Introdução ao R com aplicações em biodiversidade e conservação

2020-09-18

Contents

Chapter 1

Pré-requisitos

Chapter 2

Estatística básica

2.1 Teste T (de Student) para duas amostras independentes

2.1.1 Background da análise

Uma das perguntas mais comum em estatística é saber se há diferença entre as médias de dois grupos ou tratamentos. Para responder esta pergunta, William Sealy Gosset, químico da cervejaria Guinness, em 1908 desenvolveu o Teste T que é uma estátistica que segue uma distribuição t de Student para rejeitar ou não uma hipótese nula de médias iguais entre os grupos.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{2S_p^2}{n}}}$$

Onde:

- $\bar{X}1$ - $\bar{X}2$ = diferença entre as médias das duas amostras,
- S2p = desvio padrão das amostras,
- n = tamanho das amostras.

2.1.1.1 Premissas do Teste t :

- As amostras devem ser independentes;
- As unidades amostrais são selecionadas aleatoriamente;
- Distribuição normal (gaussiana) dos resíduos. **Observação:** Zar (2010, p. 136) indica que o Test T é robusto mesmo com moderada violação da normalidade, principalmente se o tamanho amostral for alto.

- Homogeneidade da variância. **Observação.** Caso as variâncias não sejam homogêneas, isso deve ser informado na linha de comando, pois o denominador da fórmula acima será corrigido.

2.1.1.2 Exemplo prático 1 - Teste T para duas amostras com variâncias iguais

2.1.1.2.1 Explicação dos dados Neste exemplo avaliaremos o comprimento rostro-cloacal (CRC em milímetros) de machos de *Physalaemus natteri* (Anura:Leptodactylidae) amostrados em diferentes estações do ano com armadilhas de interceptação e queda na região noroeste do estado de São Paulo (da Silva & Rossa-Feres 2010).

Pergunta:

> O CRC dos machos de *P. nattereri* é maior na estação chuvosa do que na estação seca?

Predições

> O CRC dos machos será maior na estação chuvosa porque há uma vantangem seletiva para os indivíduos maiores durante a atividade reprodutiva.

Variáveis

- Variáveis preditoras
  - Dataframe com os indivíduos (unidade amostral) nas linhas e CRC (mm - variável resposta contínua) e estação (variável preditora categórica) como colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditores nas colunas

2.1.2 Análise

Calculo do Teste T para duas amostras independentes com variâncias iguais

```
## IMPORTANDO OS DADOS
#***********************
CRC_PN_macho <- ecodados::teste_t_var_igual
# verificar se o dataframe foi lido corretamente e se não há erros
# Esses comandos são úteis para planilhas grandes
head(CRC_PN_macho) # mostra as seis primeiras linhas da planilha
```

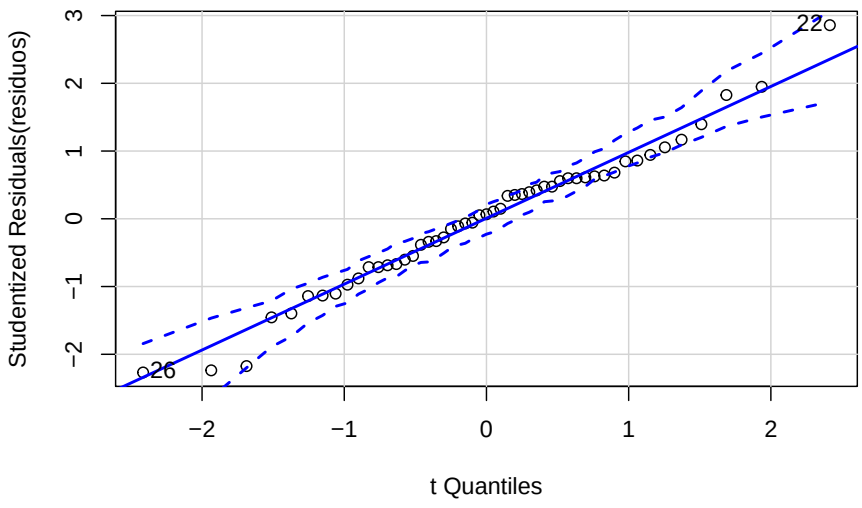
```
##    CRC Estacao
## 1 3.82 Chuvosa
```

```
## 2 3.57 Chuvosa
## 3 3.67 Chuvosa
## 4 3.72 Chuvosa
## 5 3.75 Chuvosa
## 6 3.83 Chuvosa
```

```
tail(CRC_PN_macho) # mostra as seis últimas linhas da planilha
```

```
##     CRC Estacao
## 46 3.16    Seca
## 47 3.48    Seca
## 48 3.48    Seca
## 49 3.49    Seca
## 50 3.51    Seca
## 51 3.30    Seca
```

```
# TESTE NORMALIDADE
#************************
## Verificando normalidade usando QQ-plot
## Os pontos não podem fugir da reta criando formas como U
residuos <- lm(CRC ~ Estacao, data = CRC_PN_macho)
library("car")
qqPlot(residuos)
```




```
## [1] 22 26
```

```
## Outra possibilidade é usar o teste de Shapiro-Wilk para verificar normalidade
## Hipótese nula que a distribuição é normal
## valor de p < 0.05 significa que os dados não apresentam distribuição normal
## valor de p > 0.05 significa que os dados apresentam distribuição normal
shapiro.test (CRC_PN_macho$CRC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CRC_PN_macho$CRC
## W = 0.95559, p-value = 0.05417
```

```
# TESTE DE HOMOGENEIDADE DA VARIÂNCIA
#************************************
## Hipótese nula que a variância é homogênea
## valor de p < 0.05 significa que os dados não apresentam homogeneidade
## valor de p > 0.05 significa que os dados apresentam homogeneidade
library(car)
leveneTest(CRC ~ Estacao, data = CRC_PN_macho)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.1677 0.2852
##       49
```

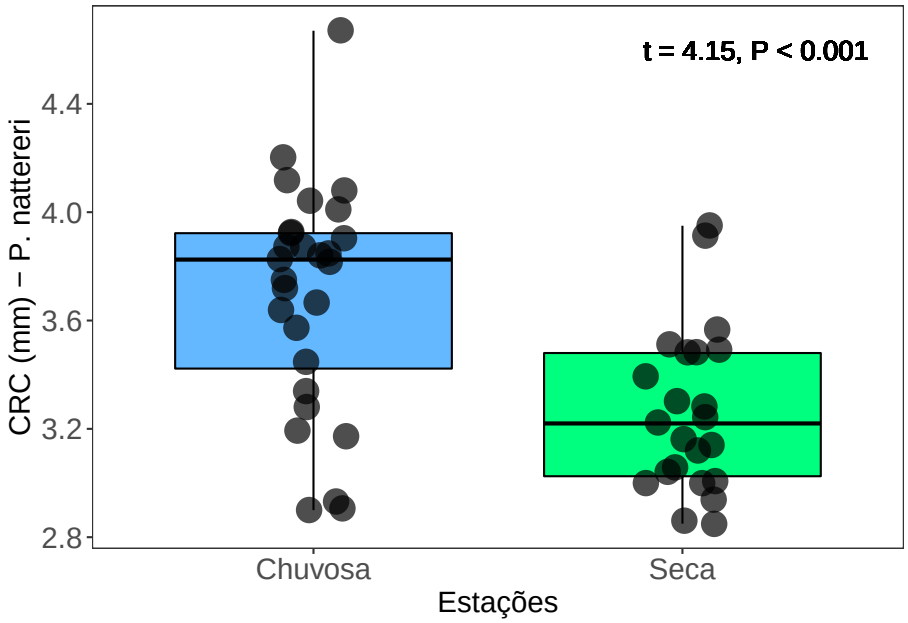
```
# TESTE T AMOSTRAS INDEPENDENTES E VARIÂNCIAS IGUAIS
#*****************************************************888
t.test(CRC ~ Estacao, data = CRC_PN_macho, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  CRC by Estacao
## t = 4.1524, df = 49, p-value = 0.000131
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2242132 0.6447619
## sample estimates:
## mean in group Chuvosa    mean in group Seca
##              3.695357              3.260870
```

Visualizar os resultados em gráfico

```
library(ggplot2)
ggplot(data = CRC_PN_macho, aes(x= Estacao, y= CRC, color = Estacao)) +
  labs(x = "Estações", y = "CRC (mm) - P. nattereri", size = 15) +
  geom_boxplot(fill=c("steelblue1", "springgreen1"), color="black", outlier.shape = NA)
  geom_jitter(shape = 16, position=position_jitter(0.1), cex = 6, alpha = 0.7) +
```

```
  scale_color_manual(values = c("black", "black")) +
  geom_text(x = 2.2, y = 4.6, label = "t = 4.15, P < 0.001", color = "black", size = 5) +
  theme_bw() +
  theme(axis.text.y = element_text(size = 15), axis.text.x = element_text(size = 15)) +
  theme(axis.title.y = element_text(size = 15), axis.title.x = element_text(size = 15)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(legend.position = "none")
```



Interpretação dos resultados

Neste exemplo, rejeitamos a hipótese nula que as médias do CRC dos machos entre as estações seca e chuvosa são iguais (t = 4,15, P < 0,001). Os resultados mostram que os machos de *P. nattereri* coletados na estação chuvosa foram em média 0,43mm maiores do que os coletados na estação seca.

2.1.2.1 Exemplo prático 2 - Teste T para duas amostras independentes com variâncias diferentes

2.1.2.1.1 Explicação dos dados Neste exemplo, avaliaremos o comprimento rostro-cloacal (CRC - milímetros) de fêmeas de *Leptodactylus podicipinus* amostradas em diferentes estações do ano com armadilhas de interceptação e queda na região noroeste do estado de São Paulo (da Silva & Rossa-Feres 2010). **Observação:** Os dados foram alterados em relação a publicação original para se enquadrarem no exemplo de amostras com variâncias diferentes.

Pergunta:

> O CRC das fêmeas de *L. podicipinus* é maior na estação chuvosa do que na estação seca?

Predições

> O CRC das fêmeas será maior na estação chuvosa porque há uma vantangem seletiva para os indivíduos maiores durante a atividade reprodutiva.

Variáveis

- Variáveis preditoras
  - Dataframe com os indivíduos (unidade amostral) nas linhas e CRC (mm - variável resposta contínua) e estação (variável preditora categórica) como colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditores nas colunas

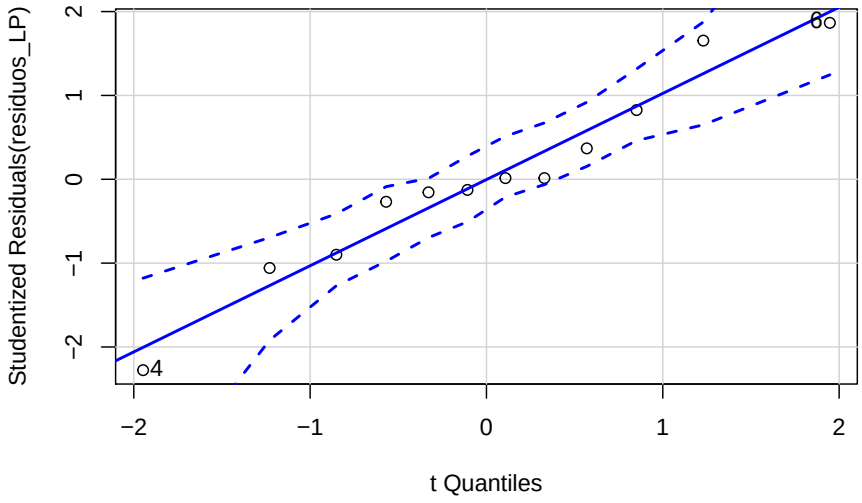
2.1.3 Análise

Calculo do Teste T para duas amostras com variâncias diferentes

```
## IMPORTANDO OS DADOS
#*************************
CRC_LP_femea <- ecodados::teste_t_var_diferente
head(CRC_LP_femea) # verificar se o dataframe foi lido corretamente
```

```
##    CRC Estacao
## 1 2.72 Chuvosa
## 2 2.10 Chuvosa
## 3 3.42 Chuvosa
## 4 1.50 Chuvosa
## 5 3.90 Chuvosa
## 6 4.00 Chuvosa
```

```
# TESTE NORMALIDADE
#************************
## Verificando normalidade usando QQ-plot
## Os pontos não podem fugir da reta criando formas como U
residuos_LP <- lm(CRC ~ Estacao, data = CRC_LP_femea)
library("car")
qqPlot(residuos_LP)
```




```
## [1] 4 6
```

```
## Outra possibilidade é usar o teste de Shapiro-Wilk para verificar normalidade
## Hipótese nula que a distribuição é normal
## valor de p < 0.05 significa que os dados não apresentam distribuição normal
## valor de p > 0.05 significa que os dados apresentam distribuição normal
shapiro.test (CRC_LP_femea$CRC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CRC_LP_femea$CRC
## W = 0.88195, p-value = 0.09284
```

```
# TESTE DE HOMOGENEIDADE DA VARIÂNCIA
#*************************************
## Hipótese nula que a variância é homogênea
## valor de p < 0.05 significa que os dados não apresentam homogeneidade
## valor de p > 0.05 significa que os dados apresentam homogeneidade
library(car)
leveneTest(CRC ~ Estacao, data = CRC_LP_femea)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  9.8527 0.01053 *
##       10
```

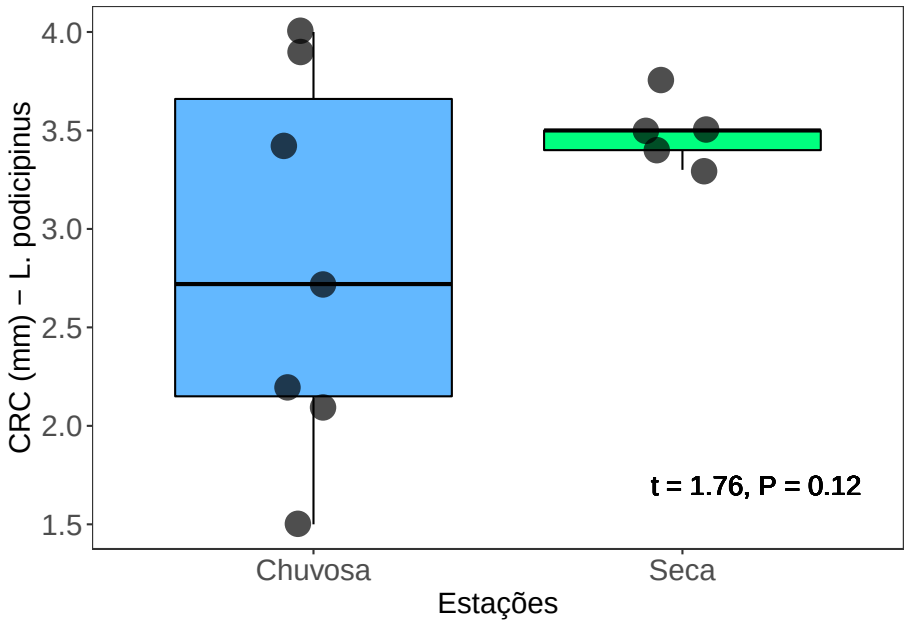
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# TESTE T COM AMOSTRAS INDEPENDENTES E VARIÂNCIS DIFERENTES
#************************************************************
## Com base no teste de Levene, avise na linha de comando que as variâncias
## não são iguais (var.equal = FALSE).
t.test(CRC ~ Estacao, data = CRC_LP_femea, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  CRC by Estacao
## t = -1.7633, df = 6.4998, p-value = 0.1245
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5489301  0.2375016
## sample estimates:
## mean in group Chuvosa    mean in group Seca
##              2.834286              3.490000
```

Visualizar os resultados em gráfico

```
library(ggplot2)
ggplot(data = CRC_LP_femea, aes(x= Estacao, y= CRC, color = Estacao)) +
  labs(x = "Estações", y = "CRC (mm) - L. podicipinus", size = 15) +
  geom_boxplot(fill=c("steelblue1", "springgreen1"), color="black", outlier.shape = NA)
  geom_jitter(shape = 16, position=position_jitter(0.1), cex = 6, alpha = 0.7) +
  scale_color_manual(values = c("black", "black")) +
  geom_text(x = 2.2, y = 1.7, label = "t = 1.76, P = 0.12", color = "black", size = 5)
  theme_bw() +
  theme(axis.text.y = element_text(size = 15), axis.text.x = element_text(size = 15)) +
  theme(axis.title.y = element_text(size = 15), axis.title.x = element_text(size = 15))
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(legend.position = "none")
```



Interpretação dos resultados

Neste exemplo, não rejeitamos a hipótese nula e consideramos que as médias do CRC das fêmeas entre as estações seca e chuvosa são iguais (t = 1,76, P = 0,12). Os resultados mostram que as fêmeas de *L. podicipinus* coletadas na estação chuvosa não são maiores do que as fêmeas coletadas na estação seca.

2.2 Teste T para amostras pareadas

2.2.1 Background da análise

O Teste T Pareado é uma estatística que usa dados medidos duas vezes na mesma unidade amostral, resultando em pares de observações para cada amostra (amostras pareadas). Ele determina se a diferença da média entre duas observações é zero.

$$t = \frac{\bar{d}}{S_{\bar{d}}}$$

Onde:

- $\bar{d}$ = média da diferença das medidas pareadas. Observe que o teste não usa as medidas originais, e sim, a diferença para cada par,

- $S\bar{d}$ = erro padrão da diferença das medidas pareadas.

2.2.1.1 Premissas do Teste t para amostras pareadas:

- As unidades amostrais são selecionadas aleatoriamente;
- Distribuição normal (gaussiana) dos valores da diferença para cada par;

2.2.1.2 Exemplo prático 1 - Teste T para amostras pareadas

2.2.1.2.1 Explicação dos dados Neste exemplo avaliaremos a diferença na riqueza de espécies de artrópodes registradas em 27 localidades. Todas as localidades foram amostradas duas vezes. A primeira amostragem foi realizada com na localidade antes da pertubação e a segunda amostragem foi realizada após a localidade ter sofrido uma queimada.

Pergunta:

A riqueza de espécies de artrópodes é prejudicada pelas queimadas?

Predições

A riqueza de espécies de artrópodes será maior antes da queimada devido a extinção local das espécies.

Variáveis

- Variáveis preditoras
  - Dataframe com as localidades nas linhas e riqueza de espécies (variável preditora contínua) e estado (Pre-queimada ou Pós-queimada - variável categórica) da localidade nas colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditores nas colunas

2.2.2 Análise

Calculo do Teste T com amostras pareadas

```
## IMPORTANDO OS DADOS
#************************
Pareado <- ecodados::teste_t_pareado
head(Pareado) # verificar se o dataframe foi lido corretamente
```

```
##   Areas Riqueza       Estado
## 1     1      92 Pre-Queimada
## 2     2      74 Pre-Queimada
## 3     3      96 Pre-Queimada
```

```
## 4     4      89 Pre-Queimada
## 5     5      76 Pre-Queimada
## 6     6      80 Pre-Queimada
```

```
tail(Pareado)
```

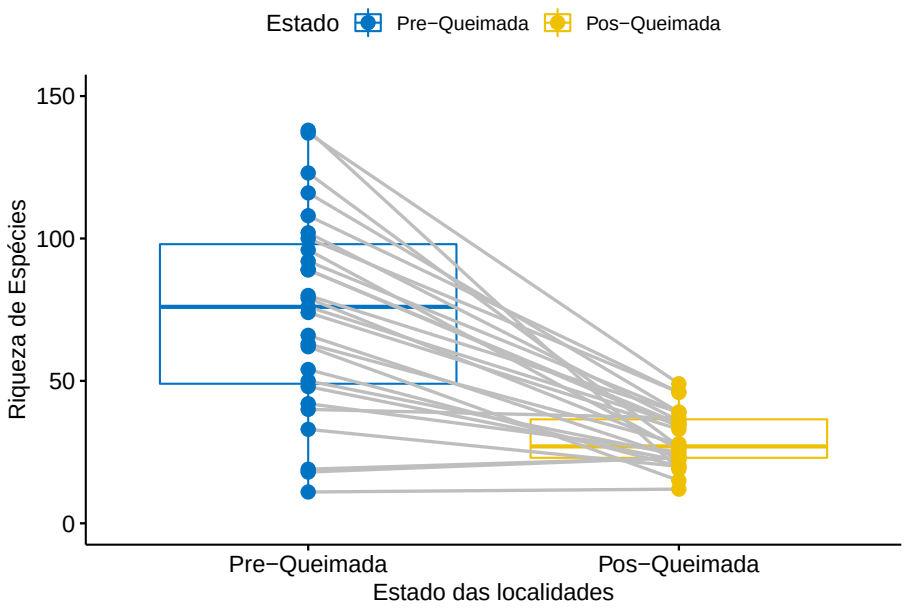
```
##    Areas Riqueza       Estado
## 49    22      37 Pos-Queimada
## 50    23      20 Pos-Queimada
## 51    24      12 Pos-Queimada
## 52    25      22 Pos-Queimada
## 53    26      27 Pos-Queimada
## 54    27      28 Pos-Queimada
```

```
# TESTE T PAREADO
#******************
# O uso do [] é para selecionar dentro do vetor/coluna *Riqueza* os 27 primeiros números [1:27] q
t.test(Pareado$Riqueza[1:27], Pareado$Riqueza[28:54], paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Pareado$Riqueza[1:27] and Pareado$Riqueza[28:54]
## t = 7.5788, df = 26, p-value = 4.803e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  32.47117 56.63994
## sample estimates:
## mean of the differences
##                44.55556
```

Visualizar os resultados em gráfico

```
library("ggpubr")
ggpaired(Pareado, x = "Estado", y = "Riqueza",
         color = "Estado", line.color = "gray", line.size = 0.8, palette = "jco", width = 0.8,
         point.size = 3, xlab = "Estado das localidades", ylab = "Riqueza de Espécies") +
  expand_limits(y=c(0,150))
```



Interpretação dos resultados

Neste exemplo, rejeitamos a hipótese nula que a riqueza de espécies de artrópodes é igual antes e depois da queimada (t = 7,57, P < 0,001). Os resultados mostram que as localidades após as queimadas apresentam em média 44,5 espécies de artrópodes a menos do que antes das queimadas.

2.3 Correlação de Pearson

2.3.1 Background da análise

É um teste que mede a o grau de associação entre duas variáveis contínuas (X e Y). Importante ressaltar que a análise de correlação não assume que a variável X influência a variável Y ou que exista uma relação de causa e efeito entre elas (Zar 2016). A análise é definida em termos da variância de X, a variância de Y, e a covariância de X e Y (i.e. como elas variam juntas).

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{\sum X^2}{n}\right)\left(\sum Y^2 - \frac{\sum Y^2}{n}\right)}}$$

Onde:

- r = coeficiente de correlação que indica a força da relação entre as duas variáveis. Seu range de valores está entre -1 $\leq$ r $\geq$ 1. A correlação positiva indica que o aumento no valor de uma das variáveis é acompanhado pelo aumento no valor da outra variável. A correlação negativa indica que um aumento no valor de uma das variáveis é acompanhado pela diminuição no valor da outra variável. Se *r* é igual a zero, não existe correlação entre as variáveis.

2.3.1.1 Premissas da Correlação de Person:

- As amostras devem ser independentes e pareadas (i.e. as duas variáveis devem ser medidas na mesma unidade amostral);
- As unidades amostrais são selecionadas aleatoriamente;
- A relação entre as variáveis tem que ser linear.

2.3.1.2 Exemplo prático 1 - Correlação de Pearson

2.3.1.2.1 Explicação dos dados Neste exemplo avaliaremos a correlação entre a altura do tronco e o tamanho da raiz medidos em 35 indivíduos de uma espécie vegetal arbustiva.

Pergunta:

Existe correlação entre a altura do tronco e o tamanho da raiz dos arbustos?

Predições

A altura do tronco é positivamente correlacionado com o tamanho da raiz.

Variáveis

- Variáveis preditoras
  - Dataframe com os indivíduos (unidade amostral) nas linhas e altura do tronco e tamanho da raiz (duas variáveis tem que ser contínuas) como colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditores nas colunas

2.3.2 Análise

Calculo do Teste de Correlação de Pearson

```
## IMPORTANDO OS DADOS
#************************
```

```
correlacao_arbustos <- ecodados::correlacao
head(correlacao_arbustos) # verificar se o dataframe foi lido corretamente
```

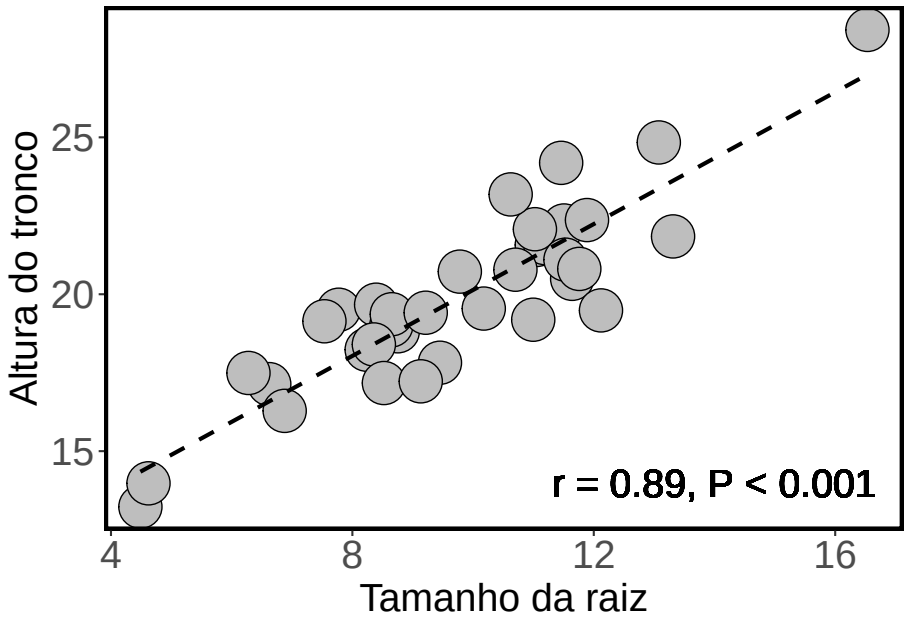
```
##   Tamanho_raiz Tamanho_tronco
## 1    10.177049       19.54383
## 2     6.622634       17.13558
## 3     7.773629       19.50681
## 4    11.055257       21.57085
## 5     4.487274       13.22763
## 6    11.190216       21.62902
```

```
# Teste de Correlação de Pearson
#*******************************
# Para outros testes de correlação como Kendall ou Spearman é só alterar na linha de c
cor.test(correlacao_arbustos$Tamanho_raiz, correlacao_arbustos$Tamanho_tronco, method =
```

```
##
##  Pearson's product-moment correlation
##
## data:  correlacao_arbustos$Tamanho_raiz and correlacao_arbustos$Tamanho_tronco
## t = 11.49, df = 33, p-value = 4.474e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7995083 0.9457816
## sample estimates:
##       cor
## 0.8944449
```

Visualizar os resultados em gráfico

```
library(ggplot2)
ggplot(data = correlacao_arbustos, aes(x= Tamanho_raiz, y= Tamanho_tronco)) +
  labs(x = "Tamanho da raiz", y = "Altura do tronco", size = 20) +
  geom_point(size = 10, shape = 21, fill = "gray") +
  geom_text(x = 14, y = 14, label = "r = 0.89, P < 0.001", color = "black", size = 7) +
  theme_bw() +
  theme(axis.title.y = element_text(size = 20), axis.title.x = element_text(size = 20)) +
  theme(axis.text.y = element_text(size = 20), axis.text.x = element_text(size = 20)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2)) +
  geom_smooth(method = lm, se = FALSE, color = "black", linetype="dashed")
```



Interpretação dos resultados

Neste exemplo, rejeitamos a hipótese nula que as variáveis não são correlacionadas (r = 0.89, P < 0,001). Os resultados mostram que o aumento na altura dos arbutos é acompanhado pelo aumento no tamanho da raiz.

2.4 Regressão Linear Simples

2.4.1 Background da análise

A regressão simples é usada para analisar a relação entre uma variável preditora (plotada no eixo-X) e uma variável resposta (plotada no eixo-Y). As duas variáveis devem ser contínuas. Diferente das correlações, a regressão assume uma relação de causa e efeito entre as variáveis. O valor da variável preditora (X) causa, direta ou indiretamente, o valor da variável resposta (Y). Assim, Y é uma função linear de X:

$$Y = \beta_0 + \beta_1 X_i + \epsilon_i$$

Onde:

- β_0 = intercepto que representa o valor da função quando X = 0,

- β_1 = inclinação (*slope*) que mede a mudança na variável Y para cada mudança de unidade da variável X. Este parâmetro é usado para testar a hipótese nula da regressão que assume que $\beta_1 = 0$.
- ϵ_1 = erro aleatório referente a variável Y que não pode ser explicado pela variável X.

2.4.1.1 Premissas da Regressão Linear Simples:

- As amostras devem ser independentes;
- As unidades amostrais são selecionadas aleatoriamente;
- Distribuição normal (gaussiana) dos resíduos;
- Homogeneidade da variância.

2.4.1.2 Exemplo prático 1 - Regressão linear simples

2.4.1.2.1 Explicação dos dados Neste exemplo, avaliaremos a relação entre o gradiente de temperatura média anual (°C) e o tamanho médio do comprimento rostro-cloacal (CRC em mm) de populações de *Dendropsophus minutus* (Anura:Hylidae) amostradas em 109 localidades no Brasil (Boaratti & da Silva 2015).

Pergunta:

> Há relação entre o tamanho do CRC das populações e a temperatura das localidades onde os indivíduos ocorrem?

Predições

> O CRC das populações serão menores em localidades mais quentes do que em localidades mais frias de acordo com a Hipótese do balanço de calor.

Variáveis

- Variáveis preditoras
  - Dataframe com as populações (unidade amostral) nas linhas e CRC médio (mm) e temperatura média anual como colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditores nas colunas

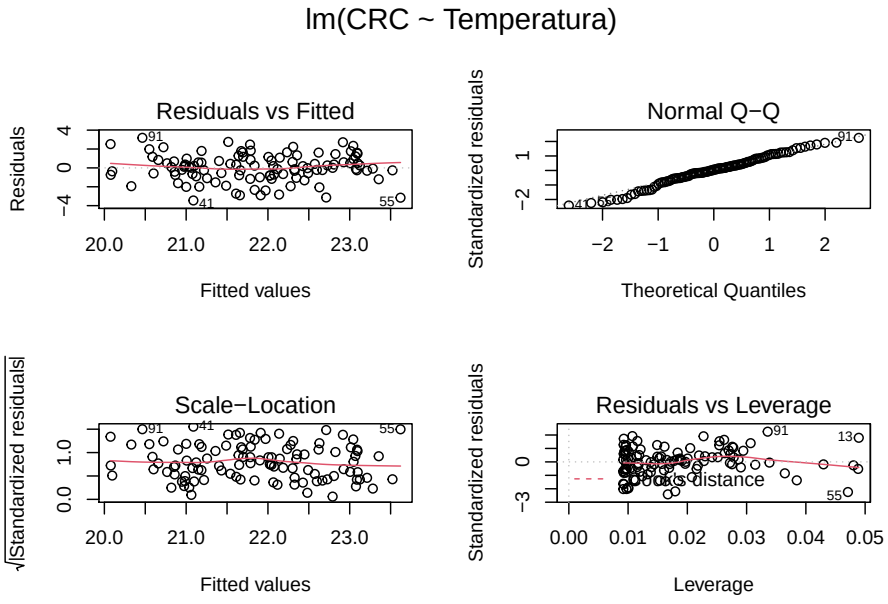
2.4.2 Análise

Calculo da regressão linear simples

```
## IMPORTANDO DADOS
#********************
dados_regressao <- ecodados::regressoes
head(dados_regressao) # verificar se o dataframe foi lido corretamente
```

```
##      Municipio      CRC Temperatura Precipitacao
## 1     Acorizal 22.98816    24.13000       1228.2
## 2  Alpinopolis 22.91788    20.09417       1487.6
## 3 Alto_Paraiso 21.97629    21.86167       1812.4
## 4    Americana 23.32453    20.28333       1266.2
## 5      Apiacas 22.83651    25.47333       2154.0
## 6  Arianopolis 20.86989    20.12167       1269.2
```

```
# ANALISE DA REGRESSÃO
#************************
modelo_regressao <- lm(CRC ~ Temperatura, data = dados_regressao)

# PRIMEIRO VAMOS VERIFICAR A NORMALIDADE E HOMOGENEIDADE DAS VARIÂNCIAS
#*************************************************************************
# Os gráficos *Residuals vs Fitted*, *Scale-Location*, e *Residual vs Leverage* estão relacionado
# O gráfico *Normal Q-Q* está relacionado com a distribuição normal dos resíduos. Neste gráfico,
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(modelo_regressao)
```



```
dev.off()
```

```
## null device
##           1
```

```
# VERIFICANDO OS RESULTADOS DA REGRESSÃO
#########################################
anova(modelo_regressao)
```

```
## Analysis of Variance Table
##
## Response: CRC
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Temperatura   1  80.931  80.931   38.92 9.011e-09 ***
## Residuals   107 222.500   2.079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

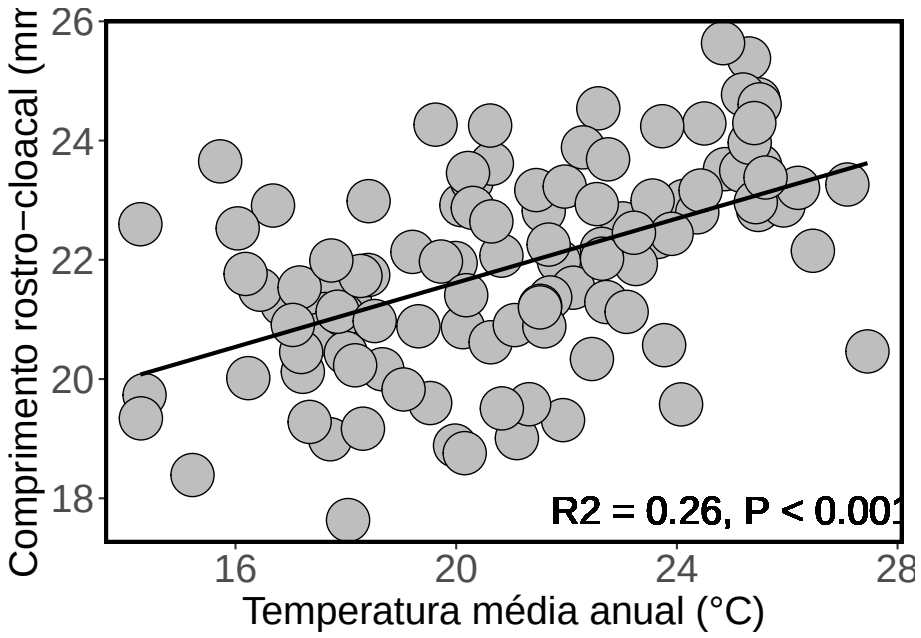
```
# ou
# esta função apresenta os resultados mais detalhados com a estimativa do intercepto,
summary(modelo_regressao)
```

```
##
## Call:
## lm(formula = CRC ~ Temperatura, data = dados_regressao)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4535 -0.7784  0.0888  0.9168  3.1868
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.23467    0.91368  17.768  < 2e-16 ***
## Temperatura  0.26905    0.04313   6.239 9.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.442 on 107 degrees of freedom
## Multiple R-squared:  0.2667,Adjusted R-squared:  0.2599
## F-statistic: 38.92 on 1 and 107 DF,  p-value: 9.011e-09
```

Visualizar os resultados em gráfico

```
library(ggplot2)
ggplot(data = dados_regressao, aes(x= Temperatura, y= CRC)) +
  labs(x = "Temperatura média anual (°C)", y = "Comprimento rostro-cloacal (mm)", size
  geom_point(size = 10, shape = 21, fill = "gray") +
  geom_text(x = 25, y = 17.8, label = "R2 = 0.26, P < 0.001", color = "black", size =
```

```
  theme_bw() +
  theme(axis.title.y = element_text(size = 20), axis.title.x = element_text(size = 20)) +
  theme(axis.text.y = element_text(size = 20), axis.text.x = element_text(size = 20)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2)) +
  geom_smooth(method = lm, se = FALSE, color = "black")
```



Interpretação dos resultados

Neste exemplo, rejeitamos a hipótese nula que não existe relação entre o tamanho do CRC das populações de *D. minutus* e a temperatura da localidade onde elas ocorrem ($F_{1,107} = 38{,}92$, $P < 0{,}001$). Os resultados mostram que o tamanho do CRC das populações tem uma relação positiva com a temperatura das localidades. Assim, populações de *D. minutus* em localidades mais quentes apresentam maior CRC do que as populações em localidades mais frias.

2.5 Regressão Linear Múltipla

2.5.1 Background da análise

A regressão múltipla é uma extensão da regressão simples. Ela é usada quando queremos determinar o valor da variável resposta (Y) com base nos valores de duas ou mais variáveis preditoras (X1, X2, Xn).

$$Y = \beta_0 + \beta_1 X_1 + \beta_n X_n + \epsilon_i$$

Onde:

- β_0 = intercepto que representa o valor da função quando X = 0;
- β_n = inclinação (*slope*) que mede a mudança na variável Y para cada mudança de unidade das variáveis Xn;
- ϵ_1 = erro aleatório referente a variável Y que não pode ser explicado pelas variáveis preditoras.

2.5.1.1 Premissas da Regressão Linear Múltipla:

- As amostras devem ser independentes;
- As unidades amostrais são selecionadas aleatoriamente;
- Distribuição normal (gaussiana) dos resíduos;
- Homogeneidade da variância.

2.5.1.2 Exemplo prático 1 - Regressão linear múltipla

2.5.1.2.1 Explicação dos dados Utilizaremos o mesmo exemplo da regressão simples. Contudo, além do gradiente de temperatura média anual (°C) incluiremos o gradiente de precipitação anual (mm) como outra variável preditora do tamanho médio do comprimento rostro-cloacal (CRC em mm) de populações de *Dendropsophus minutus* (Anura:Hylidae) amostradas em 109 localidades no Brasil (Boaratti & da Silva 2015).

Pergunta:

> O tamanho do CRC das populações de *D. minutus* é influênciado pela temperatura e precipitação das localidades onde os indivíduos ocorrem?

Predições

> O CRC das populações serão menores em localidades com clima quente e chuvoso do que em localidades com clima frio e seco.

Variáveis

- Variáveis preditoras
  - Dataframe com as populações (unidade amostral) nas linhas e CRC médio (mm) e temperatura e precipitação como colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditores nas colunas

2.5.2 Análise

Cálculo da regressão linear múltipla

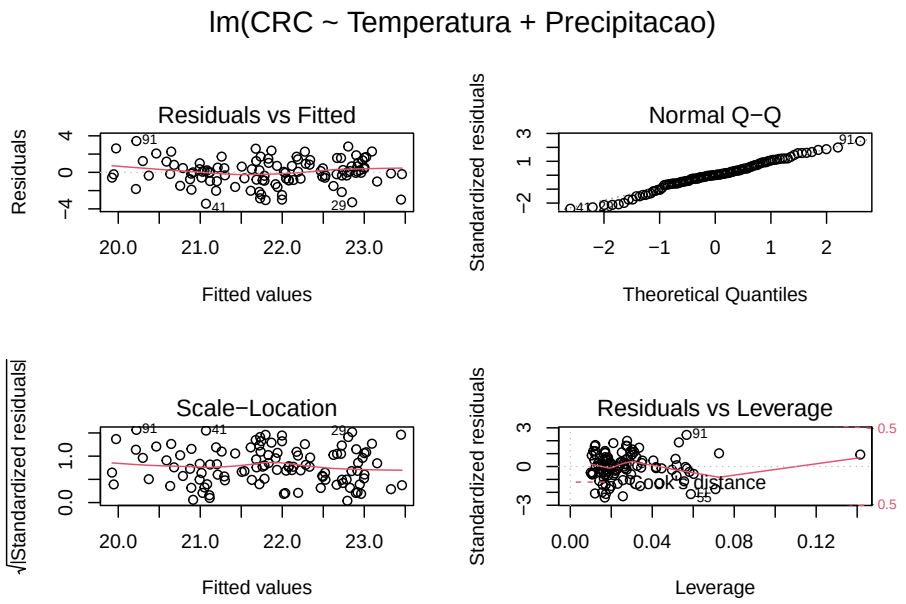
```
## IMPORTANDO DADOS
#*********************
dados_regressao_mul <- ecodados::regressoes
head(dados_regressao_mul) # verificar se o dataframe foi lido corretamente
```

```
##      Municipio      CRC Temperatura Precipitacao
## 1     Acorizal 22.98816    24.13000       1228.2
## 2  Alpinopolis 22.91788    20.09417       1487.6
## 3 Alto_Paraiso 21.97629    21.86167       1812.4
## 4    Americana 23.32453    20.28333       1266.2
## 5      Apiacas 22.83651    25.47333       2154.0
## 6  Arianopolis 20.86989    20.12167       1269.2
```

```
# ANALISE DA REGRESSÃO
#************************
modelo_regressao_mul <- lm(CRC ~ Temperatura + Precipitacao, data = dados_regressao_mul)

# MULTICOLINEARIDADE
#*********************
# Multicolinearidade ocorre quando as variáveis preditoras são correlacionadas. Essa correlação é
library(car)
vif(modelo_regressao_mul)
```

```
##  Temperatura Precipitacao
##     1.041265     1.041265
```

```
# VERIFICANDO A NORMALIDADE E HOMOGENEIDADE DAS VARIÂNCIAS
#*********************************************************************
# Os gráficos *Residuals vs Fitted*, *Scale-Location*, e *Residual vs Leverage* estão relacionado
# O gráfico *Normal Q-Q* está relacionado com a distribuição normal dos resíduos. Neste gráfico,
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(modelo_regressao_mul)
```




```
dev.off()
```

```
## null device
##           1
```

```
# VERIFICANDO OS RESULTADOS DA REGRESSÃO
#****************************************
anova(modelo_regressao_mul)
```

```
## Analysis of Variance Table
##
## Response: CRC
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## Temperatura     1  80.931  80.931 39.0028 8.94e-09 ***
## Precipitacao    1   2.549   2.549  1.2283    0.2702
## Residuals     106 219.951   2.075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ou
summary(modelo_regressao_mul)
```

```
##
## Call:
## lm(formula = CRC ~ Temperatura + Precipitacao, data = dados_regressao_mul)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.4351 -0.8026  0.0140  0.9420  3.4300
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.7162571  1.0108674  16.537  < 2e-16 ***
## Temperatura   0.2787445  0.0439601   6.341 5.71e-09 ***
## Precipitacao -0.0004270  0.0003852  -1.108     0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.44 on 106 degrees of freedom
## Multiple R-squared:  0.2751,Adjusted R-squared:  0.2614
## F-statistic: 20.12 on 2 and 106 DF,  p-value: 3.927e-08
```

Percebam que a temperatura tem uma relação significativa com o tamanho do CRC das populações ($P < 0.001$), enquanto que a precipitação não apresenta relação com o CRC ($P = 0.27$). Neste caso, é interessante saber se um modelo mais simples (e.g. contendo apenas temperatura) explicaria a distribuição tão bem ou melhor do que este modelo mais complexo considerando dois parâmetros (temperatura e precipitação).

Para isso, podemos utilizar a *LIKELIHOOD RATIO TEST (LRT)* para comparar modelos. A LRT compara dois modelos aninhados, testando se os parâmetros do modelo mais complexo diferem significativamente do modelo mais simples.Em outras palavras, ele testa se há necessidade de se incluir um parâmetro extra no modelo para explicar os dados.

```r
## CRIANDO OS MODELOS ANINHADOS
#*******************************
modelo_regressao_mul <- lm(CRC ~ Temperatura + Precipitacao, data = dados_regressao_mul)
modelo_regressao <- lm(CRC ~ Temperatura, data = dados_regressao_mul)

# LIKELIHHOD RATIO TEST (LRT)
#*******************************
# A hipótese nula é que o modelo mais simples é melhor
# Valores de p < 0.05 rejeita a hipótese nula e o modelo mais complexo é o melhor
# Valores de p > 0.05 não rejeita a hipótese nula e o modelo mais simples é o melhor
library(lmtest)
lrtest(modelo_regressao_mul, modelo_regressao)
```

```
## Likelihood ratio test
##
## Model 1: CRC ~ Temperatura + Precipitacao
## Model 2: CRC ~ Temperatura
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -192.93
```

```
## 2   3 -193.55 -1 1.2558     0.2624
```

```
# COMPARANDO COM O MODELO SOMENTE COM O INTERCEPTO
#*********************************************
# criando um modelo sem parâmetros, só o intercepto
modelo_intercepto <- lm(CRC ~ 1, data = dados_regressao_mul)
lrtest(modelo_regressao, modelo_intercepto)
```

```
## Likelihood ratio test
##
## Model 1: CRC ~ Temperatura
## Model 2: CRC ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -193.55
## 2   2 -210.46 -1 33.815  6.061e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretação dos resultados

Neste exemplo, a precipitação não está associada com a variação no tamanho do CRC das populações de *D. minutus*. Por outro lado, temperatura explicou 26% da variação do tamanho do CRC das populações.

2.6 Análises de Variância (ANOVA)

2.6.1 Background da análise

Anova refere-se a uma variedade de delineamentos experimentais nos quais a variável preditora é categórica e a variável resposta é contínua (Gotelli & Ellison 2013). Exemplos desses delineamentos experimentais são: Anova de um fator, Anova de dois fatores, Anova em blocos aleatorizados, Anova de medidas repetidas e Anova *split-splot*. De forma geral, a Anova é um teste estatístico usado para comparar a média entre grupos amostrados independentemente. Para isso, o teste leva em conta, além das médias dos grupos, a variação dos dados dentro e entre os grupos. Neste capítulo, iremos demonstrar as linhas de comandos para alguns dos principais delineamentos experimentais.

2.6.1.1 Premissas da Anova:

- As amostras devem ser independentes;
- As unidades amostrais são selecionadas aleatoriamente;
- Distribuição normal (gaussiana) dos resíduos;
- Homogeneidade da variância.

2.7 ANOVA de um fator

Este teste considera delineamentos experimentais com apenas um fator (ou tratamento) que pode ser composto por três ou mais grupos (ou níveis).

2.7.0.1 Exemplo prático 1 - Anova de um fator

2.7.0.1.1 Explicação dos dados Neste exemplo, avaliaremos se o adubo X-2020 disponibilizado recentemente no mercado melhora o crescimento dos indivíduos de *Coffea arabica* como divulgado pela empresa responsável pela venda do produto. Para isso, foi realizado um experimento com indivíduos de *C. arabica* cultivados em três grupos: i) grupo controle onde os indivíduos não receberam adubação, ii) grupo onde os indivíduos receberam a adição do adubo tradicional mais utilizado pelos produtores de *C. arabica*, e iii) grupo onde os indivíduos receberam a adição do adubo X-2020.

Pergunta:

> O crescimento dos indivíduos de *C. arabica* é melhorado pela adição do adubo X-2020?

Predições

> O crescimento dos indivíduos de *C. arabica* será maior no grupo que recebeu o adubo X-2020.

Variáveis

- Variáveis preditoras
  - Dataframe com as plantas (unidade amostral) nas linhas e o tratamento na coluna.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variável preditora na coluna.

2.7.1 Análise

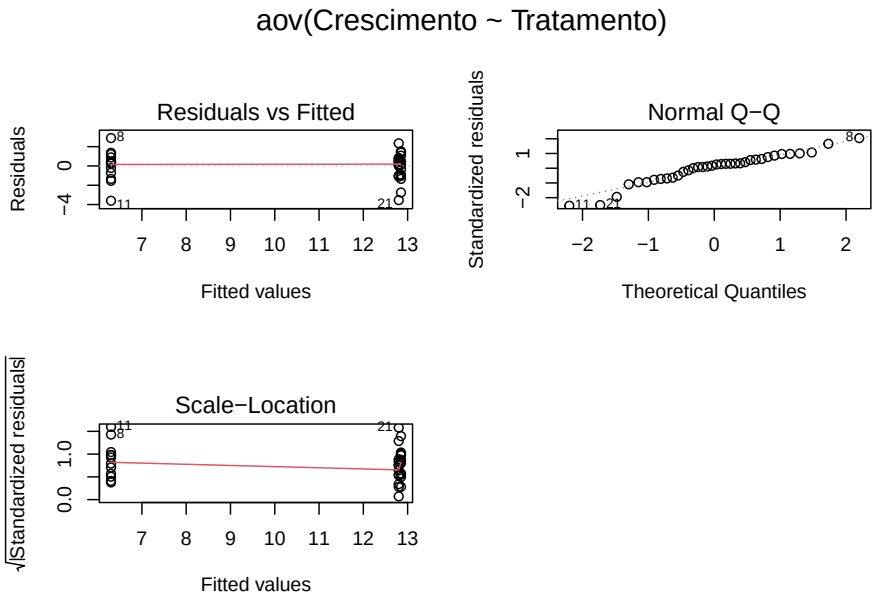
Cálculo da Anova de um fator

```
## IMPORTANDO DADOS
#*********************
dados_anova_simples <- ecodados::anova_simples
head(dados_anova_simples) # verificar se o dataframe foi lido corretamente
```

```
##   Crescimento Tratamento
## 1       7.190   Controle
## 2       6.758   Controle
## 3       6.101   Controle
## 4       4.758   Controle
```

```
## 5       6.542   Controle
## 6       7.667   Controle
```

```
# ANALISE ANOVA de um fator
#************************
Modelo_anova <- aov(Crescimento ~ Tratamento, data = dados_anova_simples)

# VERIFICANDO A NORMALIDADE E HOMOGENEIDADE DAS VARIÂNCIAS
#*********************************************************************
# Os gráficos *Residuals vs Fitted* e *Scale-Location* estão relacionados com a homoges
# O gráfico *Normal Q-Q* está relacionado com a distribuição normal dos resíduos. Nest
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(Modelo_anova)
```




```
dev.off()
```

```
## null device
##           1
```

```
# Se preferir, você pode utilizar testes estatísticos
# Teste de Shapiro-Wilk para normalidade separadamente para cada grupo
shapiro.test(dados_anova_simples$Crescimento[1:12])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados_anova_simples$Crescimento[1:12]
```

```
## W = 0.96731, p-value = 0.8806
```

```
shapiro.test(dados_anova_simples$Crescimento[13:24])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados_anova_simples$Crescimento[13:24]
## W = 0.87324, p-value = 0.07184
```

```
shapiro.test(dados_anova_simples$Crescimento[25:36])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados_anova_simples$Crescimento[25:36]
## W = 0.9294, p-value = 0.3738
```

```
# Teste de Bartlett para homogeneidade da variância
bartlett.test(Crescimento ~ Tratamento, data = dados_anova_simples)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Crescimento by Tratamento
## Bartlett's K-squared = 0.61835, df = 2, p-value = 0.7341
```

```
# VERIFICANDO OS RESULTADOS DA ANOVA
#*************************************
anova(Modelo_anova)
```

```
## Analysis of Variance Table
##
## Response: Crescimento
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Tratamento  2 340.32 170.160  77.989 3.124e-13 ***
## Residuals  33  72.00   2.182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

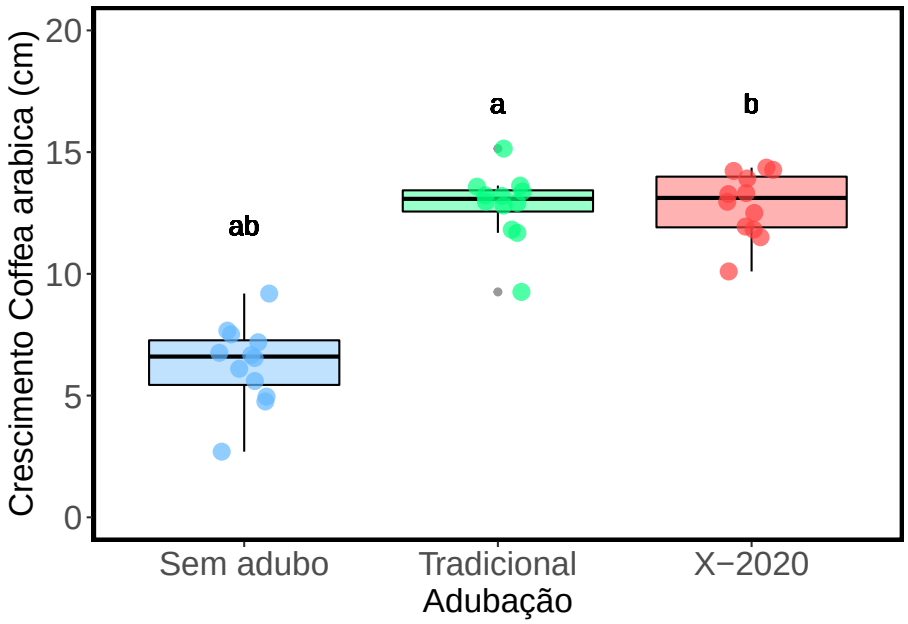
Percebam que o resultado da Anova ($\Pr(>F) < 0.001$) indica que devemos rejeitar a hipótese nula que não há diferença entre as médias dos grupos. Contudo, os resultados não mostram quais são os grupos que apresentam diferenças. Para isso, temos que realizar testes de comparações múltiplas *post-hoc* para detectar os grupos que apresentam diferenças significativas entre as médias. **Observação** Os testes *post-hoc* só devem ser utilizados quando rejeitamos a hipótese nula ($P < 0.05$) no teste da Anova.

```
# Diferenças entre os tratamentos
#**********************************
# Teste de Tuckey's honest significant difference
TukeyHSD(Modelo_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Crescimento ~ Tratamento, data = dados_anova_simples)
##
## $Tratamento
##                                       diff       lwr       upr     p adj
## Adubo_X-2020-Adubo_Tradicional  0.04991667 -1.429784  1.529617 0.9962299
## Controle-Adubo_Tradicional     -6.49716667 -7.976867 -5.017466 0.0000000
## Controle-Adubo_X-2020          -6.54708333 -8.026784 -5.067383 0.0000000
```

Visualizar os resultados em gráfico

```
# Reordenando a ordem que os grupos irão aparecer no gráfico
dados_anova_simples$Tratamento <- factor(dados_anova_simples$Tratamento ,
                                         levels=c("Controle", "Adubo_Tradicional", "Adubo_

# Gráfico
library(ggplot2)
ggplot(data = dados_anova_simples, aes(x= Tratamento, y= Crescimento, color = Tratament
  labs(x = "Adubação", y = "Crescimento Coffea arabica (cm)", size = 20) +
  geom_boxplot(fill=c("steelblue1", "springgreen1", "brown1"), color="black", show.lege
               alpha = 0.4) +
  geom_jitter(shape = 16, position=position_jitter(0.1), cex = 4, alpha = 0.7) +
  scale_color_manual(values = c("steelblue1", "springgreen1", "brown1")) +
  scale_y_continuous(limits = c(0, 20), breaks = c(0, 5, 10, 15, 20)) +
  geom_text(x = 1, y = 12, label = "ab", color = "black", size = 5) +
  geom_text(x = 2, y = 17, label = "a", color = "black", size = 5) +
  geom_text(x = 3, y = 17, label = "b", color = "black", size = 5) +
  scale_x_discrete(labels=c("Sem adubo","Tradicional","X-2020")) +
  theme_bw() +
  theme(axis.title.y = element_text(size = 17), axis.title.x = element_text(size = 17))
  theme(axis.text.y = element_text(size = 17), axis.text.x = element_text(size = 17)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2)) +
  theme(legend.position = "none")
```



Interpretação dos resultados

Neste exemplo, os indivíduos de *C. arabica* que receberam adubação (tradicional e X-2020) apresentaram maior crescimento do que os indivíduos que não receberam adubação. Contudo, diferente do que foi divulgado pela empresa, o adubo X-2020 não apresentou melhor desempenho que o adubo tradicional já utilizado pelos produtores.

2.8 ANOVA de dois fatores ou Anova fatorial

Este teste considera delineamentos amostrais com dois fatores (ou tratamento) que podem ser compostos por dois ou mais grupos (ou níveis). Esta análise tem uma vantagem, pois permite avaliar o efeito da interação entre os fatores na variável resposta. Quando a interação está presente, o impacto de um fator depende do nível (ou grupo) do outro fator.

2.8.0.1 Exemplo prático 1 - Anova de dois fatores

2.8.0.1.1 Explicação dos dados Neste exemplo, avaliaremos se o tempo que o corpo leva para elimiar uma droga utilizada em exames de ressonância magnética está relacionado com o sistema XY de determinação do sexo e/ou com a idade dos pacientes. Para isso, foi realizado um experimento com 40 pacientes distribuídos da seguinte maneira: i) 10 indivíduos XX - jovens, ii) 10

indivíduos XX - idosas, iii) 10 indivíduos XY - jovens, e iv) 10 indivíduos XY - idosos.

Pergunta:

> O tempo de eliminação da droga é dependente do sistema XY de determinação do sexo e idade dos pacientes?

Predições

> O tempo de eliminação da droga vai ser mais rápido nas pacientes XX e jovens.

Variáveis

- Variáveis preditoras
  - Dataframe com os pacientes (unidade amostral) nas linhas e os tratamentos nas colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e as variáveeis preditoras nas colunas.

2.8.1 Análise

Cálculo da Anova de dois fator

```
## IMPORTANDO DADOS
#*********************
dados_dois_fatores <- ecodados::anova_dois_fatores
head(dados_dois_fatores)
```

```
##    Tempo Pessoas Idade
## 1 18.952      XX Jovem
## 2 16.513      XX Jovem
## 3 17.981      XX Jovem
## 4 21.371      XX Jovem
## 5 14.470      XX Jovem
## 6 19.130      XX Jovem
```

```
# Análise anova de dois fatores
#*******************************
# A interação entre os fatores é representada por *
Modelo1 <- aov(Tempo ~ Pessoas * Idade, data = dados_dois_fatores)

# Olhando os resultados
anova(Modelo1)
```

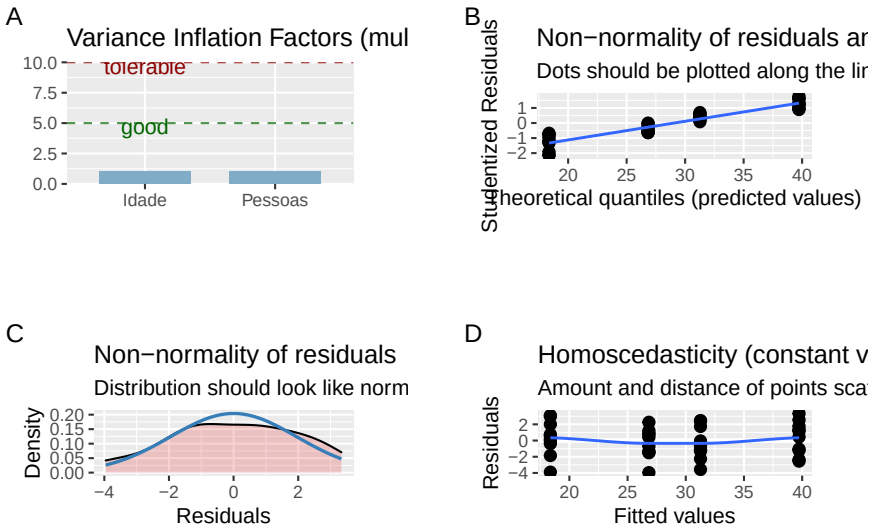
```
## Analysis of Variance Table
##
```

```
## Response: Tempo
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## Pessoas        1  716.72  716.72 178.8538  1.56e-15 ***
## Idade          1 1663.73 1663.73 415.1724 < 2.2e-16 ***
## Pessoas:Idade  1    4.77    4.77   1.1903    0.2825
## Residuals     36  144.26    4.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Percebam que a interação não apresenta um efeito significativo (P > 0.05). Assim, iremos retira
Modelo2 <- aov(Tempo ~ Pessoas + Idade, data = dados_dois_fatores)

# A hipótese nula é que o modelo mais simples é melhor
# Valores de p < 0.05 rejeita a hipótese nula e o modelo mais complexo é o melhor
# Valores de p > 0.05 não rejeita a hipótese nula e o modelo mais simples é o melhor
library(emmeans)
lrtest(Modelo1, Modelo2)
```

```
## Likelihood ratio test
##
## Model 1: Tempo ~ Pessoas * Idade
## Model 2: Tempo ~ Pessoas + Idade
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -82.413
## 2   4 -83.063 -1 1.3012      0.254
```

```r
# VERIFICANDO A NORMALIDADE E HOMOGENEIDADE DAS VARIÂNCIAS
#*****************************************************************
# Esta função mostra os resultados para multicolinearidade (a), dois gráficos avaliando a normali
library(sjPlot)
plot_grid(plot_model(Modelo2, type = "diag"))
```




```
# VERIFICANDO OS RESULTADOS DA ANOVA
#****************************************
anova(Modelo2)
```

```
## Analysis of Variance Table
##
## Response: Tempo
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Pessoas    1  716.72  716.72  177.94 1.041e-15 ***
## Idade      1 1663.73 1663.73  413.05 < 2.2e-16 ***
## Residuals 37  149.03    4.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Percebam que o resultado da Anova ($\Pr(>F) < 0.001$) indica que devemos rejeitar a hipótese nula que não há diferença entre as médias dos sistema XY e idade dos pacientes. Neste caso, não precisamos realizar testes de comparações múltiplas *post-hoc* porque os fatores apresentam apenas dois níveis. Contudo, se no seu delineamento experimental um dos fatores apresentar três ou mais níveis, você deverá utilizar os testes de comparações *post-hoc* para determinar as diferenças entre os grupos. **Observação** Os testes *post-hoc* só devem ser utilizados quando rejeitamos a hipótese nula ($P < 0.05$) no teste da Anova.

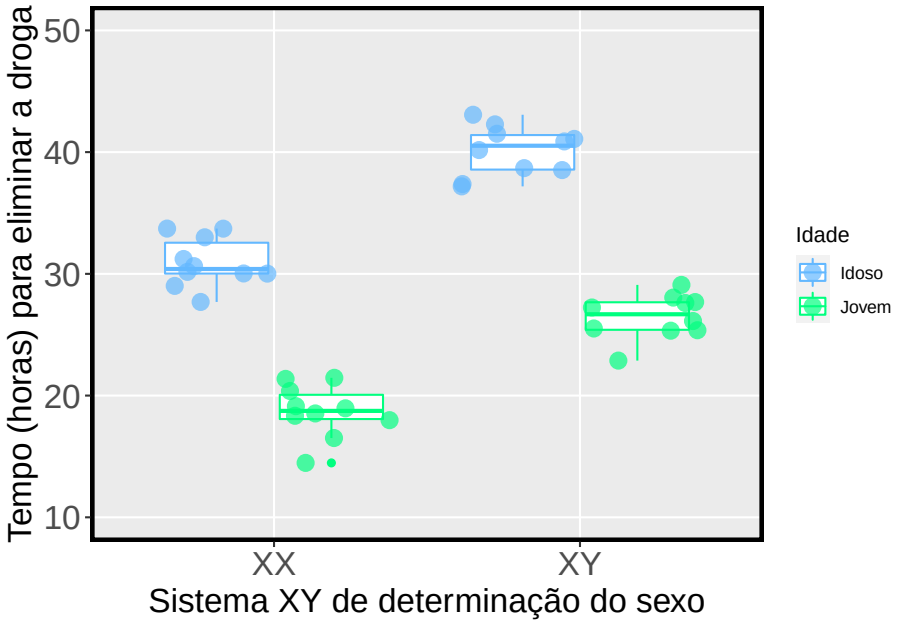
```
# Diferenças entre os tratamentos
#*********************************
# Teste de Tuckey's honest significant difference
```

```
TukeyHSD(Modelo2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Tempo ~ Pessoas + Idade, data = dados_dois_fatores)
##
## $Pessoas
##          diff      lwr      upr p adj
## XY-XX 8.46595 7.180008 9.751892     0
##
## $Idade
##                   diff       lwr       upr p adj
## Jovem-Idoso -12.89855 -14.18449 -11.61261     0
```

Visualizar os resultados em gráfico

```
# Gráfico
library(ggplot2)
ggplot(data = dados_dois_fatores, aes(y= Tempo, x= Pessoas, color = Idade)) +
  geom_boxplot() +
  labs(x = "Sistema XY de determinação do sexo", y = "Tempo (horas) para eliminar a droga",
       size = 20) +
  geom_jitter(shape = 16, position=position_jitterdodge(), cex = 4, alpha = 0.7) +
  scale_color_manual(values = c("steelblue1", "springgreen1")) +
  scale_y_continuous(limits = c(10, 50), breaks = c(10, 20, 30, 40, 50)) +
  theme(axis.title.y = element_text(size = 17), axis.title.x = element_text(size = 17)) +
  theme(axis.text.y = element_text(size = 17), axis.text.x = element_text(size = 17)) +
  theme(panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2))
```



Interpretação dos resultados

Neste exemplo, O sitesma XY de determinação do sexo e a idade dos pacientes tem um efeito no tempo de eliminação da droga do organismo. Os pacientes XX e jovens apresentam eliminação mais rápida da droga do que pacientes XY e idosos.

2.8.1.1 Exemplo prático 2 - Anova de dois fatores com efeito da interação

2.8.1.1.1 Explicação dos dados Neste exemplo usaremos os mesmos dados do exemplo anterior. Neste caso, alteremos os dados para que a interação seja significativa.

```
## IMPORTANDO DADOS
#####################
dados_dois_fatores_interacao <- ecodados::anova_dois_fatores_interacao1
head(dados_dois_fatores_interacao)
```

```
##    Tempo Pessoas Idade
## 1 31.092      XX Jovem
## 2 27.331      XX Jovem
## 3 24.874      XX Jovem
## 4 27.369      XX Jovem
## 5 28.125      XX Jovem
```

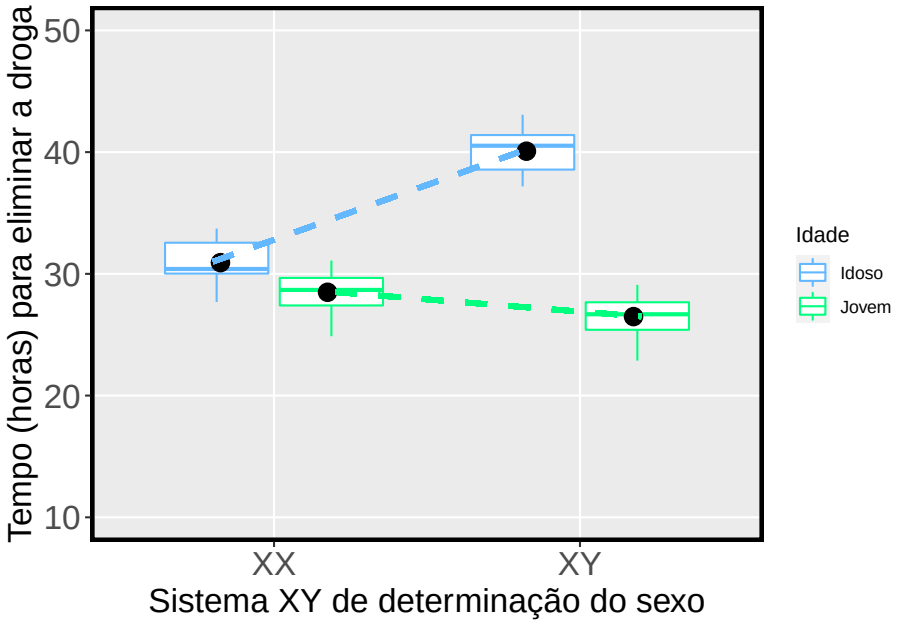
```
## 6 29.607      XX Jovem
```

```r
# Análise anova de dois fatores
#*********************************
# A interação entre os fatores é representada por *
Modelo_interacao1 <- aov(Tempo ~ Pessoas * Idade, data = dados_dois_fatores_interacao)

# Olhando os resultados
anova(Modelo_interacao1)
```

```
## Analysis of Variance Table
##
## Response: Tempo
##               Df Sum Sq Mean Sq F value    Pr(>F)
## Pessoas        1 128.04  128.04  34.841 9.377e-07 ***
## Idade          1 641.75  641.75 174.623 2.236e-15 ***
## Pessoas:Idade  1 311.17  311.17  84.672 5.463e-11 ***
## Residuals     36 132.30    3.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Percebam que a interação é significativa ($P < 0.05$). Agora nossa interpretação precisa ser baseada na interação entre os fatores. Vamos visualizar os resultados em gráfico.

```r
# Gráfico
library(ggplot2)
library(ggforce)
ggplot(data = dados_dois_fatores_interacao, aes(y= Tempo, x= Pessoas, color = Idade)) +
  geom_boxplot() +
  stat_summary(fun = mean, geom="point", aes(group=Idade, x = Pessoas), color = "black",
               position = position_dodge(0.7), size  = 4) +
  geom_link(aes(x = 0.8, y = 31, xend = 1.8, yend = 40), color = "steelblue1",
            lwd  = 1.3, linetype = 2) +
  geom_link(aes(x = 1.2, y = 28.5, xend = 2.2, yend = 26.5), color = "springgreen1",
            lwd  = 1.3, linetype = 2) +
  labs(x = "Sistema XY de determinação do sexo", y = "Tempo (horas) para eliminar a droga",
       size = 20) +
  scale_color_manual(values = c("steelblue1", "springgreen1", "steelblue1",
                                "springgreen1")) +
  scale_y_continuous(limits = c(10, 50), breaks = c(10, 20, 30, 40, 50)) +
  theme(axis.title.y = element_text(size = 17), axis.title.x = element_text(size = 17)) +
  theme(axis.text.y = element_text(size = 17), axis.text.x = element_text(size = 17)) +
  theme(panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2))
```



Interpretação dos resultados

Percebam que para saber a resposta do fator idade (jovem ou idoso) na eliminação da droga, você precisa saber com qual pessoa (XX ou XY) ele está associado. Isso porque a resposta de um fator depende do outro fator. Jovens eliminam o droga do corpo mais rápido nas pessoa XY enquanto os idosos eliminam a droga mais rápido nas pessoas XX.

2.8.1.2 Exemplo prático 3 - Anova de dois fatores com efeito da interação

2.8.1.2.1 Explicação dos dados Neste exemplo usaremos os mesmos dados do exemplo anterior. Neste caso, alteremos os dados para que a interação seja significativa.

```
## IMPORTANDO DADOS
#********************
dados_dois_fatores_interacao2 <- ecodados::anova_dois_fatores_interacao2
head(dados_dois_fatores_interacao2)
```

```
##    Tempo Pessoas Idade
## 1 18.952      XX Jovem
## 2 16.513      XX Jovem
## 3 17.981      XX Jovem
## 4 21.371      XX Jovem
```

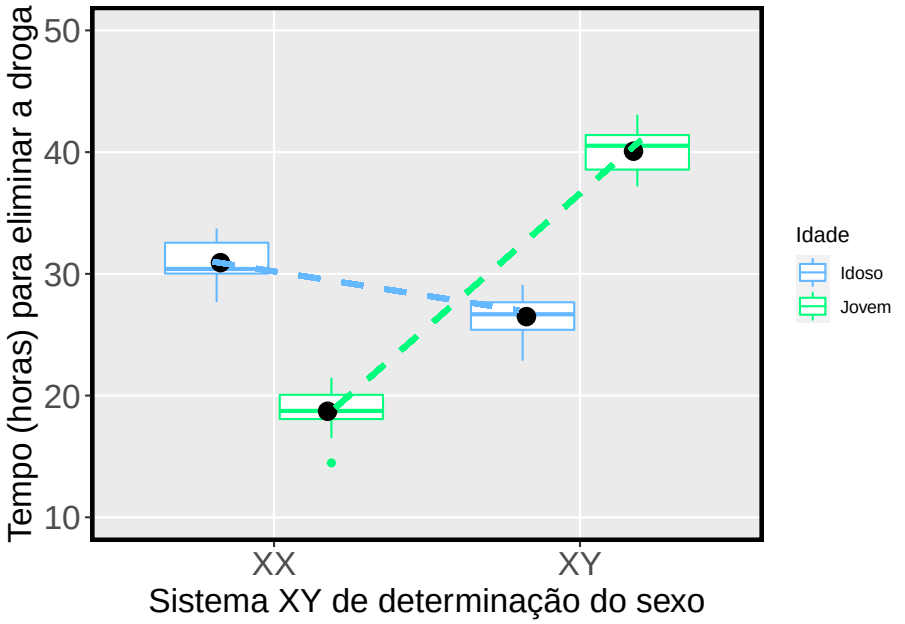
```
## 5 14.470      XX Jovem
## 6 19.130      XX Jovem
```

```
# Análise anova de dois fatores
#*********************************
# A interação entre os fatores é representada por *
Modelo_interacao2 <- aov(Tempo ~ Pessoas * Idade, data = dados_dois_fatores_interacao2)

# Olhando os resultados
anova(Modelo_interacao2)
```

```
## Analysis of Variance Table
##
## Response: Tempo
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## Pessoas        1  716.72  716.72 178.8538  1.56e-15 ***
## Idade          1    4.77    4.77   1.1903    0.2825
## Pessoas:Idade  1 1663.73 1663.73 415.1724 < 2.2e-16 ***
## Residuals     36  144.26    4.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Percebam que a interação é significativa ($P < 0.05$), mas a idade não é significativa. Nossa interpretação precisa ser baseada na interação entre os fatores. Vamos visualizar os resultados em gráfico.

```
# Gráfico
library(ggplot2)
library(ggforce)
ggplot(data = dados_dois_fatores_interacao2, aes(y= Tempo, x= Pessoas, color = Idade)) +
  geom_boxplot() +
  stat_summary(fun = mean, geom="point", aes(group=Idade, x = Pessoas), color = "black",
               position = position_dodge(0.7), size  = 4) +
  geom_link(aes(x = 0.8, y = 31, xend = 1.8, yend = 27), color = "steelblue1",
            lwd  = 1.3, linetype = 2) +
  geom_link(aes(x = 1.2, y = 19, xend = 2.2, yend = 41), color = "springgreen1",
            lwd  = 1.3, linetype = 2) +
  labs(x = "Sistema XY de determinação do sexo", y = "Tempo (horas) para eliminar a droga",
       size = 20) +
  scale_color_manual(values = c("steelblue1", "springgreen1", "steelblue1", "springgreen1")) +
  scale_y_continuous(limits = c(10, 50), breaks = c(10, 20, 30, 40, 50)) +
  theme(axis.title.y = element_text(size = 17), axis.title.x = element_text(size = 17)) +
  theme(axis.text.y = element_text(size = 17), axis.text.x = element_text(size = 17)) +
  theme(panel.grid.minor = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2))
```



Interpretação dos resultados

Percebam que as linhas se cruzam. Esse é exemplo clássico de interação. Novamente, para saber a resposta do fator idade (jovem ou idoso), você precisa saber com qual pessoa (XX ou XY) ele está associado. Jovens são mais rápidos para eliminar a droga em pessoas XX enquanto os idosos são mais rápidos para eliminar a droga nas pessoas XY.

2.9 ANOVA em blocos aleatorizados

No delineamento experimental com blocos aleatorizados, cada fator é agrupado em blocos, com réplicas de cada nível do fator representado em cada bloco (Gotelli & Elisson 2013). O bloco é uma área ou período de tempo dentro do qual as condições ambientais são relativamente homogêneas. O objetivo do uso dos blocos é controlar fontes de variações indesejadas na variável dependente que não são de interesse do pesquisador. Desta maneira, podemos retirar dos resíduos, os efeitos das variações indesejadas que não são do nosso interesse, e testar com maior poder estatístico os efeitos dos tratamentos de interesse. Importante, os blocos devem ser arranjados de forma que as condições ambientais sejam mais similares dentro dos blocos do que entre os blocos.

2.9.0.1 Exemplo prático 1 - Anova em blocos aleatorizados

2.9.0.1.1 Explicação dos dados Neste exemplo, avaliaremos a riqueza de espécies de anuros amostradas em poças artificiais instaladas a diferentes distâncias de seis fragmentos florestais no sudeste do Brasil (da Silva et al. 2020). Os fragmentos florestais apresentam diferenças entre si que não do interesse do pesquisador. Por isso, eles foram incluídos como blocos nas análises. As poças artificiais foram instaladas em todos os fragmentos florestais basedo no seguinte delineamento experimental (da Silva et al. 2012): i) 4 poças no interior do fragmento a 100m de distância da borda do fragmento; ii) 4 poças no interior no fragmento a 50m de distância da borda do fragmento; iii) 4 poças no na borda do fragmento; iv) 4 poças na matriz de pastagem a 50m de distância da borda do fragmento; e v) 4 poças na matriz de pastagem a 100m de distância da borda do fragmento. Percebam que todos os tratamentos foram instalados em todos os blocos.

Pergunta:

> A distância da poça artifical ao fragmento florestal influência a riqueza de espécies anuros?

Predições

> Poças na borda do fragmento florestal apresentarão maior riqueza de espécies do que poças distantes da borda.

Variáveis

- Variáveis preditoras
  - Dataframe com as poças (unidade amostral) nas linhas e o tratamento e bloco nas colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditoras nas colunas.

2.9.1 Análise

Cálculo da Anova em blocos aleatorizados

```
## IMPORTANDO DADOS
#*********************
dados_bloco <- ecodados::anova_bloco
str(dados_bloco) # verificar se o dataframe foi lido corretamente
```

```
## 'data.frame': 30 obs. of  3 variables:
##  $ Riqueza: int  90 95 107 92 89 92 81 92 93 80 ...
##  $ Blocos : chr  "A" "A" "A" "A" ...
##  $ Pocas  : chr  "Int-50m" "Int-100m" "Borda" "Mat-50m" ...
```

```r
# ANALISE ANOVA em blocos aleatorizados
#**************************************
# Há duas maneiras para incluir os efeitos dos blocos
model_bloco1 <- aov(Riqueza ~ Pocas + Blocos, data = dados_bloco)

model_bloco2 <- aov(Riqueza ~ Pocas + Error(Blocos), data = dados_bloco)

# Percebam que as duas formas apresentam os mesmo resultados para o efeito da distânci
anova(model_bloco1)
```

```
## Analysis of Variance Table
##
## Response: Riqueza
##           Df Sum Sq Mean Sq F value Pr(>F)
## Pocas      4 1504.5  376.12  2.9071 0.0478 *
## Blocos     5 1089.0  217.79  1.6834 0.1846
## Residuals 20 2587.5  129.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

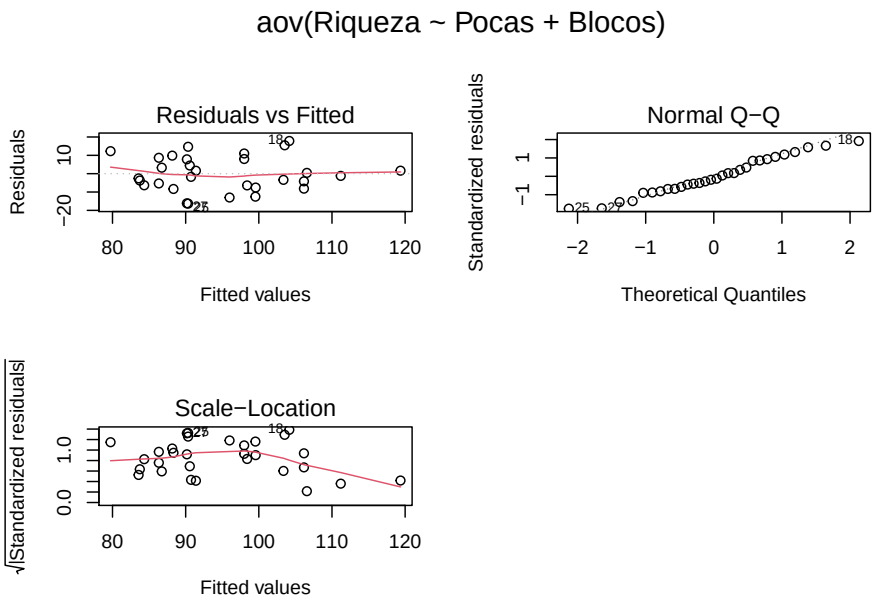
```r
summary(model_bloco2)
```

```
##
## Error: Blocos
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  5   1089   217.8
##
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## Pocas      4   1504   376.1   2.907 0.0478 *
## Residuals 20   2588   129.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# O que não pode acontecer é ignorar o efeito do bloco que é incorporado pelos resíduo

modelo_errado <- aov(Riqueza ~ Pocas, data = dados_bloco)
anova(modelo_errado)
```

```
## Analysis of Variance Table
##
## Response: Riqueza
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Pocas      4 1504.5  376.12  2.5576 0.06359 .
## Residuals 25 3676.5  147.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# VERIFICANDO A NORMALIDADE E HOMOGENEIDADE DAS VARIÂNCIAS
#************************************************************************
# Os gráficos *Residuals vs Fitted* e *Scale-Location* estão relacionados com a homogeneidade da
# O gráfico *Normal Q-Q* está relacionado com a distribuição normal dos resíduos. Neste gráfico,
par(mfrow = c(2, 2), oma = c(0, 0, 2, 0))
plot(model_bloco1)
```




```
dev.off()
```

```
## null device
##           1
```

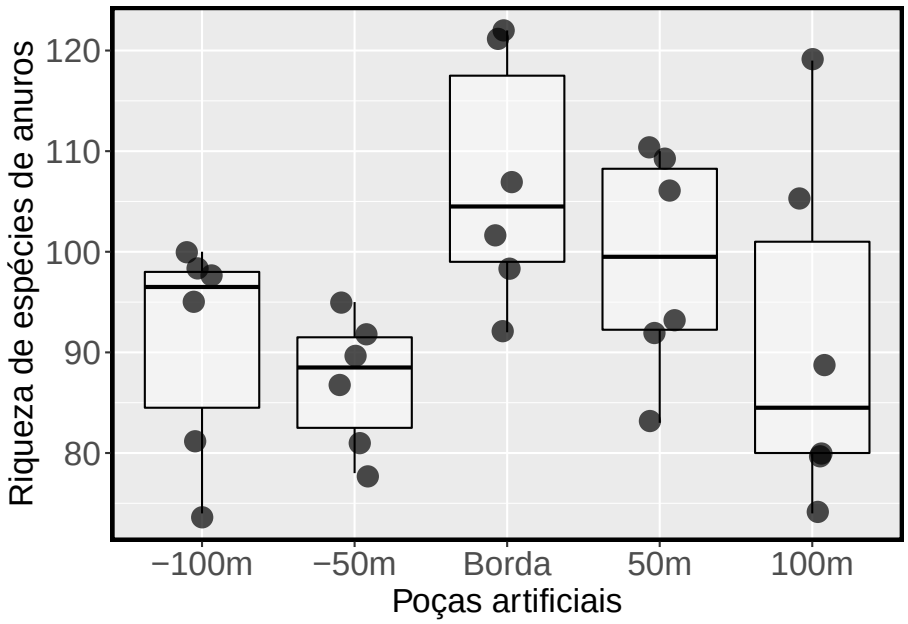
Percebam que o resultado da Anova ($\Pr(>F) < 0.001$) indica que devemos rejeitar a hipótese nula que não há diferença entre as médias dos grupos. Contudo, os resultados não mostram quais são os grupos que apresentam diferenças. Para isso, temos que realizar testes de comparações múltiplas *post-hoc* para detectar os grupos que apresentam diferenças significativas entre as médias. **Observação** Os testes *post-hoc* só devem ser utilizados quando rejeitamos a hipótese nula ($P < 0.05$) no teste da Anova.

```
# Diferenças entre os tratamentos
#*********************************
# Teste de Tuckey's honest significant difference
library(lsmeans)
pairs(lsmeans(model_bloco1, "Pocas"), adjust = "tukey")
```

```
##  contrast                estimate   SE df t.ratio p.value
##  Borda - (Int-100m)        16.000 6.57 20  2.436  0.1463
##  Borda - (Int-50m)         19.833 6.57 20  3.020  0.0472
##  Borda - (Mat-100m)        15.833 6.57 20  2.411  0.1531
##  Borda - (Mat-50m)          8.167 6.57 20  1.244  0.7269
##  (Int-100m) - (Int-50m)     3.833 6.57 20  0.584  0.9760
##  (Int-100m) - (Mat-100m)   -0.167 6.57 20 -0.025  1.0000
##  (Int-100m) - (Mat-50m)    -7.833 6.57 20 -1.193  0.7553
##  (Int-50m) - (Mat-100m)    -4.000 6.57 20 -0.609  0.9720
##  (Int-50m) - (Mat-50m)    -11.667 6.57 20 -1.777  0.4135
##  (Mat-100m) - (Mat-50m)    -7.667 6.57 20 -1.167  0.7692
##
## Results are averaged over the levels of: Blocos
## P value adjustment: tukey method for comparing a family of 5 estimates
```

Visualizar os resultados em gráfico

```
# Reordenando a ordem que os grupos irão aparecer no gráfico
dados_bloco$Pocas <- factor(dados_bloco$Pocas,
                            levels=c("Int-100m", "Int-50m", "Borda", "Mat-50m", "Mat

# Gráfico
library(ggplot2)
ggplot(data = dados_bloco, aes(x= Pocas, y= Riqueza)) +
  labs(x = "Poças artificiais", y = "Riqueza de espécies de anuros", size = 20) +
  geom_boxplot(color="black", show.legend = FALSE,
               alpha = 0.4) +
  geom_jitter(shape = 16, position=position_jitter(0.1), cex = 5, alpha = 0.7) +
  scale_x_discrete(labels=c("-100m","-50m","Borda", "50m", "100m")) +
  theme(axis.title.y = element_text(size = 17), axis.title.x = element_text(size = 17))
  theme(axis.text.y = element_text(size = 17), axis.text.x = element_text(size = 17)) +
  theme(panel.border = element_rect(colour = "black", fill=NA, size = 2))
```



Interpretação dos resultados

Neste exemplo, rejeitamos a hipótese nula que a distância da poças artificiais até as bordas dos fragmentos florestais não influência a riqueza de espécies de anuros. As poças artificiais instaladas nas bordas dos fragmentos florestais apresentaram maior riqueza de espécies do que as poças distantes.

2.10 Análise de covariância (ANCOVA)

A ANCOVA pode ser compreendida como uma extensão da Anova com a adição de variável contínua (covariável) medida em todas as unidades amostrais (Gotelli & Ellison 2013). A ideia é que a covariável também afete os valores da variável resposta. Não incluir a covariável irá fazer com que a variação não explicada pelo modelo concentre-se nos resíduos. Incluindo a covariável, o tamanho do resíduo é menor, e o teste para avaliar as diferenças nos tratamentos terá mais poder estatístico.

2.10.0.1 Exemplo prático 1 - ANCOVA

2.10.0.1.1 Explicação dos dados Neste exemplo, avaliaremos o efeito da herbivoria na biomassa dos frutos de uma espécie de árvore na Mata Atlântica. O delineamento experimental permitiu que alguns indivíduos sofressem

herbivoria e outros não. Os pesquisadores também mediram o tamanho da raiz dos indíviduos para inseri-la como uma covariável no modelo.

Pergunta:

> A herbivoria diminiu a biomssa dos frutos?

Predições

> Os indivíduos que sofreram herbivoria irão produzir frutos com menor biomassa do que os indivíduos sem herbivoria.

Variáveis

- Variáveis preditoras
  - Dataframe com as indivíduos da espéci de planta (unidade amostral) nas linhas e o tratamento e a covariável nas colunas.

Checklist

- Verificar se o seu dataframe está com as unidades amostrais nas linhas e variáveis preditoras nas colunas.

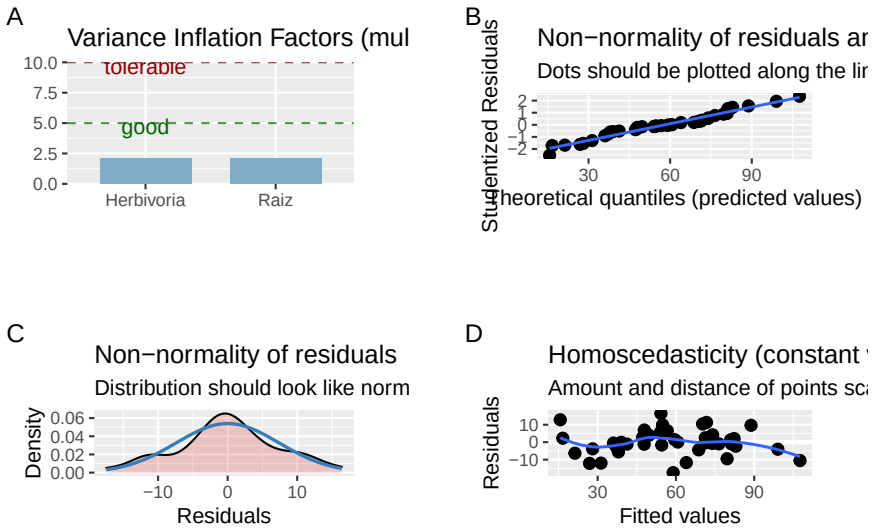
2.10.1 Análise

Cálculo da ANCOVA

```
## IMPORTANDO DADOS
#********************
dados_ancova <- ecodados::ancova
str(dados_ancova) # verificar se o dataframe foi lido corretamente
```

```
## 'data.frame': 40 obs. of  3 variables:
##  $ Raiz      : num  6.22 6.49 4.92 5.13 5.42 ...
##  $ Biomassa  : num  59.8 61 14.7 19.3 34.2 ...
##  $ Herbivoria: chr  "Sem_herb" "Sem_herb" "Sem_herb" "Sem_herb" ...
```

```
# ANALISE ANCOVA
#***************************************
modelo_ancova <- lm(Biomassa ~ Herbivoria + Raiz, data = dados_ancova)

# VERIFICANDO A NORMALIDADE E HOMOGENEIDADE DAS VARIÂNCIAS
#*************************************************************************
# Esta função mostra os resultados para multicolinearidade (a), dois gráficos avaliand
library(sjPlot)
plot_grid(plot_model(modelo_ancova, type = "diag"))
```




```
# OLHANDO OS RESULTADOS
#*********************
anova(modelo_ancova)
```

```
## Analysis of Variance Table
##
## Response: Biomassa
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Herbivoria  1  1941.9  1941.9  33.759 1.135e-06 ***
## Raiz        1 17434.1 17434.1 303.075 < 2.2e-16 ***
## Residuals  37  2128.4    57.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Percebam que o resultado da ANCOVA (Pr(>F) < 0.001) indica que tanto a herbivoria como a o tamanho da raiz (covariável) tem efeitos significativos na biomassa dos frutos. Contudo, a interação entre as variáveis não foi signigicativa. Vamos usar o Likelihood ratio test (LRT) para ver se podemos seguir com um modelo mais simples (sem interação).

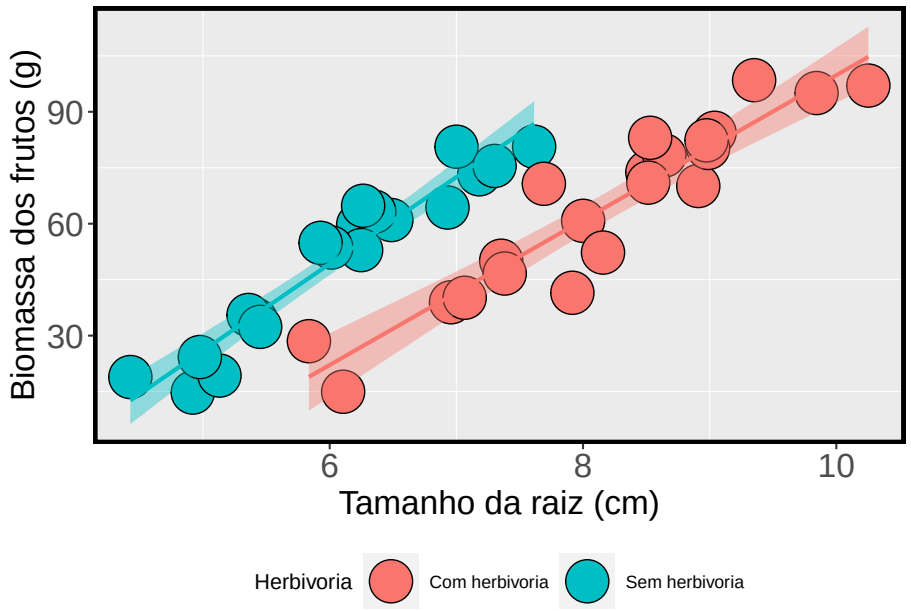
```
modelo_ancova <- lm(Biomassa ~ Herbivoria * Raiz, data = dados_ancova)
modelo_ancova2 <- lm(Biomassa ~ Herbivoria + Raiz, data = dados_ancova)

# LRT
library(lmtest)
# A hipótese nula é que o modelo mais simples é melhor
```

```
# Valores de p < 0.05 rejeita a hipótese nula e o modelo mais complexo é o melhor
# Valores de p > 0.05 não rejeita a hipótese nula e o modelo mais simples é o melhor
lrtest(modelo_ancova, modelo_ancova2)
```

```
## Likelihood ratio test
##
## Model 1: Biomassa ~ Herbivoria * Raiz
## Model 2: Biomassa ~ Herbivoria + Raiz
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -134.91
## 2   4 -136.24 -1 2.6554     0.1032
```

Visualizar os resultados em gráfico

```
# Gráfico
library(ggplot2)
ggplot(data = dados_ancova, aes(x= Raiz, y= Biomassa, fill = Herbivoria)) +
  labs(x = "Tamanho da raiz (cm)", y = "Biomassa dos frutos (g)", size = 20) +
  geom_point(size = 10, shape = 21) +
  theme(axis.title.y = element_text(size = 17), axis.title.x = element_text(size = 17))
  theme(axis.text.y = element_text(size = 17), axis.text.x = element_text(size = 17)) +
  theme(panel.grid.major = element_blank(),
        panel.border = element_rect(colour = "black", fill=NA, size = 2)) +
  theme(legend.position="bottom") +
  scale_fill_discrete(name = "Herbivoria", labels = c("Com herbivoria", "Sem herbivoria
  geom_smooth(aes(color=Herbivoria), method="lm", show.legend = FALSE)
```



Interpretação dos resultados

Neste exemplo, o tamanho da raiz (covariável) tem uma relação positiva com a biomassa dos frutos. Quanto maior o tamanho da raiz, maior a biomassa dos frutos. Usando a ANCOVA e controlando o efeito da covariável, percebemos que a herbivoria também afeta a biomassa dos frutos. Os indivíduos que não sofreram herbivoria produziram frutos com maior biomassa do que os indivíduos com herbivoria.

2.10.2 Para se aprofundar

- Recomendamos aos interessados os livros: i) Zar (2010) Biostatiscal analysis; ii) Gotelli & Ellison (2013) A primer of ecological statistics; e iii) Quinn & Keough (2002) Experimental design and data analysis for biologists.

Chapter 3

Introdução à Análises Multidimensionais

Neste módulo iremos aprender como implementar no R as análises multivariadas mais comumente utilizadas em ecologia de comunidades. Para isso precisaremos dos pacotes `vegan`,`labdsv` e `ade4`. Procuraremos explicar brevemente a lógica por trás de cada teste, a sua aplicação em problemas comumente encontrados em estudos ecológicos, mas não destrinchar detalhadamente como cada método funciona e o seu componente matemático.

Em geral, análises multivariadas têm três principais utilidades: reduzir a dimensionalidade dos dados e encontrar a principal direção de variação dos dados, testar relações entre matrizes, ou ainda encontrar diferenças entre grupos. Apesar dessas análises também serem utilizadas como análises exploratórias e para descrever padrões em estudos ecológicos, a necessidade de se ter hipóteses, ou ao menos expectativas a priori, não pode ser ignorada. Se quiser saber mais sobre aspectos teóricos e filosóficos das análises, sugerimos consultar James & McCulloch (1990). Antes de entrar de cabeça nas análises multivariadas também, sugerimos fortemente o estudo de métodos de amostragem e como fazer boas perguntas. Não vamos nos extender muito nesses tópicos porque eles foram abordados nas aulas disponíveis no YouTube.

Análises multivariadas podem ser divididas, grosseiramente, em dois tipos: agrupamento e ordenação. Análises de agrupamento em geral tentam agrupar objetos (observações) em grupos de maneira que objetos do mesmo grupo sejam mais semelhantes entre si do que objetos de outros grupos. Mais formalmente, o agrupamento de objetos (ou descritores) é uma operação pela qual um conjunto de objetos (ou descritores) é particionado em dois ou mais subconjuntos, usando regras pré-estabelecidas de aglomeração ou divisão (Legendre & Legendre, 2012). Por outro lado, a análise de ordenação é uma operação pela qual os objetos (ou descritores) são posicionados num espaço que contém menos dimen-

sões que o conjunto de dados original; a posição dos objetos ou descritores em relação aos outros também podem ser usadas para agrupá-los.

Vamos começar com análises de agrupamento. Aqui vamos exemplificar dois métodos: uma técnica de agrupamento hierarquica (dendrograma) e outra não-hierarquica (k-means.

3.1 Backgorund da análise

O objetivo da análise de agrupamento é agrupar objetos admitindo que haja um grau de similaridade entre eles. Esta análise pode ser utilizada ainda para classificar uma população em grupos homogêneos de acordo com uma característica de interesse. A grosso modo, uma análise de agrupamento tenta resumir uma grande quantidade de dados e apresentála de maneira fácil de visualizar e entender (em geral, na forma de um dendrograma). No entanto, os resultados da análise podem não refletir necessariamente toda a informação originalmente contida na matriz de dados. Para avaliar o quão bem uma análise de agrupamento representa os dados originais existe uma métrica — o coeficiente de correlação cofenético — o qual discutiremos em detalhes mais adiante.

Antes de considerar algum método de agrupamento, pense porque você esperaria que houvesse uma descontinuidade nos dados; ou ainda, considere se existe algum ganho prático em dividir uma nuvem de objetos contínuos em grupos. O padrão apresentado pelo dendograma depende do protocolo utilizado (método de agrupamento e índice de dissimilaridade); os grupos formados dependem do nível de corte escolhido. O leitor interessado é remetido à duas referências: Legendre & Legendre (2012) e Borcard et al. (2018).

3.2 Exemplo 1:

Pergunta:

> Existem grupos de espécies de anfíbios anuros com padrões de ocorrência similar ao longo de poças?

Predições

- 1: Iremos encontrar ao menos dois grupos de espécies: aquelas que ocorrem em poças dentro de floresta vs. aquelas que ocorrem em poças de áreas abertas.

Variáveis

- Variáveis preditoras
  - 1. A nossa matriz de dados contém a abundância das espécies nas linhas e locais (poças) nas colunas.

3.2.1 Explicação da análise

A matriz deve conter os objetos a serem agrupados (e.g., espécies) nas linhas e as variáveis (e.g., locais de coleta ou medidas morfológicas) nas colunas. A escolha do método de agrupamento é crítico para a escolha de um coeficiente de associação. É importante compreender as propriedades dos métodos de agrupamento para interpretar corretamente a estrutura ecológica que eles evidenciam (Legendre & Legendre, 2012). De acordo com a classificação de Sneath & Sokal (1973) existem cinco tipos de métodos: 1) seqüenciais ou simultâneos; 2) aglomerativo ou divisivo ;3) monotéticos ou politéticos; 4) hierárquico ou não hierárquicos e 5) probabilístico.

Métodos hierárquicos podem ser divididos naqueles que consideram o centróide ou amédia aritmética entre os grupos. O principal método hierárquico que utiliza a média aritmética é o UPGMA (Agrupamento pelas médias aritméticas não ponderadas), e o principal método que utiliza centróides é a Distância mínima de Ward.

O UPGMA funciona da seguinte forma: a maior similaridade (ou menor distância) identifica os próximos agrupamentos a serem formados. Após esse evento, o método calcula a média aritmética das similaridades ou distâncias entre um objeto e cada um dos membros do grupo ou, no caso de um grupo previamente formado, entre todos os membros dos dois grupos. Todos os objetos recebem pesos iguais no cálculo.

O método de Ward é baseado no critério de quadrados mínimos (OLS), o mesmo utilizado para ajustar um modelo linear. O objetivo é definir os grupos de maneira que a soma de quadrados (i.e. similar ao erro quadrado da ANOVA) dentro dos grupos seja minimizada (Borcard et al. 2018).

Checklist

- Verifique se não há espaço nos nomes das colunas e linhas
- Se os dados forem de abundância, recomenda-se realizar a transformação de Hellinger (Legendre & Gallagher, 2001).
- Se a matriz original contiver muitos valores discrepantes (e.g., uma espécie muito mais ou muito menos abundante que outras) é necessário transformar os dados usando `log1p`.
- Se as variáveis forem medidas tomadas em diferentes escalas (metros, graus celcius etc), é necessário padronizar cada variável para ter a média 0 e desvio padrão 1. Isso pode ser feito utulizando a função `decostand` do pacote `vegan`.

3.2.2 Análise

Para começar, vamos primeiro importar os dados e depois calcular a matriz de distância que seja adequada para o tipo de dado que temos (abundância de espécies - dados de contagem)

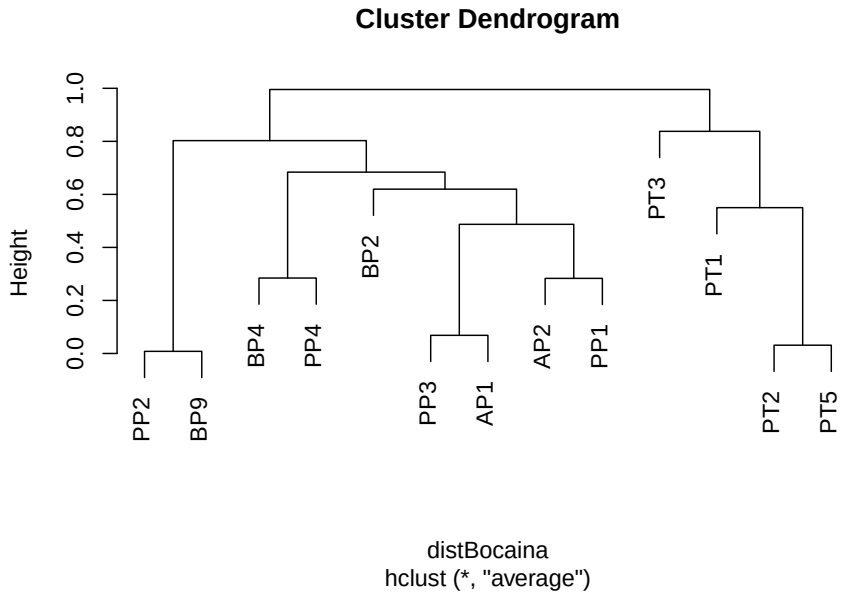
```
library(ecodados) # Carrega o arquivo multivar_bocaina
library(vegan)


#sp_compos <- read.table("bocaina.txt", h=TRUE)
sp_compos  <- multivar_bocaina
head(sp_compos)
```

```
##          BP4  PP4 PP3  AP1 AP2 PP1 PP2 BP9  PT1 PT2 PT3 BP2 PT5
## Aper       0    3   0    0   2   0   0   0    0   0   0 181   0
## Bahe     859   14  14    0  87 312 624 641    0   0   0  14   0
## Rict    1772 1517 207  573 796   0   0   0    0   0   0   0   0
## Cleuco     0    0   0    0   0   0   0   0    0  29 369   0  84
## Dmic       0    0   6   60   4   0   0   0 2758 319  25   0 329
## Dmin       0   84 344 1045  90   0   0   0    8   0   0   0   0
```

```
bocaina <- t(sp_compos)#transpondo a matriz para obter a classificação por linhas
distBocaina <- vegdist(bocaina, method="horn")#produz uma matriz de similaridade com o
dendro <- hclust(distBocaina, method="average")#produz um agrupamento com a função hcl
```

Visualizar os resultados

```
plot(dendro)
```



3.2.2.1 Interpretação dos resultados

Antes de começarmos a interpretar os resultados precisamos verificar que o agrupamento reduziu a dimensionalidade da matiz de forma eficiente, de maneira a não distorcer a informação. Fazemos isso calculando o **Coeficiente de correlação cofenética (CCC)**

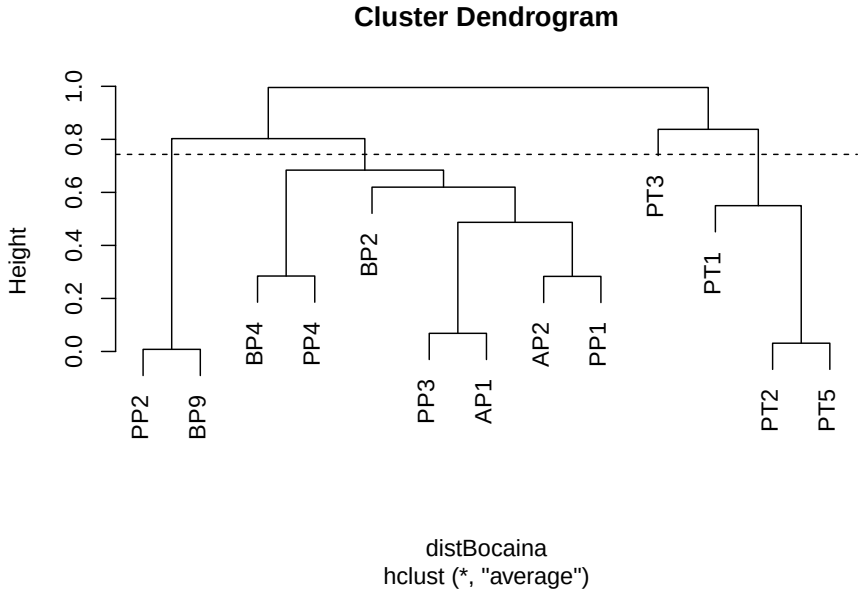
```
cofresult <- cophenetic(dendro)
cor(cofresult, distBocaina)
```

```
## [1] 0.8819701
```

Um CCC $>$ .7 indica uma boa representação. Portanto, o nosso resultado de 0.8819701 é bastante alto, garantindo que o dendrograma é adequado.

No entanto, para interpretar os resultados precisamos antes definir um nível de corte, que vai nos dizer quantos grupos existem. Há vários métodos para definir grupos, desde os heurísticos aos que utilizam bootstrap. Se quisermos interpretar este dendrograma, podemos por exemplo estabelecer um nível de corte de 50% de distância (ou seja, grupos cujos objetos tenham ao menos 50% de similaridade entre si).

```
plot(dendro)
k = 4
n = nrow(bocaina)
MidPoint = (dendro$height[n-k] + dendro$height[n-k+1]) / 2
abline(h = MidPoint, lty=2)
```



Nesse caso teremos a formação de cinco grupos, representados pelos nós que estão abaixo da linha de corte.

3.3 Exemplo 2:

A seguir, vamos utilizar o pacote `pvclust` que calcula automaticamente o nível de corte de similaridade baseado no Bootstrap de cada nó. Uma desvantagem deste método é que ele somente aceita índices de similaridade da função `dist` que possui apenas a distância Euclidiana, Manhattan e Canberra. Uma maneira de contornarmos essa limitação é utilizar transformações dos dados disponíveis na função `disttransform` no pacote `BiodiversityR` ou o `decostand` do pacote `vegan`. Também é possível utilizar a transformação de Box-Cox para dados multivariados, disponível no material suplementar de Legendre & Borcard (2018) aqui

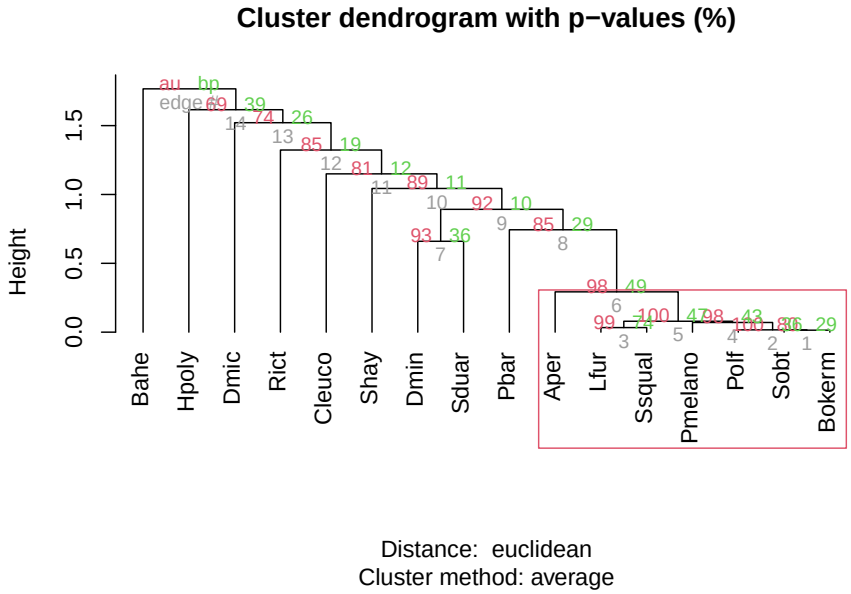
```
library(pvclust)
library(BiodiversityR)
```

Aqui vamos utilizar a distância de Chord para calcular a matriz de distância. Se transformarmos uma matriz usando a transformação Chord e depois calcularmos a distância Euclidiana, isso equivale à calcular diretamente a distância de Chord:

```
bocaina_transf <- disttransform(bocaina, "chord")
analise <- pvclust(bocaina_transf, method.hclust="average", method.dist="euclidean")
```

```
## Bootstrap (r = 0.46)... Done.
## Bootstrap (r = 0.54)... Done.
## Bootstrap (r = 0.69)... Done.
## Bootstrap (r = 0.77)... Done.
## Bootstrap (r = 0.85)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.08)... Done.
## Bootstrap (r = 1.15)... Done.
## Bootstrap (r = 1.23)... Done.
## Bootstrap (r = 1.38)... Done.
```

```
plot(analise, hang=-1)
pvrect(analise)
```

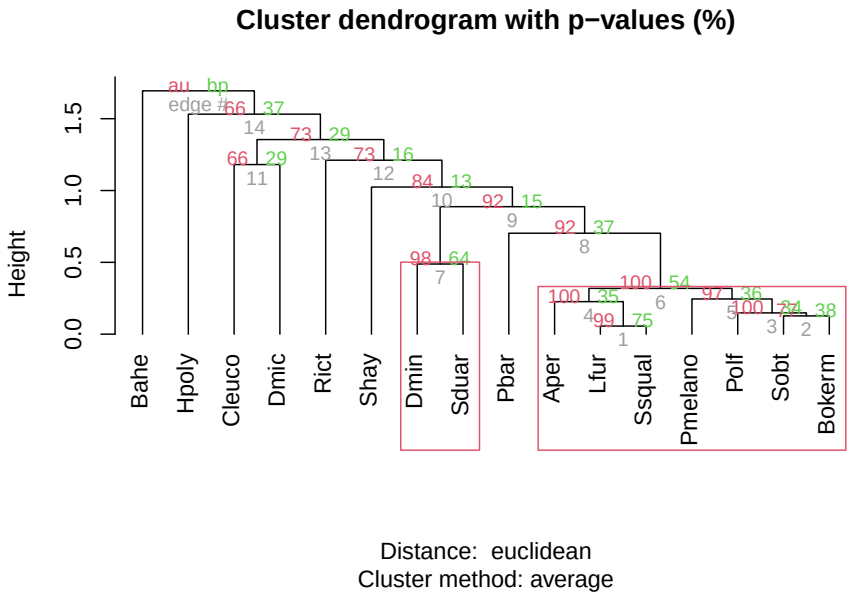


É possível notar que existe um único grupo com BS $> 95\%$. Agora vamos tentar usar a distância de Hellinger:

```
bocaina_transf2 <- disttransform(bocaina, "hellinger")
analise2 <- pvclust(bocaina_transf2, method.hclust="average", method.dist="euclidean")
```

```
## Bootstrap (r = 0.46)... Done.
## Bootstrap (r = 0.54)... Done.
## Bootstrap (r = 0.69)... Done.
## Bootstrap (r = 0.77)... Done.
## Bootstrap (r = 0.85)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.08)... Done.
## Bootstrap (r = 1.15)... Done.
## Bootstrap (r = 1.23)... Done.
## Bootstrap (r = 1.38)... Done.
```

```
plot(analise2, hang=-1)
pvrect(analise2)
```



Notem que se mudarmos o coeficiente de associação, o resultado também muda. Agora temos 1 grupo a mais, composto por *Dendropsophus minutus* e *Scinax duartei* que não apareciam antes. Isso se deve ao fato de que a distância de Hellinger dá menos peso para espécies raras do que a Chord.

Chapter 4

K-means e agrupamentos não-hierarquicos

4.1 Backgorund da análise

K-means é um tipo de agrupamento não hierarquico porque não busca obter grupos menores que por sua vez pertencem a grupos maiores. Resumidamente, podemos calcular o K-means apartir de uma matriz quadrada ou de distância. Essa técnica procura particionar os objetos em k grupos de maneira a minimizar a soma de quadrados entre grupos e maximizá-la dentro dos grupos. Um critério similar ao de uma ANOVA.

4.2 Exemplo 1:

Pergunta:

> Qual é o número de grupos que melhor sumariza o padrão de ocorrência de espécies de peixes ao longo de um riacho?

Predições

- 1: O agrupamento ideal para explicar a variância no padrão de ocorrência de espécies é 4.

Variáveis

- Variáveis resposta
  - 1. Para este exemplo iremos utilizar um conjunto de dados disponível no pacote `ade4` que contém dados de 27 espécies de peixes coletados em 30 pontos ao longo do Rio Doubs, na fronteira entre a França e Suiça.

4.2.1 Explicação da análise

Um diferencial do K-means em relação aos agrupamentos hierarquicos (=clusters) é que o usuário pode escolher antecipadamente o número de grupos que quer formar.

Checklist

- Vamos normalizar os dados de abundância antes de entrar na análise propriamente, já que existem muitos zeros na matriz.

4.2.2 Análise

```
library(ade4)
data(doubs)
head(doubs$fish)
```

```
##   Cogo Satr Phph Neba Thth Teso Chna Chto Lele Lece Baba Spbi Gogo Eslu Pefl
## 1    0    3    0    0    0    0    0    0    0    0    0    0    0    0    0
## 2    0    5    4    3    0    0    0    0    0    0    0    0    0    0    0
## 3    0    5    5    5    0    0    0    0    0    0    0    0    0    1    0
## 4    0    4    5    5    0    0    0    0    0    1    0    0    1    2    2
## 5    0    2    3    2    0    0    0    0    5    2    0    0    2    4    4
## 6    0    3    4    5    0    0    0    0    1    2    0    0    1    1    1
##   Rham Legi Scer Cyca Titi Abbr Icme Acce Ruru Blbj Alal Anan
## 1    0    0    0    0    0    0    0    0    0    0    0    0
## 2    0    0    0    0    0    0    0    0    0    0    0    0
## 3    0    0    0    0    0    0    0    0    0    0    0    0
## 4    0    0    0    0    1    0    0    0    0    0    0    0
## 5    0    0    2    0    3    0    0    0    5    0    0    0
## 6    0    0    0    0    2    0    0    0    1    0    0    0
```

```
spe <- doubs$fish[-8,]# retiro a linha 8, pois não há dados
spe.norm <- decostand(spe, "normalize") # função do pacote vegan, ela faz várias padro
```

O argumento `centers` na função abaixo indica o número de grupos que se quer formar. Neste exemplo estamos utilizando `centers=4`.

```
spe.kmeans <- kmeans(spe.norm, centers=4, nstart=100)
spe.kmeans
```

```
## K-means clustering with 4 clusters of sizes 6, 3, 12, 8
##
## Cluster means:
##         Cogo        Satr       Phph       Neba        Thth        Teso
## 1 0.06167791 0.122088022 0.26993915 0.35942538 0.032664966 0.135403325
## 2 0.00000000 0.000000000 0.00000000 0.00000000 0.000000000 0.000000000
## 3 0.10380209 0.542300691 0.50086515 0.43325916 0.114024105 0.075651573
```

```
## 4 0.00000000 0.006691097 0.02506109 0.06987391 0.006691097 0.006691097
##         Chna       Chto       Lele      Lece       Baba      Spbi       Gogo
## 1 0.06212775 0.21568957 0.25887226 0.2722562 0.15647062 0.1574388 0.16822286
## 2 0.05205792 0.00000000 0.07647191 0.3166705 0.00000000 0.0000000 0.20500174
## 3 0.00000000 0.00000000 0.06983991 0.1237394 0.02385019 0.0000000 0.05670453
## 4 0.10687104 0.09377516 0.14194394 0.2011411 0.24327992 0.1326062 0.28386032
##         Eslu       Pefl      Rham       Legi       Scer       Cyca       Titi
## 1 0.12276089 0.17261621 0.0793181 0.06190283 0.04516042 0.06190283 0.14539027
## 2 0.07647191 0.00000000 0.0000000 0.05205792 0.07647191 0.00000000 0.00000000
## 3 0.04722294 0.02949244 0.0000000 0.00000000 0.00000000 0.00000000 0.03833408
## 4 0.20630360 0.16920496 0.2214275 0.19066542 0.13171275 0.16019126 0.26230024
##         Abbr      Icme       Acce       Ruru       Blbj      Alal       Anan
## 1 0.01473139 0.0000000 0.03192175 0.32201597 0.01473139 0.1095241 0.04739636
## 2 0.00000000 0.0000000 0.18058775 0.31667052 0.05205792 0.7618709 0.00000000
## 3 0.00000000 0.0000000 0.00000000 0.01049901 0.00000000 0.0000000 0.00000000
## 4 0.19561641 0.1331835 0.26713081 0.32103755 0.22883055 0.3326939 0.18873077
##
## Clustering vector:
##  1  2  3  4  5  6  7  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
##  3  3  3  3  1  3  3  1  3  3  3  3  3  3  1  1  1  1  4  4  4  2  2  2  4  4
## 28 29 30
##  4  4  4
##
## Within cluster sum of squares by cluster:
## [1] 1.7361453 0.3560423 2.5101386 0.4696535
##  (between_SS / total_SS =  66.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

O objeto que fornece o resultado contém: 1) o tamanho (número de objetos) em cada um dos 4 grupos; 2) o centroid de cada grupo e o pertencimento de cada espécie a cada grupo; e 3) o quando da Soma de Quadrados dos dados é explicada por esta conformação de grupos.

No entanto, não é possível saber a priori qual o número *ideal* de grupos. Para descobrir isso repetimos o k-means com uma série de valores de **K**. Isso pode ser feito na função `cascadeKM`.

```
spe.KM.cascade <- cascadeKM(spe.norm, inf.gr=2, sup.gr=10, iter=100, criterion="ssi")
```

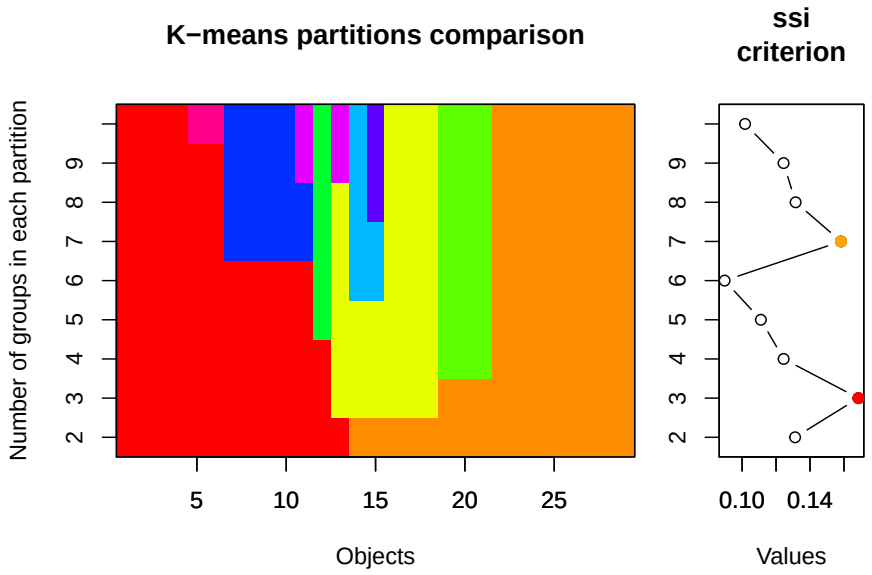
Tanto **calinski** quando **ssi** são bons critérios para encontrar o número ideal de grupos. Quanto maior o valor de **ssi** melhor (veja `?cascadeKM` mais detalhes).

```
# summary
spe.KM.cascade$results
```

```
##      2 groups  3 groups  4 groups  5 groups   6 groups  7 groups  8 groups
## SSE 8.2149405 6.4768108 5.0719796 4.3015573 3.58561200 2.9523667 2.4840549
## ssi 0.1312111 0.1685126 0.1244603 0.1110514 0.08971007 0.1582497 0.1315299
##      9 groups 10 groups
## SSE 2.0521888  1.759929
## ssi 0.1244271  0.101762
```

SSE: critério utilizado pelo algorítimo para achar o agrupamento ótimo dos objetos.

```
plot(spe.KM.cascade, sortg=TRUE)
```



Este resultado nos mostra que o número ideal de grupos é 3, vejam que o SSI máximo é alcançado neste número de grupos 0.1685126 (também indicado pela bola vermelha no plot).

#Espécies indicadoras

4.3 Backgorund da análise

Uma pergunta normalmente feita por ecólogos é: qual espécie pode ser indicadora de uma determinada condição ambiental?

O índice IndVal mede dois aspectos das espécies: Especificidade e fidelidade. Uma alta fidelidade significa que espécies ocorrem em todos os locais do grupo e uma alta especificidade significa que as espécies ocorrem somente naquele grupo. Uma boa espécie indicadora é aquela na qual todos os indivíduos ocorrem em

todas a amostras referentes a um grupo específico. A Especificidade é dada pela divisão da abundancia média da espécie no grupo pela somatória das abundancias médias dos grupos. Fidelidade é igual ao número de lugares no grupo onde a espécie está presente dividido pelo número total de lugares do grupo (Dufrêne & Legendre, 1997).

Espécies raras podem receber o mesmo valor de IndVal das espécies indicadoras e são chamadas de indicadoras assimétricas, i.e., contribuem com a especificidade do habitat mas não servem para predizer grupos. Ao contrário, as espécies indicadoras são verdadeiros indicadores simétricos e podem ser usadas para predizer grupos.

4.4 Exemplo 1:

Pergunta:

> Qual espécie de anfíbio anuro na fase larval pode ser indicadora da fitofisionomia onde é encontrada?

Predições

- 1: Espécies terrestres serão indicadoras de área aberta, enquanto espécies arborícolas serão indicadoras de áreas florestais.

Variáveis

- Variáveis resposta
  - 1. Mesma matriz já utilizada contendo a abundância de girinos ao longo de poças na Serra da Bocaina.

4.4.1 Explicação da análise

A análise procede da seguinte forma:

- 1. Uma matriz de distância é construída e as unidades amostrais são classificadas com alguma análise de agrupamento, hierárquico ou não;
- 2. A variável ambiental para a qual se deseja classificar os grupos é inserida;
- 3. As espécies indicadoreas de cada grupo são formadas através do cálculo da especificidade e fidelidade, obtendo-se o valor de IndVal para cada espécie;
- 4. Por fim, o conjunto de dados originais é comparado para ver se análise faz sentido.

O cálculo da significância do índice de IndVal é feito por aleatorização de Monte Carlo. Assim, o valor do índice é aleatorizado 999 vezes (ou o número de vezes que você optar) dentro dos tratamentos e o valor de P é dado pelo número de vezes em que o índice observado foi igual ou maior que os valores aleatorizados.

4.4.2 Análise

O IndVal está disponível tanto no pacote `indicspecies` quando no `labdsv`. Para este exemplo iremos usar o labdsv.

```
library(labdsv)
```

Primeiro vamos agrupar as unidades amostrais (poças) que informa os grupos de fitofisionomias onde as poças se localizam e para os quais deseja-se encontrar espécies indicadoras:

```
fitofis <- c(rep(1,4), rep(2,4), rep(3,4), rep(4,4), rep(5,4))
```

```
resultado <- indval(bocaina, fitofis)
summary(resultado)#só exibe o resultado para as espécies indicadoras
```

```
##       cluster indicator_value probability
## Rict        1          0.8364       0.011
## Sduar       1          0.7475       0.045
##
## Sum of probabilities                   =  8.077
##
## Sum of Indicator Values                =  7.3
##
## Sum of Significant Indicator Values  =  1.58
##
## Number of Significant Indicators     =  2
##
## Significant Indicator Distribution
##
## 1
## 2
```

Para apresentar uma tabela dos resultados para todas as espécies temos de processar os dados:

```
resultado$maxcls
```

```
##    Aper    Bahe    Rict  Cleuco    Dmic    Dmin   Hpoly    Lfur    Pbar    Polf
##       3       2       1       3       3       1       2       3       1       3
## Pmelano   Sduar    Shay    Sobt  Ssqual  Bokerm
##       2       1       3       2       3       2
```

```
resultado$indcls
```

```
##      Aper      Bahe      Rict    Cleuco      Dmic      Dmin     Hpoly      Lfur
## 0.2432796 0.6487329 0.8363823 0.4128631 0.6645244 0.7032145 0.6208711 0.2279412
##      Pbar      Polf   Pmelano     Sduar      Shay      Sobt    Ssqual    Bokerm
## 0.2813725 0.2437500 0.2500000 0.7474527 0.4930269 0.2222222 0.2500000 0.4583333
```

```
resultado$pval
```

```
##    Aper    Bahe    Rict  Cleuco    Dmic    Dmin   Hpoly    Lfur    Pbar    Polf
##   1.000   0.063   0.011   0.444   0.200   0.075   0.262   1.000   0.585   1.000
## Pmelano   Sduar    Shay    Sobt  Ssqual  Bokerm
##   1.000   0.045   0.450   0.715   1.000   0.227
```

```
tab.resultado=cbind(resultado$maxcls,resultado$indcls,resultado$pval)
colnames(tab.resultado)<-c("maxgrp", "ind. value","P")
tab.resultado
```

```
##         maxgrp ind. value     P
## Aper         3  0.2432796 1.000
## Bahe         2  0.6487329 0.063
## Rict         1  0.8363823 0.011
## Cleuco       3  0.4128631 0.444
## Dmic         3  0.6645244 0.200
## Dmin         1  0.7032145 0.075
## Hpoly        2  0.6208711 0.262
## Lfur         3  0.2279412 1.000
## Pbar         1  0.2813725 0.585
## Polf         3  0.2437500 1.000
## Pmelano      2  0.2500000 1.000
## Sduar        1  0.7474527 0.045
## Shay         3  0.4930269 0.450
## Sobt         2  0.2222222 0.715
## Ssqual       3  0.2500000 1.000
## Bokerm       2  0.4583333 0.227
```

No resultado podemos ver que temos duas espécies indicadoras da fitofisionimia 1: *Rhinella icterica* (Rict) e *Scinax duartei* (Sduar). Nenhuma espécie foi indicadora dos outros grupos neste exemplo.

4.4.3 Para se aprofundar

- Agrupamento de espécies e locais baseado em modelos
- Numerical Ecology with R

Chapter 5

Rarefação

5.1 Background da análise

Uma das dificuldades na comparação da riqueza de espécies entre comunidades é decorrente da diferença no esforço amostral (e.g.diferença no número de indivíduos, discrepância na quantidade de unidades amostrais ou área amostrada) que inevitavelmente influenciará no número de espécies observadas (Gotelli & Chao 2013). O método de rarefação nos permite comparar o número de espécies entre comunidades quando o tamanho da amostra ou a abundância de indivíduos não são iguais. A rarefação calcula o número esperado de espécies em cada comunidade tendo como base comparativa um valor em que todas as amostras atinjam um tamanho padrão, ou comparações baseadas na comunidade com menor número de amostragens ou com menos indivíduos. O teste foi formulado considerando seguinte pergunta: Se considerarmos n indivíduos ou amostras (n < N) para cada comunidade, quantas espécies registraríamos nas comunidades considerando o mesmo número de indivíduos ou amostras?

$$E(S) = \sum 1 - \frac{(N - N_1)/n}{N/n}$$

Onde:

- E(S) = Número de espécies esperado,
- N = Número total de indivíduos na amostra,
- Ni = Número de indivíduos da iésima espécie,
- n = tamanho da amostra padronizada (menor amostra).

Gotelli & Collwel (2001) descrevem este método e discutem em detalhes as restrições sobre seu uso na ecologia:

- As amostras a serem comparados devem ser consistentes do ponto de vista taxonômico, ou seja, todos os indivíduos devem pertencer ao mesmo grupo taxonômico;

- As comparações devem ser realizadas somente entre amostras com as mesmas técnicas de coleta;

- Os tipos de hábitat onde as amostras são obtidas devem ser semelhantes;

- É um método para estimar a riqueza de espécies em uma amostra menor – não pode ser usado para extrapolar e estimar riqueza.

Contudo, é importante ressaltar que esta última restrição foi superada por Colwell et al. (2012) e Chao & Jost (2012) que desenvolveram uma nova abordagem onde os dados podem ser interpolados (rarefeito) para amostras menores e extrapolados para amostras maiores.

5.2 Exemplo prático 1 - Morcegos

5.2.1 Explicação

Explicação dos dados

Neste exemplo usaremos os dados de espécies de morcegos amostradas em três fragmentos florestais (Breviglieri 2008): i) Mata Ciliar do Córrego Talhadinho com 12 hectares inserida em uma matriz de pastagem; ii) Mata Ciliar do Córrego dos Tenentes com 10 hectares inserida em uma matriz de cultivo de cana-de-açucar e pastagem; e iii) Fazenda Experimental de Pindorama com 128 hectares inserida uma matriz de cana-de-açúcar e pastagem.

Pergunta:

> A riqueza de espécies de morcegos é maior na Fazenda Experimental do que nos fragmentos florestais menores?

Predições

> O número de espécies será maior em fragmentos florestais maiores.

Variáveis

- Variáveis preditoras
  - matriz ou dataframe com as abundâncias das espécies de morcegos registradas nos três fragmentos florestais

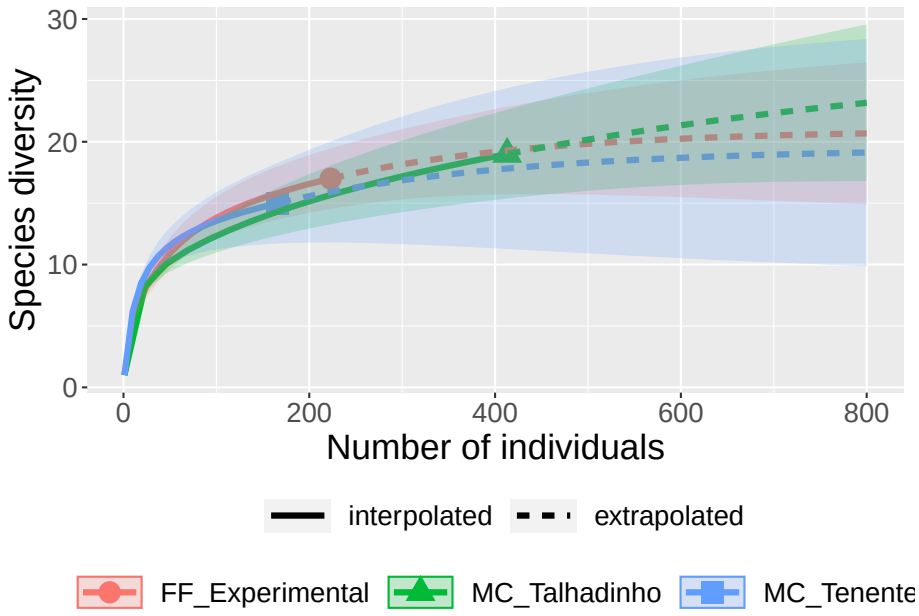
Checklist

- Verificar se a sua matriz ou dataframe estão com as espécies nas linhas e os fragmentos florestais nas colunas

5.2.2 Análise

Calculo da rarefação

```
library(iNEXT)
library(devtools)
devtools::install_github("paternogbc/ecodados")
library(ecodados)

dados_rarefacao <- rarefacao_morcegos
resultados_morcegos <- iNEXT(dados_rarefacao, q = 0, datatype = "abundance", endpoint = 800)
# q refere-se a família *Hill-numbers* (Hill 1973) onde 0 = riqueza de espécies, 1 =  diversidade
# Veja mais detalhes sobre os números de Hill no Capítulo 7 onde tratamos de extrapolações.
# datatype refere-se ao tipo de dados que você vai analisar (e.g. abudância, incidência).
# endpoint refere-se ao valor de referência que você determina para a extrapolação.

# Visualizar os resultados
ggiNEXT(resultados_morcegos, type = 1)
```



5.2.3 Interpretação dos resultados

Neste exemplo, foram registrados 166 indivíduos na MC_Tenentes, 413 na MC_Talhadinho e 223 na FF_Experimental. Lembrando, você não pode comparar a riqueza de espécies observada diretamente: 15 espécies na MC_Tenentes, 17 espécies na MC_Talhadinho, e 13 espécies no FF_Experimental. A comparação da riqueza de espécies entre as comunidades

deve ser feita com base na riqueza de espécies estimada que é calculada com base no número de indivíduos da comunidade com menor abundância (166 indivíduos). Olhando o gráfico é possível perceber que a riqueza de espécies de morcegos estimada não é diferente entre os três fragmentos florestais quando corrigimos o problema da abundância pela rarefação. A interpretação é feita com base no intervalo de confiança de 95%. As curvas serão diferentes quando os intervalos de confiança não se sobreporem (Chao et al. 2014). Percebam que está abordagem, além da interpolação (rarefação), também realiza extrapolações que podem ser usadas para estimar o número de espécies caso o esforço de coleta fosse maior. Este é o assunto do nosso próximo capítulo.

5.3 Exemplo prático 2 - Rarefação

5.3.1 Explicação

Explicação dos dados

Neste exemplo iremos comparar o número de espécies de anuros e répteis (serptentes e lagartos) usando informações dos indivíduos depositados em coleções científicas e coletas de campo (da Silva et al. 2017).

Pergunta:

> A riqueza de espécies de anuros e répteis é maior em coleções científicas do que nas coletas de campo?

Predições

> O número de espécies será maior em coleções científicas devido ao maior esforço amostral (i.e. maior variação temporal para depositar os indíviduos e maior número de pessoas contribuindo com as informações de diferentes estudos e/ou coletas esporádicas).

Variáveis

- Variáveis preditoras
  - matriz ou dataframe com as abundâncias das espécies de anuros e répteis (planilhas separadas) registradas em coleções científicas e coletas de campo.

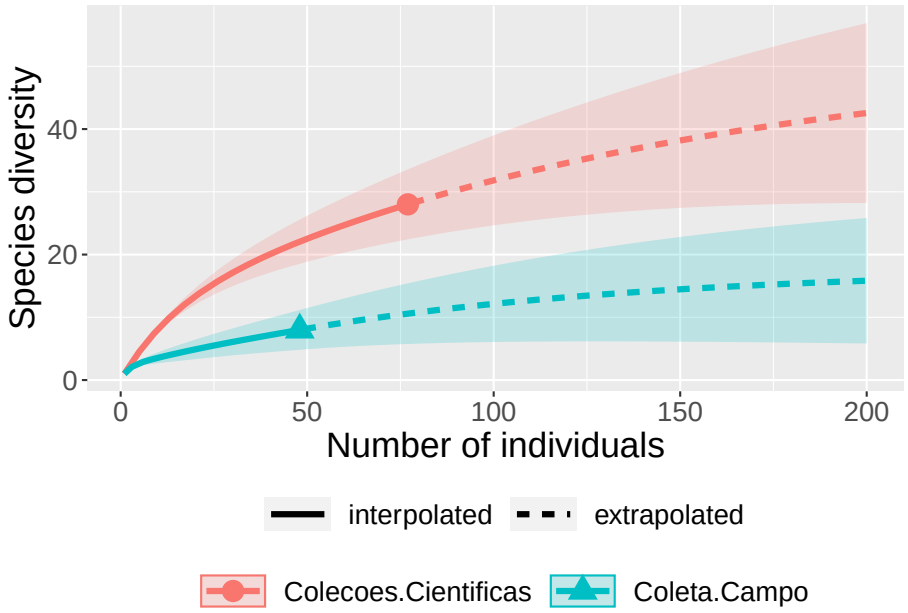
Checklist

- Verificar se a sua matriz ou dataframe estão com as espécies nas linhas e a fonte dos dados nas colunas.

5.3.2 Análise

Calculo da rarefação para os dados de répteis

```
library(iNEXT)

rarefacao_repteis <- rarefacao_repteis
resultados_repteis <- iNEXT(rarefacao_repteis, q = 0, datatype = "abundance", endpoint = 200)

# Visualizar os resultados
ggiNEXT(resultados_repteis, type = 1)
```

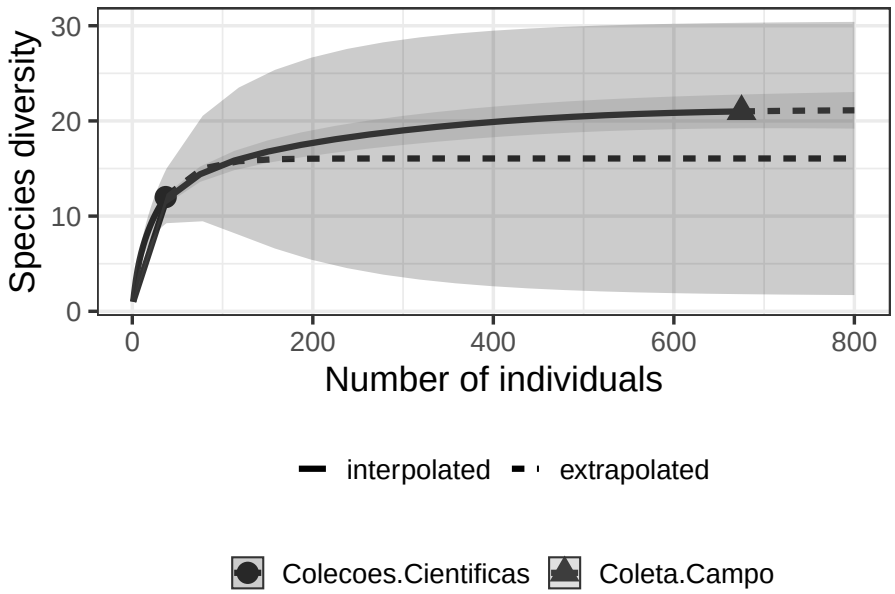


5.3.3 Interpretação dos resultados

Neste exemplo,foram registradas oito espécies de répteis nas coletas de campo (40 indivíduos) e 28 espécies nas coleções científicas (77 indivíduos). Com base na rarefação, concluímos que a riqueza de espécies de répteis obtida nas coleções científicas é 2,5 vezes maior do que a obtida em coletas de campo.

Calculo da rarefação para os dados dos anuros

```
library(iNEXT)

rarefacao_anuros <- rarefacao_anuros
resultados_anuros <- iNEXT(rarefacao_anuros, q = 0, datatype = "abundance", endpoint = 800)

# Visualizar os resultados
ggiNEXT(resultados_anuros, type = 1, grey = TRUE)
```



Interpretação dos resultados

Neste exemplo,foram registradas 21 espécies de anuros nas coletas de campo (709 indivíduos) e 12 espécies nas coleções científicas (37 indivíduos). Com base na rarefação, concluímos que não há diferença entre a riqueza de espécies de anuros obtida em coletas de campo e coleções científicas.

5.4 Para se aprofundar

- Recomendamos aos interessados que olhem a página do EstimateS software e baixem o manual do usuário que contém informações detalhadas sobre os índices de rarefação.Este site foi criado e é mantido pelo Dr. Robert K. Colwell, um dos maiores especialistas do mundo em estimativas da biodiversidade
- Recomendamos também o livro Magurran & McGill (2010) - Biological Diversity Frontiers in Measurement and Assessment.

Chapter 6

Estimadores de Riqueza

6.1 Backgorund da análise

Uma vez que determinar o número total de espécies numa área é praticamente impossível, principalmente em regiões com alta riqueza de espécies, os estimadores são úteis para extrapolar a riqueza observada e tentar estimar a riqueza total através de uma amostra incompleta de uma comunidade biológica (Walther & Moore 2005). Neste capítulo serão considerados os estimadores não paramétricos que usam informações da frequencia de espécies raras na comunidade (Gotelli & Chao 2013). Isto porque tanto os testes paramétricos que tentam determinar os parâmetros de uma curva usando o formato da curva de acumulação de espécies (e.g. equação logística, Michaelis-Menten) quanto os testes que usam a frequencia do número de indivíduos para enquadrá-las em uma das distribuições de abundância das espécies (e.g. distribuições log-séries, log-normal) não funcionam muito bem com dados empíricos (Gotelli & Chao 2013). Para mais detalhes sobre os testes paramétricos veja Magurran (2004) e Colwell (2019).

6.1.0.1 Quatro características para um bom estimador de riqueza (Chazdon et al. 1998; Horter et al. 2006):

- Independência do tamanho da amostra (quantidade de esforço amostral realizado);
- Insensibilidade a diferentes padrões de distribuições (diferentes equitabilidades);
- Insensibilidade em relação à ordem das amostragens;
- Insensibilidade à heterogeneidade entre as amostras usadas entre estudos.

6.2 Estimadores baseados na abundância das espécies

6.2.1 CHAO 1 - (Chao 1984, 1987):

Estimador simples do número absoluto de espécies em uma comunidade. É baseado no número de espécies raras dentro de uma amostra.

$$Chao_1 = S_{obs} + \left(\frac{n-1}{n}\right)\frac{F_1(F_1-1)}{2(F_2+1)}$$

onde:

- Sobs = o número de espécies na comunidade,
- n = número de amostras,
- F1 = número de espécies observadas com abundância de um indivíduo (espécies *singleton*),
- F2 = número de espécies observadas com abundância de dois indivíduos (espécies *doubletons*).

O valor de Chao 1 é máximo quando todas as espécies menos uma são únicas (*singleton*). Neste caso, a riqueza estimada é aproximadamente o dobro da riqueza observada.

6.2.1.1 Exemplo prático - Chao 1

6.2.1.1.1 Explicação dos dados Neste exemplo usaremos os dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com as abundâncias das espécies de anuros registradas em um habitat reprodutivo

Checklist

- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas
- Verificar se os dados são de abundância e não presença e ausência

6.2.2 Análise

Calculo do estimador de riqueza - Chao 1

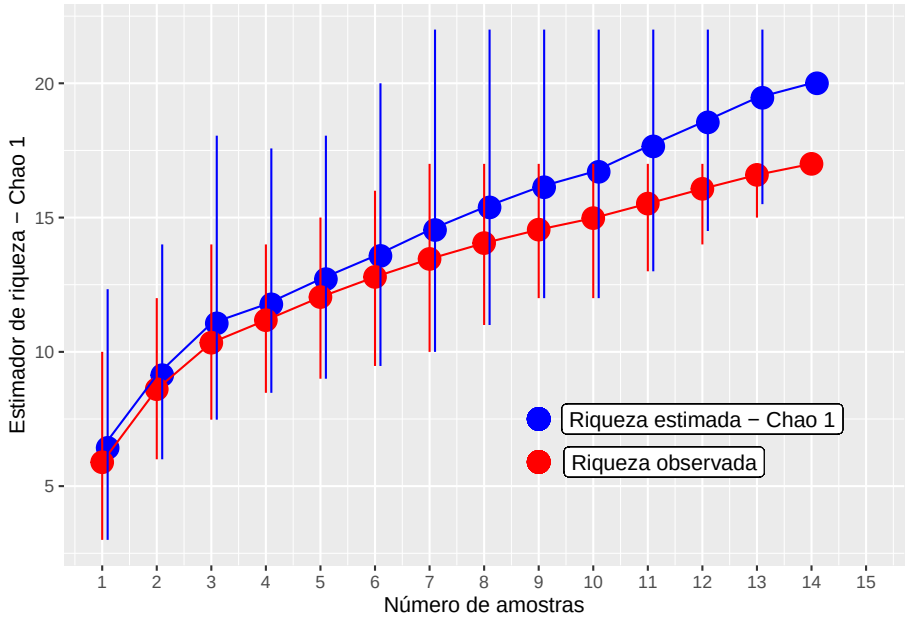
```
library(ecodados)
library(vegan)
dados_coleta <- poca_anuros
est_chao1 <- estaccumR(dados_coleta, permutations = 100)
summary(est_chao1, display = "chao")
```

```
## $chao
##          N      Chao   2.5%    97.5%  Std.Dev
## Dia_2    1  6.425000  3.000 12.33333 2.602541
## Dia_11   2  9.132357  6.000 14.00000 2.603968
## Dia_7    3 11.061167  7.475 18.05000 2.643445
## Dia_6    4 11.770000  8.475 17.57500 2.419151
## Dia_4    5 12.708333  9.000 18.05000 2.510290
## Dia_12   6 13.573333  9.475 20.00000 2.598057
## Dia_14   7 14.536667 10.000 22.00000 2.921684
## Dia_10   8 15.378333 11.000 22.00000 2.939996
## Dia_13   9 16.130000 12.000 22.00000 2.870335
## Dia_8   10 16.703333 12.000 22.00000 2.753895
## Dia_5   11 17.650000 13.000 22.00000 2.520121
## Dia_1   12 18.545000 14.500 22.00000 2.225989
## Dia_9   13 19.460000 15.500 22.00000 1.626299
## Dia_3   14 20.000000 20.000 20.00000 0.000000
##
## attr(,"class")
## [1] "summary.poolaccum"
```

Visualizar os resultados com intervalo de confiança de 95%.

```
library(ggplot2)
# preparando os dados para fazer o gráfico
resultados <- summary(est_chao1, display = c("S", "chao"))
res_chao <- cbind(resultados$chao[,1:4], resultados$S[,2:4])
res_chao <- as.data.frame(res_chao)
colnames(res_chao) <- c("Amostras", "Chao", "C_inferior", "C_superior", "Riqueza",
                        "R_inferior", "R_superior")

# comando para o gráfico
ggplot(res_chao, aes(y = Riqueza, x = Amostras)) +
```

```
  geom_point(aes(y = Chao, x = Amostras + 0.1), size = 5, color = "blue", alpha = 1) +
  geom_point(aes(y = Riqueza, x = Amostras), size = 5, color = "red", alpha = 1) +
  geom_line (aes(y = Chao, x = Amostras), color = "blue") +
  geom_line (aes(y = Riqueza, x = Amostras), color = "red") +
  geom_linerange(aes(ymin = C_inferior, ymax = C_superior, x = Amostras + 0.1),
color = "blue") +
  geom_linerange(aes(ymin = R_inferior, ymax = R_superior, x = Amostras), color = "red
  ylab ("Estimador de riqueza - Chao 1") +
  xlab ("Número de amostras") +
  scale_x_continuous(limits = c(1,15), breaks=seq(1,15,1)) +
  geom_point(y= 7.5, x = 9, size = 5, color = "blue", alpha = 1) +
  geom_point(y= 5.9, x = 9, size = 5, color = "red", alpha = 1) +
  geom_label( y = 7.5, x = 12, label = "Riqueza estimada - Chao 1") +
  geom_label( y = 5.9, x = 11.3, label = "Riqueza observada")
```



6.2.2.1 Interpretação dos resultados

Com base no número de espécies raras (*singletons* e *doubletons*), o estimador Chao 1 indica a possibilidade de encontrarmos mais três espécies caso o esforço amostral fosse maior e não mostra tendência de estabilização da curva em uma assíntota.

6.2.3 ACE - *Abundance-based Coverage Estimador* (Chao & Lee 1992, Chao et al. 2000):

Este método trabalha com a abundância das espécies raras (i.e. abundância baixa). Entretanto, diferente do estimador anterior, esse método permite ao pesquisador determinar os limites para os quais uma espécie seja considerada rara. Em geral, são consideradas raras espécies com abundância entre 1 e 10 indivíduos. A riqueza estimada pode variar conforme se aumente ou diminua o limiar de abundância, e infelizmente não existem critérios biológicos definidos para a escolha do melhor intervalo.

$$ACE = S_{abund} + \frac{S_{rare}}{C_{ace}} + \frac{F_1}{C_{ace}} Y^2_{ace}$$

onde:

$$Y^2_{ace} = max\left[\frac{S_{rare}}{C_{ace}} \frac{\sum_{i=i}^{10} i(i-1)F1}{(N_{rare})(N_{rare}-1)} - 1, 0\right]$$

$$C_{ace} = 1 - \frac{F1}{N_{rare}}$$

$$N_{rare} = \sum_{i=1}^{10} iF_i$$

Não precisa fazer cara feia, é óbvio que iremos usar o programa para fazer esses cálculos.

6.2.3.1 Exemplo prático - ACE

6.2.3.1.1 Explicação dos dados Usaremos os mesmos dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com as abundâncias das espécies de anuros registradas em um habitat reprodutivo

Checklist

- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas
- Verificar se os dados são de abundância e não presença e ausência

6.2.4 Análise

Calculo do estimador de riqueza - ACE

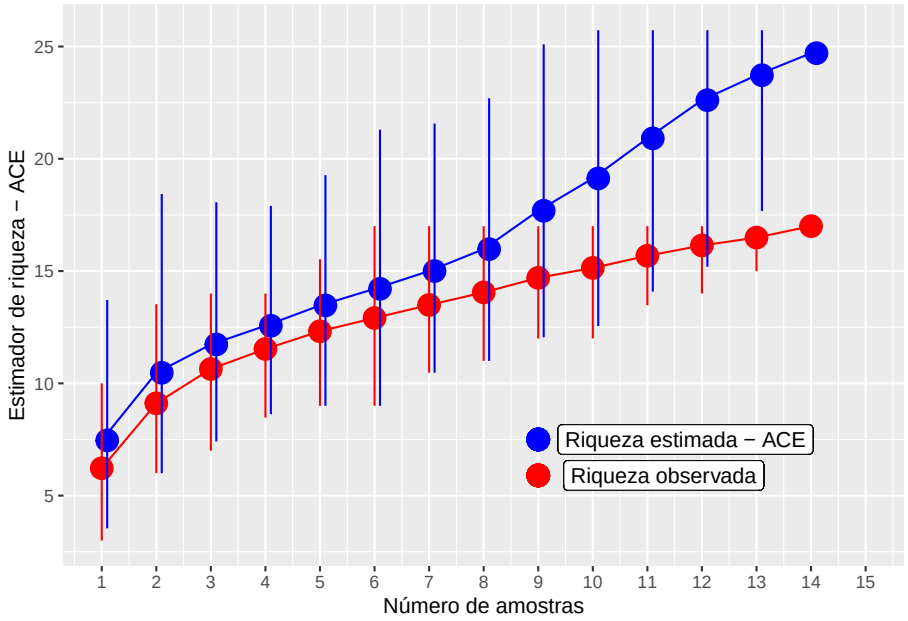
```
library(vegan)
dados_coleta <- poca_anuros
est_ace <- estaccumR(dados_coleta, permutations = 100)
summary(est_ace, display = "ace")
```

```
## $ace
##          N       ACE      2.5%    97.5%  Std.Dev
## Dia_8    1  7.460536  3.545190 13.71429 2.912522
## Dia_7    2 10.470854  6.000000 18.42880 3.272935
## Dia_1    3 11.730845  7.414357 18.06351 2.711759
## Dia_5    4 12.564230  8.623204 17.90292 2.614339
## Dia_14   5 13.470064  9.000000 19.27114 2.644979
## Dia_12   6 14.208073  9.000000 21.29764 2.926479
## Dia_4    7 14.999757 10.475000 21.56822 2.856993
## Dia_13   8 15.976322 11.000000 22.69653 3.127611
## Dia_9    9 17.684920 12.052778 25.09679 3.737648
## Dia_11  10 19.125599 12.551350 25.72368 4.041629
## Dia_10  11 20.908207 14.083617 25.72368 3.934751
## Dia_3   12 22.609058 15.192940 25.72368 3.347806
## Dia_6   13 23.720305 17.676471 25.72368 2.483438
## Dia_2   14 24.703704 24.703704 24.70370 0.000000
##
## attr(,"class")
## [1] "summary.poolaccum"
```

Visualizar os resultados com intervalo de confiança de 95%

```
library(ggplot2)
# preparando os dados para fazer o gráfico
resultados_ace <- summary(est_ace, display = c("S", "ace"))
res_ace <- cbind(resultados_ace$ace[,1:4], resultados_ace$S[,2:4])
res_ace <- as.data.frame(res_ace)
colnames(res_ace) <- c("Amostras", "ACE", "ACE_inferior", "ACE_superior", "Riqueza",
                       "R_inferior", "R_superior")
```

```
# comando para o gráfico
ggplot(res_ace, aes(y = Riqueza, x = Amostras)) +
  geom_point(aes(y = ACE, x = Amostras + 0.1), size = 5, color = "blue", alpha = 1) +
  geom_point(aes(y = Riqueza, x = Amostras), size = 5, color = "red", alpha = 1) +
  geom_line (aes(y = ACE, x = Amostras), color = "blue") +
  geom_line (aes(y = Riqueza, x = Amostras), color = "red") +
  geom_linerange(aes(ymin = ACE_inferior, ymax = ACE_superior, x = Amostras + 0.1),
 color = "blue") +
  geom_linerange(aes(ymin = R_inferior, ymax = R_superior, x = Amostras), color = "red") +
  ylab ("Estimador de riqueza - ACE") +
  xlab ("Número de amostras") +
  scale_x_continuous(limits = c(1,15), breaks=seq(1,15,1)) +
  geom_point(y= 7.5, x = 9, size = 5, color = "blue", alpha = 1) +
  geom_point(y= 5.9, x = 9, size = 5, color = "red", alpha = 1) +
  geom_label( y = 7.5, x = 11.7, label = "Riqueza estimada - ACE") +
  geom_label( y = 5.9, x = 11.3, label = "Riqueza observada")
```



6.2.4.1 Interpretação dos resultados

Com base no número de espécies raras (abundância menor que 10 indivíduos - *default*), o estimador ACE indica a possibilidade de encontrarmos mais sete espécies caso o esforço amostral fosse maior e não mostrou tendência de estabilição da curva em uma assíntota.

6.3 Estimadores baseados na incidência das espécies

6.3.1 CHAO 2 - (Chao 1987):

De acordo com Anne Chao, o estimador Chao 1 pode ser modificado para uso com dados de presença/ausência levando em conta a distribuição das espécies entre amostras. Neste caso é necessário conhecer o número de espécies encontradas em somente uma amostra e o número de espécies encontradas exatamente em duas amostras. Essa variação ficou denominada como Chao 2:

$$Chao_2 = S_{obs} + \left(\frac{m-1}{m}\right)\left(\frac{Q_1(Q_1-1)}{2(Q_2+1}\right)$$

onde:

- Sobs = o número de espécies na comunidade,
- m = número de amostragens,
- Q1 = número de espécies observadas em uma amostragem (espécies *uniques*),
- Q2 = número de espécies observadas em duas amostragens (espécies *duplicates*).

O valor de Chao2 é máximo quando as espécies menos uma são únicas (*uniques*). Neste caso, a riqueza estimada é aproximadamente o dobro da riqueza observada. Colwell & Coddington (1994) encontraram que o valor de Chao 2 mostrou ser o estimador menos enviesado para amostras com tamanho pequeno.

6.3.1.1 Exemplo prático - Chao 2

6.3.1.1.1 Explicação dos dados Usaremos os mesmos dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com a incidência das espécies de anuros registradas em um habitat reprodutivo

Checklist

- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas

6.3.2 Análise

Calculo do estimador de riqueza - Chao 2

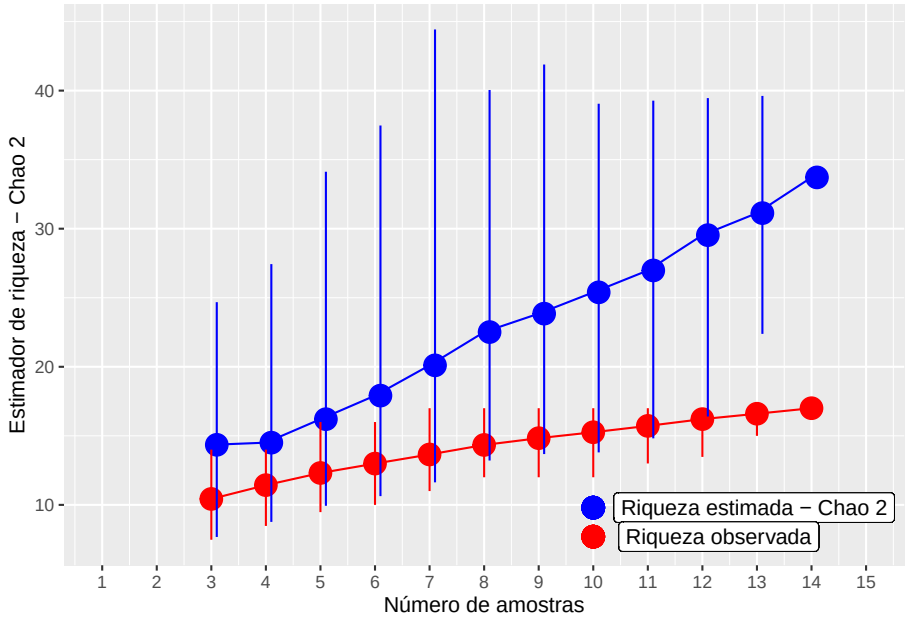
```
library(vegan)
dados_coleta <- poca_anuros
est_chao2 <- poolaccum(dados_coleta, permutations = 100)
summary(est_chao2, display = "chao")
```

```
## $chao
##        N     Chao      2.5%    97.5%  Std.Dev
##  [1,]  3 14.34910  7.676389 24.67917 5.656480
##  [2,]  4 14.50504  8.767188 27.43594 4.262175
##  [3,]  5 16.20640  9.933000 34.12500 6.038519
##  [4,]  6 17.91264 10.633854 37.47292 6.788475
##  [5,]  7 20.10500 11.632143 44.42857 7.800545
##  [6,]  8 22.51469 13.211458 40.04531 7.310359
##  [7,]  9 23.84852 13.680556 41.88889 7.577925
##  [8,] 10 25.39175 13.800000 39.05000 7.640206
##  [9,] 11 26.97682 14.818182 39.27273 7.011164
## [10,] 12 29.53375 16.396875 39.45833 6.515418
## [11,] 13 31.12923 22.384615 39.61538 4.894901
## [12,] 14 33.71429 33.714286 33.71429 0.000000
##
## attr(,"class")
## [1] "summary.poolaccum"
```

Visualizar os resultados com intervalo de confiança de 95%

```
library(ggplot2)
# preparando os dados para fazer o gráfico
resultados_chao2 <- summary(est_chao2, display = c("S", "chao"))
res_chao2 <- cbind(resultados_chao2$chao[,1:4], resultados_chao2$S[,2:4])
res_chao2 <- as.data.frame(res_chao2)
colnames(res_chao2) <- c("Amostras", "Chao2", "C_inferior", "C_superior", "Riqueza",
                         "R_inferior", "R_superior")

# comando para o gráfico
```

```
ggplot(res_chao2, aes(y = Riqueza, x = Amostras)) +
  geom_point(aes(y = Chao2, x = Amostras + 0.1), size = 5, color = "blue", alpha = 1) +
  geom_point(aes(y = Riqueza, x = Amostras), size = 5, color = "red", alpha = 1) +
  geom_line (aes(y = Chao2, x = Amostras), color = "blue") +
  geom_line (aes(y = Riqueza, x = Amostras), color = "red") +
  geom_linerange(aes(ymin = C_inferior, ymax = C_superior, x = Amostras + 0.1),
 color = "blue") +
  geom_linerange(aes(ymin = R_inferior, ymax = R_superior, x = Amostras), color = "red"
  ylab ("Estimador de riqueza - Chao 2") +
  xlab ("Número de amostras") +
  scale_x_continuous(limits = c(1,15), breaks=seq(1,15,1)) +
  geom_point(y= 9.8, x = 10, size = 5, color = "blue", alpha = 1) +
  geom_point(y= 7.7, x = 10, size = 5, color = "red", alpha = 1) +
  geom_label( y = 9.8, x = 12.95, label = "Riqueza estimada - Chao 2") +
  geom_label( y = 7.7, x = 12.3, label = "Riqueza observada")
```



6.3.2.1 Interpretação dos resultados

Com base no número de espécies raras (*uniques* e *duplicates*), Chao 2 estimou a possibilidade de encontrarmos mais dezesseis espécies caso o esforço amostral fosse maior e não mostrou tendência de estabilização da curva em uma assíntota.

6.3.3 JACKKNIFE 1 (Burnham & Overton 1978, 1979):

Este estimador baseia-se no número de espécies que ocorrem em somente uma amostra (Q1).

$$S_{jack1} = S_{obs} + Q1\left(\frac{m-1}{m}\right)$$

onde:

- Sobs = o número de espécies na comunidade,
- Q1 = número de espécies observadas em uma amostragem (espécies *uniques*),
- m = número de amostragens.

Palmer (1990) verificou que Jackknife 1 foi o estimador mais preciso e menos enviesado comparado a outros métodos de extrapolação.

6.3.3.1 Exemplo prático - Jackknife 1

6.3.3.1.1 Explicação dos dados Usaremos os mesmos dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com as abundâncias das espécies de anuros registradas em um habitat reprodutivo

Checklist

- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas

6.3.4 Análise

Calculo do estimador de riqueza - Jackknife 1

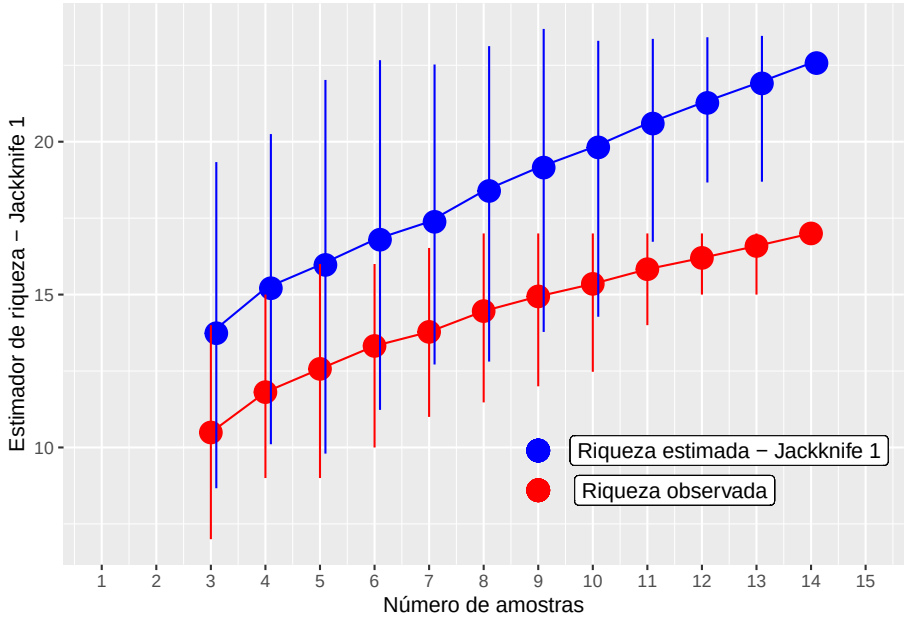
```
library(vegan)
dados_coleta <- poca_anuros
est_jack1 <- poolaccum(dados_coleta, permutations = 100)
summary(est_jack1, display = "jack1")
```

```
## $jack1
##        N Jackknife 1      2.5%    97.5%  Std.Dev
##  [1,]  3    13.73667  8.666667 19.33333 2.872641
##  [2,]  4    15.20750 10.106250 20.25000 2.977834
##  [3,]  5    15.97000  9.800000 22.02000 3.019181
##  [4,]  6    16.79500 11.229167 22.66667 2.868112
##  [5,]  7    17.38000 12.714286 22.52500 2.623037
##  [6,]  8    18.38875 12.809375 23.12500 2.524870
##  [7,]  9    19.15333 13.777778 23.68889 2.574723
##  [8,] 10    19.81400 14.275000 23.30000 2.400548
##  [9,] 11    20.59364 16.727273 23.36364 1.955787
## [10,] 12    21.26917 18.666667 23.41667 1.592361
## [11,] 13    21.90692 18.692308 23.46154 1.331956
## [12,] 14    22.57143 22.571429 22.57143 0.000000
##
## attr(,"class")
## [1] "summary.poolaccum"
```

Visualizar os resultados com 95% intervalo de confiança

```
library(ggplot2)
# preparando os dados para fazer o gráfico
resultados_jack1 <- summary(est_jack1, display = c("S", "jack1"))
res_jack1 <- cbind(resultados_jack1$jack1[,1:4], resultados_jack1$S[,2:4])
res_jack1 <- as.data.frame(res_jack1)
colnames(res_jack1) <- c("Amostras", "JACK1", "JACK1_inferior", "JACK1_superior", "Riqu
                         "R_inferior", "R_superior")

# comando para o gráfico
ggplot(res_jack1, aes(y = Riqueza, x = Amostras)) +
  geom_point(aes(y = JACK1, x = Amostras + 0.1), size = 5, color = "blue", alpha = 1) +
  geom_point(aes(y = Riqueza, x = Amostras), size = 5, color = "red", alpha = 1) +
  geom_line (aes(y = JACK1, x = Amostras), color = "blue") +
  geom_line (aes(y = Riqueza, x = Amostras), color = "red") +
  geom_linerange(aes(ymin = JACK1_inferior, ymax = JACK1_superior, x = Amostras + 0.1)
 color = "blue") +
  geom_linerange(aes(ymin = R_inferior, ymax = R_superior, x = Amostras), color = "red"
  ylab ("Estimador de riqueza - Jackknife 1") +
  xlab ("Número de amostras") +
```

```
  scale_x_continuous(limits = c(1,15), breaks=seq(1,15,1)) +
  geom_point(y= 9.9, x = 9, size = 5, color = "blue", alpha = 1) +
  geom_point(y= 8.6, x = 9, size = 5, color = "red", alpha = 1) +
  geom_label( y = 9.9, x = 12.5, label = "Riqueza estimada - Jackknife 1") +
  geom_label( y = 8.6, x = 11.5, label = "Riqueza observada")
```



6.3.4.1 Interpretação dos resultados

Com base no número de espécies raras, o estimador Jackknife 1 calculou a possibilidade de encontrarmos mais seis espécies caso o esforço amostral fosse maior e não mostrou tendência de estabilização da curva em uma assíntota.

6.3.5 JACKKNIFE 2 (Burnham & Overton 1978, 1979, Palmer 1991):

Este método basea-se no número de espécies que ocorrem em apenas uma amostra e no número de espécies que ocorrem em exatamente duas amostras.

$$S_{jack2} = S_{obs} + \left[\frac{Q_1(2m-3)}{m} - \frac{Q_2(m-2)^2}{m(m-1)} \right]$$

onde:

- Sobs = o número de espécies na comunidade,

- m = número de amostragens,
- Q1 = número de espécies observadas em uma amostragem (espécies *uniques*),
- Q2 = número de espécies observadas em duas amostragens (espécies *duplicates*).

6.3.5.1 Exemplo prático - Jackknife 2

6.3.5.1.1 Explicação dos dados Usaremos os mesmos dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com as abundâncias das espécies de anuros registradas em um habitat reprodutivo

Checklist

- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas

6.3.6 Análise

Calculo do estimador de riqueza - Jackknife 2

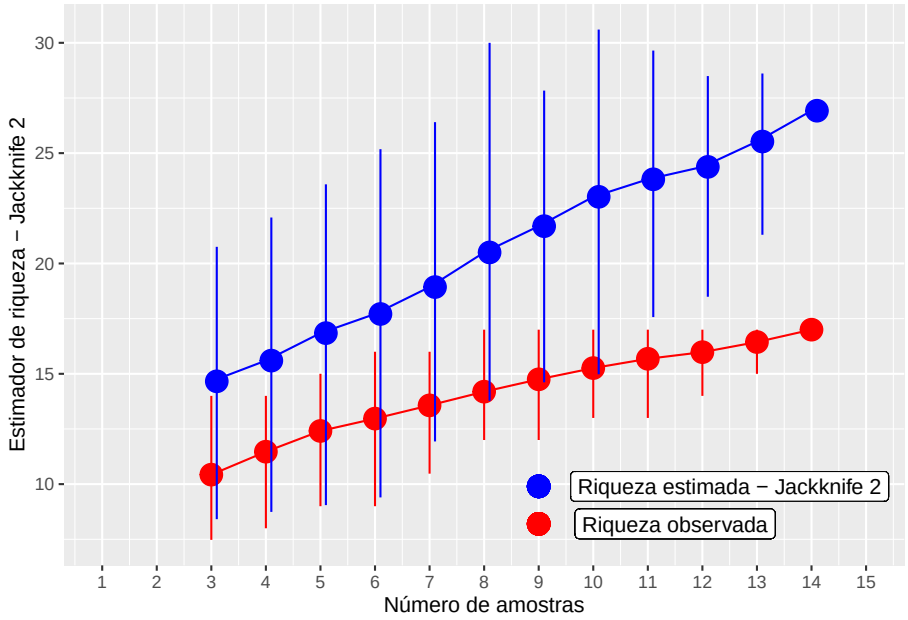
```
library(vegan)
dados_coleta <- poca_anuros
est_jack2 <- poolaccum(dados_coleta, permutations = 100)
summary(est_jack2, display = "jack2")
```

```
## $jack2
##        N Jackknife 2      2.5%    97.5%  Std.Dev
##  [1,]  3    14.66000  8.412500 20.75417 3.194492
##  [2,]  4    15.59833  8.741667 22.08333 3.635801
```

```
##  [3,]  5    16.84400  9.050000 23.58625 3.947466
##  [4,]  6    17.71067  9.398333 25.18000 4.019302
##  [5,]  7    18.93095 11.935119 26.40476 4.087603
##  [6,]  8    20.50018 13.776786 30.00000 4.230534
##  [7,]  9    21.69083 14.614931 27.83437 3.749235
##  [8,] 10    23.02044 14.977778 30.60000 3.899725
##  [9,] 11    23.81473 17.570682 29.64795 3.669697
## [10,] 12    24.37674 18.492424 28.49242 3.192388
## [11,] 13    25.52551 21.301282 28.60897 2.255938
## [12,] 14    26.92308 26.923077 26.92308 0.000000
##
## attr(,"class")
## [1] "summary.poolaccum"
```

Visualizar os resultados com intervalo de confiança de 95%

```
library(ggplot2)
# preparando os dados para fazer o gráfico
resultados_jack2 <- summary(est_jack2, display = c("S", "jack2"))
res_jack2 <- cbind(resultados_jack2$jack2[,1:4], resultados_jack2$S[,2:4])
res_jack2 <- as.data.frame(res_jack2)
colnames(res_jack2) <- c("Amostras", "JACK2", "JACK2_inferior", "JACK2_superior", "Riqueza",
                         "R_inferior", "R_superior")

# comando para o gráfico
ggplot(res_jack2, aes(y = Riqueza, x = Amostras)) +
  geom_point(aes(y = JACK2, x = Amostras + 0.1), size = 5, color = "blue", alpha = 1) +
  geom_point(aes(y = Riqueza, x = Amostras), size = 5, color = "red", alpha = 1) +
  geom_line (aes(y = JACK2, x = Amostras), color = "blue") +
  geom_line (aes(y = Riqueza, x = Amostras), color = "red") +
  geom_linerange(aes(ymin = JACK2_inferior, ymax = JACK2_superior, x = Amostras + 0.1),
 color = "blue") +
  geom_linerange(aes(ymin = R_inferior, ymax = R_superior, x = Amostras), color = "red") +
  ylab ("Estimador de riqueza - Jackknife 2") +
  xlab ("Número de amostras") +
  scale_x_continuous(limits = c(1,15), breaks=seq(1,15,1)) +
  geom_point(y= 9.9, x = 9, size = 5, color = "blue", alpha = 1) +
  geom_point(y= 8.2, x = 9, size = 5, color = "red", alpha = 1) +
  geom_label( y = 9.9, x = 12.5, label = "Riqueza estimada - Jackknife 2") +
  geom_label( y = 8.2, x = 11.5, label = "Riqueza observada")
```



6.3.6.1 Interpretação dos resultados

Com base no número de espécies raras, o estimador Jackknife 2 calculou a possibilidade de encontrarmos mais dez espécies caso o esforço amostral fosse maior e não mostrou tendência estabilização da curva em uma assíntota.

6.3.7 BOOTSTRAP (Smith & van Belle 1984):

Este método difere dos demais por utilizar dados de todas as espécies coletadas para estimar a riqueza total, não se restringindo às espécies raras. Ele requer somente dados de incidência. A estimativa pelo bootstrap é calculada somando-se a riqueza observada à soma do inverso da proporção de amostras em que cada espécie ocorre.

$$S_{boot} = S_{obs} + \sum_{k=1}^{S_{obs}} (1 - P_k)^m$$

onde:

- Sobs = o número de espécies na comunidade,
- m = número de amostragens,
- Pk = proporção do número de amostras em que cada espécie foi registrada.

6.3.7.1 Exemplo prático - Bootstrap

6.3.7.1.1 Explicação dos dados Usaremos os mesmos dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com as abundâncias das espécies de anuros registradas em um habitat reprodutivo

Checklist

- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas

6.3.8 Análise

Calculo do estimador de riqueza - Bootstrap

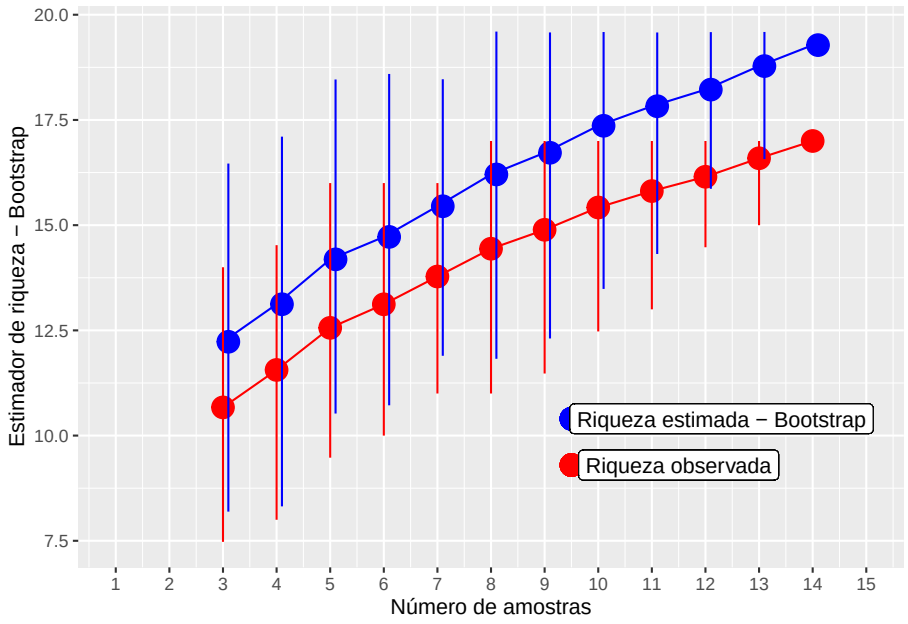
```
library(vegan)
dados_coleta <- poca_anuros
est_boot <- poolaccum(dados_coleta, permutations = 100)
summary(est_boot, display = "boot")
```

```
## $boot
##        N Bootstrap      2.5%    97.5%   Std.Dev
##  [1,]  3  12.22926  8.193519 16.46389 2.3711167
##  [2,]  4  13.12133  8.318262 17.10312 2.3562103
##  [3,]  5  14.18726 10.523888 18.46186 2.1958469
##  [4,]  6  14.72106 10.719349 18.59313 2.1560773
##  [5,]  7  15.44956 11.896422 18.46717 1.8842047
##  [6,]  8  16.20524 11.824978 19.60012 1.8142375
##  [7,]  9  16.72008 12.307939 19.57946 1.7966583
##  [8,] 10  17.36141 13.485395 19.58909 1.6611224
##  [9,] 11  17.82640 14.315556 19.57877 1.4617363
## [10,] 12  18.22062 15.864304 19.58719 1.3302413
## [11,] 13  18.78008 16.570376 19.59107 0.9924151
```

```
## [12,] 14  19.27832 19.278321 19.27832 0.0000000
##
## attr(,"class")
## [1] "summary.poolaccum"
```

Visualizar os resultados com intervalo de confiança de 95%

```
library(ggplot2)
# preparando os dados para fazer o gráfico
resultados_boot <- summary(est_boot, display = c("S", "boot"))
res_boot <- cbind(resultados_boot$boot[,1:4], resultados_boot$S[,2:4])
res_boot <- as.data.frame(res_boot)
colnames(res_boot) <- c("Amostras", "BOOT", "BOOT_inferior", "BOOT_superior", "Riqueza
                        "R_inferior", "R_superior")

# comando para o gráfico
ggplot(res_boot, aes(y = Riqueza, x = Amostras)) +
  geom_point(aes(y = BOOT, x = Amostras + 0.1), size = 5, color = "blue", alpha = 1) +
  geom_point(aes(y = Riqueza, x = Amostras), size = 5, color = "red", alpha = 1) +
  geom_line (aes(y = BOOT, x = Amostras), color = "blue") +
  geom_line (aes(y = Riqueza, x = Amostras), color = "red") +
  geom_linerange(aes(ymin = BOOT_inferior, ymax = BOOT_superior, x = Amostras + 0.1),
 color = "blue") +
  geom_linerange(aes(ymin = R_inferior, ymax = R_superior, x = Amostras), color = "red
  ylab ("Estimador de riqueza - Bootstrap") +
  xlab ("Número de amostras") +
  scale_x_continuous(limits = c(1,15), breaks=seq(1,15,1)) +
  geom_point(y= 10.4, x = 9.5, size = 5, color = "blue", alpha = 1) +
  geom_point(y= 9.3, x = 9.5, size = 5, color = "red", alpha = 1) +
  geom_label( y = 10.4, x = 12.3, label = "Riqueza estimada - Bootstrap") +
  geom_label( y = 9.3, x = 11.5, label = "Riqueza observada")
```



6.3.8.1 Interpretação dos resultados

Com base na frequencia de ocorrência das espécies, o estimador bootstrap calculou a possibilidade de encontrarmos mais duas espécies caso o esforço amostral fosse maior e não mostrou tendência de estabilização da curva em uma assíntota.

6.3.9 Interpolação e Extrapolação baseadas em rarefação usando amostragens de incidência ou abundância (Chao & Jost 2012, Colwell et al. 2012):

Este método utiliza teoria de amostragem (e.g. modelos multinomial, Poisson e Bernoulli) para conectar rarefação (interpolação) e predição (extrapolação) com base no tamanho da amostra. Contudo, é importante enfatizar que a extrapolação torna-se altamente incerta quando extendida para o dobro do tamanho da amostragem. Este método utiliza uma abordagem com bootstrap para calcular o intervalo de confiança de 95%. Uma das vantagens desta abordagem é que ela permite além da riqueza de espécies, interpolar e extrapolar os índices de diversidade de Shannon entropy (i.e. quantifica a incerteza da identidade da espécie baseado na amostragem aleatória de um indivíduo da comunidade) e Gini-Simpson (i.e. quantifica a probabilidade que dois indivíduos escolhidos aleatoriamente sejam de diferentes espécies). Contudo, estes índices de diversidades são transformados e apresentados em unidades de riqueza de espécies (*Números de Hill*, Hill 1973). Hill percebeu que poderíamos calcular a riqueza de espécies máxima usando os índices de Shannon entropy e Gini-Simpson quando consideramos que todas as espécies na comunidade possuem abundâncias idên-

ticas (máxima equitabilidade). Então, eles propos a transformação dos indíces de diversidade determinando qual seria o número de espécies equivalente da nossa comunidade (observado) se todas as espécies fossem igualmente abundantes (teórico). Desta maneira, os indíces podem ser comparavéis pois estão representados pela mesma unidade - riqueza de espécies. Os números de Hill são representados pelo parâmetro q que controla a sensibilidide do índice em relação a abudância relativa das espécies (Gotelli & Chao 2013). Quando q = 0 ele não considera a abundância das espécies e é igual a riqueza de espécies. Quando q = 1 ele é o exponencial da diversidade de Shannon, que dá uma peso maior para as espécies raras. Quando q = 2 ele é a diversidade de Simpson, que dá um peso maior para as espécies mais comuns na comunidade.

6.3.9.1 Exemplo prático

6.3.9.1.1 Explicação dos dados Usaremos os mesmos dados de 17 espécies de anuros amostradas em 14 dias de coletas de campo em um habitat reprodutivo localizado na região noroeste do estado de São Paulo, Brasil.

Pergunta:

> Quantas espécies a mais poderiam ser amostradas caso aumentassemos o esforço amostral?

Predições

- O número de espécies estimadas é similar ao número de espécies observada;
- O número de espécies estimadas é maior do que o número de espécies observada.

Variáveis

- Variáveis preditoras
  - matriz ou vetor com as abundâncias das espécies de anuros registradas em um habitat reprodutivo

Checklist

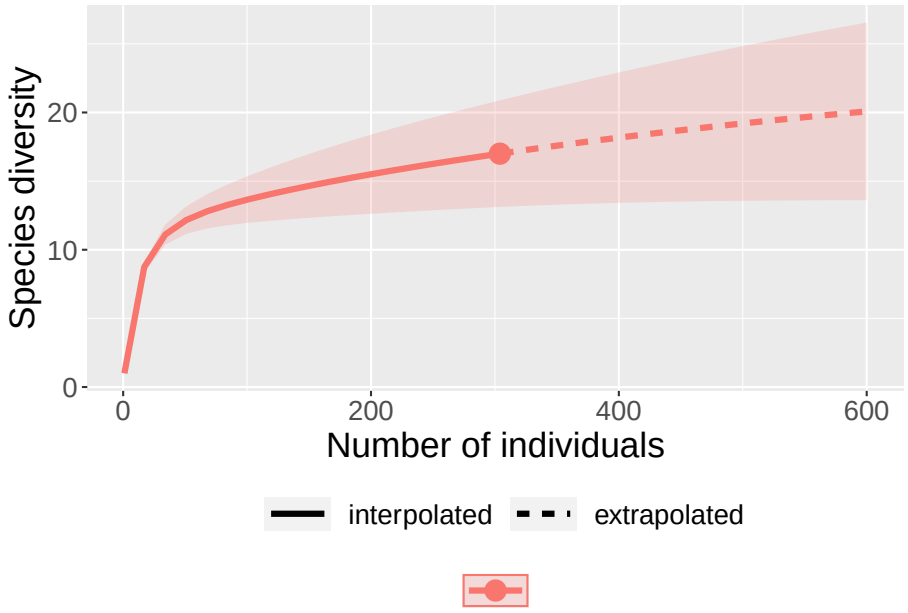
- Verificar se a sua matriz está com as espécies nas colunas e as amostragens nas linhas.

6.3.10 Análise

Calculo da extrapolação da riqueza com base no número de indivíduos

```
library(iNEXT)
dados_coleta <- poca_anuros

# preparando os dados para análises considerando a abundância
dados_inext_abu <- colSums(dados_coleta)
```

```
resultados_abundancia <- iNEXT(dados_inext_abu, q = 0, datatype = "abundance",
                         endpoint = 600)

# Visualizar os dados no gráfico
ggiNEXT(resultados_abundancia, type = 1)
```



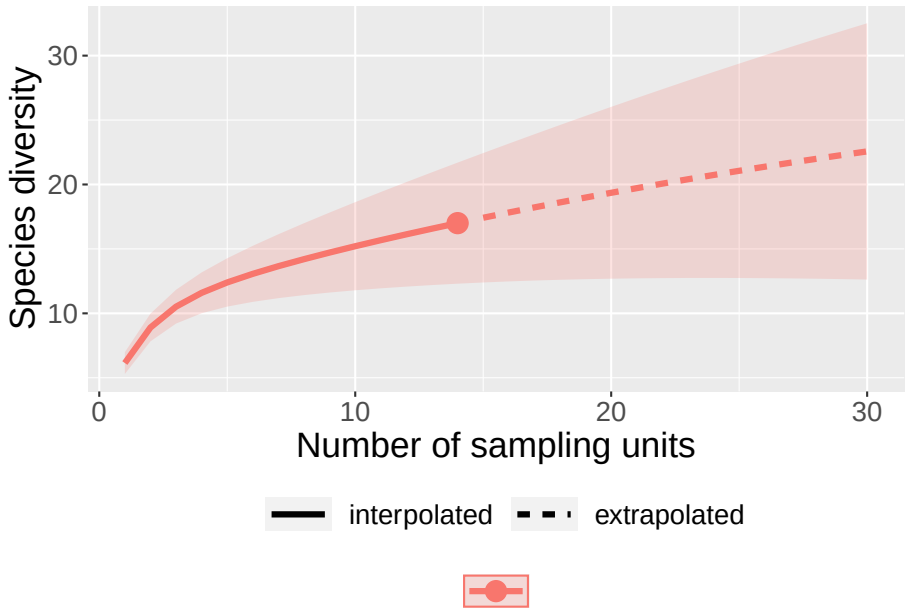
6.3.10.1 Interpretação dos resultados

Veja que o ponto no final da linha contínua representa as 17 espécies de anuros (eixo Y) observadas entre os 304 individuos (eixo X). A extrapolação máxima (600 indivíduos no nosso exemplo), estima um aumento de até oito espécies (intervalo de confiança) caso amostrássemos mais 300 indivíduos.

Calculo da extrapolação da riqueza com base no número de amostras

```
library(iNEXT)
dados_coleta <- poca_anuros

# preparando os dados para análises considerando a incidência
dados_inext <- as.incfreq(t(dados_coleta)) # preciso transpor o dataframe

resultados_incidencia <- iNEXT(dados_inext, q = 0, datatype = "incidence_freq",
                         endpoint = 30)

# Visualizar os dados no gráfico
```

```
ggiNEXT(resultados_incidencia, type = 1)
```



6.3.10.2 Interpretação dos resultados

Veja que o ponto no final da linha contínua representa as 17 espécies de anuros (eixo Y) observadas nos 14 dias de coleta (eixo X - amostras). A extrapolação máxima (30 dias de coleta no nosso exemplo), estima um aumento de até 13 espécies (intervalo de confiança) caso amostrássemos mais 16 dias.

6.3.11 Para se aprofundar

- Recomendamos aos interessados que olhem a página do EstimateS software e baixem o manual do usuário que contém informações detalhadas sobre os índices de rarefação e estimadores de riqueza.Este site foi criado e é mantido pelo Dr. Robert K. Colwell, um dos maiores especialistas do mundo em estimativas da biodiversidade
- Recomendamos também o livro Magurran & McGill (2010) - Biological Diversity Frontiers in Measurement and Assessment.