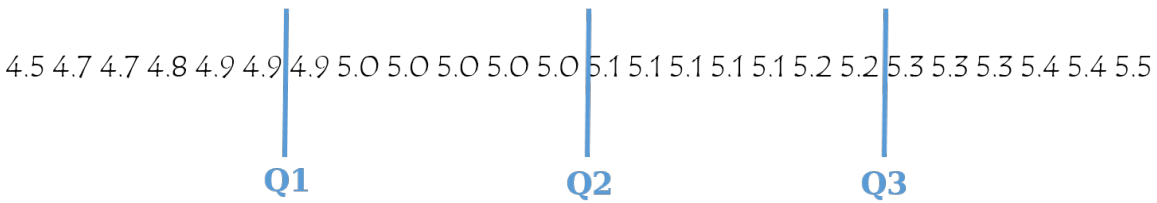
- A **amplitude entre quartis** (ou **distância interquartílica**) é definida como a **diferença entre o terceiro e o primeiro quartil**:

$$DIQ = Q_3 - Q_1.$$

- Note que entre Q_1 e Q_3 estão 50% dos valores mais centrais da distribuição.

Amplitude entre quartis

- **Exemplo:**



- Temos que $Q_1 = 4,9$, $Q_3 = 5,25$, e portanto, a amplitude entre quartis é

$$Q_3 - Q_1 = 5,25 - 4,9 = 0,35.$$

Diagrama de caixa

- Existe uma forma de apresentar graficamente uma distribuição de uma variável contínua através de algumas medidas resumo, tais como: o mínimo, o máximo e os quartis.
- Tal gráfico é conhecido como **diagrama de caixa**[3].

[3] Também chamado pelo nome em inglês, o *boxplot*.

Diagrama de caixa

Para construir o diagrama de caixa, siga os seguintes passos:

1. Desenhe um retângulo (caixa, ou *box*) com comprimento igual a amplitude entre quartis (distância interquartílica).
2. Trace uma linha para representar a mediana, na posição (medida pela distância) que ela ocupa entre o primeiro e o terceiro quartil.
3. Trace linhas perpendiculares à mediana, saindo do meio do retângulo; a linha abaixo da caixa terá comprimento igual à distância entre o primeiro quartil e o valor mínimo; a linha acima da caixa, comprimento igual à distância entre o terceiro quartil e o valor máximo[4].

[4] Existem variações na construção destas linhas do diagrama de caixa, também conhecidas por cerquilhas. Esta forma apresentada não é a mais usual. É comum traçar a linha abaixo da caixa partindo de $Q_1 - 1,5 \times DIQ$ e chegando em Q_1, e a linha acima da caixa partindo de Q_3 e chegando em $Q_3 + 1,5 \times DIQ$. Quando o diagrama de caixas é construído desta forma, alguns autores sugerem que os valores fora destas cerquilhas sejam consideradas discrepantes (*outliers*). É claro que este é apenas um critério prático, e classificar uma observação como discrepante, ou anômala, depende do contexto.

Diagrama de caixa

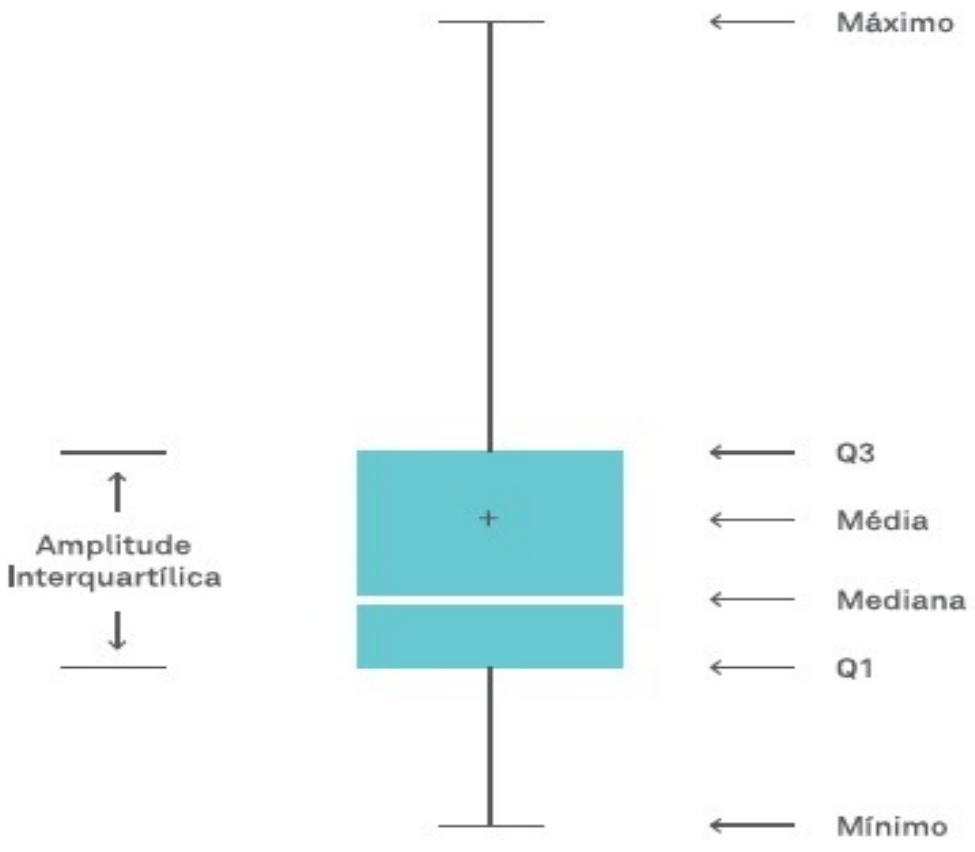


Diagrama de caixa

- O diagrama de caixa é bastante utilizado para comparar a distribuição de uma certa variável em diferentes grupos.

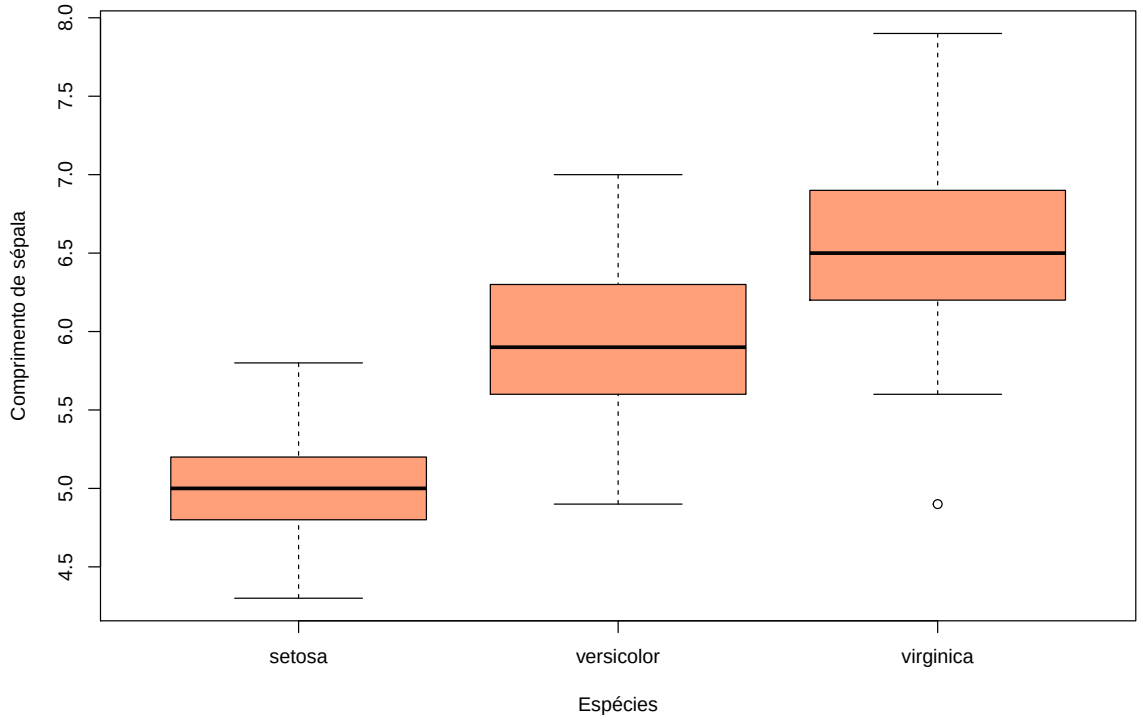
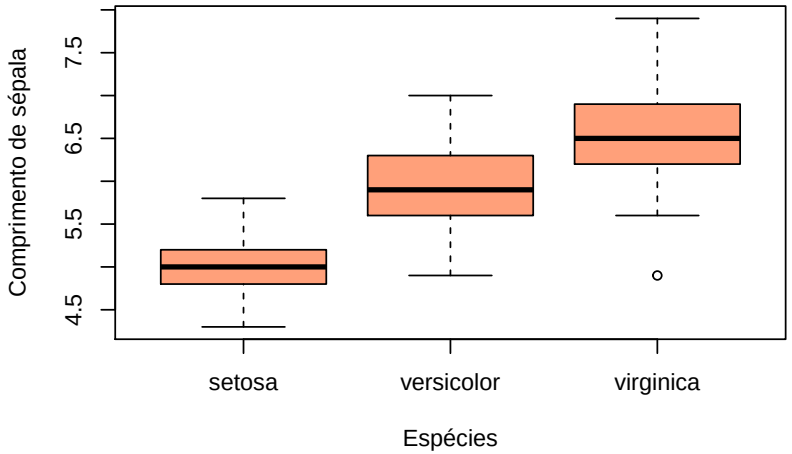
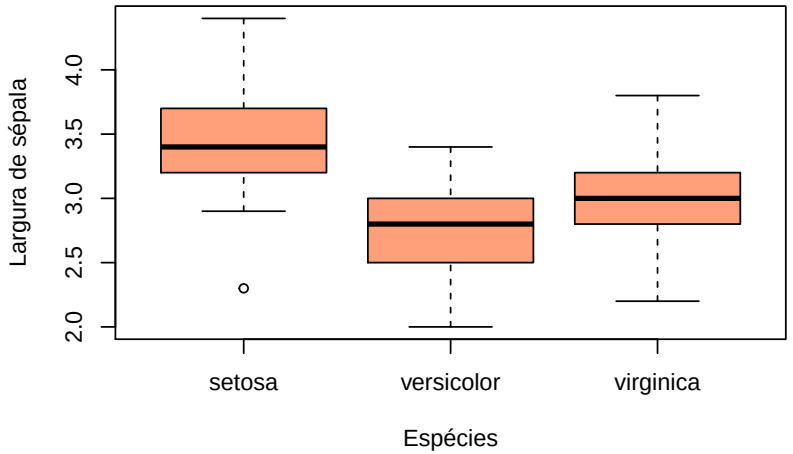


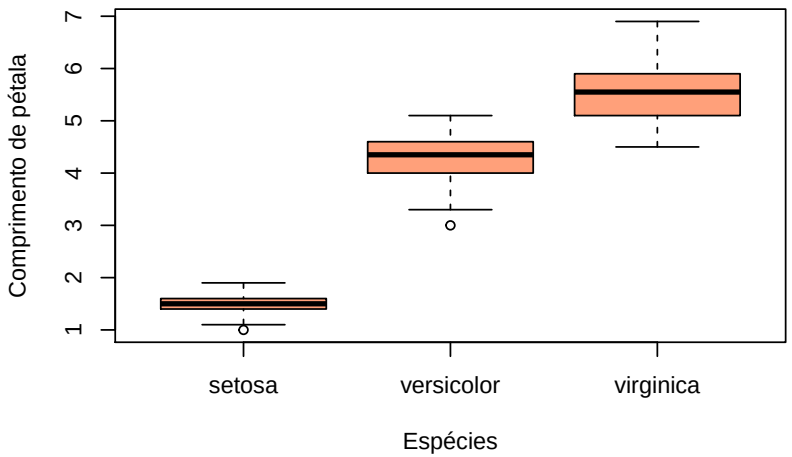
Diagrama de caixa

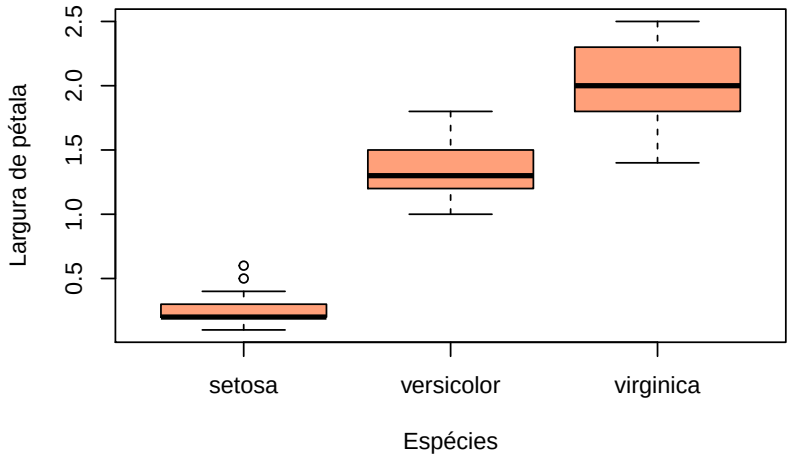
- **Sua vez:** para cada gráfico a seguir (cada gráfico apresenta o diagrama de caixa de uma certa variável para três espécies de íris), aponte o grupo de maior e menor variabilidade.
  - Como você chegou nestas conclusões?

Diagrama de caixa









ComplementaR

ComplementaR

Esta seção é complementar. São apresentadas algumas poucas funções em R relacionadas a discussão da aula. Para tal, vamos utilizar o exemplo original de (BUSSAB; MORETTIN, 2017) sobre os dados dos empregados da seção de orçamentos da Companhia MB. A planilha eletrônica correspondente encontra-se no arquivo `companhia_mb.xlsx`. Vamos começar carregando os dados para o R. Existem várias formas de se carregar **arquivos de dados** em diferentes no R. Como arquivo de interesse encontra-se no formato do Excel (xlsx), vamos utilizar a função `read_excel` do pacote `readxl`[5].

[5] Caso você não tenha o pacote, instale-o: `install.packages("readxl")`.

ComplementaR

```r
# install.packages("readxl")
library(readxl)

dados <- read_excel(path = "companhia_mb.xlsx")
```

```r
class(dados) # classe do objeto dados
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```r
dim(dados) # dimensão do objeto dados
```

```
## [1] 36  7
```

ComplementaR

- Note que o objeto dados é uma tabela de dados bruto.

```
head(dados) # apresenta as primeiras linhas do objeto dados
```

```
## # A tibble: 6 x 7
##       N `Estado Civil` `Grau de Instrução` `N de Filhos` `Salario (x Sa~` Id
##   <dbl> <chr>          <chr>                       <dbl>            <dbl> <d
## 1     1 solteiro       ensino fundamental             NA             4
## 2     2 casado         ensino fundamental              1             4.56
## 3     3 casado         ensino fundamental              2             5.25
## 4     4 solteiro       ensino médio                   NA             5.73
## 5     5 solteiro       ensino fundamental             NA             6.26
## 6     6 casado         ensino fundamental              0             6.66
```

ComplementaR

- As funções `min`, `max` e `range` retornam, respectivamente, o mínimo, o máximo e a variação de um conjunto de dados.

```
min(dados$Idade)
```

```
## [1] 20
```

```
max(dados$Idade)
```

```
## [1] 48
```

```
range(dados$Idade)
```

```
## [1] 20 48
```

ComplementaR

- Utilizando a função `diff` podemos computar a amplitude de variação do conjunto de dados.

```
diff(range(dados$Idade))
```

```
## [1] 28
```

- O argumento `na.rm = TRUE` para remover os dados ausentes do conjunto antes de calcular os extremos.

```
min(dados$`N de Filhos`, na.rm = TRUE)
```

```
## [1] 0
```

```
max(dados$`N de Filhos`, na.rm = TRUE)
```

```
## [1] 5
```

ComplementaR

- A variância de um conjunto de dados pode ser obtida utilizando a função `var`. O desvio padrão pode ser computado de duas formas (pelo menos): através da função `sd`, ou computando a raiz quadrada da variância utilizando a função `sqrt`.

```
var(dados$Idade)
```

```
## [1] 45.39286
```

```
sd(dados$Idade)
```

```
## [1] 6.737422
```

```
sqrt(var(dados$Idade))
```

```
## [1] 6.737422
```

ComplementaR

- O coeficiente de variação pode ser computado dividindo o desvio padrão pela média

```
sd(dados$Idade)/mean(dados$Idade)
```

```
## [1] 0.194817
```

ComplementaR

- Os quartis (e demais separatrizes) podem ser obtidos pela função `quantile` da seguinte forma.

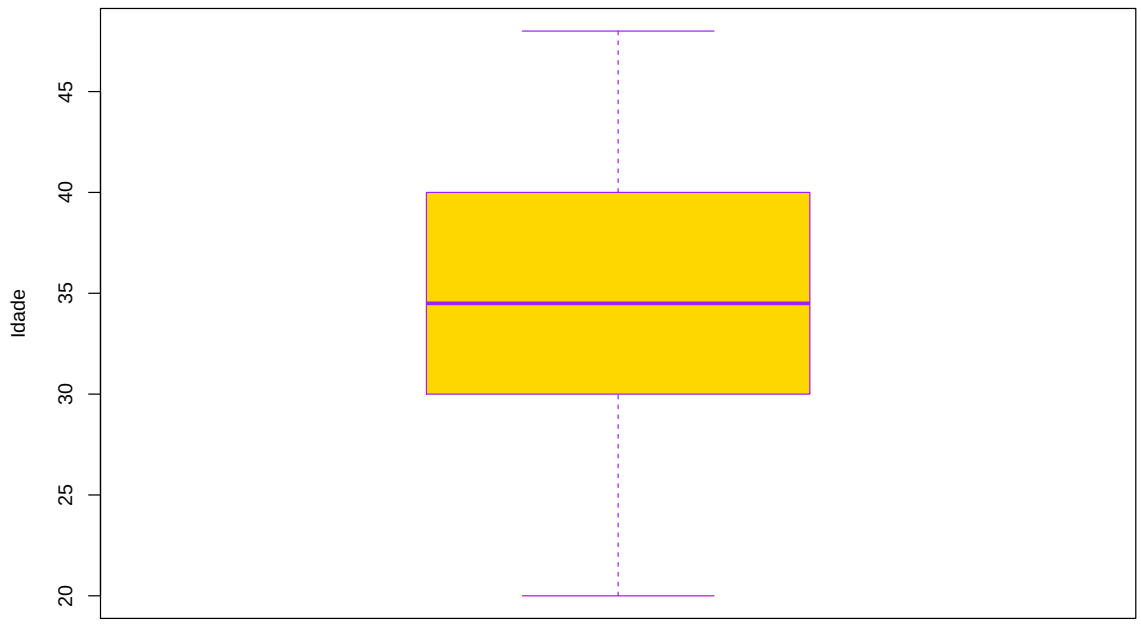
```
quantile(dados$Idade,
         probs = c(0.25, 0.5, 0.75))
```

```
##  25%  50%  75%
## 30.0 34.5 40.0
```

ComplementaR

- O diagrama de caixa (boxplot) pode ser obtido pela função `boxplot`

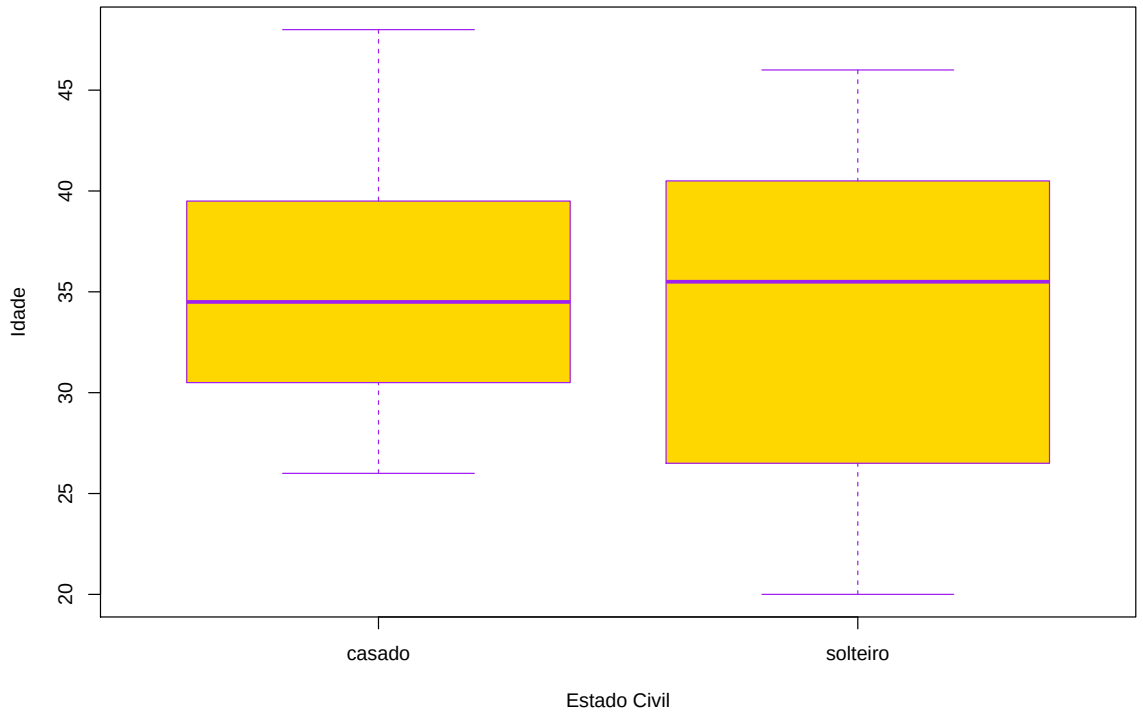
```
boxplot(dados$Idade,
        ylab = "Idade", col = "gold", border = "purple")
```



ComplementaR

- Se quisermos comparar as idades de solteiros e casados utilizando o boxplot, podemos utilizar a seguinte sintaxe:

```
boxplot(Idade ~ `Estado Civil`, data = dados,
        ylab = "Idade", col = "gold", border = "purple")
```



Para casa

1. Resolver os exercícios 5, 9 e 10 do Capítulo 9.7 do livro **Fundamentos de Estatística**[6] (disponível no Sabi+).
2. Para o seu levantamento estatístico, calcule os quartis, a amplitude entre quartis e construa um diagrama de caixa, de acordo com a classificação das variáveis. Compartilhe no Fórum Geral do Moodle.

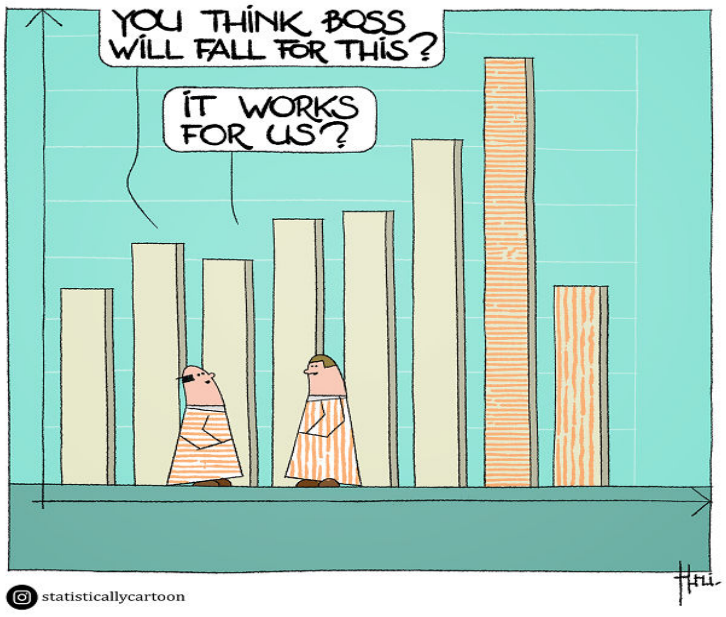
[6] Vieira, S. **Fundamentos de Estatística**, Atlas, 2019, p. 153-154.

Próxima aula

- Medidas de forma;
- Propriedades de média e variância;
- Valores padronizados (z-escores).

Por hoje é só!

Bons estudos!



BUSSAB, W. de O.; MORETTIN, P. A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.