Introdução à Estatística usando o R: Seja bem-vind@ ao tidyverse

Profa Carolina & Prof Gilberto

Instituto de Matemática e Estatística
Universidade Federal da Bahia

Conceitos básicos

Começamos com alguns conceitos básicos, que usaremos durante todo esse curso.

- **População**: Todos os elementos ou indivíduos alvo do estudo;
- **Amostra**: Parte da população;
- **Parâmetro:** característica da população (grandeza);
- **Estimativa:** característica da amostra. Usamos a estimativa para aproximar o parâmetro;
- **Variável:** característica de um elemento da população. Geralmente usamos uma letra maiúscula do alfabeto latino para representar uma variável (mensurando ou analito), e uma letra minúscula do alfabeto latino para representar o valor de uma variável para um elemento (indicação) da população.Por exemplo, podemos representar a variável "altura" por X e uma altura $x = 175$ cm de uma pessoa.

Classificação de variáveis

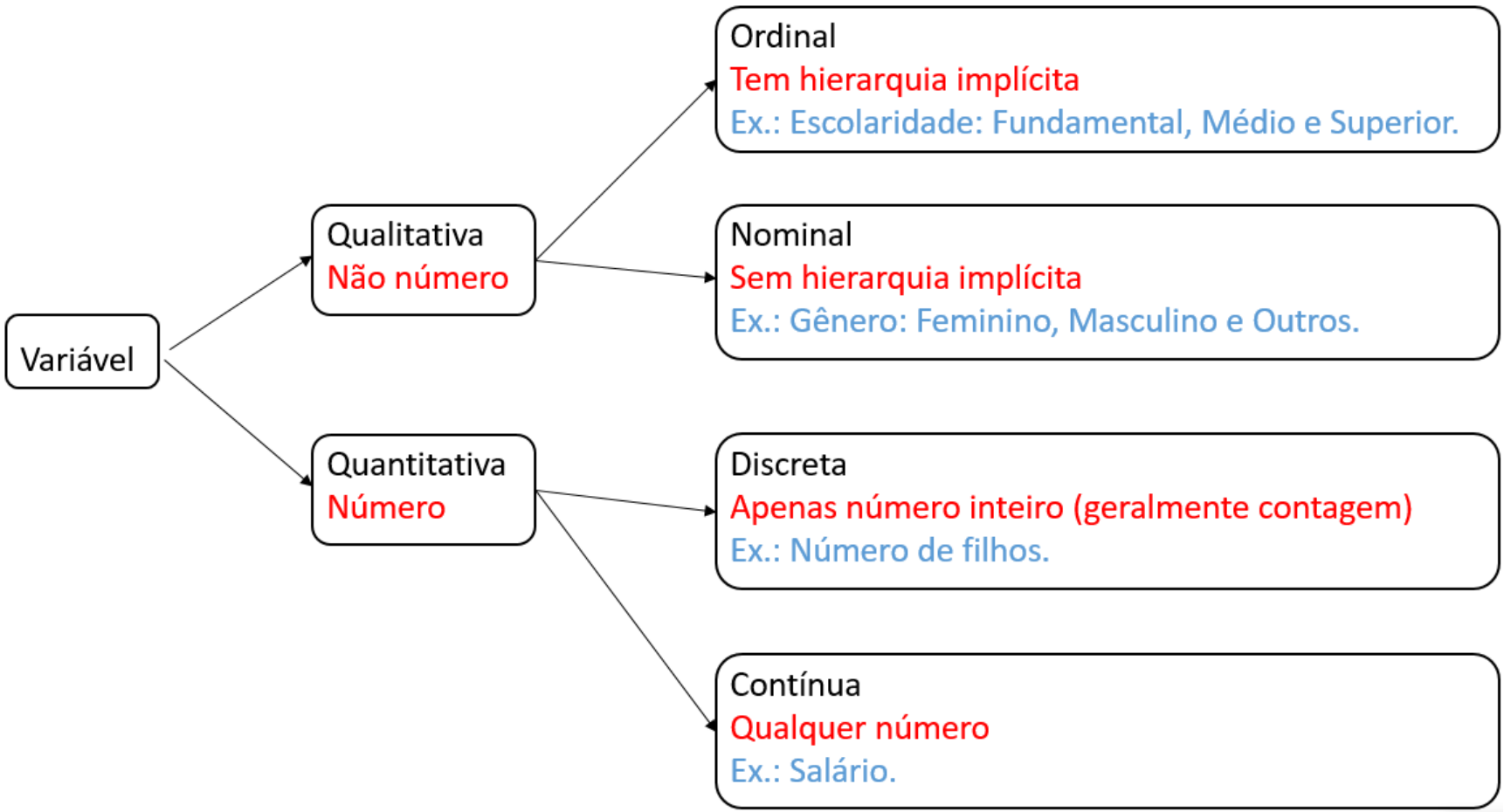


Figura 1: Classificação de variáveis.

Tabela de distribuição de frequência – Variável qualitativa

A primeira coisa que fazemos é contar!

Seja X uma variável qualitativa com valores possíveis $B_1, \ldots, B_k$, então construímos a tabela de distribuição de frequências como ilustrado na Tabela 1.

Tabela 1: Tabela de distribuição de frequências – variável qualitativa.

X	Frequência	Frequência Relativa	Porcentagem
B_1	n_1	$f_1 = \frac{n_1}{n}$	$100 \cdot f_1$
B_2	n_2	$f_2 = \frac{n_2}{n}$	$100 \cdot f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_k	n_k	$f_k = \frac{n_k}{n}$	$100 \cdots f_k$
Total	n	1	100%

Em que $n_i,\ i = 1, \ldots, k$ é o número de indivíduos com valor de X igual a B_i.

Tabela de distribuição de frequência – Variável qualitativa

A primeira coisa que podemos fazer é construir a tabela de distribuição de frequência.

```
df_ciaMB <- read_xlsx('dados.xlsx', sheet = 'companhia_MB')

tab_freq <- df_ciaMB %>%
  group_by(escolaridade) %>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
         porcentagem = 100 * frequencia_relativa)

tab_freq %>%
  add_case(escolaridade = 'Total',
           frequencia=sum(tab_freq$frequencia),
           frequencia_relativa = sum(tab_freq$frequencia_relativa),
           porcentagem = sum(tab_freq$porcentagem))
```

```
## # A tibble: 4 x 4
##   escolaridade      frequencia frequencia_relativa porcentagem
##   <chr>                  <int>                <dbl>       <dbl>
## 1 ensino fundamental        12                0.333        33.3
## 2 ensino médio              18                0.5          50
## 3 superior                   6                0.167        16.7
## 4 Total                     36                1          100
```

Gráfico no R

Vamos construir o gráfico de barras para a variável `especie`.

Vamos usar o pacote `ggplot2` já incluso no pacote `tidyverse`.

O gráficos usando ggplot tem o seguinte formato:

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

Gráfico de barras – variável qualitativa

Para a variável `especie`, temos que

```
ggplot(data = df_ciaMB) +
  geom_bar(mapping = aes(x = escolaridade, y = ..prop..,
                         group = 1),
           fill = 'blue')+
  labs(x = 'Espécie', y = 'Frequência Relativa',
       title = 'Gráfico de Barras') +
  theme_minimal()
```

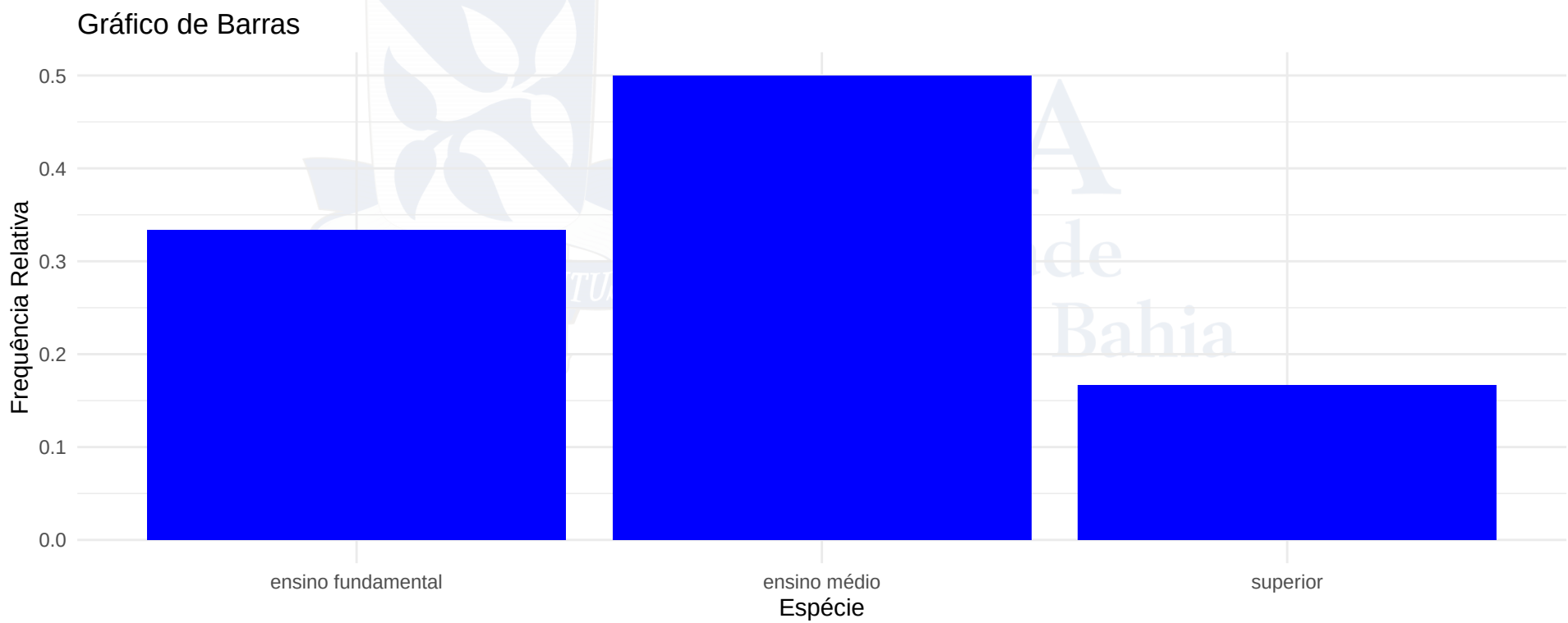


Tabela de distribuição de frequências – variável quantitativa discreta

```
## `summarise()` ungrouping output (override with `.groups` argument)
## # A tibble: 6 x 4
##   numero_filhos frequencia frequencia_relativa porcentagem
##   <chr>              <int>               <dbl>       <dbl>
## 1 0                     20               0.556        55.6
## 2 1                      5               0.139        13.9
## 3 2                      7               0.194        19.4
## 4 3                      3               0.0833        8.33
## 5 5                      1               0.0278        2.78
## 6 Total                 36               1           100
```

Gráfico de barras – variável quantitativa discreta

Gráfico de barras.

```
ggplot(data = df_companhia_MB)+
  geom_bar(mapping = aes(x = numero_filhos),
           fill = "blue")+
  theme_minimal()+
  scale_x_continuous(breaks = 0:5) +
  labs(x = "Número de sementes germinadas", y = "Frequência",
       title = "Gráfico de barras")
```

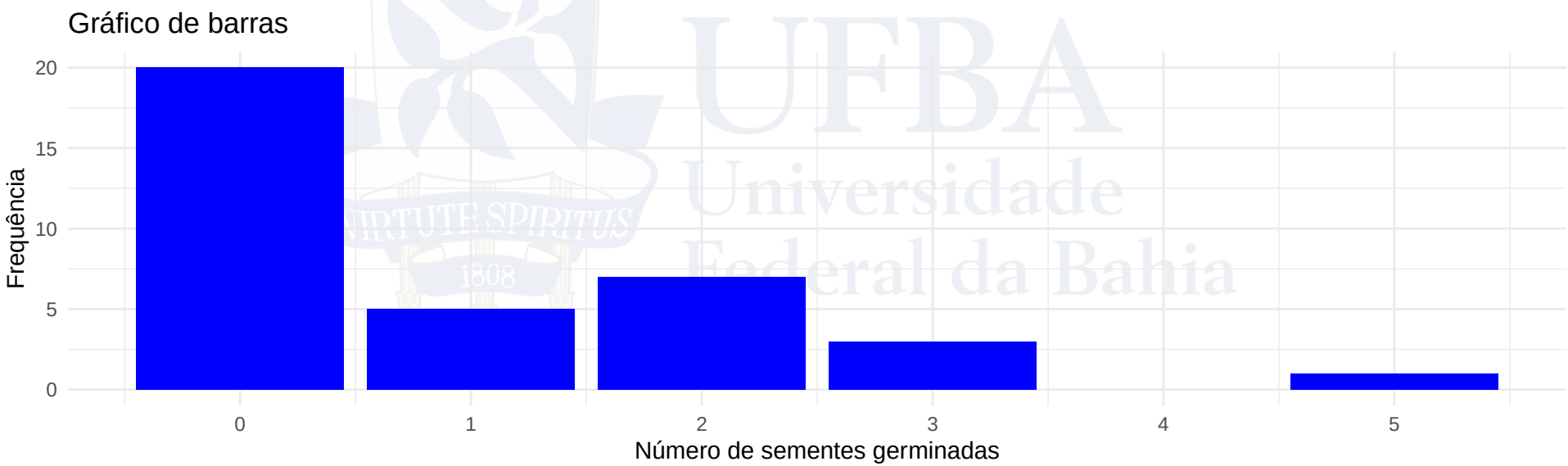


Tabela de distribuição de frequências – variável quantitativa contínua

Vamos construir um histograma para a comprimento de pétala para a espécie `versicolor`.

```
df_iris <- read_xlsx('dados.xlsx', sheet = 'Iris')
df_versicolor <- df_iris %>% filter(especie %in% 'versicolor')
k <- (1 + nrow(df_versicolor) %>% log2()) %>% ceiling()
tabela <- df_versicolor %>%
  group_by(petala_qual = cut(petala_comp, breaks = k,
                             include.lowest = T, right = F)) %>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
         porcentagem = frequencia_relativa * 100)
tabela %>% add_row(petala_qual = 'Total',
                   frequencia = sum(tabela$frequencia),
                   frequencia_relativa =
                     sum(tabela$frequencia_relativa),
                   porcentagem = sum(tabela$porcentagem))
```

Tabela de distribuição de frequências – variável quantitativa contínua

Vamos construir um histograma para a comprimento de pétala para a espécie `versicolor`.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 8 x 4
##   petala_qual frequencia frequencia_relativa porcentagem
##   <chr>           <int>                <dbl>       <dbl>
## 1 [3,3.3)             1                 0.02           2
## 2 [3.3,3.6)           4                 0.08           8
## 3 [3.6,3.9)           3                 0.06           6
## 4 [3.9,4.2)          11                 0.22          22
## 5 [4.2,4.5)          10                 0.2           20
## 6 [4.5,4.8)          15                 0.3           30
## 7 [4.8,5.1]           6                 0.12          12
## 8 Total              50                 1            100
```

Histograma – variável quantitativa contínua

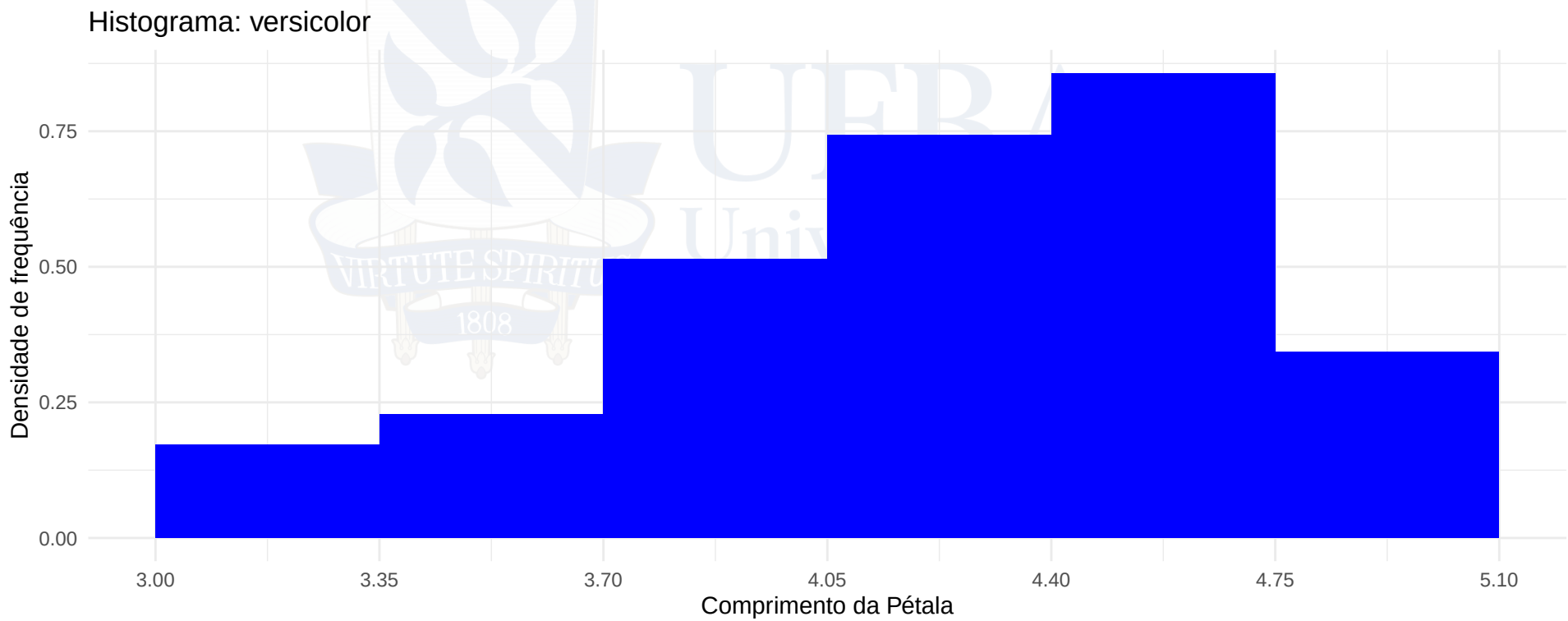
Nos gráficos de barras, a frequência (ou frequência relativa ou porcentagem) está no eixo y, ou seja, na altura da barra.

O histograma tem uma interpretação ligeiramente diferente: a área da barra é a frequência relativa.

- Para variável quantitativa contínua, dividimos os valores em faixas de valores e calculamos a frequência relativa para cada faixa.
- Para a barra correspondente à faixa $[a, b)$ a altura da barra precisa ser $\frac{f}{b-a}$, em que f é a frequência relativa da faixa $[a, b)$.
- Chamamos a razão $\frac{f}{b-a}$ de densidade de frequência.
- Número de faixas, podemos usar a regra de Sturge: $[1 + log_2(n)]$.

Histograma – variável quantitativa contínua

```
limites <- with(df_versicolor, seq(from = min(petala_comp),
                                   to = max(petala_comp), length.out = k))
ggplot(data = df_versicolor)+
  geom_histogram(mapping = aes(x = petala_comp, y = ..density..),
                 breaks = limites, fill = 'blue') +
  scale_x_continuous(breaks = limites) +
  theme_minimal()+
  labs(x = 'Comprimento da Pétala', y = 'Densidade de frequência',
       title = 'Histograma: versicolor')
```



Medidas de Resumo (variável quantitativa)

A ideia é encontrar um ou alguns valores que sintetizem todas as indicações.

Medidas de posição (tendência central)

A ideia é encontrar um valor que representa "bem" todas as indicações.

- **Média:** $\overline{x} = \dfrac{x_1 + \cdots + x_n}{n}$
- **Mediana:** valor que divide a sequência ordenada de valores em duas partes iguais.

$$\begin{cases} x_{(\frac{n+1}{2})}, & \text{n é ímpar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{n é par} \end{cases}$$

em que $x_{(j)}$ é o j-ésimo menor valor da variável quantitativa X.

Medidas de dispersão

A ideia é medir a homogeneidade das indicações.

- **Variância:** $s^2 = \dfrac{(x_1 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2}{n-1}$;
- **Desvio padrão:** $s = \sqrt{s^2}$ (mesma unidade dos dados);
- **coeficiente de variação** $cv = \dfrac{s}{\overline{x}} \cdot 100\%$ (adimensional, ou seja, "sem unidade")

Medidas de Resumo: exemplo

Podemos usar a função `summarise` do pacote `dplyr` (incluso no pacote `tidyverse`).

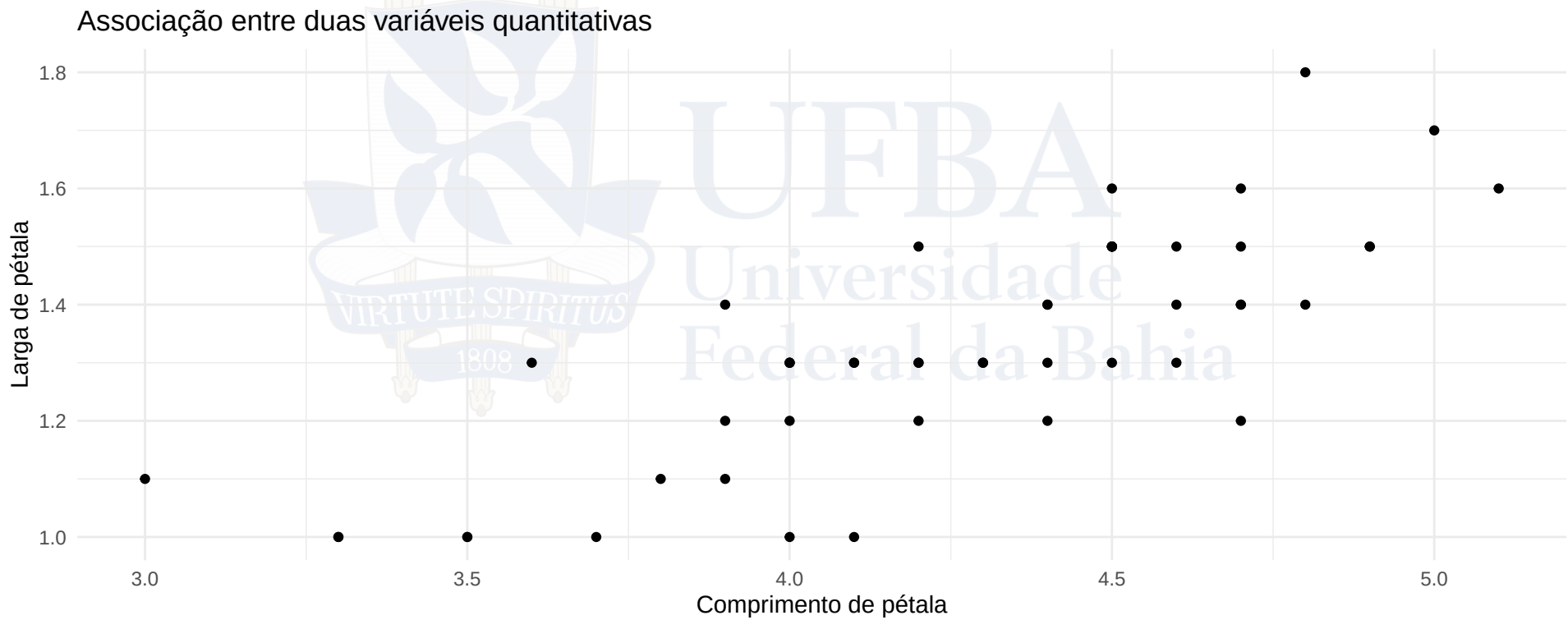
```
# Média para o comprimento de pétala para a espécie versicolor
df_versicolor %>% summarise(media = mean(petala_comp),
                            s2 = var(petala_comp),
                            s = sd(petala_comp),
                            mediana = median(petala_comp),
                            cv = s * 100 / media)
```

```
## # A tibble: 1 x 5
##   media    s2     s mediana    cv
##   <dbl> <dbl> <dbl>   <dbl> <dbl>
## 1  4.26 0.221 0.470    4.35  11.0
```

Associação entre duas variáveis quantitativas

Para duas variáveis quantitativas, estudamos a associação entre as duas variáveis usando o gráfico de dispersão. Além disso, podemos calcular o coeficiente de correlação linear de Pearson.

```
ggplot(data = df_versicolor)+
  geom_point(aes(x=petala_comp, y = petala_larg)) +
  theme_minimal()+
  labs(x = 'Comprimento de pétala', y = 'Larga de pétala',
       title = 'Associação entre duas variáveis quantitativas')
```



Associação entre duas variáveis quantitativas

Também podemos calcular o coeficiente de correlação linear de Pearson. Lembre que se X e Y são duas variáveis quantitativas com valores

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

Então, o coeficiente de correlação linear é dado por

$$r = \left(\frac{(x_1 - \overline{x})}{s_x} \cdot \frac{(y_1 - \overline{y})}{s_y} \right) + \cdots + \left(\frac{(x_n - \overline{x})}{s_x} \cdot \frac{(y_n - \overline{y})}{s_y} \right)$$

```
#No R, o cálculo é bem simples
with(df_versicolor, cor(petala_comp, petala_larg))
```

```
## [1] 0.7866681
```

Associação entre duas variáveis qualitativas

Objetivo

Sejam X e Y duas variáveis qualitativas com valores possíveis:

- X: $A_1, A_2, \cdots, A_r$;
- Y: $B_1, B_2, \cdots, B_s$.

Desejamos estudar a associação entre X e Y.

O que é associação entre X e Y?

Suponha que $f_i \cdot 100\%$ dos elementos da população tenham valor de X igual a A_i. Então, X e Y são

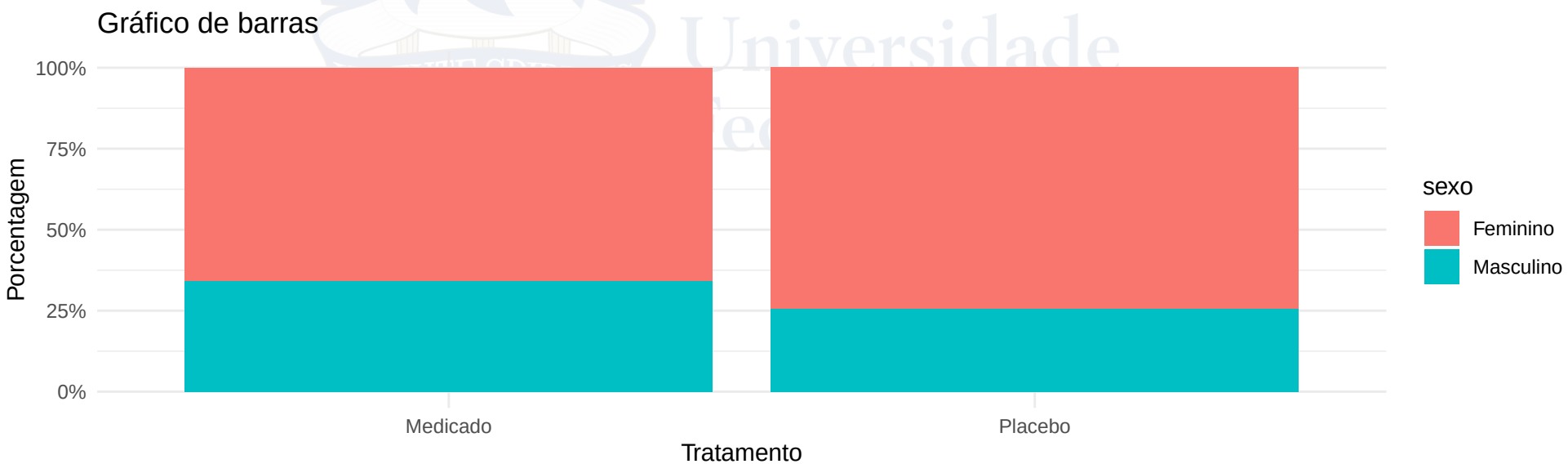
- **não associados** se ao conhecermos o valor de Y para um elemento da população, continuamos com o valor $f_i \cdot 100\%$ de chance do indivíduo ter valor de X igual a A_i;
- **associados** se ao conhecermos o valor de Y para um elemento da população, alteramos o valor $f_i \cdot 100\%$ de chance do indivíduo ter valor de X igual a A_i;

Associação entre duas variáveis qualitativas

Para duas variáveis quantitativas, estudamos a associação entre as duas variáveis usando uma tabela de contingência e o gráfico de barras.

Exemplo didático sem associação: gráfico de barras e teste qui-quadrado.

```
df_trat <- read_xlsx("dados.xlsx", sheet = "tratamento")

with(df_trat, DescTools::ContCoef(tratamento, sexo))

## [1] 0.09323073

ggplot(df_trat) +
  geom_bar(aes(x = tratamento, fill = sexo), position = "fill") +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Tratamento", y = "Porcentagem", title = "Gráfico de barras")
```



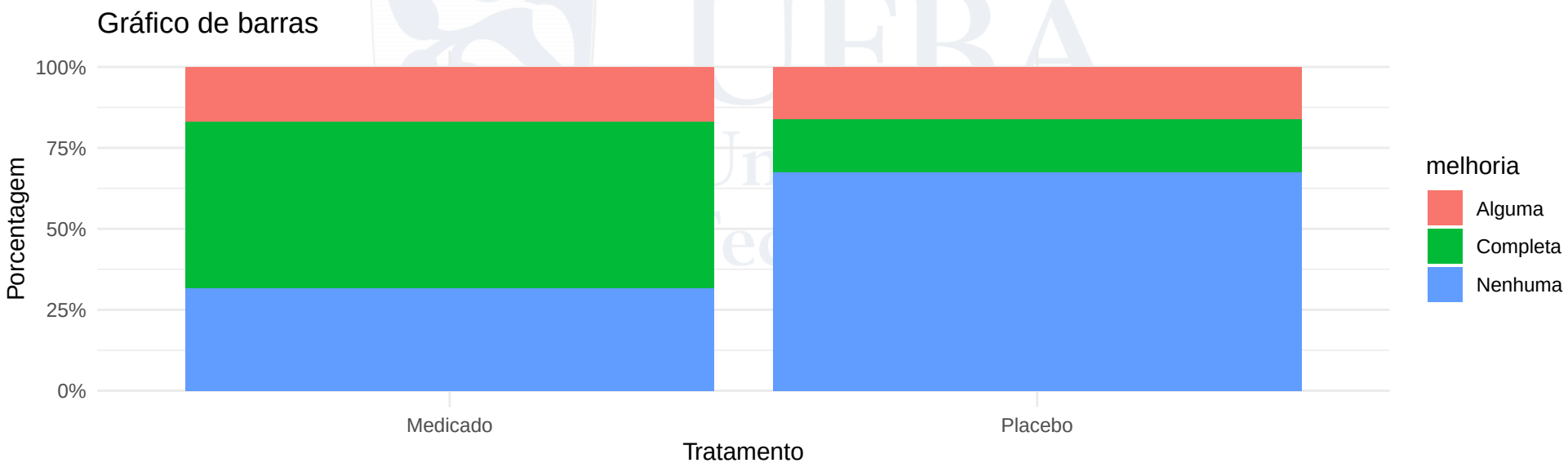
Associação entre duas variáveis qualitativas

Exemplo didático sem associação: gráfico de barras e teste qui-quadrado.

```
with(df_trat, DescTools::ContCoef(tratamento, melhoria))
```

```
## [1] 0.3667581
```

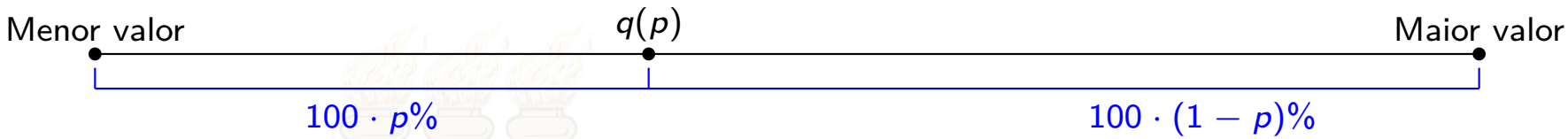
```
ggplot(df_trat) +
  geom_bar(aes(x = tratamento, fill = melhoria), position = "fill") +
  theme_minimal() +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Tratamento", y = "Porcentagem", title = "Gráfico de barras")
```



Quantis e quartis

O *quantil de ordem* $p \in (0,1)$, $q(p)$, é um valor de uma variável `x` que divide os valores amostrados em duas partes: $100 \cdot p\%$ dos valores estão entre o `min(x)` e $q(p)$, e $100 \cdot (1-p)\%$ dos valores estão entre $q(p)$ e `max(x)`.

Figura 2:



- Quando $p = \frac{1}{4}$, dizemos que $q(p)$ é o primeiro quartil e usamos a notação q_1;
- Quando $p = \frac{2}{4}$, dizemos que $q(p)$ é o primeiro quartil e usamos a notação q_2;
- Quando $p = \frac{3}{4}$, dizemos que $q(p)$ é o primeiro quartil e usamos a notação q_3;

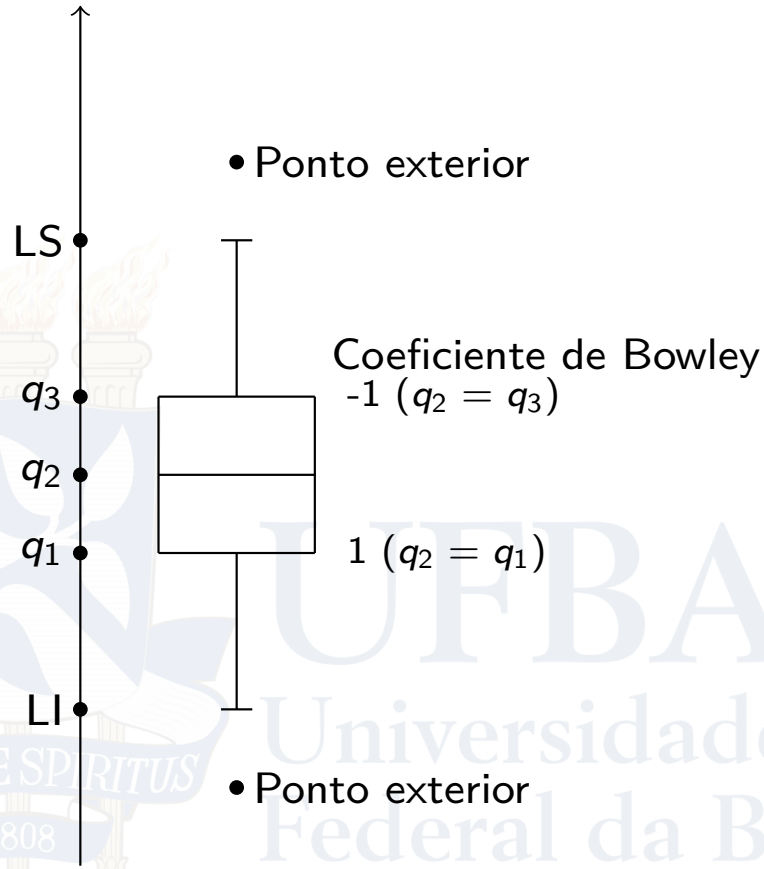
```
dados <- read_xlsx("dados.xlsx", sheet = "companhia_MB")

dados %>% group_by(escolaridade) %>%
  summarise(q1 = quantile(salario, prob = 0.25),
            q2 = quantile(salario, prob = 0.50),
            q3 = quantile(salario, prob = 0.75))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 4
##   escolaridade          q1    q2    q3
##   <chr>              <dbl> <dbl> <dbl>
## 1 ensino fundamental  6.01  7.12  9.16
## 2 ensino médio        8.84 10.9  14.4
```

Diagrama de caixa ou *boxplot*

O diagrama de caixa tem o seguinte aspecto



- **dq** $= \mathbf{q_3} - \mathbf{q_1}$ é o intervalo interquartil ou amplitude interquartil e é interpretada como medida de dispersão;
- **Limite Superior** $LS = q_3 + 1,5 \cdot dq$;
- **Limite Inferior** $LI = q_1 - 1,5 \cdot dq$;
- **Ponto Adjacente** Todos os valores da variável entre LI e LS;
- **Ponto Exterior** Todos os valores da variável que não estão entre LI e LS. Estes valores da variável são provavelmente destoantes que precisam de atenção do pesquisador;

Diagrama de caixa

```
dados <- read_xlsx("dados.xlsx", sheet = "companhia_MB")

ggplot(dados) +
  geom_boxplot(aes(x = escolaridade, y = salario), color = "blue")+
  theme_minimal() +
  scale_x_discrete(labels = c("Ens Fundamental", "Ens Médio", "Ens Sup"))
  labs(x = "Escolaridade", y = "Salário")
```

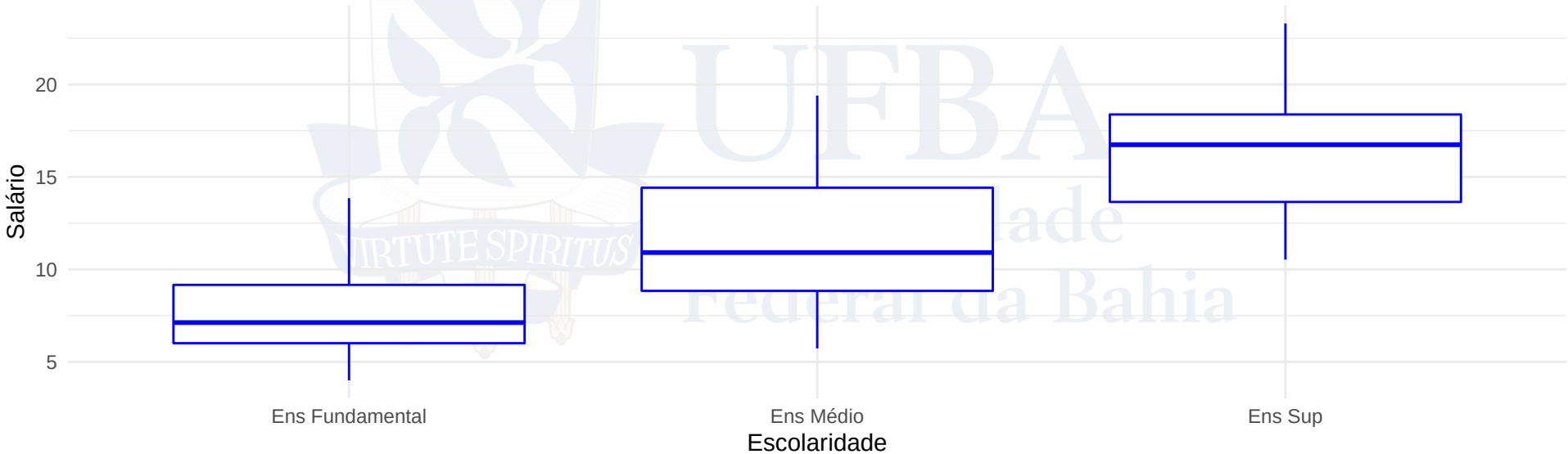


Diagrama de caixa e simetria

- Se q_2 perto de q_1: temos assimetria à direita ou positiva e valores tendem a ser menores;
- Se q_2 perto de q_3: temos assimetria à esquerda ou negativa e valores tendem a ser maiores;
- Se q_2 está entre q_1 e q_3: temos simetria.

Exemplo de simetria.

```
dados <- read_xlsx("dados.xlsx", sheet = "companhia_MB")

quartis <- with(dados, quantile(idade, probs = c(0.25, 0.5, 0.75)))
(quartis[3] - 2 * quartis[2] + quartis[1]) / (quartis[3] - quartis[1])
```

```
## 75%
## 0.1
```

```
dados <- read_xlsx("dados.xlsx", sheet = "notas")

with(dados, bowley_coeff(turma_1))
```

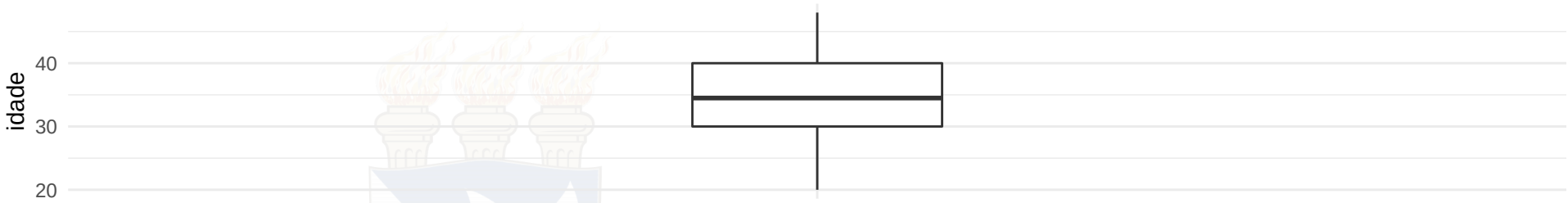
```
## [1] 0.4042553
```

```
with(dados, bowley_coeff(turma_2))
```

```
## [1] -0.4285714
```

Diagrama de caixa e simetria

```
dados <- read_xlsx("dados.xlsx", sheet = "companhia_MB")

ggplot(dados) +
  geom_boxplot(aes(x = "", y = idade), width = 0.2) + theme_minimal()+
  labs(x = "")
```




```
k <- (1 + nrow(dados) %>% log2()) %>% round()
limites <- with(dados,
                seq(from = min(idade), to = max(idade), length.out = k))
ggplot(dados) +
  geom_histogram(aes(x = idade, y = ..density..),
                 breaks = limites, fill = "white", color = "blue") +
  scale_x_continuous(breaks = limites) +
  theme_minimal() + labs(x = "Idade", y = "Função densidade")
```

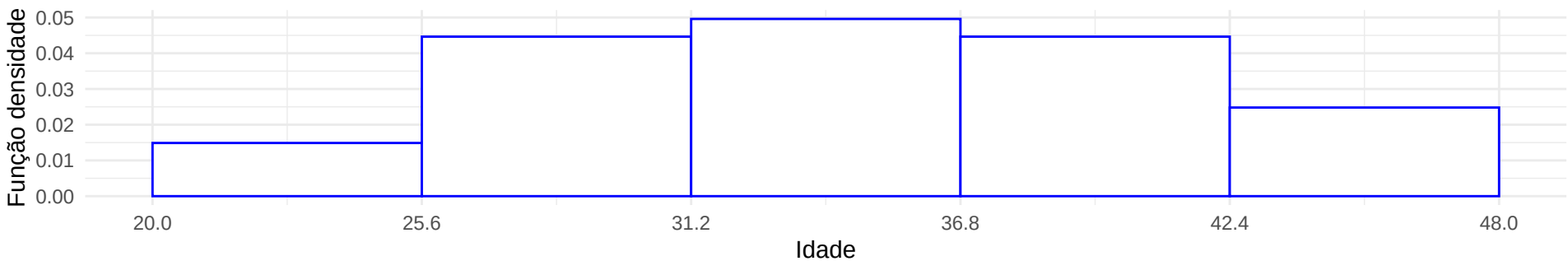
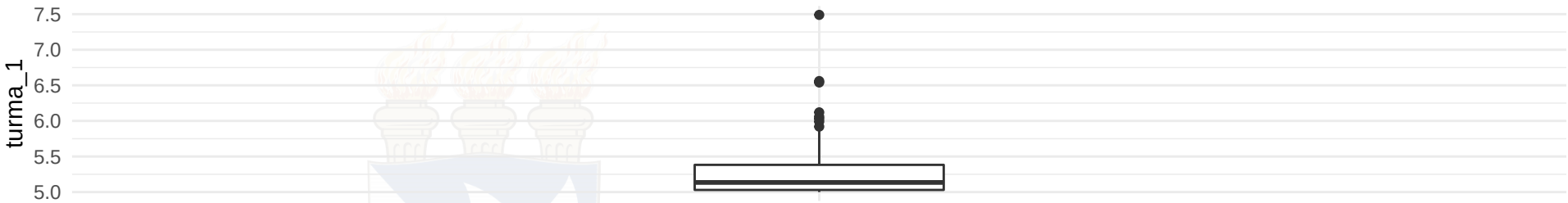


Diagrama de caixa e assimetria à direita ou positiva

```
dados <- read_xlsx("dados.xlsx", sheet = "notas")

ggplot(dados) +
  geom_boxplot(aes(x = "", y = turma_1), width = 0.2) + theme_minimal()+
  labs(x = "")
```




```
k <- (1 + nrow(dados) %>% log2()) %>% round()
limites <- with(dados,
                seq(from = min(turma_1), to = max(turma_1), length.out = k
ggplot(dados) +
  geom_histogram(aes(x = turma_1, y = ..density..), breaks = limites,
                 fill = "blue")+
  theme_minimal() + labs(x = "Turma 1", y = "Função densidade")
```

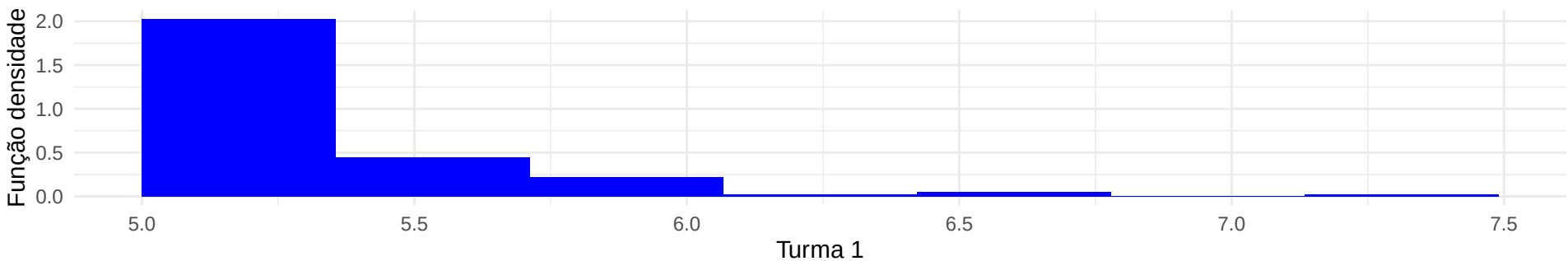
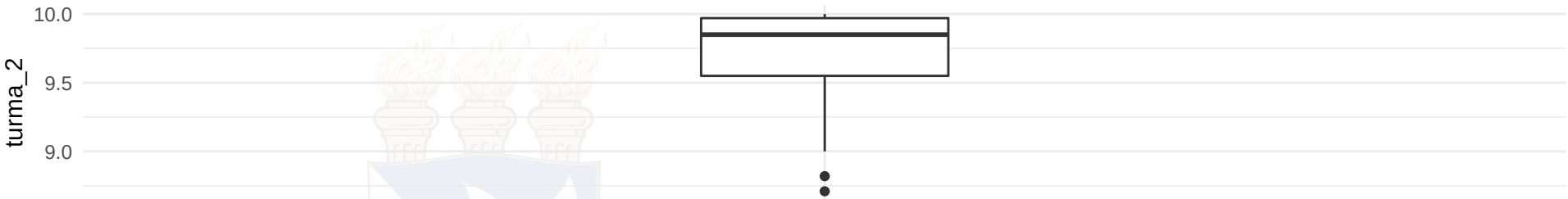


Diagrama de caixa e assimetria à esquerda ou negativa

```
dados <- read_xlsx("dados.xlsx", sheet = "notas")

ggplot(dados) +
  geom_boxplot(aes(x = "", y = turma_2), width = 0.2) + theme_minimal()+
  labs(x = "")
```




```
k <- (1 + nrow(dados) %>% log2()) %>% round()
limites <- with(dados,
                seq(from = min(turma_2), to = max(turma_2), length.out = k
ggplot(dados) +
  geom_histogram(aes(x = turma_2, y = ..density..), breaks = limites,
                 fill = "blue")+
  theme_minimal() + labs(x = "Turma 2", y = "Função densidade")
```