Introdução ao uso do software R

Fernando de Pol Mayer[1]   Rodrigo Sant'Ana[2]

[1]Laboratório de Estatística Ambiental (LEA)
Instituto de Matemática, Estatística e Física (IMEF)
Universidade Federal do Rio Grande (FURG)
fernando.mayer@furg.br

[2]Instituto Albatroz
oc.rodrigosantana@gmail.com

Sumário

Sumário

Distribuições de probabilidade

A maioria das distribuições de probabilidade tradicionais estão implementadas no R, e podem ser utilizadas para substituir as tabelas estatísticas tradicionais. Existem 4 itens fundamentais que podem ser calculados para cada distribuição:

- d* Calcula a densidade de probabilidade ou probabilidade pontual
- p* Calcula a função de probabilidade acumulada
- q* Calcula o quantil correspondente a uma dada probabilidade
- r* Gera números aleatórios (ou "pseudo-aleatórios")

Distribuições de probabilidade

As distribuições de probabilidade mais comuns são:

Distribuição	Nome no R	Parâmetros
Binomial	`*binom`	`size, prob`
χ^2	`*chisq`	`df`
Normal	`*norm`	`mean, sd`
Poisson	`*pois`	`lambda`
t	`*t`	`df`
Uniforme	`*unif`	`min, max`

Distribuições de probabilidade

Alguns exemplos:

```
# valores críticos de z com alfa = 0,05 (bilateral)
qnorm(0.025)

[1] -1.959964

qnorm(0.975)

[1] 1.959964

# valores críticos de t com diferentes G.L.
qt(0.025, df = 9)

[1] -2.262157

qt(0.025,df = 900)

[1] -1.962603
```

Distribuições de probabilidade

Intervalos de confiança: suponha uma amostra de $n = 5$, com $\bar{x} = 83$ e $s = 12$. Um intervalo de 95% de confiança ($\alpha = 0.05$) para μ pode ser calculado como:

```
## Dados
xbarra <- 83
desvio <- 12
n <- 5
## Erro padrão
erro <- desvio/sqrt(n)
## Média - erro
xbarra + erro * qt(0.025, df = n)

[1] 69.20481

## Média + erro
xbarra + erro * qt(0.975, df = n)

[1] 96.79519
```

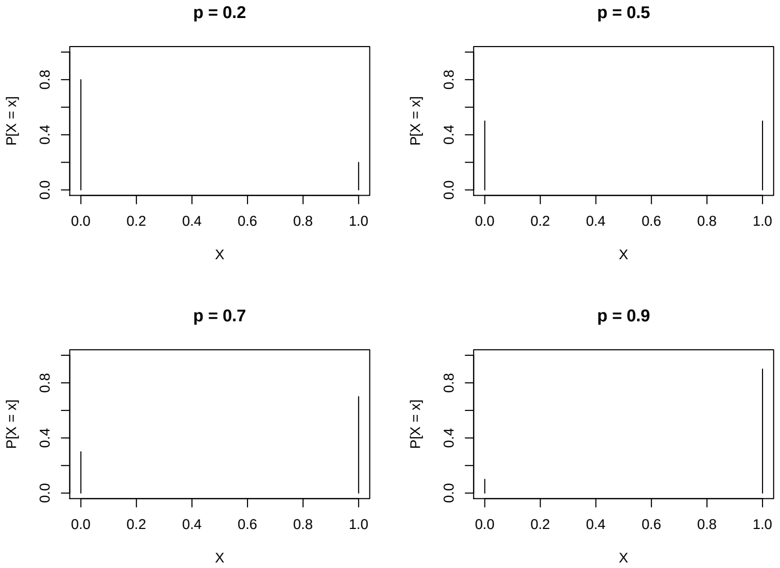
Modelo Bernoulli


```r
dbinom(x = 0:1, size = 1, prob = .2)

[1] 0.8 0.2

dbinom(x = 0:1, size = 1, prob = .5)

[1] 0.5 0.5

dbinom(x = 0:1, size = 1, prob = .7)

[1] 0.3 0.7

dbinom(x = 0:1, size = 1, prob = .9)

[1] 0.1 0.9
```

Modelo Bernoulli

```
par(mfrow=c(2,2))
plot(0:1, dbinom(x = 0:1, size = 1, prob = .2), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "p = 0.2",
     ylim = c(0,1))
plot(0:1, dbinom(x = 0:1, size = 1, prob = .5), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "p = 0.5",
     ylim = c(0,1))
plot(0:1, dbinom(x = 0:1, size = 1, prob = .7), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "p = 0.7",
     ylim = c(0,1))
plot(0:1, dbinom(x = 0:1, size = 1, prob = .9), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "p = 0.9",
     ylim = c(0,1))
```

Modelo Bernoulli



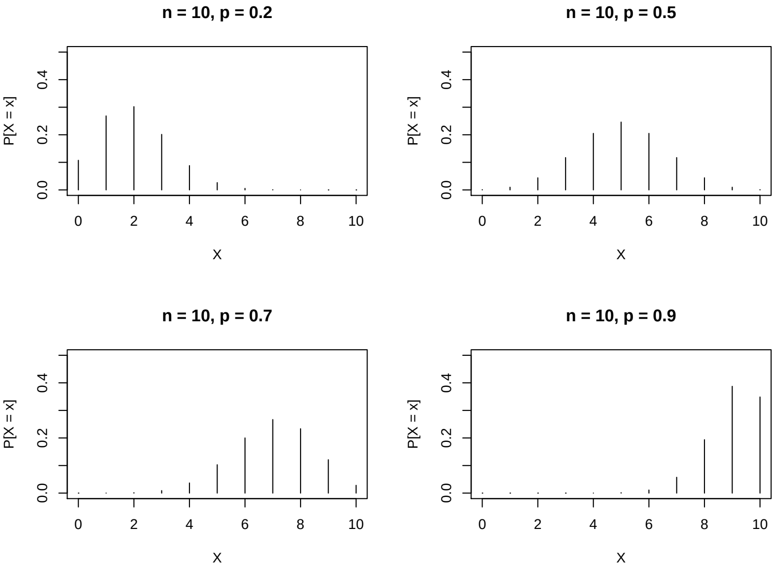
Modelo binomial


```r
dbinom(x = 0:5, size = 5, prob = .2)

[1] 0.32768 0.40960 0.20480 0.05120 0.00640 0.00032

dbinom(x = 0:5, size = 5, prob = .5)

[1] 0.03125 0.15625 0.31250 0.31250 0.15625 0.03125

dbinom(x = 0:5, size = 5, prob = .7)

[1] 0.00243 0.02835 0.13230 0.30870 0.36015 0.16807

dbinom(x = 0:5, size = 5, prob = .9)

[1] 0.00001 0.00045 0.00810 0.07290 0.32805 0.59049
```

Modelo binomial

```
par(mfrow=c(2,2))
plot(0:10, dbinom(x = 0:10, size = 10, prob = .2), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "n = 10, p = 0.2",
     ylim = c(0,.5))
plot(0:10, dbinom(x = 0:10, size = 10, prob = .5), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "n = 10, p = 0.5",
     ylim = c(0,.5))
plot(0:10, dbinom(x = 0:10, size = 10, prob = .7), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "n = 10, p = 0.7",
     ylim = c(0,.5))
plot(0:10, dbinom(x = 0:10, size = 10, prob = .9), type = "h",
     xlab = "X", ylab = "P[X = x]", main = "n = 10, p = 0.9",
     ylim = c(0,.5))
```

Modelo binomial



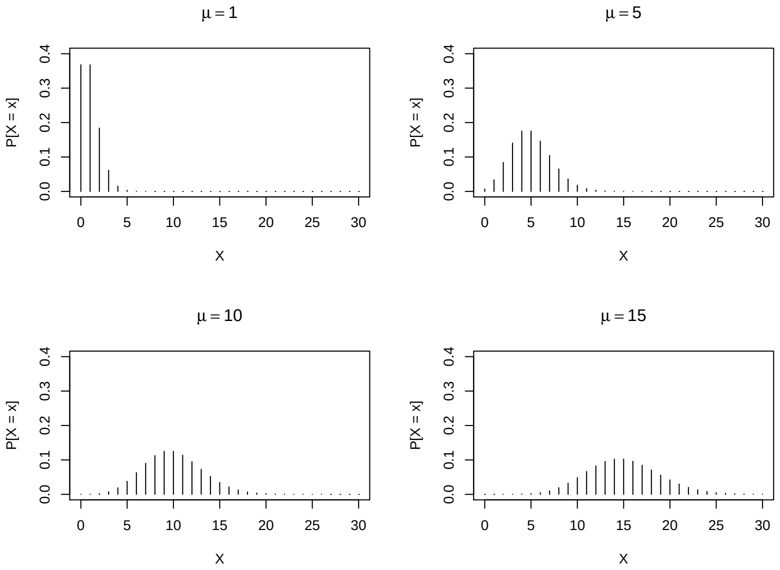
Modelo Poisson

```
dpois(x = 0:10, lambda = 1)

 [1] 3.678794e-01 3.678794e-01 1.839397e-01 6.131324e-02
 [5] 1.532831e-02 3.065662e-03 5.109437e-04 7.299195e-05
 [9] 9.123994e-06 1.013777e-06 1.013777e-07

dpois(x = 0:10, lambda = 5)

 [1] 0.006737947 0.033689735 0.084224337 0.140373896 0.1754673
 [6] 0.175467370 0.146222808 0.104444863 0.065278039 0.0362655
[11] 0.018132789

dpois(x = 0:10, lambda = 10)

 [1] 4.539993e-05 4.539993e-04 2.269996e-03 7.566655e-03
 [5] 1.891664e-02 3.783327e-02 6.305546e-02 9.007923e-02
 [9] 1.125990e-01 1.251100e-01 1.251100e-01

dpois(x = 0:10, lambda = 15)
```

Modelo Poisson

```
par(mfrow=c(2,2))
plot(0:30, dpois(x = 0:30, lambda = 1), type = "h",
     xlab = "X", ylab = "P[X = x]", main = expression(mu == 1),
     ylim = c(0,.4))
plot(0:30, dpois(x = 0:30, lambda = 5), type = "h",
     xlab = "X", ylab = "P[X = x]", main = expression(mu == 5),
     ylim = c(0,.4))
plot(0:30, dpois(x = 0:30, lambda = 10), type = "h",
     xlab = "X", ylab = "P[X = x]", main = expression(mu == 10)
     ylim = c(0,.4))
plot(0:30, dpois(x = 0:30, lambda = 15), type = "h",
     xlab = "X", ylab = "P[X = x]", main = expression(mu == 15)
     ylim = c(0,.4))
```

Modelo Poisson



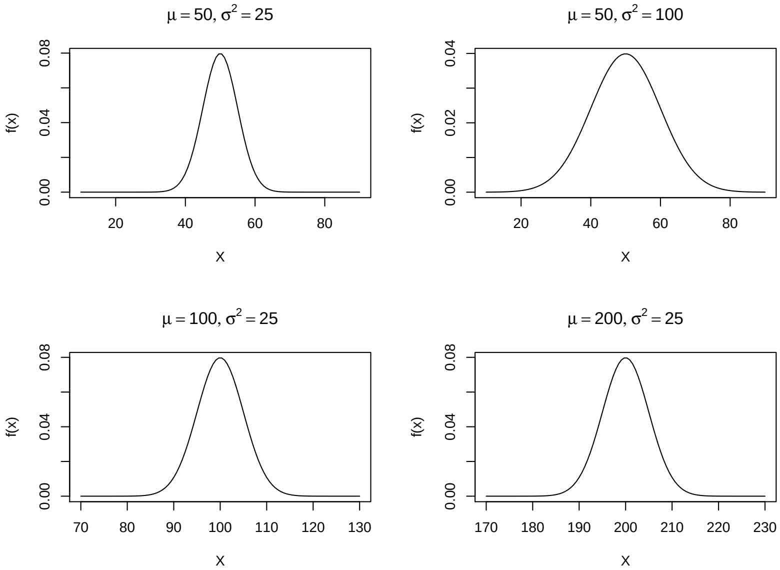
Modelo normal

```
dnorm(x = 40:60, mean = 50, sd = 5)

 [1] 0.01079819 0.01579003 0.02218417 0.02994549 0.03883721
 [6] 0.04839414 0.05793831 0.06664492 0.07365403 0.07820854
[11] 0.07978846 0.07820854 0.07365403 0.06664492 0.05793831
[16] 0.04839414 0.03883721 0.02994549 0.02218417 0.01579003
[21] 0.01079819

dnorm(x = 40:60, mean = 50, sd = 10)

 [1] 0.02419707 0.02660852 0.02896916 0.03122539 0.03332246
 [6] 0.03520653 0.03682701 0.03813878 0.03910427 0.03969525
[11] 0.03989423 0.03969525 0.03910427 0.03813878 0.03682701
[16] 0.03520653 0.03332246 0.03122539 0.02896916 0.02660852
[21] 0.02419707

dnorm(x = 90:110, mean = 100, sd = 5)

 [1] 0.01079819 0.01579003 0.02218417 0.02994549 0.03883721
 [6] 0.04839414 0.05793831 0.06664492 0.07365403 0.07820854
[11] 0.07978846 0.07820854 0.07365403 0.06664492 0.05793831
[16] 0.04839414 0.03883721 0.02994549 0.02218417 0.01579003
[21] 0.01079819
```

Modelo normal

```r
par(mfrow=c(2,2))
plot(seq(10, 90, length=100), type = "l", xlab = "X", ylab = "f(x)",
     y = dnorm(x = seq(10, 90, length=100), mean = 50, sd = 5),
     main = expression(list(mu == 50, sigma^2 == 25)))
plot(seq(10, 90, length=100), type = "l", xlab = "X", ylab = "f(x)",
     y = dnorm(x = seq(10, 90, length=100), mean = 50, sd = 10),
     main = expression(list(mu == 50, sigma^2 == 100)))
plot(seq(70, 130, length=100), type = "l", xlab = "X", ylab = "f(x)",
     y = dnorm(x = seq(70, 130, length=100), mean = 100, sd = 5),
     main = expression(list(mu == 100, sigma^2 == 25)))
plot(seq(170, 230, length=100), type = "l", xlab = "X", ylab = "f(x)",
     y = dnorm(x = seq(170, 230, length=100), mean = 200, sd = 5),
     main = expression(list(mu == 200, sigma^2 == 25)))
par(mfrow=c(1,1))
```

Modelo normal



Sumário

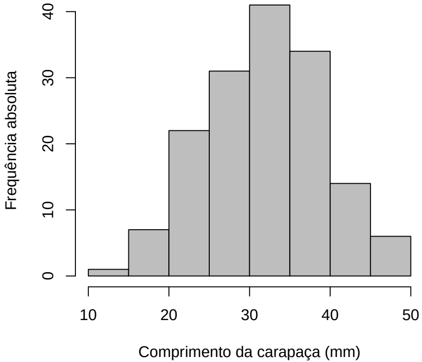
Base de dados

```
dados <- read.table("../dados/crabs.csv", header = T,
                    sep = ";", dec = ",")
str(dados)

'data.frame': 156 obs. of  7 variables:
 $ especie: Factor w/ 2 levels "azul","laranja": 1 1 1 1 1 1 1
 $ sexo   : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ FL     : num  8.1 8.8 9.2 9.6 10.8 11.6 11.8 12.3 12.6 12.8
 $ RW     : num  6.7 7.7 7.8 7.9 9 9.1 10.5 11 10 10.9 ...
 $ CL     : num  16.1 18.1 19 20.1 23 24.5 25.2 26.8 27.7 27.4
 $ CW     : num  19 20.8 22.4 23.1 26.5 28.4 29.3 31.5 31.7 31.
 $ BD     : num  7 7.4 7.7 8.2 9.8 10.4 10.3 11.4 11.4 11 ...
```

Testes de hipótese

Teste-t para uma amostra

```r
hist(dados$CL, main = "", ylab = "Frequência absoluta",
     xlab = "Comprimento da carapaça (mm)", col = "grey")
```



Procedimentos gerais para um teste de hipótese

(1) Definir a hipótese nula (H_0) e a alternativa (H_1)

(2) Definir um nível de **significância** α (ex.: $\alpha = 0,05$), que irá determinar o nível de **confiança** $100(1 - \alpha)\%$ do teste

(3) Determinar a **região de rejeição** com base no nível de significância $\rightarrow t_{crit}$

(4) Calcula a **estatística de teste**, sob a hipótese nula

$$t_{calc} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

(5) Rejeitar a hipótese nula se a estatística de teste calculada estiver dentro da região de rejeição ($t_{calc} > t_{crit}$)

- Alternativamente, calcula-se o p-valor, que é a probabilidade de se obter um valor de t igual ou maior do que t_{calc}

Testes de hipótese

Teste-t para uma amostra

- Testar a hipótese de que a média (μ) de CL é igual a 30 mm (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu = 30$$
$$H_1 : \mu \neq 30$$

Testes de hipótese

Teste-t para uma amostra

```
t.test(dados$CL, mu = 30, alternative = "two.sided",
       conf.level = 0.95)


One Sample t-test

data:  dados$CL
t = 3.4627, df = 155, p-value = 0.0006913
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 30.86071 33.14698
sample estimates:
mean of x
 32.00385
```

Fazendo manualmente

```
## Dados
xbarra <- mean(dados$CL)
mu0 <- 30
dp <- sd(dados$CL)
n <- nrow(dados)
# t calculado
(tcalc <- (xbarra - mu0)/(dp/sqrt(n)))

[1] 3.462731

# t critico (não é apresentado no resultado)
qt(0.025, df = n - 1, lower.tail = FALSE)

[1] 1.975387

# valor p (multiplicado por 2 pois o teste é bilateral)
pt(tcalc, df = n - 1, lower.tail = FALSE) * 2

[1] 0.000691346
```

Detalhe: O teste pode ser armazenado em um objeto para futuras referências

```
teste <- t.test(dados$CL, mu = 30, alternative = "two.sided",
                conf.level = 0.95)
names(teste)

[1] "statistic"   "parameter"   "p.value"     "conf.int"
[5] "estimate"    "null.value"  "alternative" "method"
[9] "data.name"

teste$statistic

       t
3.462731

teste$p.value

[1] 0.000691346
```

- Testar a hipótese de que a média (μ) de CL é menor ou igual a 30 mm (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu \leq 30$$
$$H_1 : \mu > 30$$

Módulo III
Inferência e Modelagem
IMEF 2014
Distribuições de probabilidade
Inferência
Regressão e correlação
Regressão
Estimação
Correlação
ANOVA
MLGs

```
t.test(dados$CL, mu = 30, alternative = "greater",
       conf.level = 0.95)


One Sample t-test

data:  dados$CL
t = 3.4627, df = 155, p-value = 0.0003457
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
 31.04626      Inf
sample estimates:
mean of x
 32.00385
```

- Testar a hipótese de que a média (μ) de CL é maior ou igual a 30 mm (com 95% de confiança)
- As hipóteses são

$$H_0 : \mu \geq 30$$
$$H_1 : \mu < 30$$

Testes de hipótese

Teste-t para uma amostra

```r
t.test(dados$CL, mu = 30, alternative = "less",
       conf.level = 0.95)
```

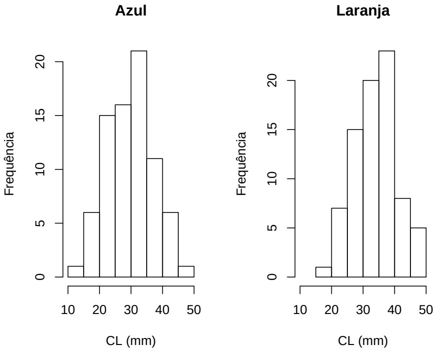
```
One Sample t-test

data:  dados$CL
t = 3.4627, df = 155, p-value = 0.9997
alternative hypothesis: true mean is less than 30
95 percent confidence interval:
     -Inf 32.96143
sample estimates:
mean of x
 32.00385
```

Faça dois histogramas lado-a-lado da medida CL para cada uma das espécies.

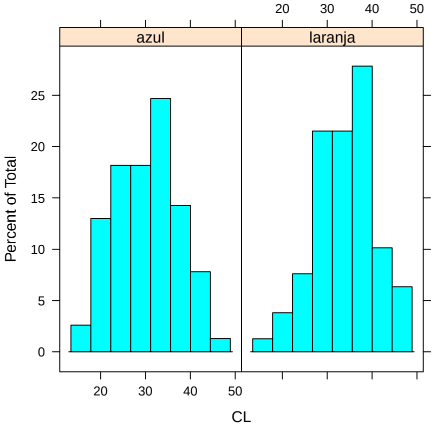
Testes de hipótese

Teste-t para duas amostras

```
par(mfrow = c(1,2))
hist(dados$CL[dados$especie == "azul"], main = "Azul",
     xlab = "CL (mm)", ylab = "Frequência", xlim = c(10, 50))
hist(dados$CL[dados$especie == "laranja"], main = "Laranja",
     xlab = "CL (mm)", ylab = "Frequência", xlim = c(10, 50))
par(mfrow = c(1,1))
```



```r
require(lattice) # pacote para gráficos avançados
histogram(~CL | especie, data = dados)
```



```
with(dados, tapply(CL, especie, summary))

$azul
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.70   24.60   30.10   29.87   34.50   47.10

$laranja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.70   29.40   34.50   34.08   39.25   47.60
```

Existem evidências de que uma espécie é maior do que a outra?

Testes de hipótese

Teste-t para duas amostras

- Testar a hipótese de que a **diferença** entre a média de CL da espécie azul (μ_A) e a média de CL da espécie laranja (μ_L) é igual a 0 (zero) (com 95% de confiança)
- As hipóteses são

$$\begin{aligned} H_0 : \mu_A - \mu_L = 0 \quad &\Rightarrow \quad \mu_A = \mu_L \\ H_1 : \mu_A - \mu_L \neq 0 \quad &\Rightarrow \quad \mu_A \neq \mu_L \end{aligned}$$

```
t.test(CL ~ especie, data = dados, mu = 0,
       alternative = "two.sided", conf.level = 0.95)


Welch Two Sample t-test

data:  CL by especie
t = -3.7935, df = 152.732, p-value = 0.0002135
alternative hypothesis: true difference in means is not equal t
95 percent confidence interval:
 -6.411592 -2.020366
sample estimates:
   mean in group azul mean in group laranja
             29.86883              34.08481
```

- Testar a hipótese de que a **diferença** entre a média de CL da espécie azul (μ_A) e a média de CL da espécie laranja (μ_L) é **menor** ou iual a 0 (zero) (com 95% de confiança)
- Em outras palavras: "O CL médio é menor para a espécie azul?"
- As hipóteses são

$$\begin{aligned} H_0 : \mu_A - \mu_L \leq 0 \quad &\Rightarrow \quad \mu_A \leq \mu_L \\ H_1 : \mu_A - \mu_L > 0 \quad &\Rightarrow \quad \mu_A > \mu_L \end{aligned}$$

```
t.test(CL ~ especie, data = dados, mu = 0,
       alternative = "greater", conf.level = 0.95)


Welch Two Sample t-test

data:  CL by especie
t = -3.7935, df = 152.732, p-value = 0.9999
alternative hypothesis: true difference in means is greater tha
95 percent confidence interval:
 -6.055151       Inf
sample estimates:
   mean in group azul mean in group laranja
             29.86883              34.08481
```

Como você faria para calcular a diferença observada das médias de CL entre as duas espécies?

Exercícios

Com base no objeto dados:

(1) Faça um histograma de CW

(2) Com base no histograma, construa uma hipótese para a média de CW
   (a) Teste a igualdade dessa hipótese
   (b) Teste uma desigualdade dessa hipótese

   Em ambos os casos use um nível de confiança de 90%, e escreva uma frase com a sua conclusão.

(3) Faça um histograma de CW para cada sexo

(4) Com base nesses histogramas, construa uma hipótese para a diferença média de CW entre os sexos
   (a) Teste a igualdade dessa hipótese
   (b) Teste uma desigualdade dessa hipótese

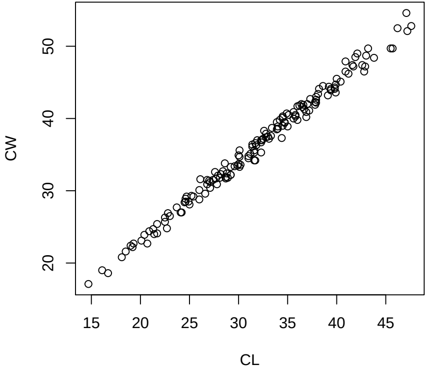
   Em ambos os casos use um nível de confiança de 90%, e escreva uma frase com a sua conclusão.

Sumário

Regressão e correlação

Vamos analisar a relação que existe entre CL e CW

```r
plot(CW ~ CL, data = dados)
```



Regressão e correlação

Um **modelo linear** entre duas variáveis X e Y, é definido matematicamente como uma equação com dois parâmetros desconhecidos,

$$Y = \beta_0 + \beta_1 X$$

A **análise de regressão** é a técnica estatística que analisa as relações existentes entre uma única variável **dependente**, e uma ou mais variáveis **independentes**

O objetivo é estudar as relações entre as variáveis, a partir de um **modelo matemático**, permitindo **estimar** o valor de uma variável a partir da outra

- Exemplo: sabendo a altura podemos determinar o peso de uma pessoa, se conhecemos os parâmetros do modelo anterior

Regressão linear

O problema da análise de regressão consiste em definir a **forma** de relação existente entre as variáveis.

Por exemplo, podemos ter as seguintes relações

$Y = \beta_0 + \beta_1 X$	linear
$Y = \beta_0 X^{\beta_1}$	potência
$Y = \beta_0 e^{\beta_1 X}$	exponencial
$Y = \beta_0 + \beta_1 X + \beta_2 X^2$	polinomial

Em todos os casos, a variável **dependente** é Y, aquela que será **predita** a partir da relação e da variável **independente** X

Sumário

Regressão linear

Em uma **análise de regressão linear** consideraremos apenas as variáveis que possuem uma **relação linear** entre si.

Uma análise de regressão linear **múltipla** pode associar k variáveis independentes (X) para "explicar" uma única variável dependente (Y),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

Uma análise de regressão linear **simples** associa uma única variável independente (X) com uma variável dependente (Y),

$$Y = \beta_0 + \beta_1 X + e$$

Regressão linear

Assim, dados n pares de valores, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, se for admitido que Y é função linear de X, pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, 2, \ldots, n$$

onde:

- Y é a variável **resposta** (ou **dependente**)
- X é a variável **explicativa** (ou **independente**)
- β_0 é o **intercepto** da reta (valor de Y quando $X = 0$)
- β_1 é o **coeficiente angular** da reta (**efeito** de X sobre Y)
- $e_i \sim \mathrm{N}(0, \sigma^2)$ é o **erro**, ou **desvio**, ou **resíduo**

O problema agora consiste em **estimar** os parâmetros β_0 e β_1.

Regressão linear

Interpretação dos parâmetros:

β_0 representa o ponto onde a reta corta o eixo Y (na maioria das vezes não possui interpretação prática)

β_1 representa a variabilidade em Y causada pelo aumento de uma unidade em X. Além disso,

- $\beta_1 > 0$ mostra que com o aumento de X, também há um aumento em Y
- $\beta_1 = 0$ mostra que **não há efeito** de X sobre Y
- $\beta_1 < 0$ mostra que com a aumento de X, há uma diminuição em Y

Estimação dos parâmetros

Como através de uma amostra obtemos uma estimativa da verdadeira equação de regressão, denominamos

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

ou seja, $\hat{Y}_i$ é o valor **estimado** de Y_i, através das **estimativas** de β_0 e β_1, que chamaremos de $\hat{\beta}_0$ e $\hat{\beta}_1$.

Para cada valor de Y_i, temos um valor $\hat{Y}_i$ estimado pela equação de regressão,

$$Y_i = \hat{Y}_i + e_i$$

Estimação dos parâmetros

Portanto, o erro (ou desvio) de cada observação em relação ao modelo adotado será

$$e_i = Y_i - \hat{Y}_i$$
$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Devemos então adotar um modelo cujos parâmetros β_0 e β_1, tornem esse diferença a menor possível.

Isso equivale a **minimizar** a **soma de quadrados dos resíduos** (SQR), ou do erro,

$$SQR = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

Estimação dos parâmetros

O método de minimizar a soma de quadrados dos resíduos é denominado de **método dos mínimos quadrados**.

Para se encontrar o ponto mínimo de uma função, temos que obter as derivadas parciais em relação a cada parâmetro,

$$\frac{\partial SQR}{\partial \beta_0} = 2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X_i](-1)$$

$$\frac{\partial SQR}{\partial \beta_1} = 2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X_i](-X_i)$$

e igualar os resultados a zero

$$\hat{\beta}_0 = \frac{\partial SQR}{\partial \beta_0} = 0 \qquad \text{e} \qquad \hat{\beta}_1 = \frac{\partial SQR}{\partial \beta_1} = 0$$

Estimação dos parâmetros

Dessa forma, chegamos às **estimativas de mínimos quadrados** para os parâmetros β_0 e β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

onde

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{e} \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ajustando um modelo linear no R

```
mod <- lm(CW ~ CL, data = dados)
mod


Call:
lm(formula = CW ~ CL, data = dados)

Coefficients:
(Intercept)           CL
      1.187        1.097
```

Módulo III
Inferência e Modelagem

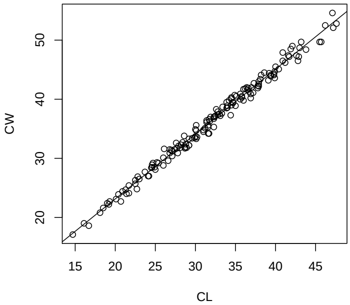
IMEF 2014

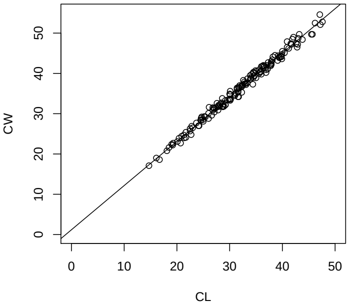
Distribuições de probabilidade
Inferência
Regressão e correlação
Regressão
Estimação
Correlação
ANOVA
MLGs

```
summary(mod)


Call:
lm(formula = CW ~ CL, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7762 -0.5699  0.1098  0.4629  1.8273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.186950   0.285340    4.16 5.28e-05 ***
CL          1.097451   0.008698  126.17  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7827 on 154 degrees of freedom
Multiple R-squared:  0.9904,Adjusted R-squared:  0.9904
F-statistic: 1.592e+04 on 1 and 154 DF,  p-value: < 2.2e-16
```

```
anova(mod)

Analysis of Variance Table

Response: CW
           Df Sum Sq Mean Sq F value    Pr(>F)
CL          1 9752.6  9752.6   15919 < 2.2e-16 ***
Residuals 154   94.3     0.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

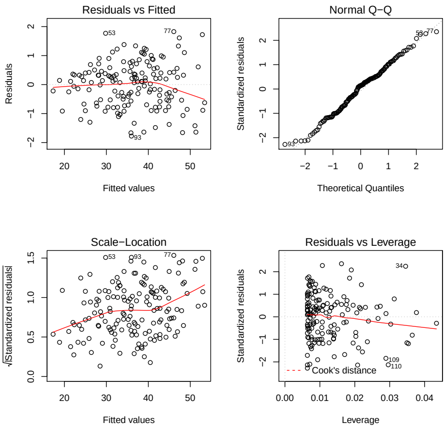
```r
plot(CW ~ CL, data = dados)
abline(mod)
plot(CW ~ CL, data = dados, xlim = c(0,50), ylim = c(0,55))
abline(mod)
```





Regressão
Análise dos resíduos

```
par(mfrow = c(2,2))
plot(mod)
par(mfrow = c(1,1))
```



Regressão

Acessando os componentes do objeto mod:

```
names(mod)

 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"

names(summary(mod))

 [1] "call"          "terms"         "residuals"
 [4] "coefficients"  "aliased"       "sigma"
 [7] "df"            "r.squared"     "adj.r.squared"
[10] "fstatistic"    "cov.unscaled"

names(anova(mod))

[1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

Regressão

Veja que o `Residual standard error:  0.7827` é o estimador do desvio-padrão residual $\hat{\sigma}_e^2 = \frac{\text{SQRes}}{n-2}$, ou seja,

```r
sqrt(anova(mod)$Sum[2]/anova(mod)$Df[2])

[1] 0.7827079
```

e que `F-statistic:  1.592e+04` (15920) é o mesmo valor de

```r
anova(mod)$F[1]

[1] 15919.11
```

que testa a mesma hipótese da ANOVA. De fato, o valor de t^2 para β_1 no sumário do modelo é

```r
summary(mod)$coef[2,3]^2

[1] 15919.11
```

Sumário

Correlação

Até agora o interesse estava em estudar qual a influência de uma V.A. X sobre uma V.A. Y, por meio de uma **relação linear**.

Assim, em uma análise de regressão é indispensável identificar qual variável é dependente.

Na **análise de correlação** isto não é necessário, pois queremos estudar o **grau de relacionamento** entre as variáveis X e Y, ou seja, uma medida de **covariabilidade** entre elas.

A correlação é considerada como uma medida de **influência mútua** entre variáveis, por isso não é necessário especificar quem influencia e quem é influenciado.

Correlação

O **grau de relação** entre duas variáveis pode ser medido através do **coeficiente de correlação linear** (r), dado por

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}}}$$

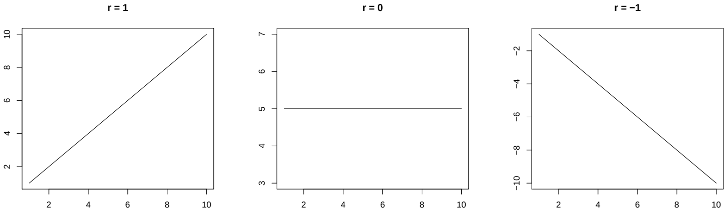
onde

$$-1 \leq r \leq 1$$

Portanto,

- $r = 1$ correlação **positiva** perfeita entre as variáveis
- $r = 0$ **não há** correlação entre as variáveis
- $r = -1$ correlação **negativa** perfeita entre as variáveis

Correlação



Correlação

O **coeficiente de determinação** (r^2) é o quadrado do coeficiente de correlação, por consequência

$$0 \leq r^2 \leq 1$$

O r^2 nos dá a **porcentagem de variação em Y que pode ser explicada pela variável independente X.**

Quanto mais próximo de 1, maior é a explicação da variável Y pela variável X.

Correlação

```
Error in eval(expr, envir, enclos):  cannot open file
'/home/fernando/GitHub/R-rautu/modulo_III/cache/unnamed-chunk-3
No such file or directory
Warning:  restarting interrupted promise evaluation
Error in eval(expr, envir, enclos):  cannot open file
'/home/fernando/GitHub/R-rautu/modulo_III/cache/unnamed-chunk-3
No such file or directory
Warning:  restarting interrupted promise evaluation
Error in eval(expr, envir, enclos):  cannot open file
'/home/fernando/GitHub/R-rautu/modulo_III/cache/unnamed-chunk-3
No such file or directory
Error in do.call(expression, lis):  cannot open file
'/home/fernando/GitHub/R-rautu/modulo_III/cache/unnamed-chunk-3
No such file or directory
Warning:  restarting interrupted promise evaluation
Error in eval(expr, envir, enclos):  cannot open file
'/home/fernando/GitHub/R-rautu/modulo_III/cache/unnamed-chunk-3
No such file or directory
Warning:  restarting interrupted promise evaluation
Error in eval(expr, envir, enclos):  cannot open file
```

Exercícios

Com as colunas BD e CL do objeto `dados`

(1) Faça um gráfico da relação entre estas variáveis

(2) Faça um teste de correlação

(3) Ajuste um modelo linear

  (a) Veja o sumário

  (b) Ajuste a linha do modelo no gráfico

  (c) Verifique os resíduos

Qual sua conclusão?

- Existe correlação significativa? De que tipo (positiva, negativa)?
- O modelo linear descreve bem a relação entre estas duas variáveis (verifique com o valor de `Pr(>|t|)` e do R^2)
- O modelos foi bem ajustado aos dados (observe os resíduos)

Sumário

Análise de Variância

Definição: y_{ij} representa a observação j do grupo i; $\bar{y}_i$ é a média do grupo i; $\bar{y}$ é a média geral de todas as observações. As observações podem ser decompostas em

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

que corresponde ao modelo

$$y_{ij} = \theta + \mu_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

A hipótese a ser testada de que todos os grupos são iguais (*i.e* médias iguais) implica que todos os μ_i são iguais:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$
$$H_1 : \text{pelo menos um } \mu_i \text{ é diferente dos demais}$$

Análise de Variância

Voltando ao exemplo da diferença de CL entre as duas espécies:
$\bar{y}_A = 29.9$ e $\bar{y}_L = 34.1$

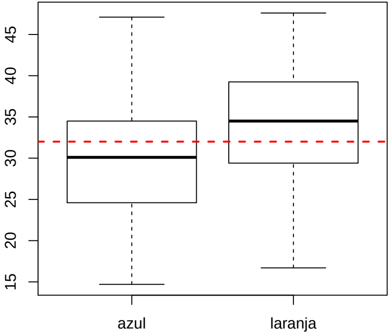
```
with(dados, tapply(CL, especie, summary))

$azul
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.70   24.60   30.10   29.87   34.50   47.10

$laranja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.70   29.40   34.50   34.08   39.25   47.60
```

Média geral $\bar{y} = 32$

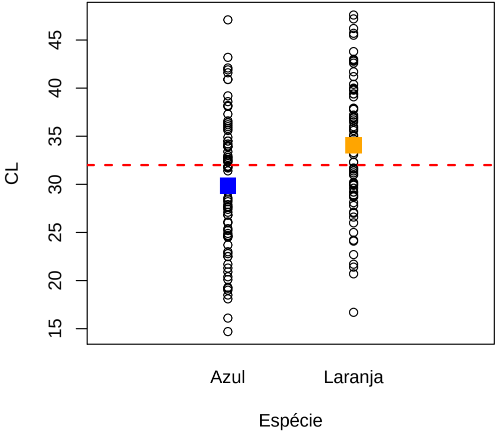
```
mean(dados$CL)

[1] 32.00385
```

Análise de Variância

```r
boxplot(CL ~ especie, data = dados)
abline(h = mean(dados$CL), lty = 2, col = "red", lwd = 2)
```



Geometricamente



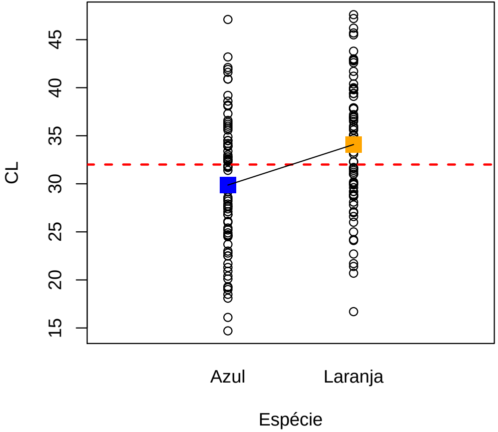
Análise de Variância

Podemos ajustar um modelo linear entre CL e espécie

```
mod <- lm(CL ~ especie, data = dados)
summary(mod)
```

```
Call:
lm(formula = CL ~ especie, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3848  -5.0188   0.2732   5.0192  17.2312

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     29.8688     0.7902  37.799  < 2e-16 ***
especielaranja   4.2160     1.1104   3.797  0.00021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.934 on 154 degrees of freedom
Multiple R-squared:  0.08559,Adjusted R-squared:  0.07966
F-statistic: 14.42 on 1 and 154 DF,  p-value: 0.0002104
```

Ajustando o modelo



Você lembra do teste-t feito anteriormente?

```
teste <- t.test(CL ~ especie, data = dados, mu = 0,
                alternative = "two.sided", conf.level = 0.95)
teste


Welch Two Sample t-test

data:  CL by especie
t = -3.7935, df = 152.732, p-value = 0.0002135
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.411592 -2.020366
sample estimates:
   mean in group azul mean in group laranja
             29.86883              34.08481
```

Módulo III
Inferência e Modelagem

IMEF 2014

Distribuições de probabilidade
Inferência
Regressão e correlação
Regressão
Estimação
Correlação
ANOVA
MLGs

Notou a relação?

```
summary(mod)$coefficients

                 Estimate Std. Error  t value     Pr(>|t|)
(Intercept)     29.868831  0.7902012 37.79902 8.192364e-80
especielaranja   4.215979  1.1104178  3.79675 2.104221e-04

teste$p.value

[1] 2.135202e-04

teste$estimate

   mean in group azul mean in group laranja
             29.86883              34.08481

diff(teste$estimate)

mean in group laranja
             4.215979
```

Análise de Variância

A ANOVA vai testar apenas a hipótese inicial

$$H_0 : \mu_A = \mu_L$$
$$H_1 : \mu_A \neq \mu_L$$

```
anova(mod)

Analysis of Variance Table

Response: CL
           Df Sum Sq Mean Sq F value    Pr(>F)
especie     1  693.1  693.09  14.415 0.0002104 ***
Residuals 154 7404.3   48.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aqui a única conclusão é de que os μ_i não são iguais (mas você não sabe quanto e nem quais!)

Análise de Variância

Se olharmos apenas o resultado da ANOVA, podemos prosseguir com a análise fazendo um teste *a posteriori* para verificarmos quais são os grupos que diferem entre si. Um deles é o teste de Tukey

```
mod.anova <- aov(CL ~ especie, data = dados)
TukeyHSD(mod.anova)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = CL ~ especie, data = dados)

$especie
                 diff      lwr      upr     p adj
laranja-azul 4.215979 2.022362 6.409596 0.0002104
```

Análise de Variância

Porque então fazer uma ANOVA???

- Quando formos comparar a média de mais de 2 grupos
- Não é possível fazer um teste-t para mais de 2 grupos
- Por exemplo, com 3 grupos (A, B, C) teríamos que fazer 3 comparações (A:B, A:C, B:C)
  - Com um nível de confiança de 95% ($\alpha = 0.05$) para cada teste, os 3 testes teriam um nível de confiança $(1 - \alpha)^3$
  - Portanto $(1 - 0.05)^3 = (0.95)^3 = 0.85$
  - Isso implica que quanto mais comparações forem feitas, menor será seu nível de confiança no resultado dos testes.

Sumário

Modelos Lineares Generalizados

Nelder e Wedderburn (1972) mostraram que uma série de técnicas estatísticas podem ser formuladas de forma unificada, como uma classe de modelos de regressão. A essa teoria, uma extensão dos modelos clássicos de regressão, deram o nome de **Modelos Lineares Generalizados**.

Teste-t $\subset$ ANOVA $\subset$ ANCOVA* $\subset$ ML $\subset$ ML-MULT* $\subset$ MLG

- Teste-t: compara uma ou duas médias
- ANOVA: compara 2 ou mais médias (fator)
- ANCOVA: compara 2 ou mais médias (fator) + variáveis numéricas
- ML: regressão de y (numérico) em função de um único x (numérico ou fator)
- ML-MULT: regressão de y (numérico) em função de mais de um x (numéricos ou fatores)
- MLG: Similar ao ML-MULT, mas extende o modelo para que y possa ser um fator ou ter uma distribuição diferente da normal.

Modelos Lineares Generalizados

Os MLGs são formados por três componentes:

- **Componente aleatório**: a variável resposta do modelo, com distribuição pertencente à família de distribuições exponencial.
- **Componente sistemático**: as variáveis explicativas, que entram na forma de uma estrutura linear.
- **Função de ligação**: função que liga os componentes aleatório e sistemático.

De maneira geral, os MLGs descrevem a relação entre a variável resposta y_i $(i = 1, \ldots, n)$ através de preditores x_i. A média de y_i condicionada aos preditores x_i é

$$E(y_i|x_i) = \mu_i$$

e existe uma transformação de μ_i de forma que

$$g(\mu_i) = x_i^T \beta$$

onde $g(\cdot)$ é uma função de ligação conhecida, e β é o vetor de parâmetros a ser estimado.

Módulo III
Inferência e
Modelagem

IMEF 2014

Distribuições de probabilidade
Inferência
Regressão e correlação
Regressão
Estimação
Correlação
ANOVA
MLGs

Distribuições da família exponencial e funções de ligação (P = link padrão)

Link	binomial	poisson	negative binomial	Gamma	gaussian	inverse gaussian
logit	P					
probit	•					
cloglog	•					
identity		•	•	•	P	
inverse				P		
log		P	P	•		
1/mu^2						P
sqrt		•	•			

Para ajustar um MLG usamos a função `glm()`

```
mod.glm <- glm(CL ~ especie, data = dados,
               family = gaussian(link = "identity"))
summary(mod.glm)


Call:
glm(formula = CL ~ especie, family = gaussian(link = "identity"),
    data = dados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-17.3848   -5.0188    0.2732    5.0192   17.2312

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     29.8688     0.7902  37.799  < 2e-16 ***
especielaranja   4.2160     1.1104   3.797  0.00021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 48.08018)
```

Modelos Lineares Generalizados

Quando existe mais de uma variável resposta (y)? **Métodos multivariados**!

Exercícios

Com o objeto dados

(1) Faça um boxplot de CW por sexo

(2) Faça um teste-t para testar se existe diferença entre as médias de CW para machos e fêmeas

(3) Ajuste um modelo linear para testar essa mesma hipótese

(4) Faça uma ANOVA e o teste de Tukey

Qual sua conclusão?