Estatística Computacional

Universidade Federal da Bahia

Gilberto Pereira Sassi

Tópico 2

Pacotes que iremos usar na semana 3

```
library(readxl)
library(readODS)
library(writexl)
library(ggthemes)
library(lvplot)
library(tidyverse)
```

Estatística Descritiva no **R**

gráficos e tabelas

Tabela de Distribuição de Frequências

Tabela de Distribuição de Frequências

- X: variável qualitativa ou variável quantitativa discreta

Tabela de distribuição de Frequências.

X	Frequência	Frequência relativa	Porcentagem
B_1	n_1	$f_1 = \frac{n_1}{n_1+\cdots+n_k}$	$p_1 = f_1 \cdot 100$
B_2	n_2	$f_2 = \frac{n_2}{n_1+\cdots+n_k}$	$p_2 = f_2 \cdot 100$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_k	n_k	$f_k = \frac{n_k}{n_1+\cdots+n_k}$	$p_k = f_k \cdot 100$

Em que os valores possíveis de X são $B_1, \ldots, B_k$, n_i é a frequência da categoria B_i, $i = 1, \ldots, k$, f_i é a frequência relativa da categoria B_i, $i = 1, \ldots, k$, e p_i é a porcentagem da categoria B_i, $i = 1, \ldots, k$.

Estatística Descritiva no R

Tabela de Distribuição de Frequências

```
dados_iris <- read_xlsx("data/raw/iris.xlsx")

tab <- dados_iris |>
  group_by(Species) |>
  summarise(`Frequência` = n()) |>
  mutate(`Frequência relativa` = `Frequência` / sum(`Frequência`),
         Porcentagem = `Frequência relativa` * 100)
tab
```

```
## # A tibble: 3 × 4
##   Species    Frequência `Frequência relativa` Porcentagem
##   <chr>           <int>                 <dbl>       <dbl>
## 1 setosa             50                 0.333        33.3
## 2 versicolor         50                 0.333        33.3
## 3 virginica          50                 0.333        33.3
```

```
write_ods(tab, "data/processed/tabela_frequencias_especies.ods")
write_xlsx(tab, "data/processed/tabela_frequencias_especies.xlsx")
```

Gráficos no `tidyverse`

- **Pacote:** `ggplot`
- Permite gráficos personalizados com uma sintaxe simples e rápida, e iterativa *por camadas*
- Começamos com um camada com os dados `ggplot(dados)`, e vamos adicionando as camadas de anotações, e sumários estatísticos
- Usa a *gramática de gráficos* proposta por Leland Wilkinson: Grammar of Graphics
- Ideia desta gramática: delinear os atributos estéticos das figuras geométricas (incluindo transformações nos dados e mudança no sistema de coordenadas)
- Para mais detalhes, você pode consultar `ggplot2`: elegant graphics for data analysis e documentação do `ggplot2`

Gráficos no `tidyverse`

Estrutura básica de `ggplot2`:

```
ggplot(data = <data possible tibble>) +
  <Geom functions>(mapping = aes(<MAPPINGS>)) +
  <outras camadas>
```

Você pode usar diversos temas e extensões que a comunidade cria e criou para melhorar a aparência e facilitar a construção de `ggplot2`.

- Lista com extensões do `ggplot`: extensões do `ggplots`

Eu já usei as seguinte extensões:

- Temas adicionais para o pacote `ggplot2`: `ggthemes`
- Gráfico de matriz de correlação: `ggcorrplot`
- Gráfico quantil-quantil: `qqplotr`

Gráficos no `tidyverse`

Gráfico de Barras no `ggplot2`

- **função:** `geom_bar()`. Para porcentagem: `geom_bar(x = <variável no eixo x>, y = ..prop.. * 100)`.
- Argumentos adicionais:
  - **`fill`**: mudar a cor do preenchimento das figuras geométricas
  - **`color`**: mudar a cor da figura geométrica

Rótulos dos eixos

- **Mudar os rótulos:** `labs(x = <rótulo do eixo x>, y = <rótulo do eixo y>)`
- **Trocar o eixo-x pelo eixo-y:** `coord_flip()`

```
ggplot(dados_iris) +
  geom_bar(mapping = aes(x = Species, y = ..prop.. * 100, group = 1),
  fill = "blue", color = "red") +
  labs(x = "Espécies", y = "Porcentagem") +
  theme_gdocs() +
  coord_flip()
```

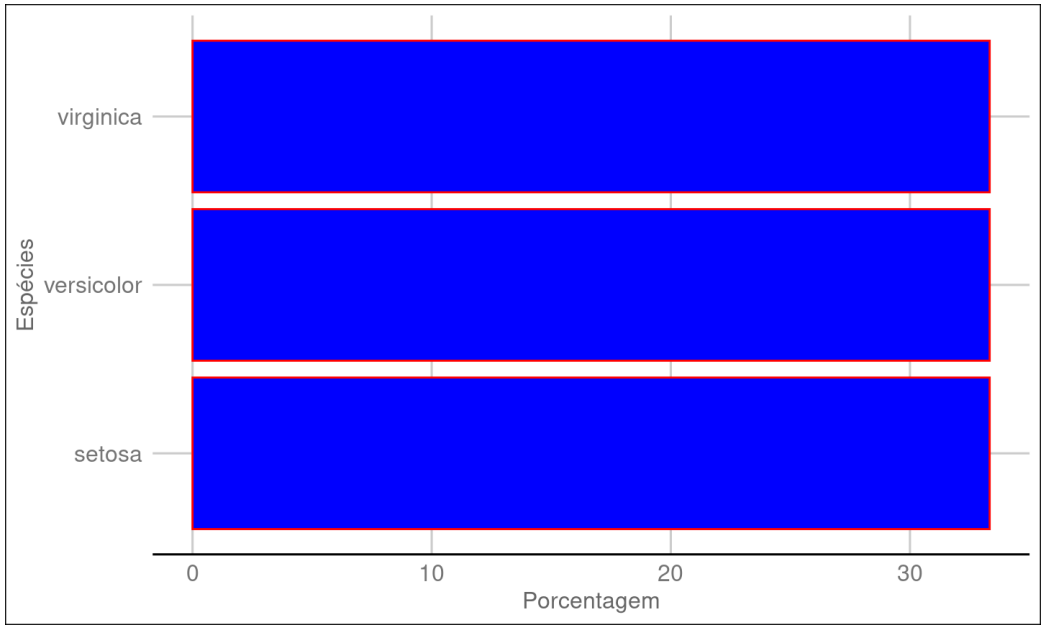


Tabela de Distribuição de Frequências

Tabela de Distribuição de Frequências

- X: variável quantitativa contínua

Tabela de Distribuição de Frequências para a variável quantitativa contínua.

x	Frequência	Frequência relativa	Porcentagem
$[l_0, l_1)$	n_1	$f_1 = \frac{n_1}{n_1+\cdots+n_k}$	$p_1 = f_1 \cdot 100$
$[l_1, l_2)$	n_2	$f_2 = \frac{n_2}{n_1+\cdots+n_k}$	$p_2 = f_2 \cdot 100$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[l_{k-1}, l_k]$	n_k	$f_k = \frac{n_k}{n_1+\cdots+n_k}$	$p_k = f_k \cdot 100$

Em que $\mathbf{min} = l_0 \leq l_1 \leq \cdots \leq l_{k-1} \leq l_k = \mathbf{max}$ ($\mathbf{min}$ é o menor valor do suporte da variável X e $\mathbf{max}$ é o maior valor do suporte da variável X), n_i é número de valores de X entre l_{i-1} e l_i, e $l_0, l_1, \ldots, l_k$ quebram o suporte da variável X (*breakpoints*).

$l_0, l_1, \cdots, l_k$ são escolhidos de acordo com a teoria por trás da análise de dados (ou pelo regulador). Se você está em uma nova área, use $l_0, l_1, \cdots, l_k$ igualmente espaçados, e use a regra de Sturges para determinar o valor de k: $k = 1 + \log 2(n)$ onde n é tamanho da amostra. Se $1 + \log 2(n)$ não é um número inteiro, usamos $k = \lceil 1 + \log 2(n) \rceil$.

Tabela de Distribuição de Frequências

Tabela de Distribuição de Frequências

```
k <- ceiling(1 + log2(nrow(dados_iris))) # regra de Sturges

dados_iris <- dados_iris |>
  mutate(sepal_length_intervalo = cut(Sepal.Length, breaks = k, include.lowest = T, right = F))

tab <- dados_iris |>
  group_by(sepal_length_intervalo) |>
  summarise(`Frequência` = n()) |>
  mutate(`Frequência relativa` = `Frequência` / sum(`Frequência`),
         Porcentagem = `Frequência relativa` * 100)
head(tab, n = 3)
```

```
## # A tibble: 3 × 4
##   sepal_length_intervalo Frequência `Frequência relativa` Porcentagem
##   <fct>                       <int>                 <dbl>       <dbl>
## 1 [4.3,4.7)                       9                 0.06          6
## 2 [4.7,5.1)                      23                 0.153        15.3
## 3 [5.1,5.5)                      20                 0.133        13.3
```

```
write_ods(tab, "data/processed/tabela_frequencias_sepal_length.ods")
write_xlsx(tab, "data/processed/tabela_frequencias_sepal_length.xlsx")
```

Gráficos no `tidyverse`

Histograma no `ggplot2`

- **função:** `geom_histogram()`. Para densidade de frequência: `geom_bar(x = <variável no eixo x>, y = ..density..)`. É necessário fornecer ou *bins* ou *breaks*:
  - Se *bins* é fornecido como um númeo inteiro, esta função cria faixas de tamanhos iguais para calcular a densidade de frequências
  - Se *breaks* é fornecido como um vetor
- Argumentos adicionais:
  - **`fill`**: mudar a cor do preenchimento das figuras geométricas
  - **`color`**: mudar a cor da figura geométrica

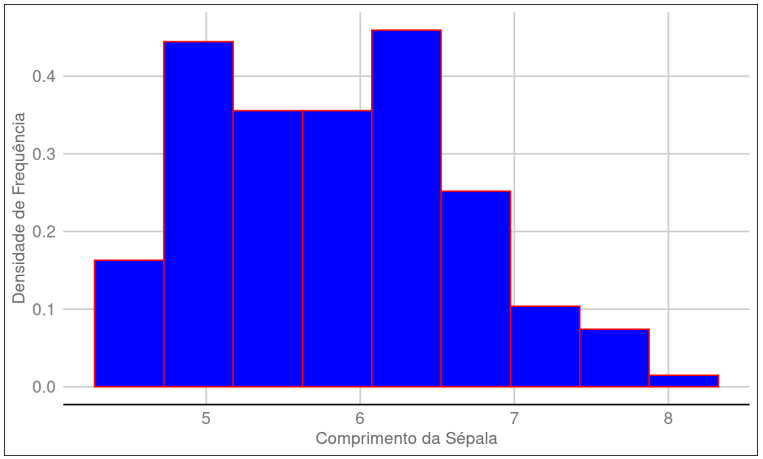
Rótulos dos eixos

- **Mudar os rótulos:** `labs(x = <rótulo do eixo x>, y = <rótulo do eixo y>)`
- **Trocar o eixo-x pelo eixo-y:** `coord_flip()`

```r
k <- ceiling(1 + log2(nrow(dados_iris))) # regra de Sturges

dados_iris <- dados_iris |>
  mutate(sepal_length_intervalo = cut(Sepal.Length, breaks = k, include.lowest = T,
                                      right = F))

ggplot(dados_iris) +
  geom_histogram(aes(x = Sepal.Length, y = ..density..),
                 bins = k, closed = 'left', fill = "blue", color = "red") +
  theme_gdocs() +
  labs(x = "Comprimento da Sépala", y = "Densidade de Frequência")
```



Gráficos no `tidyverse`

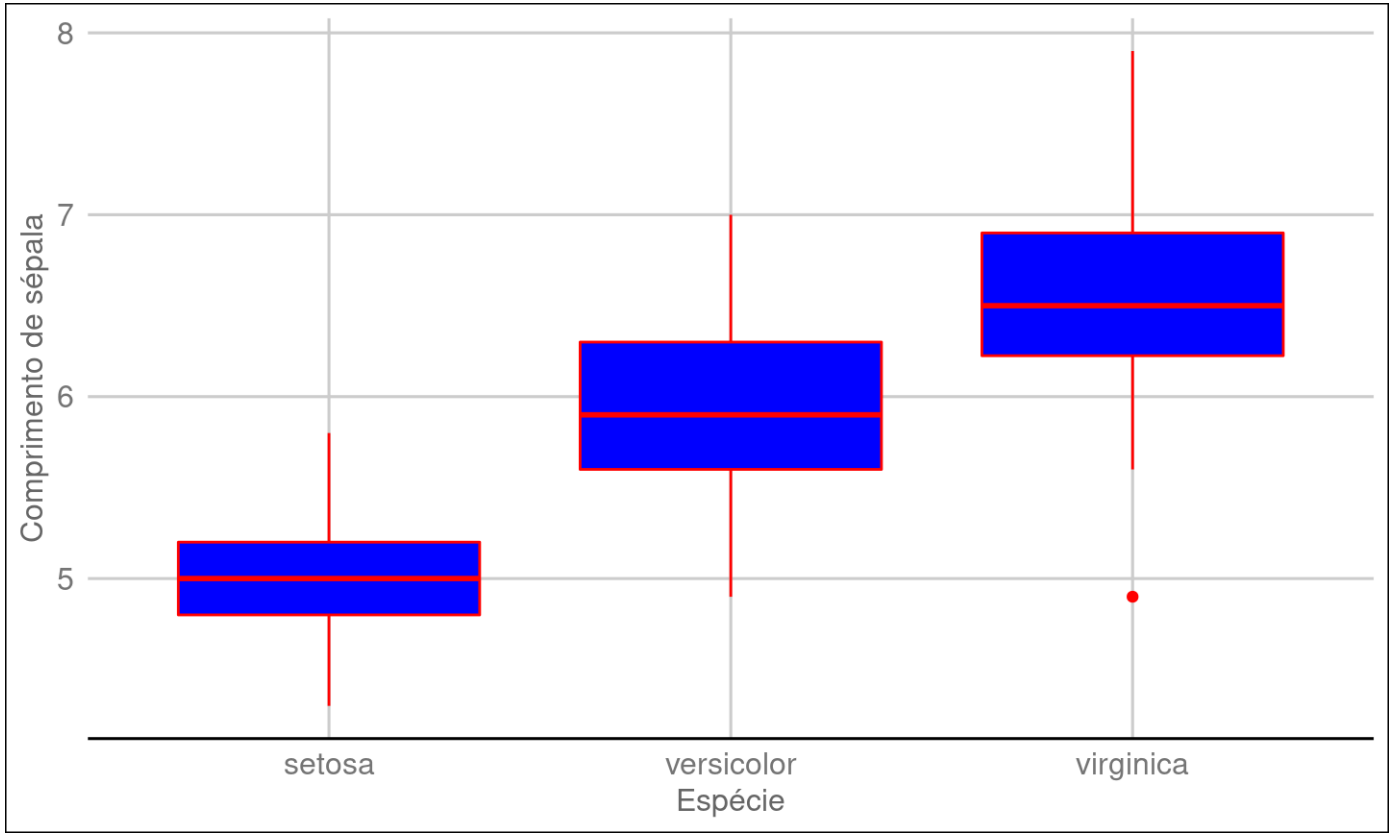
Diagrama de caixa (boxplot) no `ggplot2`

- **função:** `geom_boxplot()`. É necessário fornecer a variável *y*:
- Argumentos adicionais:
  - **`fill`**: mudar a cor do preenchimento das figuras geométricas
  - **`color`**: mudar a cor da figura geométrica

Rótulos dos eixos

- **Mudar os rótulos:** `labs(x = <rótulo do eixo x>, y = <rótulo do eixo y>)`
- **Trocar o eixo-x pelo eixo-y:** `coord_flip()`

```
ggplot(dados_iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length), fill = "blue", color = "red") +
  labs(x = "Espécie", y = "Comprimento de sépala") +
  theme_gdocs()
```



Medidas de resumo

- Usamos a função `summarise` do pacote `dplyr`
- Usamos a função `group_by` do pacote `dplyr` (medidas de resumo por categoria)

```
tab <- dados_iris |>
  group_by(Species) |>
  summarise(`Média` = mean(Sepal.Length), `Variância` = var(Sepal.Length),
            `Desvio Padrão` = sd(Sepal.Length), Mediana = median(Sepal.Length),
            q1 = quantile(Sepal.Length, probs = 0.25),
            q3 = quantile(Sepal.Length, probs = 0.75))

tab
```

```
## # A tibble: 3 × 7
##   Species    Média Variância `Desvio Padrão` Mediana    q1    q3
##   <chr>      <dbl>     <dbl>           <dbl>   <dbl> <dbl> <dbl>
## 1 setosa      5.01     0.124           0.352     5    4.8    5.2
## 2 versicolor  5.94     0.266           0.516     5.9  5.6    6.3
## 3 virginica   6.59     0.404           0.636     6.5  6.22   6.9
```

```
write_ods(tab, 'data/processed/medidas_resumo_sepal_length.ods')
write_xlsx(tab, 'data/processed/medidas_resumo_sepal_length.xlsx')
```

Tabela de contingência

- `empresa.xlsx`: conjunto de dados com informações socio-econômicas de `36` funcionários de uma determinada empresa. Este conjunto de dados é um exemplo didátivo do livro `Estatística Bussab` de Bussab e Morettin. Este conjunto de dados pode ser baixado na página pessoal de Pedro Morettin: Pedro Morettin.

```
dados_funcionarios <- read_xlsx("data/raw/empresa.xlsx")

tab <- dados_funcionarios |>
  group_by(`Estado Civil`, `Grau de Escolaridade`) |>
  summarise(`Frequência` = n(), .groups = 'rowwise') |>
  pivot_wider(names_from = `Grau de Escolaridade`, values_from = `Frequência`)
tab
```

```
## # A tibble: 2 × 4
##   `Estado Civil` `ensino fundamental` `ensino médio` superior
##   <chr>                         <int>          <int>    <int>
## 1 casado                            5             12        3
## 2 solteiro                          7              6        3
```

```
write_ods(tab, "data/processed/contingencia_estado_civil_escolaridade.ods")
write_xlsx(tab, "data/processed/contingencia_estado_civil_escolaridade.xlsx")
```

Gráfico de barras (duas variáveis)

```
ggplot(dados_funcionarios) +
  geom_bar(aes(x = `Estado Civil`, fill = Procedencia), position = 'fill') +
  labs(x = "Estado civil", y = "Porcentagem", fill = "Procedência") +
  scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1),
                     labels = c(0, 25, 50, 75, 100)) +
  theme_gdocs()
```

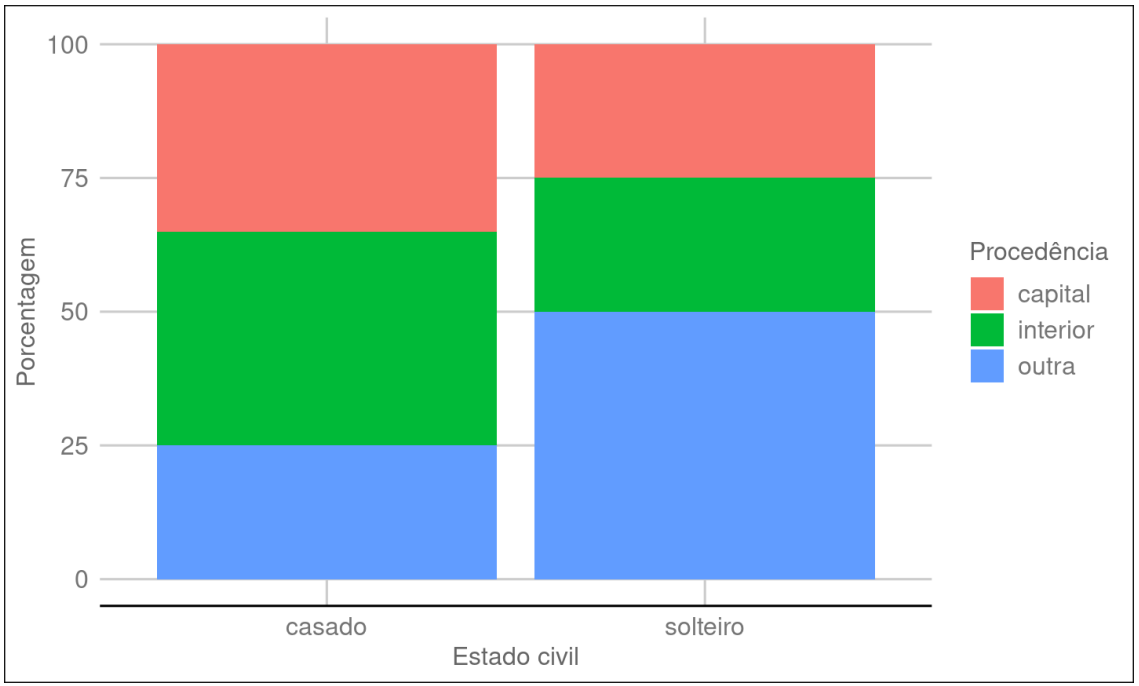


Gráfico de dispersão (duas variáveis)

```
ggplot(dados_iris) +
  geom_point(aes(x = Sepal.Length, y = Petal.Length, color = Species)) +
  labs(x = "Comprimento de sépala", y = "Comprimento de pétala",
      color = "Espécies") +
  theme_gdocs()
```

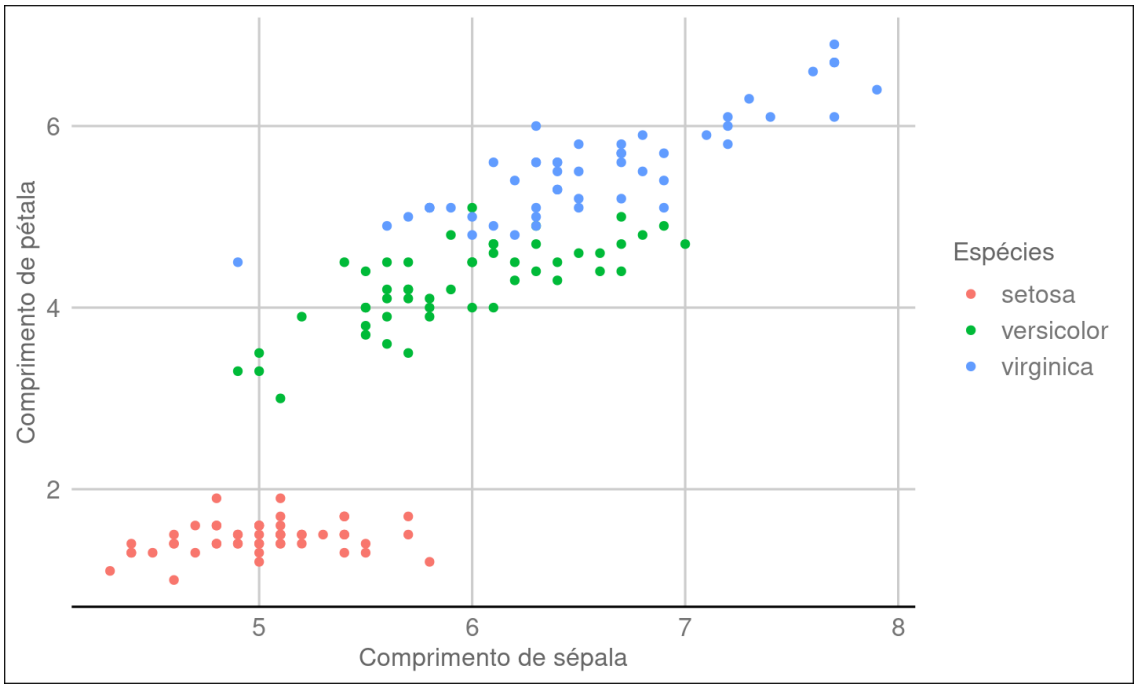


Gráfico *violin*

- Generalização do diagrama de caixa. Para mais detalhes, consulte: `geom_violin`

```
ggplot(dados_iris) +
  geom_violin(aes(x = Species, y = Sepal.Length, fill = Species),
              show.legend = FALSE) +
  labs(x = "Espécies", y = "Comprimento de sépala") +
  theme_gdocs()
```

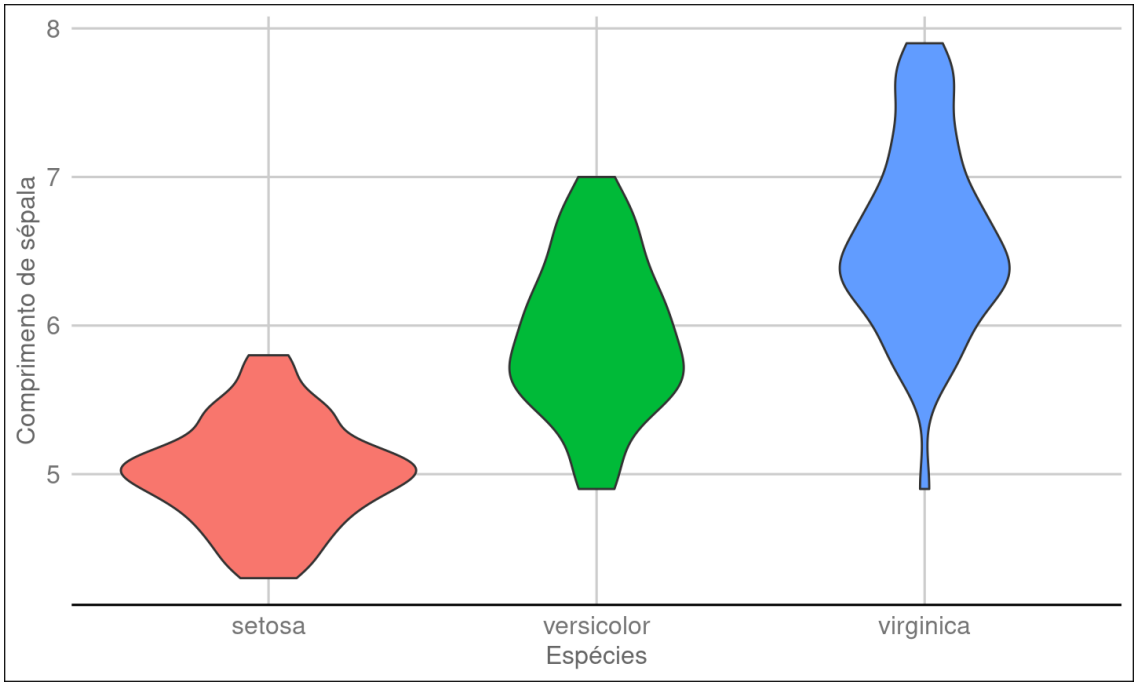


Gráfico *lvplot*

- Generalização do diagrama de caixa. Para mais detalhes, consulte: `lvplot`
- **Pacote:** `lvplot`

```
ggplot(dados_iris) +
  geom_lv(aes(x = Species, y = Sepal.Length, fill = Species),
              show.legend = FALSE) +
  labs(x = "Espécies", y = "Comprimento de sépala") +
  theme_gdocs()
```