Introdução à Estatística usando o R com Aplicação em Análises Laboratoriais

Profa Carolina & Prof Gilberto

Instituto de Matemática e Estatística
Universidade Federal da Bahia

30 de novembro de 2019

Análise de variância com um fator

É uma extensão do teste de comparação de médias para duas populações independentes.

A ANOVA é um procedimento para comparar três ou mais médias populacionais baseado na análise das variâncias amostrais.

Na ANOVA, os dados amostrais são separados em grupos segundo uma característica (fator, tratamento, grupo).

Conceitos básicos:

- O **fator (tratamento ou grupo)** é uma característica (qualidade) que permite distinguir diferentes populações umas das outras.
- Cada fator contém dois ou mais **níveis** (classificações).
- A **variável resposta** é aquela que estamos comparando. Exemplos: comparações de média de idade por cor/raça (Branca, Preta, Amarela, Parda, Indígena, Sem declaração).

ANOVA com um fator

Problema:

Seja Y uma v.a. quantitativa de interesse.

Suponha k populações normais e independentes tal que, em cada população, a média de Y é μ_i e a variância de Y é σ^2, $i = 1, 2, \ldots, k$.

Isto é,

$$Y_i \sim N(\mu_i, \sigma^2); \quad i = 1, 2, \ldots, k.$$

ANOVA com um fator

Seja $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ uma a.a. de tamanho n_i de $Y_i \sim N(\mu_i, \sigma^2)$, para $i = 1, 2, \ldots, k$.

Os dados são da forma:

Tabela 1: Estrutura geral dos dados em ANOVA com um fator.

Tratamento	Amostra
1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
$\vdots$	$\vdots$
k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$

Suposições para a ANOVA com um fator

- As populações são normalmente distribuídas.
- As populações têm a mesma variância (ou mesmo desvio padrão).
- As amostras são aleatórias e mutuamente independentes.

Modelo com um classificação

Sejam $Y_1, Y_2, \ldots, Y_k$, k populações normais com médias $\mu_1, \mu_2, \ldots, \mu_k$ com mesma variância σ^2.

Isto é, $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, k$.

Seja $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$, para $i = 1, 2, \ldots, k$, uma a.a. de tamanho n_i de Y_i.

Equivalentemente, podemos escrever

$$Y_{ij} = \mu_i + \epsilon_{ij}, \ \epsilon_{ij} \sim N(0, \sigma^2), \ i = 1, \ldots, k, \ j = 1, \ldots, n_i.$$

Modelo com um classificação

Interesse:

- **Hipótese nula:** as médias de todas as populações são iguais. Isto é, o tratamento (fator) não tem efeito (nenhuma na variação da média entre os grupos).
- **Hipótese alternativa:** nem todas a médias populacionais são iguais. Isto é, pelo menos uma média é diferente (existe efeito do tratamento).

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1: \ \mu_i \neq \mu_j \ \text{ para algum par } \ (i,j), \quad i \neq j.$$

Modelo com uma classificação

Se as médias das k populações forem iguais, esperaríamos que as k médias da amostra estivessem próximas.

Em outras palavras:

- se a variabilidade entre as médias amostrais for "pequena", temos evidências a favor de H_0;
- se a variabilidade entre as médias amostrais for "grande", temos evidências contra H_0.

Modelo com um classificação

Tratamento	Amostra
1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
$\vdots$	$\vdots$
k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$

Temos que:

Tamanho amostral: $n = n_1 + \ldots + n_k$.

Média amostral para o tratamento i: $\overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$.

Média amostral geral: $\overline{Y}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$.

Desvio das observações em relação à média geral: $Y_{ij} - \overline{Y}_{..}$.

Desvio das médias dos tratamentos em relação à média geral: $\overline{Y}_{i.} - \overline{Y}_{..}$.

Desvio das observações em relação à média do tratamento: $Y_{ij} - \overline{Y}_{i.}$.

Decomposição da variância

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{..})^2.$$

SQT: medida da variação das observações em torno da a média geral.

$$SQTr = \sum_{i=1}^{k} n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

$SQTr$: medida da variação entre as médias dos tratamentos e a média geral.

$$SQR = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

SQR: medida da variação dentro dos grupos (o que deixou de ser explicado pelo fator).

Decomposição da variância

Segue que,

$$SQT = SQTr + SQR.$$

SQT tem $n - 1$ graus de liberdade (temos n observações para estimar a média geral).

$SQTr$ tem $k - 1$ graus de liberdade (a soma dos desvios é zero e temos k desvios, então temos $k - 1$ desvios independentes)

SQR tem $n - k$ graus de liberdade (temos n observações para estimar k médias).

Teste F

Sob H_0, temos que

$$F_0 = \frac{QMTr}{QMR} = \frac{\frac{SQTr}{k-1}}{\frac{SQR}{n-k}} \sim F_{k-1,n-k}.$$

A um nível de significância $\alpha \in (0,1)$, rejeita-se H_0 se

$$F_0 > F_{\alpha,k-1,n-k}.$$

Em resumo, temos a tabela da ANOVA com um fator:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
Tratamento	$SQTr$	$k-1$	$QMTr$	$\frac{QMTr}{QMR}$
Resíduo	SQR	$n-k$	QMR	
Total	SQT	$n-1$		

Comparações múltiplas

Anova com um fator:

Em linhas gerais, estamos interessados em determinar se as médias de mais de duas populações ou grupos são iguais ou não.

Para testar se a diferença nas médias é estatisticamente significativa, podemos realizar a análise de variância usando o teste F.

Se o teste F da Anova mostrar que há uma diferença significativa nas médias entre os grupos, poderemos querer realizar comparações múltiplas entre todos as médias de pares para determinar como (e quais) diferem.

Comparações múltiplas

Testes de comparações múltiplas:

Para cada par (i, j), com $i \neq j$, estamos interessados no teste de hipótese:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j, \quad i \neq j.$$

Análise de resíduos

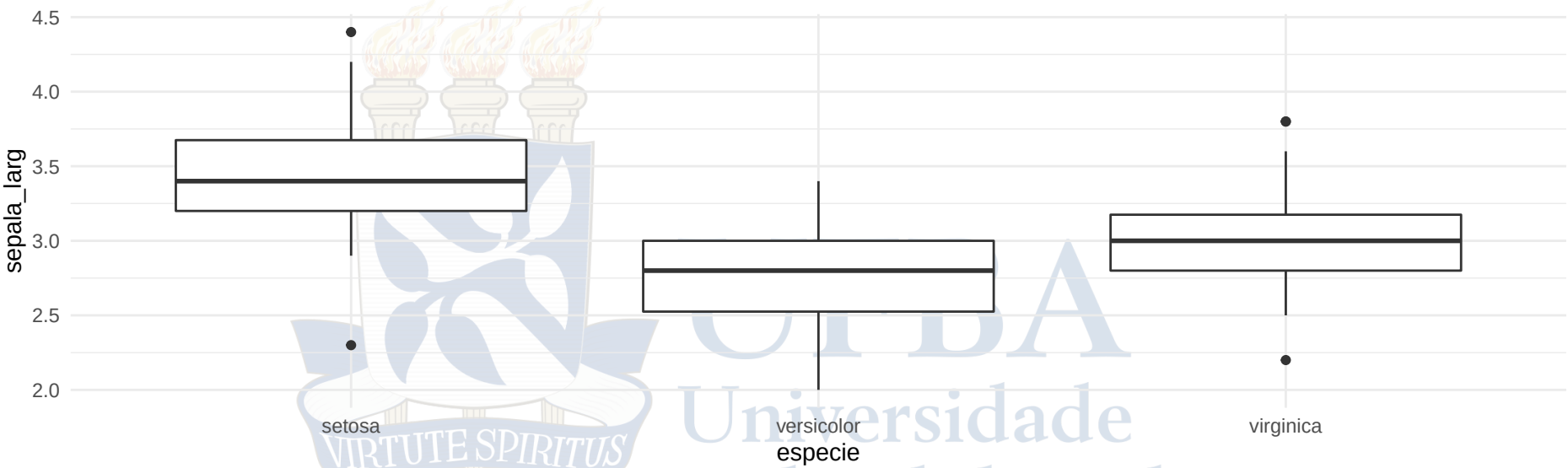
As principais suposições feitas até agora no estudo de análise de regressão linear simples foram as seguintes:

1. Linearidade: a relação entre a resposta Y e as variáveis regressoras é linear, pelo menos aproximadamente.
2. O termo de erro ϵ tem média zero.
3. Homoscedasticidade: o termo de erro ϵ tem variância constante σ^2.
4. Independência: os erros são não correlacionados.
5. Normalidade: os erros são normalmente distribuídos.

Exemplo

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

dados <- dados %>%
  mutate(especie = as_factor(especie))

ggplot(dados, aes(x = especie, y = sepala_larg)) +
  geom_boxplot()

ajuste <- aov(sepala_larg ~ especie, data = dados)

anova <- anova(ajuste)
```

Exemplo



Exemplo

```
## Call:
##    aov(formula = sepala_larg ~ especie, data = dadc
##
## Terms:
##                   especie Residuals
## Sum of Squares   11.34493  16.96200
## Deg. of Freedom         2       147
##
## Residual standard error: 0.3396877
## Estimated effects may be unbalanced
```

Exemplo

```
## Analysis of Variance Table
##
## Response: sepala_larg
##             Df Sum Sq Mean Sq F value    Pr(>F)
## especie      2 11.345  5.6725  49.16 < 2.2e-16 ***
## Residuals 147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
```

Exemplo

```
with(dados,
     pairwise.t.test(sepala_larg, especie,
                     alternative = "two.sided",
                     p.adjust.method = "bonferroni"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  sepala_larg and especie
##
##            setosa  versicolor
## versicolor < 2e-16 -
## virginica  1.4e-09 0.0094
##
## P value adjustment method: bonferroni
```

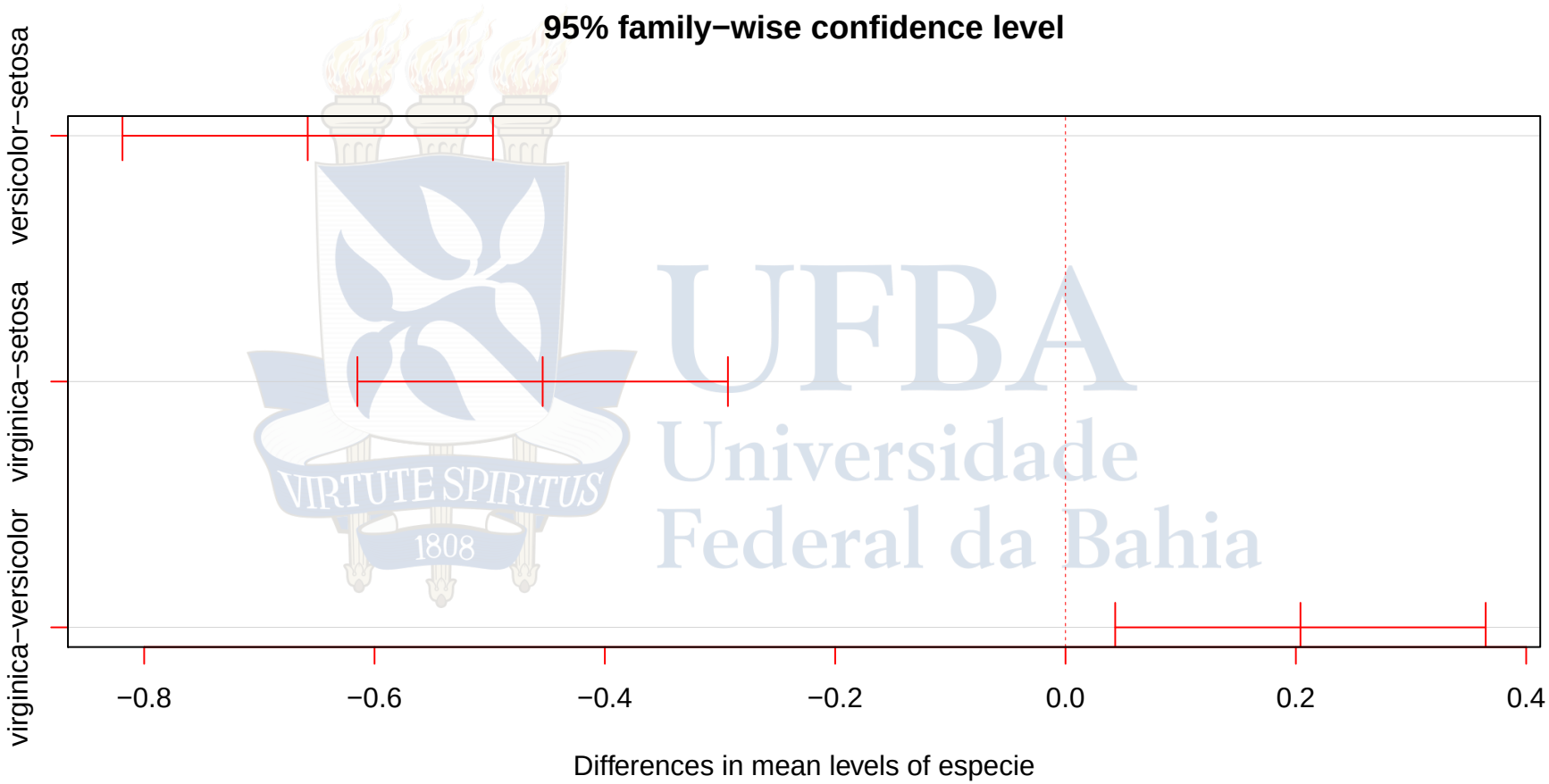
Exemplo

```
tukey <- TukeyHSD(ajuste, conf.level = 0.95)

tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sepala_larg ~ especie, data = dados)
##
## $especie
##                       diff        lwr        upr     p adj
## versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
## virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
## virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

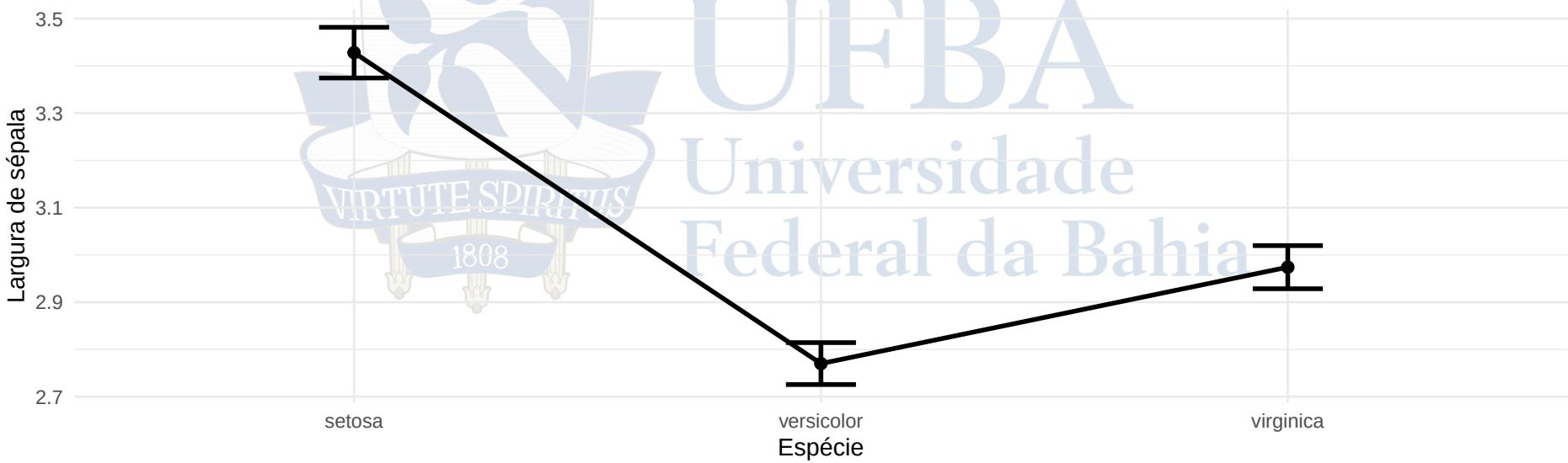
Exemplo

```
plot(tukey, col = "red")
```



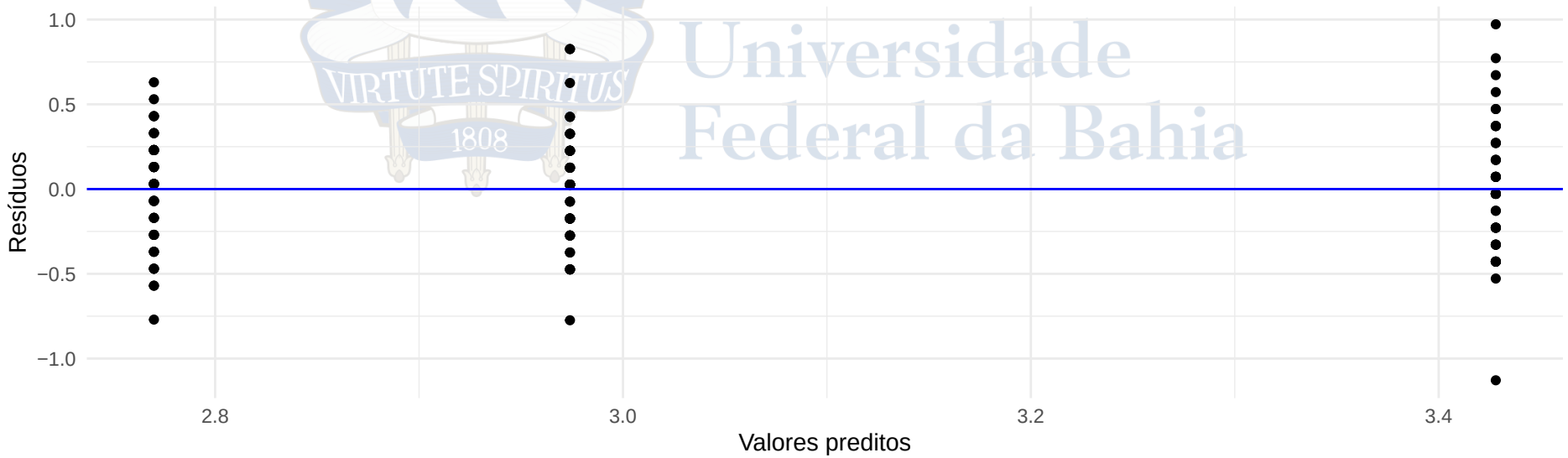
Exemplo

```
ggline(dados, x = "especie", y = "sepala_larg",
       add = "mean_se",
       size = 1,
       ylab = "Largura de sépala", xlab = "Espécie") +
  theme_minimal()
```



Exemplo

```
tab <- augment(ajuste)

ggplot(tab, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 0,
              color = 'blue') +
  labs(x = "Valores preditos", y = "Resíduos") +
  theme_minimal()
```



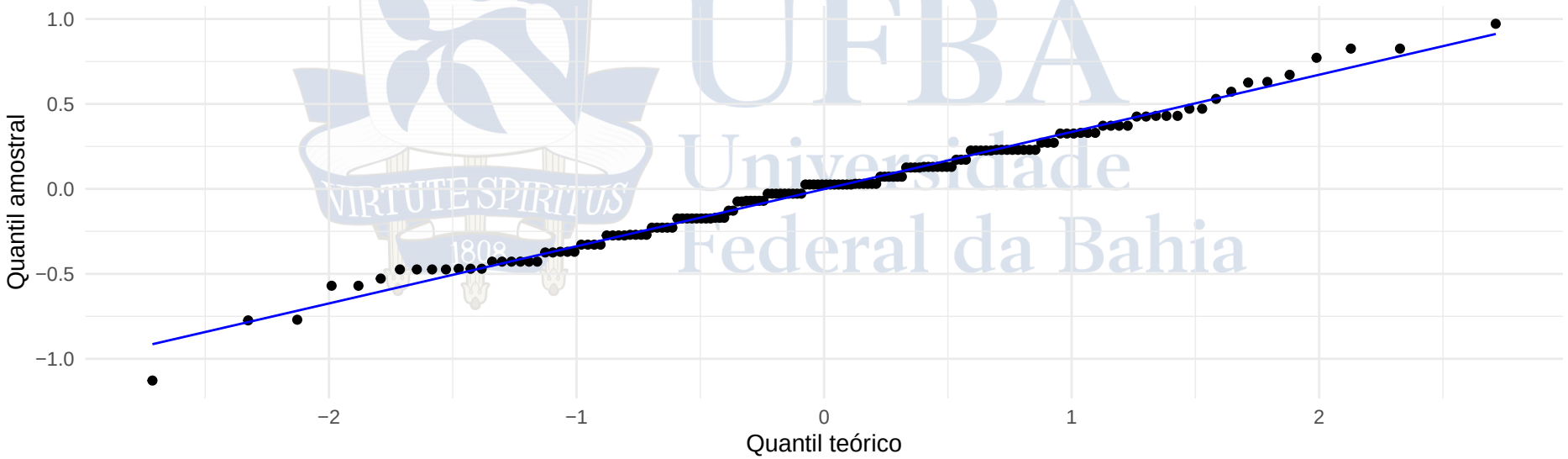
Exemplo

```
shapiro.test(ajuste$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ajuste$residuals
## W = 0.98948, p-value = 0.323
```

Exemplo

```
ggplot(tab, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line(color = "blue") +
  labs(x = "Quantil teórico", y = "Quantil amostral")
  theme_minimal()
```



Exemplo

```
bartlett.test(sepala_larg ~ especie, data = dados)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  sepala_larg by especie
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

```
leveneTest(sepala_larg ~ especie, data = dados)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.5902 0.5555
##       147
```