



Introdução ao uso do software R

Fernando de Pol Mayer¹ Rodrigo Sant'Ana²

¹Universidade Federal de Santa Catarina (UFSC)
Departamento de Ecologia e Zoologia (ECZ/CCB)
fernando.mayer@gmail.com

²Instituto Albatroz
oc.rodigosantana@gmail.com

26 e 27 de Novembro, 2012



Sumário

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Correlação e regressão
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



Sumário

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Correlação e regressão
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



Distribuições de probabilidade

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

As distribuições de probabilidade mais comuns são:

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



Distribuições de probabilidade

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Alguns exemplos:

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

```
[1] -1.96
```

```
> qnorm(0.975)
```

```
[1] 1.96
```

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

```
[1] -2.2622
```

```
> qt(0.025, df = 900)
```

```
[1] -1.9626
```



Sumário

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Correlação e regressão
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



Base de dados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

```
> dados <- read.table("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 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 31.9 ...  
 $ 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

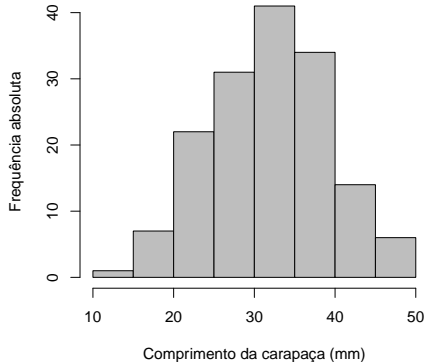
Inferência

Correlação e
regressão

ANOVA

MLGs

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





Testes de hipótese

Teste-t para uma amostra

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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 Calcular a **estatística de teste**, sob a hipótese nula

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

- 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

```
> 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:

30.861 33.147

sample estimates:

mean of x

32.004



Testes de hipótese

Teste-t para uma amostra

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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  
3.4627  
  
> teste$p.value  
  
[1] 0.00069135
```



Testes de hipótese

Teste-t para uma amostra

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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

$$H_0 : \mu \leq 30$$

$$H_1 : \mu > 30$$



Testes de hipótese

Teste-t para uma amostra

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

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:

31.046 Inf

sample estimates:

mean of x

32.004



Testes de hipótese

Teste-t para uma amostra

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- Testar a hipótese de que a média (μ) de CL é maior do que 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

```
> 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 32.961

sample estimates:

mean of x

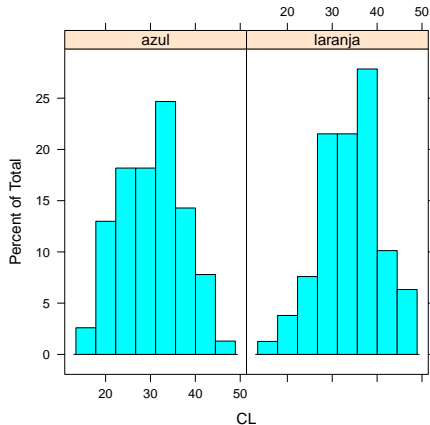
32.004



Testes de hipótese

Teste-t para duas amostras

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





Testes de hipótese

Teste-t para duas amostras

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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

\$azul

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14.7	24.6	30.1	29.9	34.5	47.1

\$laranja

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.7	29.4	34.5	34.1	39.2	47.6

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



Testes de hipótese

Teste-t para duas amostras

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 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

$$H_0 : \mu_A - \mu_L = 0 \quad \Rightarrow \quad \mu_A = \mu_L$$

$$H_1 : \mu_A - \mu_L \neq 0 \quad \Rightarrow \quad \mu_A \neq \mu_L$$



Testes de hipótese

Teste-t para duas amostras

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

```
> 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 = 152.73, p-value = 0.0002135

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-6.4116 -2.0204

sample estimates:

mean in group azul mean in group laranja

29.869

34.085



Testes de hipótese

Teste-t para duas amostras

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 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** do que 0 (zero) (com 95% de confiança)
- Em outras palavras: “O CL médio é menor para a espécie azul?”
- As hipóteses são

$$H_0 : \mu_A - \mu_L \leq 0 \quad \Rightarrow \quad \mu_A \leq \mu_L$$

$$H_1 : \mu_A - \mu_L > 0 \quad \Rightarrow \quad \mu_A > \mu_L$$



Testes de hipótese

Teste-t para duas amostras

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

```
> 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 = 152.73, p-value = 0.9999

alternative hypothesis: true difference in means is greater than

95 percent confidence interval:

-6.0552 Inf

sample estimates:

mean in group azul mean in group laranja

29.869

34.085

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



Exercícios

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Com base no objeto dados:

- ❶ Faça um histograma de CW
- ❷ Com base no histograma, construa uma hipótese para a média de CW

- ❶ Teste a igualdade dessa hipótese
- ❷ Teste uma desigualdade dessa hipótese

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

- ❸ Faça um histograma de CW para cada sexo
 - ❹ Com base nesses histogramas, construa uma hipótese para a diferença média de CW entre os sexos
- ❶ Teste a igualdade dessa hipótese
 - ❷ 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA
MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Correlação e regressão
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



Correlação e regressão

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

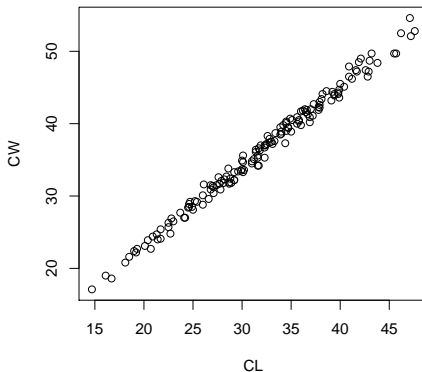
Correlação e
regressão

ANOVA

MLGs

Vamos analisar a correlação que existe entre CL e CW

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

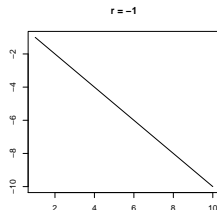
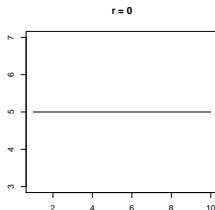
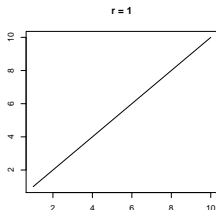




Correlação

A correlação entre duas variáveis é simbolizada por ρ (para a população) e r (para a amostra), e varia no intervalo $[-1, 1]$

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



Portanto, um **teste de correlação** tem as seguintes hipóteses

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$



Correlação

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Teste de correlação entre CL e CW

```
> cor(dados$CL, dados$CW)
```

```
[1] 0.9952
```

```
> cor.test(dados$CL, dados$CW)
```

Pearson's product-moment correlation

data: dados\$CL and dados\$CW

t = 126.17, df = 154, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.99341 0.99650

sample estimates:

cor

0.9952



O modelo linear é definido por:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

onde

- y é a variável resposta
- x é a variável explicativa
- β_0 é o intercepto da reta (valor de y quando $x = 0$)
- β_1 é a inclinação da reta (**efeito** de x sobre y)
- $i = 1, 2, \dots, n$ observações
- $\epsilon \sim N(0, \sigma^2)$



Regressão

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Os parâmetros β_0 e β_1 são estimados pelo **método dos mínimos quadrados**. Os resíduos são

$$\epsilon_i = y_i - (\beta_0 - \beta_1 x_i)$$

Portanto, a **soma dos quadrados dos resíduos (SQR)** é

$$SQR = \sum_{i=1}^n (y_i - (\beta_0 - \beta_1 x_i))^2$$

Através das derivadas da SQR em relação à β_0 e β_1 chega-se aos resultados

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Regressão

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade
Inferência

Correlação e
regressão

ANOVA

MLGs

Ajustando um modelo linear no R

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

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

```
Coefficients:
(Intercept)          CL
      1.19         1.10
```



Regressão

Sumário

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

```
> summary(mod)
```

Call:

```
lm(formula = CW ~ CL, data = dados)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.776	-0.570	0.110	0.463	1.827

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.1869	0.2853	4.16	5.3e-05	***
CL	1.0975	0.0087	126.17	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.783 on 154 degrees of freedom

Multiple R-squared: 0.99, Adjusted R-squared: 0.99

F-statistic: 1.59e+04 on 1 and 154 DF, p-value: <2e-16



Regressão

Ajuste gráfico

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

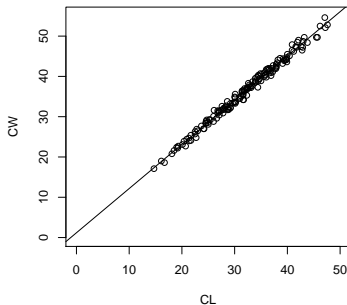
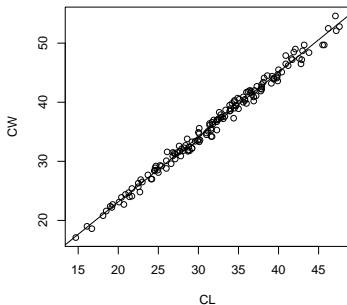
Inferência

Correlação e
regressão

ANOVA

MLGs

```
> 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

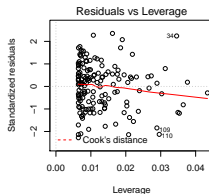
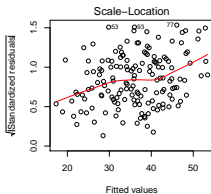
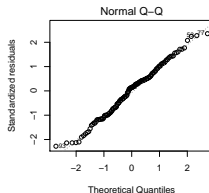
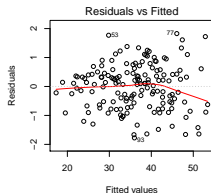
Inferência

Correlação e
regressão

ANOVA

MLGs

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





Regressão

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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"
```



Exercícios

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Com as colunas BD e CL do objeto dados

- ❶ Faça um gráfico da relação entre estas variáveis
- ❷ Faça um teste de correlação
- ❸ Ajuste um modelo linear
 - ❶ Veja o sumário
 - ❷ Ajuste a linha do modelo no gráfico
 - ❸ 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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Correlação e regressão
- 4 Análise de Variância**
- 5 Modelos Lineares Generalizados



Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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

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

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

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n$$

$$H_1 : \text{pelo menos um } \mu_i \text{ é diferente dos demais}$$



Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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.
14.7	24.6	30.1	29.9	34.5	47.1

```
$laranja
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.7	29.4	34.5	34.1	39.2	47.6

Média geral $\bar{y} = 32$

```
> mean(dados$CL)
```

```
[1] 32.004
```



Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

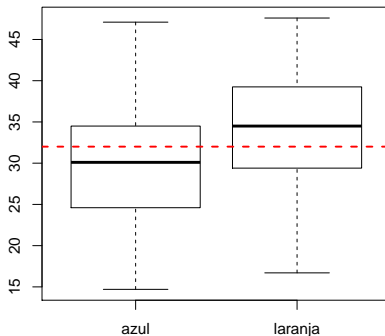
Inferência

Correlação e
regressão

ANOVA

MLGs

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





Análise de Variância

Geometricamente

Módulo III
Inferência e
Modelagem

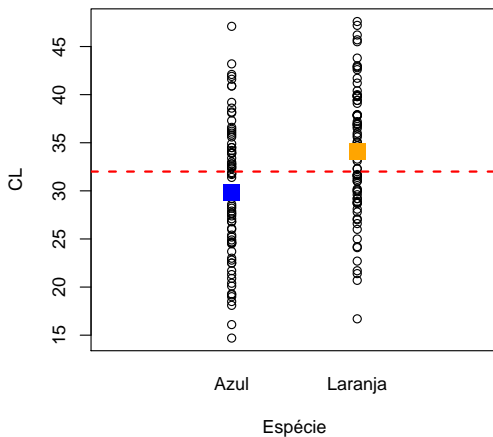
Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs





Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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
-17.385	-5.019	0.273	5.019	17.231

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	29.87	0.79	37.8	< 2e-16 ***
especieleranja	4.22	1.11	3.8	0.00021 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.93 on 154 degrees of freedom

Multiple R-squared: 0.0856, Adjusted R-squared: 0.0797

F-statistic: 14.4 on 1 and 154 DF, p-value: 0.00021



Análise de Variância

Ajustando o modelo

Módulo III
Inferência e
Modelagem

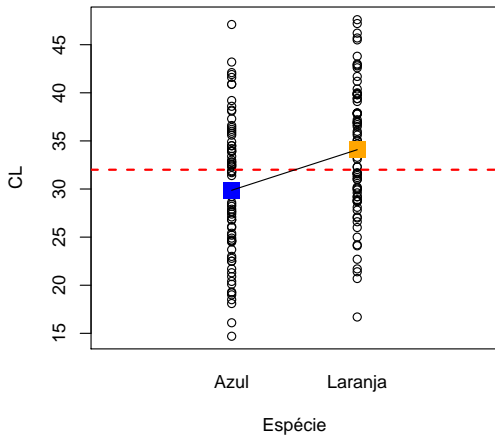
Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs





Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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 = 152.73, p-value = 0.0002135  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -6.4116 -2.0204  
sample estimates:  
 mean in group azul mean in group laranja  
      29.869      34.085
```



Análise de Variância

Notou a relação?

```
> summary(mod)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	29.869	0.7902	37.7990	8.1924e-80
especiellaranja	4.216	1.1104	3.7968	2.1042e-04

```
> teste$p.value
```

```
[1] 2.1352e-04
```

```
> teste$estimate
```

mean in group azul	mean in group laranja
29.869	34.085

```
> diff(teste$estimate)
```

mean in group laranja
4.216

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs



Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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	693	14.4	0.00021 ***
Residuals	154	7404	48		

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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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	4.216	2.0224	6.4096	0.00021



Análise de Variância

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

- 1 Distribuições de probabilidade
- 2 Inferência
- 3 Correlação e regressão
- 4 Análise de Variância
- 5 Modelos Lineares Generalizados



Modelos Lineares Generalizados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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**.

$\text{Teste-t} \subset \text{ANOVA} \subset \text{ANCOVA}^* \subset \text{ML} \subset \text{ML-MULT}^* \subset \text{MLG}$

- 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 estende o modelo para que y possa ser um fator ou ter uma distribuição diferente da normal.



Modelos Lineares Generalizados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Os MLGs são formados por três componentes:

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



Modelos Lineares Generalizados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

De maneira geral, os MLGs descrevem a relação entre a variável resposta y_i ($i = 1, \dots, n$) 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.



Modelos Lineares Generalizados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

Distribuições da família exponencial e funções de ligação ($P = \text{link padrão}$)

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		•	•			



Modelos Lineares Generalizados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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
-17.385	-5.019	0.273	5.019	17.231

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	29.87	0.79	37.8	< 2e-16 ***
especiellaranja	4.22	1.11	3.8	0.00021 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 48.08)



Modelos Lineares Generalizados

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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



Exercícios

Módulo III
Inferência e
Modelagem

Distribuições
de
probabilidade

Inferência

Correlação e
regressão

ANOVA

MLGs

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?