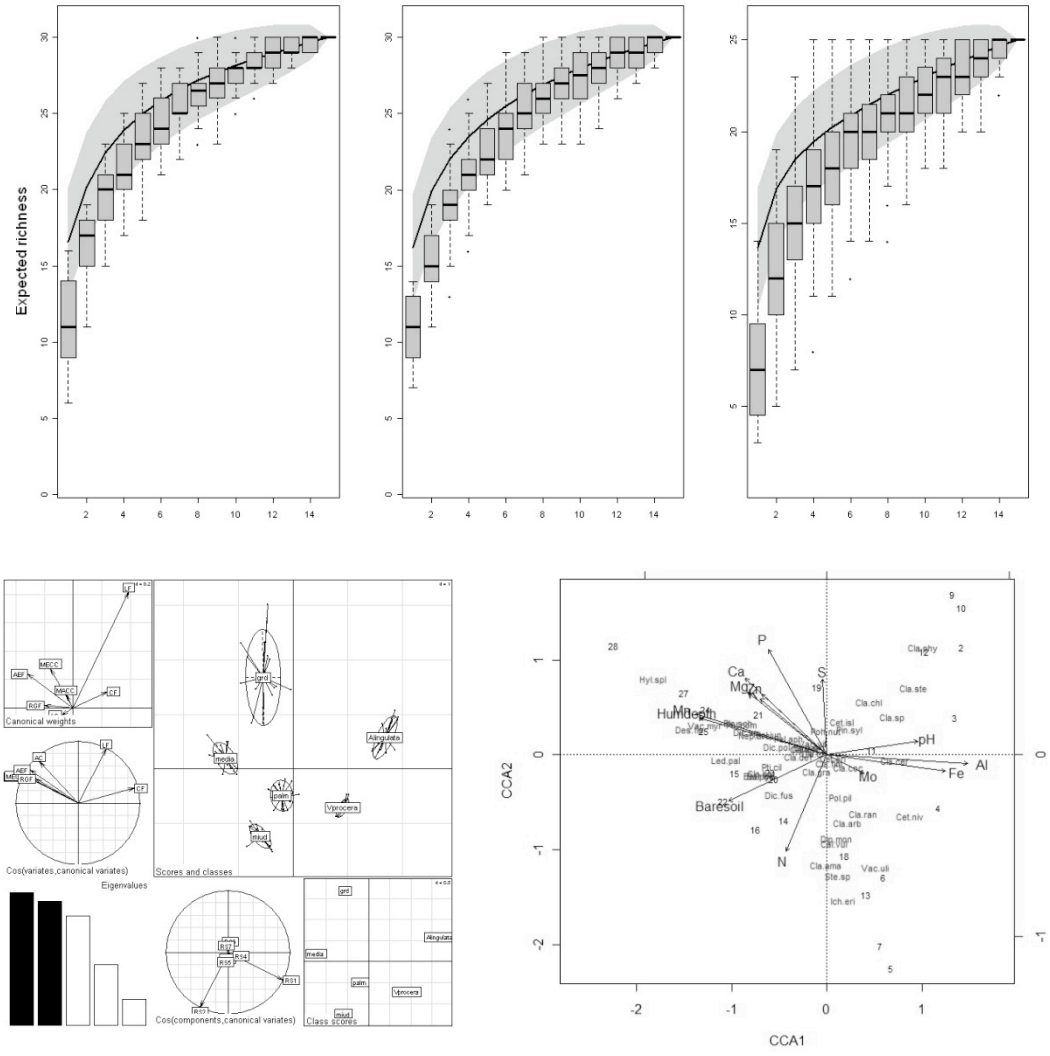
Universidade Estadual Paulista

Programa de Pós-Graduação Biologia Animal

Estatística aplicada à ecologia usando o R



Professores responsáveis:

Diogo Borges Provete (dbprovete@gmail.com)

Fernando Rodrigues da Silva (bigosbio@yahoo.com.br)

Thiago Gonçalves Souza (tgoncalves.souza@gmail.com)

São José do Rio Preto, SP

Abril, 2011

SUMÁRIO

OBJETIVO DO CURSO

Esta apostila foi elaborada para servir como material de apoio para um curso ministrado no PPG Biolgia Animal da UNESP de S.J. Rio Preto. Nossa proposta com o curso e com esta apostila é de traçar o melhor caminho (pelo menos em nosso ponto de vista) entre questões ecológicas e os métodos estatísticos mais robustos para testá-las. Guiar seus passos nesse caminho (nem sempre linear) necessita que você utilize um requisito básico: o de utilizar seu esforço para caminhar. O nosso esforço, em contrapartida, será o de segurar suas mãos, mantê-lo de pé e indicar as melhores direções para que adquira certa independência em análises ecológicas. Todo o material utilizado durante este curso, incluindo scripts e pdf das aulas está disponível em: https://sites.google.com/site/diogoprovetepage/teaching. Um dos nossos objetivos é mostrar que o conhecimento de teorias ecológicas e a utilização de questões apropriadas são o primeiro passo na caminha rumo à compreensão da lógica estatística. Não deixe que a estatística se torne a "pedra no seu caminho". Em nossa opinião, programas com ambiente de programação favorecem o entendimento da lógica estatística, uma vez que cada passo (lembre-se de que você está caminhado em uma estrada desconhecida) precisa ser coordenado, ou seja, as linhas de comando (detalhes abaixo) precisam ser compreendidas para que você teste suas hipóteses.

A primeira parte desta apostila pretende utilizar uma estratégia que facilita a escolha do teste estatístico apropriado, por meio da seleção de questões/hipóteses claras e da ligação dessas hipóteses com a teoria e o método. Posteriormente à escolha de suas questões é necessário transferir o contexto ecológico para um contexto meramente estatístico (hipótese nula/alternativa). A partir da definição de sua hipótese nula partiremos para a aplicação de cada teste estatístico (de modelos lineares generalizados à análises multivariadas) utilizando como plataforma o programa R. Antes de detalhar cada análise estatística, apresentaremos os comandos básicos para a utilização do R e os tipos de distribuição estatística que são essenciais para o desenvolvimento do curso. Para isso, organizamos um esquema que chamamos de "estrutura lógica" que facilita a compreensão dos passos necessários para testar suas hipóteses (Fig. 1).

É sempre bom ter em mente que "é muito importante saber aonde se quer chegar para poder escolher o que fazer".

O QUE VOCÊ NÃO ENCONTRARÁ NESTA APOSTILA

Aprofundamento teórico, detalhes matemáticos, e explicação dos algoritmos são informações que infelizmente não serão abordadas neste curso. O foco do curso é a explicação de como cada teste funciona (teoria e procedimentos matemáticos básicos) e sua aplicação em testes ecológicos usando o programa R. Para tanto, o livro dos irmãos Pierre e Louis Legendre (Legendre & Legendre 1998) é uma leitura que permite o aprofundamento de cada uma das análises propostas aqui. Além disso, são de fundamental importância para o amadurecimento em análises ecológicas as seguintes leituras: Manly (1991), Pinheiro & Bates (2000), Scheiner & Gurevitch (2001), Quinn & Keough (2002), Venables & Ripley (2002), Magurran (2004) e Gotelli & Ellison (2004).

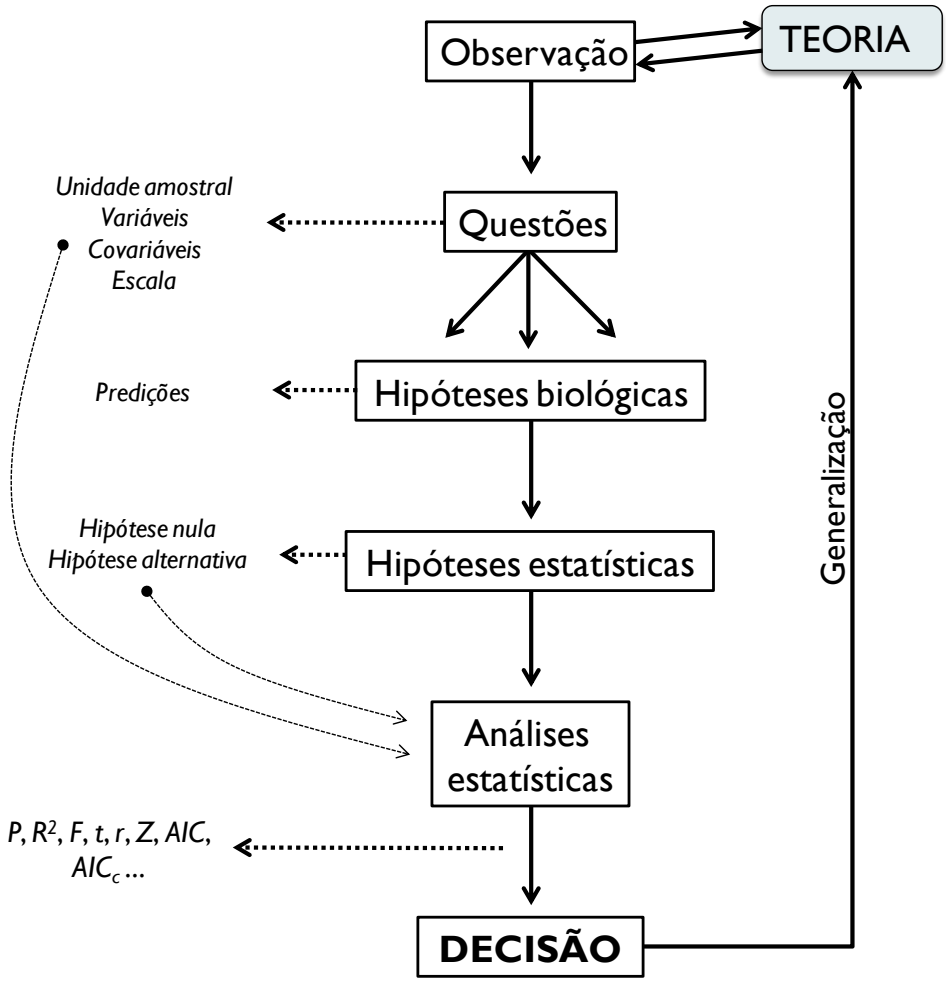


Figura 1. Estrutura lógica para integrar teorias/questões ecológicas com análises estatísticas (e vice-versa). Lembre-se de que omitimos etapas importantes desta estrutura lógica, como o delineamento experimental, a coleta e organização dos dados, que estão além do objetivo desta apostila.

INTRODUÇÃO – INTEGRANDO QUESTÕES ECOLÓGICAS E ANÁLISES ESTATÍSTICAS

Para a grande maioria dos estudantes [e professores] de biologia a palavra "estatística" traz certa vertigem e aversão. Em geral, alunos e professores consideram este passo um dos mais (se não o mais) problemáticos da pesquisa científica. Para ecologia e, especialmente, ecologia de comunidades, métodos analíticos complexos e que consomem muito tempo para serem realizados tornam a estatística uma tarefa ainda mais distante de ser alcançada (e compreendida). Infelizmente, a maioria opta por não cumprir esta tarefa. Em nossa opinião, muito dessa aversão à estatística se deve às disciplinas introdutórias do curso de graduação em Ciências Biológicas (a maioria, é claro) estarem baseados em um contexto puramente estatístico e com exemplos não-biológicos, sem um programa que integre a ferramenta analítica a um "problema de pesquisa". De fato, entender exemplos estatísticos com uma lógica puramente estatística não parece uma tarefa trivial para alunos que buscam entender, por exemplo, como processos populacionais, de comunidades e ecossistêmicos determinam a distribuição das espécies. Uma alternativa que pode facilitar a compreensão das análises estatísticas para biólogos (e para todos os cientistas!) é a utilização da lógica do método científico tomando como fator de decisão os resultados estatísticos. Ao final do curso, ou da leitura desta apostila, gostaríamos de que você refletisse um pouco sobre as seguintes questões: (1) qual a principal teoria do meu trabalho? (2) Qual a principal pergunta do meu trabalho? (3) Qual é a unidade amostral, a variável dependente e independente do meu trabalho? A seguir, apresentamos a seqüência lógica que sugerimos que seja aplicada a todo e qualquer teste que utilize estatística frequentista (interpretação objetiva da probabilidade baseada no critério de falseamento de Karl R. Popper). Esta interpretação é, por sua vez, diferente da interpretação subjetiva da probabilidade utilizada no arcabouço da estatística Bayesiana e da Maxima Verossimilhança. É importante ressaltar ainda que a probabilidade (o fator de decisão dos frequentistas, i.e., o tão sonhado "$p < 0,05$") representa uma classe de eventos (observados) comparados com uma série de repetições, e portanto o grau de incerteza relacionada a eventos. Todo este arcabouço dos testes de hipóteses estatísticas foi desenvolvido por Jerzy Neyman e Egon S. Pearson (Neyman & Pearson, 1933) adotando a visão Popperiana de que uma observação não fornece confirmação para uma teoria, devido ao problema da indução (para uma discussão mais detalhada veja os cap. 2 e 3 de Godfrey-Smith, 2003). Ao contrário, um teste deveria procurar refutar uma teoria, somente desta forma haveria ganhado conhecimento. Então, segundo o arcabouço de Neyman-Pearson, o teste estatístico procura rejeitar a hipótese nula, e não a confirmação da hipótese alternativa. Numa regressão, por exemplo, se o teste verificar que o coeficiente β é significativo, isto quer dizer que a inclinação da reta é diferente de zero, no entanto a interpretação biológica de uma relação linear entre as duas variáveis deve ser feita à luz das predições da teoria que se pretende testar. Por outro lado, os testes de modelos lineares generalizados em mistos utiliza a

lógica da estatística Bayesiana e da Maxima Verossimilhança. Estes arcabouços utilizam a interpretaçãoo subjetiva da probabilidade. Como uma analogia, o arcabouço frequentista presume que a "verdade" ou todo o universo amostral está numa nuvem, distante e inalcançável, e que somente temos acesso a pequenas amostras de dados, que nesta metáfora, seriam um monte, com o qual chegaríamos o mais próximo possível da nuvem. Seguindo esta metáfora, a estatística Bayesiana e Maxima Verossimilhança assumem que j que a "nuvem" é algo inatingível não devemos considerá-la na análise e que a melhor estimativa que temos são os dados reais que coletamos. Portanto, neste contexto, devemos considerar nossos dados como o universo amostral total.

Ao definir a questão de pesquisa é essencial conhecer como a teoria pode ser usada e como e porque ela pode explicar ou ser aplicada à sua questão (Ford 2000). Os modelos gerados pelas teorias podem ser aproveitados para criar suas hipóteses e predições. As hipóteses [científicas] são definidas como explicações potenciais que podem ser retiradas de observações do mundo externo (processo indutivo) ou de componentes de uma teoria (processo dedutivo). Uma hipótese científica, do ponto de vista de Popper, *deve* ser falseável. As predições são afirmações deduzidas de uma estrutura lógica ou causal de uma teoria, ou induzidas a partir de informações empíricas; em outras palavras, a predição é a conseqüência da hipótese, o resultado esperado se a hipótese for verdadeira. Uma hipótese bem articulada *deve* ser capaz de gerar predições. Um exercício fundamental para a criação de hipóteses e articulação de suas predições se faz a partir da construção de fluxogramas (Fig. 2). No fluxograma você pode separar cada variável e a relação esperada entre cada uma delas. As setas indicam a relação esperada entre as variáveis (os sinais acima das setas mostram a direção da relação). Setas com espessuras diferentes podem ser usadas como forma de demonstrar a importância relativa esperada para cada variável.

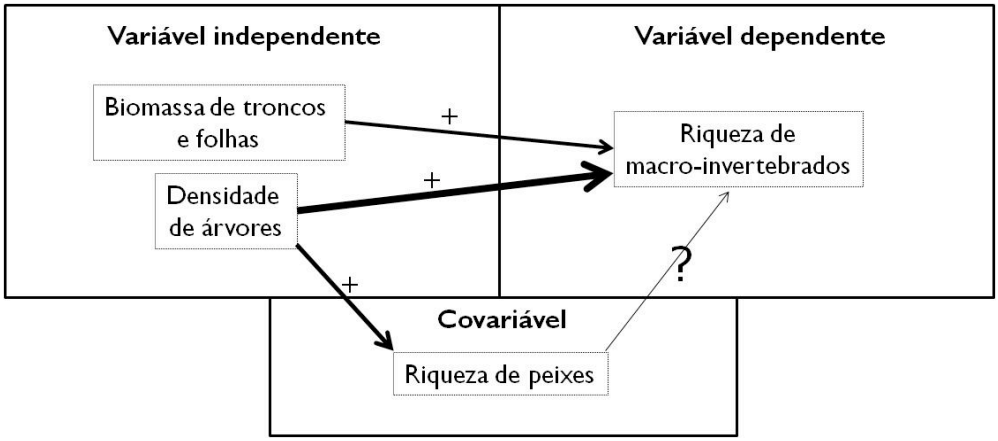


Figura 2. Fluxograma representando as predições que foram articuladas a partir da hipótese "as florestas ripárias aumentam a riqueza de macro-invertebrados".

O *MELHOR* CAMINHO PARA FAZER A PERGUNTA *CERTA*

Em geral, questões não devem ser muito gerais (e.g., qual o efeito das mudanças globais nas florestas?) por que dificultam a compreensão do que efetivamente você está testando. É preferível que suas hipóteses sejam mais gerais (teóricas) e suas questões mais específicas (referidas como operacionais daqui pra frente), para que você e o seu leitor saibam o que vai ser testado efetivamente e qual teste deverá ser empregado. Por exemplo, um pesquisador tem a seguinte hipótese: "mudanças globais afetam a dinâmica e estrutura de florestas"; para testar esta hipótese este pesquisador levantou duas questões operacionais: (1) o aumento da temperatura modifica a composição de espécies vegetais? (2) O aumento da temperatura aumenta a ocorrência de espécies exóticas? Com essas questões operacionais fica mais fácil compreender qual sua variável independente (neste caso temperatura) que representa a "mudança climática" e que afeta sua variável dependente ("dinâmica e estrutura de florestas") que foi operacionalizada em duas variáveis "composição de espécies vegetais" e "ocorrência de espécies exóticas". Além disso, é muito importante saber qual a unidade amostral do seu trabalho. No exemplo acima, o pesquisador coletou em 30 áreas de floresta em diversos pontos da América do Norte. Desse modo, os pontos seriam unidades amostrais (as linhas em sua planilha) e as variáveis dependentes e independentes seriam consideradas as colunas de sua análise. É bastante importante ter em mente o formato padrão das planilhas utilizadas na maioria das análises ecológicas (Tabela 1). Alguns pacotes ou funções do R utilizam como padrão a matriz transposta da Tabela 1.

Após a definição das hipóteses/questões e de suas predições, é preciso pensar na estatística (lembra-se que entre os dois é preciso coletar e organizar os dados!). A estatística é necessária para descrever padrões nos nossos dados e para decidir se predições das hipóteses são verdadeiras ou não. Para começar a análise estatística é preciso definir as hipóteses estatísticas, i.e., hipótese nula (H_0) e hipótese alternativa (H_1). A hipótese nula representa a "ausência de padrão" na hipótese científica (i.e., as diferenças entre grupos não é maior do que o esperado ao acaso), enquanto a hipótese alternativa mostra exatamente a existência do padrão (notem que uma hipótese nula pode ter uma ou mais hipóteses alternativas). Por exemplo, a hipótese nula da Fig. 2 é que a densidade de árvores da zona ripária não afeta a riqueza de macro-invertebrados aquáticos, enquanto a hipótese alternativa é de que a densidade de árvores afeta positivamente a riqueza desses organismos. Neste exemplo, o pesquisador comparou grupos de riachos com densidades diferentes (e.g., variando de 0 a 10 árvores/m^2) e encontrou que riachos com florestas ripárias com densidade de árvores acima de 7/m^2 possuem 20% mais macro-invertebrados aquáticos (P = 0,01). Desse modo, a hipótese nula de ausência de padrão é rejeitada. Para decidir se a hipótese nula pode ser aceita ou não, os testes estatísticos utilizam

um valor de probabilidade. Como posso dizer que a média de um grupo é diferente da média de outro grupo ou que o aumento na variável X representa um aumento na variável Y? Como posso diferenciar se essas diferenças são reais ou frutos do acaso? O fator de decisão para a maioria dos testes estatísticos é o valor de *P* (probabilidade). O valor de *P* mede a probabilidade de que a hipótese nula (a ausência de um padrão) seja verdadeira. Desse modo, valores de *P* muito próximos de zero indicam que a probabilidade de que a hipótese nula seja verdadeira é muito baixa e que é possível considerar cenários alternativos, ou seja, aceitar a hipótese alternativa. No exemplo acima, a chance de a hipótese nula (a zona ripária não afeta a riqueza de macro-invertebrados) ser verdadeira é de 1 em 100 ($P = 0,01$). Se o valor de *P* fosse 0,76 a chance de a hipótese nula ser verdadeira seria de 76 em 100. O número "mágico" considerado como valor crítico de decisão é de 0,05. Desse modo, se a probabilidade de a hipótese nula ser verdadeira em um teste específico for $\leq 0,05$ ("resultado significativo"), decidimos por rejeitá-la. Do contrário, valores maiores do que 0,05 indicam que a hipótese nula deve ser aceita. A fixação do valor de significância de 5% foi puramente nominal, um consenso que visou o equilíbrio entre o erro do tipo I e do tipo II. Para entender os porquês desse valor de corte, consulte o livro do Gotelli & Ellison (2004, p. 96). Neste mesmo livro é preciso dedicar uma atenção especial aos erros atribuídos ao teste de hipóteses (erros do tipo I e II), que têm importância fundamental no processo analítico.

Tabela 1. Planilha modelo para análises estatística, com unidades amostrais nas linhas, e variáveis dependentes e independentes nas colunas

	v. dependente1	v. dependente2	...	v. dependente *n*	v. independente1	v. independente2	...	v. independente m
unid.amostral1	2.593	3.789		n_1	2.177	3.318		m_1
unid.amostral2	2.326	1.000		n_2	2.910	2.575		m_2
unid.amostral3	2.190	1.828		n_3	5.007	3.128		m_3
unid.amostral4	2.883	3.207		n_4	5.479	4.250		m_4
unid.amostral5	1.828	1.810		n_5	1.404	3.298		m_5
unid.amostral6	3.657	2.760		n_6	2.614	3.491		m_6
...	...	...		...	...	...		...
unid.amostral n_i	n_1	n_2		n_n	m_1	m_2		m_n

INTRODUÇÃO AO AMBIENTE DE PROGRAMAÇÃO R

O objetivo desta seção é apresentar aspectos básicos para qualquer pessoa livrar-se do receio inicial e começar a usar o R para efetuar análise de dados. Todo processo de aprendizagem torna-se mais efetivo quando a teoria é combinada com a prática, então nós recomendamos fortemente que você leitor acompanhe os exercícios desta apostila ao mesmo

tempo que os executa no seu computador, e não só os leia passivamente. Ainda, por motivo de tempo e espaço não abordaremos todas as questões relacionadas ao uso do R nesta apostila. Logo, aconselhamos que o leitor ao final das aulas você consulte o material sugerido para poder se aprofundar nas questões abordadas.

BAIXANDO E INSTALANDO A VERSÃO BASE DO R

Para começarmos a trabalhar com o R é necessário baixá-lo na página do R project da internet. Então, digite http://www.r-project.org na barra de endereços do seu navegador. Em seguida, clique no link **download R** embaixo da página, que o levará à pagina do CRAN (Comprehensive R Archive Network). Escolha qualquer página espelho do Brasil para baixar o programa. Escolha o sistema operacional do seu computador e clique em **base**.

Reserve algum tempo posteriormente para explorar esta página do R-project. Existem vários livros (http://www.r-project.org/doc/bib/R-books.html) dedicados a diversos assuntos baseados no R, além disso, estão disponíveis manuais (http://cran.r-project.org/manuals.html) em diversas línguas (http://cran.r-project.org/other-docs.html) para serem baixados gratuitamente.

Como o R é um software livre, não existe a possibilidade de o usuário entrar em contato com um serviço de suporte de usuários, muito comuns em softwares pagos. Ao invés disso, existem várias listas de correio eletrônico que fornecem suporte à comunidade de usuários (http://www.r-project.org/mail.html). Nós, particularmente, recomendamos o ingresso nas seguintes listas: R-help, R-sig-ecology, e R_BR (http://www.leg.ufpr.br/doku.php/software:rbr). Este último representa um grupo de usuários brasileiro do programa R. Ainda, existem vários blogs e páginas com arquivos de ajuda e planilhas com comandos, alguns deles podem ser baixados aqui: http://www.nceas.ucsb.edu/scicomp/software/r e http://devcheatsheet.com/tag/r/.

PORQUE USAR O R?

Os criadores do R o chamam de uma linguagem e ambiente de programação estatística e gráfica. O R também é chamado de programa "orientado ao objeto" (*object oriented programming*), o que significa que utilizar o R envolve basicamente a criação e manipulação de objetos em uma tela branca em que o usuário tem de dizer exatamente o que deseja que o

programa execute ao invés de simplesmente pressionar um botão. E vem daí uma das grandes vantagens em se usar o R: o usuário tem total controle sobre o que está acontecendo e também tem de compreender totalmente o que deseja antes de executar uma análise.

Na página pessoal do Prof. Nicolas J. Gotelli existem vários conselhos para um estudante iniciante de ecologia. Dentre esses conselhos, o Prof. Gotelli menciona que o domínio de uma linguagem de programação é uma das mais importantes, porque dá liberdade ao ecólogo para executar tarefas que vão além daquelas disponíveis em pacotes comerciais. Além disso, a maioria das novas análises propostas nos mais reconhecidos periódicos em ecologia normalmente são implementadas em linguagem R, e os autores incluem normalmente o código fonte no material suplementar dos artigos, tornando a análise acessível. A partir do momento que essas análises ficam disponíveis (seja por código fornecido pelo autor ou por implementação em pacotes pré-existentes), é mais simples entendermos a lógicas de análises complexas, especialmente as multivariadas, com nossos próprios dados realizando-as passo a passo. Sem a utilização do R, normalmente temos que contatar os autores que nem sempre são acessíveis.

Uma última vantagem é que por ser um software livre, a citação do R em artigos é permitida e até aconselhável. Para saber como citar o R, digite `citation()` na linha de comando. Para citar um pacote específico, digite `citation()` com o nome do pacote entre aspas dentro dos parênteses. Neste ponto, esperamos ter convencido você leitor de que aprender a utilizar o R tem inúmeras vantágens, vai ser difícil no começo mas continue e perceberá que o investimento vai valer à pena no futuro.

O "WORKSPACE" DO R E O TINN-R

Com o R é possível manipular e analisar dados, visualizar gráficos e escrever desde pequenas linhas de comando até programas inteiros. O R é a versão em código aberto de uma linguagem de programação inventada nos anos 1980 no Bell Labs chamada de S. Essa linguagem tornou-se bastante popular e vários produtos comerciais que a usam estão disponíveis, como o S-PLUS, SPSS, STATA e SAS. Um aspecto digno de nota é que a linguagem R, ao contrário de outras linguagem como Fortran e C, é uma linguagem *interpretada*, o que a faz ser mais fácil de programar, pois processa linhas de comando e as transforma em linguagem de máquina (código binário que o computador efetivamente lê), mas isso diminui a velocidade de processamento.

Nas linhas de comandos do R haverá um sinal de >, que indica o *prompt,* representando que o R está pronto para receber comandos. Se uma linha de comando não está completa, aparecerá um sinal de +, indicando que você poderá continuar a digitar aquela linha. Para que o *prompt* apareça novamente, pressione Esc. Para que os comandos sejam executados, pressione Enter. Para criar objetos, podemos utilizar os símbolos -> ou = . Estes símbolos representam que queremos "guardar" a informação dentro do objeto.

Neste curso iremos utilizar o R em conjunto com um editor, o Tinn-R. Existem vários editores para a linguagem R, como o RStudio, Eclipse etc. (veja uma lista não exaustiva em http://en.wikipedia.org/wiki/R_(programming_language)), mas preferimos o Tinn-R por ser de mais fácil utilização e por possibilitar o destaque das sintaxes de programação, diminuindo erros de digitação tão comuns. E ainda, é possível salvar os scripts para continuar a trabalhar neles posteriormente. Para baixá-lo, vá até http://www.sciviews.org/Tinn-R/ e faça o download do programa. Assim que o instalar, somente será necessário clicar no ícone do Tinn-R e o R abrirá automaticamente. Toda vez que terminar de escrever uma linha de comando, pressione Ctrl+Enter para enviá-la para o R.

Para saber qual é o diretório de trabalho do R, ou seja, em qual pasta o programa salvará arquivos, digite:

```
>get.wd()
```

É possível mudar o diretório de trabalho do R de acordo com as necessidades do usuário. Então, como exercício para este curso, clique em Arquivo>mudar dir. e defina o diretório para uma pasta deste curso dentro de Meus documentos. Nós recomendamos mudar o diretório sempre que um novo conjunto de análises for feito como, por exemplo, quando for mudar das análises do primeiro capítulo da sua dissertação para o segundo, escolha a pasta onde estarão os dados deste capítulo como diretório de trabalho.

OS TIPOS DE OBJETOS: CRIAÇÃO E MANIPULAÇÃO

Existem cinco classes de objetos na linguagem R: vetor, matriz, data frame, funções e lista.

Vetor

Existem três tipos de vetores: o vetor de caracteres, numérico e o lógico.

Vetor numérico

```
>a<-1
>c(1,2,3,4,5)->b
>dados.campo=seq(1,10,2)#cria uma sequência de números de 1 até
10, de 2 em 2
>x=seq(3,10) #cria uma sequência de números de 3 até 10
>sample(x, 2, replace=T)
>mata.1=rep(1:2, c(10,3))#repete o número 1 dez vezes e o número
2 três vezes
>exemplo=c(1:10)
>length(exemplo)
```

A linguagem R é *case sensitive*, o que quer dizer que ele distingue entre letras minúsculas e maiúsculas. Desse modo, fique atento ao criar um objeto e digite-o exatamente como quando você o criou. Ainda, não use acentos, til, crases etc. ao dar nome aos objetos.

Vetor de caracter

Também é possível criar vetores de caracteres, ou seja, com nomes ao invés de números. No R, sequências de caracteres textuais são sempre delimitados por aspas:

```
>dados.pessoais=c(nome="seuNome", nascimento="aniversario",
estadoCivil="solteiro")
>dados.pessoais
```

Vetor lógico

Vetores lógicos são quantidades lógicas manipuladas no R. Estes vetores são bastante úteis em programação. Os elementos de um vetor lógico são TRUE, FALSE ou NA (*not available*). Abaixo estão exemplos de *condições* criadas, quando a condição é satisfeita, o R retorna o valor TRUE, quando a mesma não é satisfeita, retorna FALSE

```
>is.factor(x)
>FALSE
>is.matrix(xy)
>FALSE
>a<-1
```

```
>a<1
>a==1
>a>=1
>a!=2
```

Fator

Um fator é utilizado para criar uma variável categórica, muito comum em análises estatísticas. Para criar um fator, digite:

```
>dados=factor(c("baixo", "menos baixo","médio" ,"alto"))#notem
que utilizamos um acento em médio, isto é possível porque esta
palavra aqui é tratada como um caracter (por isso as aspas) e
não como um objeto
>is.factor(dados)#testa a conversão
```

Matriz

Uma matriz é um arranjo bi-dimensional de vetores, todos os vetores devem ser do mesmo tipo (numérico ou de caracteres). Veja um exmplo abaixo de como criar uma matriz e manipulá-la:

```
>xy=matrix(1:12, nrow=3)
>rownames(xy)=LETTERS[1:3]
>colnames(xy)=c("mata.1", "mata.2", "mata.3", "mata.4")
>xy
>t(xy)#transpõe a matriz
>class(xy)
>xy[,1] #para acessar a primeira coluna de uma matriz
>xy[1,] #para acessar a primeira linha de uma matriz. Veja que
as chaves representam [linha, coluna]
>head(xy) #para acessar as primeiras linhas de uma matriz
>tail(xy) #para acessar as últimas linhas de uma matriz
>fix(xy) #edita uma matriz ou data frame
>str(xy)#avalia a estrutura do objeto
>summary(xy)
```

Data frame

O mesmo que uma matriz, mas aceita vetores de tipos diferentes. Este é o tipo mais comum de objeto que iremos usar ao longo deste curso. Um *data frame* permite incluir num mesmo objeto vetores numéricos e de caracteres, por exemplo:

```
>comunidade<- data.frame(especies = c("D.nanus",
"S.alter","I.guentheri", "A. callipygius"), habitat =
factor(c("Folhiço", "Arbóreo", "Riacho", "Poça")), altura =
c(1.1, 0.8, 0.9, 1), distancia = c(1, 1.7, 0.6, 0.2))
>class(comunidade)
>xy=as.data.frame(xy)#converte (coerce) a matriz que criamos
acima numa data frame
>class(xy) #testa a conversão
>str(comunidade)
>fix(comunidade)
>edit(comunidade)
```

Lista

Uma lista é um objeto que consiste de um conjunto de objetos ou componentes ordenados de forma hierárquica. Por exemplo, é possível construir uma lista com uma matriz, um vetor lógico, etc.

```
> Lista.ex <- list(name="Toyoyo", wife="Rafaela", no.children=2,
child.ages=c(2,6))
```

Muitos testes produzem objetos em formato de listas como resultado. Às vezes é útil extrair partes de uma lista para que possam ser utilizados posteriormente.

```
>Lista.ex$name
```

OPERAÇÕES ARITMÉTICAS BÁSICAS

O R também pode ser utilizado como uma calculadora. Faça algumas operações aritméticas com os objetos que você acabou de criar, por exemplo:

```
>a*2
>b*3 #observe o que aconteceu? Como foi feita essa operação?
>b[1]*3 #e agora?
>b/4
>2+3
>3^3
>log(2)#observe o que aconteceu? Este é a função que calcula o
logaritmo neperiano (ln).
>log10(2) #compare o resultado anterior com este. São
diferentes?
>sqrt(3)
>sum(a)
>mean(b)
>sum(b)/length(a)
>pi
>cor(a,b)
>cor.test(a,b)
?cor.test
```

ENTENDENDO O ARQUIVO DE AJUDA

Um importante passo para ter certa intimidade com a linguagem R é aprender a usar a ajuda de cada função. Além disso, existem uma função (RSiteSearch) e um pacote (*sos*) que também auxiliam o usuário a realizar uma análise quando não se sabe qual (e se) a mesma já foi implementada no R. Para utilizar o RSiteSearch, digite um tema ou o nome de uma análise entre aspas no argumento da função, como no exemplo abaixo:

```
>RSiteSearch("analysis of variance")
```

A função irá buscar na página do R na internet qual(is) função está(ão) disponível(is) para implementar aquela dada análise.

Se o pacote *sos* estiver instalado e carregado, basta digitar:

```
>???"analysis of variance"
```

e o navegador de internet abrirá uma página mostrando qual(is) funções executam aquela análise. Também é necessário acesso à internet. Outra ferramenta de busca é a página

http://www.rseek.org na qual é possível buscar por um termo não só nos pacotes do R, mas também em listas de emails, manuais, páginas na internet e livros sobre o programa.

Vamos fazer um exercício para nos ambientarmos com a página de ajuda do R, digite:

```
>?aov
```

O arquivo de ajuda do R possui geralmente nove ou dez tópicos:

Description - resumo da função

Usage*- como utilizar a função e quais os seus argumentos

Arguments* - detalha os argumentos e como os mesmos devem ser especifidados

Details - detalhes importantes para se usar a função

Value - mostra como interpretar a saída (*output*) da função (os resultados)

Notes - notas gerais sobre a função

Authors - autores da função

References - referências bibliográficas para os métodos usados pra construir a função

See also - funções relacionadas

Examples* - exemplos do uso da função. Às vezes pode ser útil copiar esse trecho e colar no R para ver como funciona e como usar a função.

INSTALANDO E CARREGANDO PACOTES

O R é um ambiente de programação e existem atualmente mais de 3000 pacotes que desempenham funções específicas e que precisam ser instalados e carregados independentemente. Os pacotes *stats* e *base* já vêm instalados e carregados, são estes pacotes que possuem as funções para o cálculo de modelos lineares simples, como teste t, ANOVA, χ^2, glm etc. A função que instala pacotes no R é a `install.packages()`.

Ao longo deste curso utilizaremos vários pacotes, entre eles o *vegan*, para instalá-lo, utilize:

```
>install.packages("vegan")
para instalar vários pacotes ao mesmo tempo, utilize a função
c()para criar um vetor:
```

```
>install.packages(c("vegan", "sos"))
```

e para carregá-los, utilize:

```
>library(vegan)
?vegan
```

Sempre que tiver de usar as funções de um pacote será preciso carregá-lo usando a função `library()`. A maioria dos pacotes vem com bancos de dados que podem ser acessados pelo comando `data()`. Esses bancos de dados podem ser usados para testar as funções do pacote. Se estiver com dúvida na maneira como você deve preparar a planilha para realizar uma análise específica, entre no help da função e veja os conjuntos de dados que estão no exemplo desta função.

IMPORTAÇÃO E EXPORTAÇÃO DE DADOS

```
>obj=read.table(file.choose(), header=TRUE) # este comando irá abrir uma tela para que o usuário navegue nas pastas e escolha o arquivo a ser aberto.
>obj=read.table("clipboard", h=T)#importa objetos que estiverem na área de transferência
>obj=read.table("nomedoarquivo.txt", h=T) #para utilizar este argumento, o arquivo a ser importado deve estar no diretório de trabalho
>obj=read.csv(file.choose(), h=T)
>write.table(nomeDoObjeto, "NomeDoObjetoParaSerGravado", sep=" ", quote=F, dec=".")
>sink("japi-so.xls") #Exporta pra o wd o(s) objetos que forem exibidos depois, com o nome que for colocado nesta linha de comando
>japi.so1
>sink()#Fecha o dispositivo
>?tiff
>?jpeg
```

Exercícios

1) Crie 2 conjuntos de dados de 30 unidades amostrais cada com distribuição normal, média 1 e desvio padrão 2.5 e descubra como calcular um teste t para este conjunto, tentem:

```
>?rnorm
>?t.test
```

2) Crie 4 vetores numéricos de qualquer tamanho com a função `c()`, você também pode combinar as funções `seq()` e `c()` se desejar.

a) calcule o comprimento de cada um desses vetores e guarde o resultado num outro vetor.

b) calcule o somatório dos componentes de cada vetor e guarde o valor num outro vetor.

c) utilize os itens b) e c) para calcular a média dos valores de cada um dos vetores.

3) Calcule novamente a média dos vetores, agora utilizando a função `mean()`.

4) Digite `ls()` e recupere o objeto dados.campo, selecione:

a) os cinco primeiros elementos deste objeto;

b) todos os elementos MENOS os 2 primeiros;

c) o 3º elemento;

d) todos menores que 4.

5) Crie duas sequências de 1 a 20 com intervalo de 1. Atribua nomes diferentes a cada uma.

7) Utilize a função `cbind()` para unir os dois vetores. Nomeie as colunas de *a* até *u* utilizando o vetor 'letters', e as duas colunas com o vetor 'LETTERS' já disponíveis no R.

8) Recupere o objeto xy que criamos há pouco, ele é uma matriz.

a) Multiplique-o por um escalar qualquer, por exemplo 3, veja o que acontece;

b) Divida o valor encontrado por 4, observe o que acontece e tente se lembrar das aulas de álgebra de matrizes do 3º colegial.

c) acesse o elemento $a_{3,1.}$

CRIAÇÃO E MANIPULAÇÃO DE GRÁFICOS NO R

O R é uma poderosa ferramenta para criação e manipulação de gráficos. Os pacotes *graphics* e *grid*, que já vêm instalados no R, possuem a função genérica `plot()`, além de outras como `hist()`. As funções `par()` e `layout()` permitem ainda plotar vários gráficos conjuntamente, formando uma única figura.

Alguns pacotes foram desenvolvidos especialmente para manipulação de gráficos, como *lattice*, *ggplot2*, *ggobi* e *rgl*. Estes pacotes nos permitem fazer praticamente todos os tipos de gráficos, incluindo 3-D e mapas em relevo. Para visualizar uma parte das potencialidades dos pacotes, instale e carregue-os. Digite no prompt do R `demo(lattice)` e vá apertando Enter. Faça o mesmo com o *ggplot2*. Neste módulo iremos demonstrar algumas das potencialidades gráficas do R. Reiteramos que esses pacotes são um mundo em si só. Logo, convidamos o leitor a ler e explorar a literatura sugerida abaixo, consultar os quadros resumos, além de acessar as seguintes páginas da internet:

http://research.stowers-institute.org/efg/R/
http://addictedtor.free.fr/graphiques/
http://www.gnuplot.info/
http://gnuplot.sourceforge.net/demo_4.2/
http://www.statmethods.net/advgraphs/parameters.html.

As principais funções que possibilitam modificar gráficos no R são:

`plot()`#Função genérica para plotar gráficos

#utilize os argumentos *xlab* e *ylab* para adicionar legendas aos eixos, use aspas.

bty=”L” retira as molduras das partes direita e superior.

xlim e *ylim* determina os limites das escalas dos eixos.

cex modifica o tamanho dos pontos.

pch modifica o tipo do ponto

col modifica as cores dos pontos. Veja também a ajuda da função `par()`.

```
hist()# plota um histograma
barchart()# plota um gráfico de barras
```

```
locator()#localiza uma coordenda x-y no gráfico, utilize o
argumento 1, 2 etc para definir quantos pontos quer localizar
text()#adiciona um texto
arrows()#adiciona uma seta
mtext()adiciona um texto nas margens do gráfico
box()#adiciona uma moldura
segments()#adiciona uma linha
legend()#adiciona legendas no alto e embaixo
points()#adiciona pontos no gráfico
lines()#adiciona linhas no gráfico
par()#divide o layout e plota vários gráficos, utilize o
argumento mfrow=c(2,2) para especificar o número de linhas e
colunas. Neste caso a função par(mfrow=c(2,2)) cria uma janela
para que quatro gráficos sejam visualizados (i.e., duas linhas e
duas colunas)
layout()#divide o layout e plota vários gráficos, utilize o
argumento layout(matrix(1:4, ncol=2, nrow=2)) pra definir o
número de colunas e linhas.
```

O pacote *lattice* permite fazer gráficos univariados e multivariados de alto nível. Além disso, ele permite criar objetos da classe trellis que podem ser exportados e modificados.

```
xyplot()#função do lattice para gráficos univariados
bwplot()# plota um boxplotcoplot()#plota vários gráficos com
estilos diferentes
```

Exercícios

1) Carregue o pacote *lattice* e o conjunto de dados quakes, `data(quakes)`, plote os dados utilizando a função `xyplot()`.

2) Carregue o conjunto de dados melanoma e utilizando a função `plot()` faça um gráfico com o tamanho dos pontos 24, legenda do eixo x “Frequência”, legenda do eixo y “Anos” e sem as molduras da direita e superior.

3) Crie dois conjuntos de dados quaisquer e combinando as funções `abline()` e `lm()` calcule uma regressão linear simples e ajuste uma reta que indique o modelo.

4) Crie um conjunto aleatório de números com distribuição normal e dê nome a este objeto. Utilize a função `hist()` para plotar um gráfico com as barras em cor cinza.

a) Utilize a função `points()` para criar um ponto em formato de círculo no eixo x no lugar da média.

b) Agora crie dois pontos verdes em formato de triângulo verde invertido no lugar dos 2 quantis.

c) Crie uma legenda no canto superior esquerdo com os símbolos utilizados (triângulo e círculo), com os significado (média e quantil).

d) Pinte de vermelho e verde os símbolos.

Quais funções você aprendeu?

Uma linguagem de programação é uma linguagem como qualquer outra, e sua aprendizagem exige domínio de vocabulário e sintaxe. O vocabulário da linguagem R são as funções e comandos. Então, sempre que um módulo acabar, lembre-se de tomar nota das funções e comandos, bem como para que serve cada uma delas. Utilize o marcador # em frente a uma função para explicar a sua utilidade. Você se lembra de todas que aprendeu hoje?

DISTRIBUIÇÕES ESTATÍSTICAS

Uma distribuição estatística é definida como uma **função** que define uma curva. A área sob essa curva determina a **probabilidade** de ocorrência de um dado evento.

Variáveis aleatórias:

A variável aleatória (X) é uma variável que tem um valor único (determinado aleatoriamente) para cada resultado de um experimento. A palavra aleatória indica que em geral só conhecemos aquele valor depois do experimento ser realizado.

Exemplos de variáveis aleatórias:

a. Número de presas capturadas em um determinado dia;

b. Comprimento de um peixe adulto selecionado aleatoriamente.

As variáveis aleatórias podem ser **discretas** ou **contínuas**.

Variável aleatória discreta: número ou a quantidade observada na unidade experimental ou tentativa.

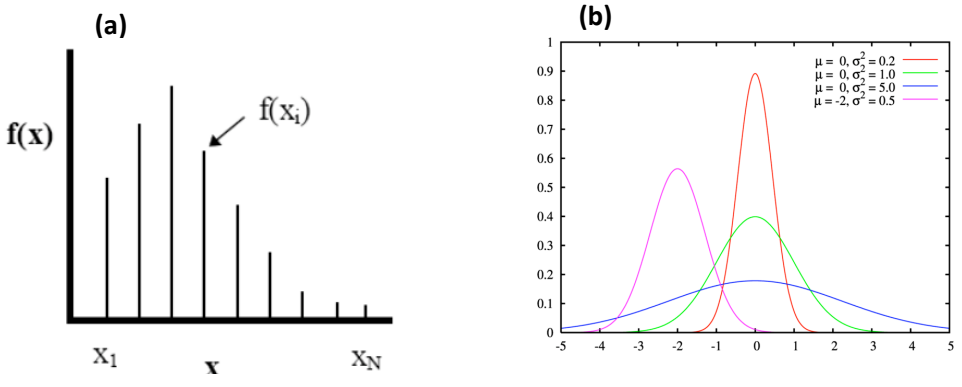
- Representada por números inteiros (0, 1, 2, 3, 4...);
- Não pode conter números negativos;
- Número finito de possibilidades;
- Podemos achar a probabilidade de cada evento.

Variável aleatória contínua: usualmente medidas contínuas como peso, altura, distância, pH, biomassa, etc.

- Representada por números não inteiros (1,3; - 1,54; - 1,7);
- Pode conter números negativos;
- Número infinito de possibilidades;
- Probabilidade de cada evento é zero.

FUNÇÕES DE PROBABILIDADE

A função probabilidade associa cada possível valor da variável aleatória (X) à sua probabilidade de ocorrência P(X). Quando conhecemos todos os valores de uma variável aleatória, juntamente com suas respectivas probabilidades, temos uma distribuição de probabilidades (Fig. 3). As distribuições de probabilidade discreta é conhecida como **função massa de probabilidade,** enquanto que distribuições de probabilidade contínua é conhecida como **função de densidade de probabilidade.** A diferença está no fato de que nas distribuições discretas temos a probabilidade para cada valor de X (Fig. 3a), enquanto que nas distribuições contínuas temos a probabilidade para um intervalo (Fig. 3b).



Função massa de probabilidade **Função densidade de probabilidade**

Figura 3. Funções de probabilidade para (a) variável discreta e (b) variável contínua.

FUNÇÕES DE DISTRIBUIÇÃO ACUMULADA

A **função de distribuição acumulada** é igual à probabilidade de que a variável aleatória *X* assuma um valor inferior ou igual a determinado *x* (Figura 4).

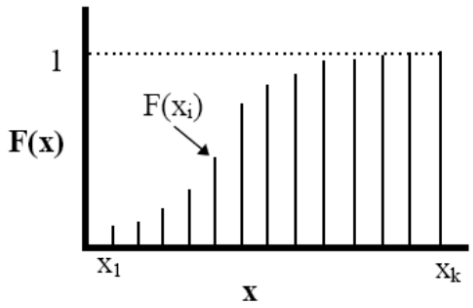


Figura 4. Função de distribuição acumulada.

DISTRIBUIÇÃO BINOMIAL

É a distribuição de probabilidade **discreta** do **número de sucessos** em uma sequência de ***n* tentativas** tal que: i) as tentativas são **independentes**; ii) cada tentativa resulta apenas em duas possibilidades, sucesso ou fracasso; e iii) a probabilidade de cada tentativa, **p, permanece constante**.

Se a variável aleatória X que contém o número de tentativas que resultam em sucesso tem uma distribuição binomial com parâmetros *n* e *p*, escrevemos $X \sim B(n, p)$. A probabilidade de se ter exatamente *k* sucessos é dada pela função de probabilidade:

$$p(X) = \binom{n}{X} q^x (1-q)^{n-X}$$

onde *q* é a probabilidade de um evento ocorrer, 1 – *q* é a probabilidade do evento não ocorrer, *X* é a freqüência de ocorrência e pode adquirir os valores 0, 1, 2, ..., *n*. Portanto, esta função fornece a probabilidade de ocorrerem *X* sucessos em *n* tentativas.

Se a $X \sim B(n, p)$, isto é, *X* é uma variável aleatória distribuída binomialmente, então o valor esperado de *X* é:

$$E[X] = np$$

e a variância é

$$var(X) = np(1-p)$$

Exemplo

Há uma probabilidade de 0,30 de um girino, ao forragear em um corpo d'água, ser predado por uma larva de odonata. Determine as probabilidades de que, dentre seis girinos que estão forrageando no corpo d'água, 0, 1, 2, 3, 5 ou 6 sejam predados. Trace um histograma dessa distribuição de probabilidade.

Solução

Admitindo que a escolha seja aleatória, fazemos n = 6, q = 0,30 e, respectivamente, X = 0, 1, 2, 3, 4, 5 e 6 na fórmula da distribuição binomial:

$$p(X) = \binom{n}{X} q^x (1-q)^{n-X}$$

$$p(0) = \binom{6}{0}(0{,}30)^0(0{,}70)^6 \approx 0{,}118 \qquad p(4) = \binom{6}{4}(0{,}30)^4(0{,}70)^2 \approx 0{,}060$$

$$p(1) = \binom{6}{1}(0{,}30)^1(0{,}70)^5 \approx 0{,}303 \qquad p(5) = \binom{6}{5}(0{,}30)^5(0{,}70)^1 \approx 0{,}010$$

$$p(2) = \binom{6}{2}(0{,}30)^2(0{,}70)^4 \approx 0{,}324 \qquad p(6) = \binom{6}{6}(0{,}30)^6(0{,}70)^0 \approx 0{,}001$$

$$p(3) = \binom{6}{2}(0{,}30)^3(0{,}70)^3 \approx 0{,}185$$

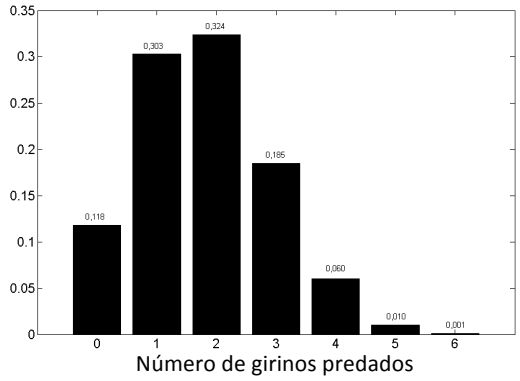


Figura 5. Histograma da distribuição binomial com n = 6 e q = 0,30.

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

Existem quatro funções que podem ser utilizadas para gerar os valores associados à distribuição binomial. Você pode obter uma lista completa das mesmas e as suas opções com o comando help:

```
>help(Binomial)
```

Quando o número de tentativas (size) e a probabilidade de sucesso são conhecidos para cada evento (prob) é possível utilizar o comando abaixo para descobrir a probabilidade para qualquer valor da variável *x*.

```
>dbinom(x, size, prob)
```

No caso do exemplo acima, para descobrirmos qual a probabilidade de dois girinos serem predados, precisamos digitar o seguinte comando:

```
>dbinom (2, size = 6, prob = 0.3)
0.324135
```

A probabilidade de três girinos serem predados

```
>dbinom (3, size = 6, prob = 0.3)
0.18522
```

Função de probabilidade acumulativa - Para descobrir a probabilidade de valores menores ou iguais a X utilizamos o comando:

```
>pbinom(q, size, prob)
```

Para descobrirmos qual a probabilidade de dois ou menos girinos (0, 1) serem predados, precisamos digitar o seguinte comando:

```
>pbinom (2, size = 6, prob = 0.3)
0.74431
```

Para descobrirmos qual a probabilidade de que cinco ou menos girinos (0, 1, 2, 3, 4) sejam predados, precisamos digitar o seguinte comando:

```
>pbinom (5, size = 6, prob = 0.3)
0.999271
```

Inverso da função de probabilidade acumulativa - Um exemplo contrário ao comando anterior é utilizado quando um valor de probabilidade é fornecido e o programa retorna o valor de X associado a ele. Para isso utiliza-se o seguinte comando:

```
>qbinom(p, size, prob)
```

Qual o valor de X (número de girinos predados) associado à probabilidade de 0,74?

```
>qbinom(0.74, size = 6, prob = 0.3)
2
```

Qual o valor de X (número de girinos predados) associado a probabilidade de 0,99?

```
>qbinom(0.99, size = 6, prob = 0.3)
5
```

Finalmente, números aleatórios podem ser gerados de acordo com a distribuição binomial com o seguinte comando:

```
>rbinom(n, size, prob)
```

Por exemplo, para gerar dez números aleatórios de uma distribuição binomial com 20 tentativas e probabilidade 0,63.

```
>rbinom(10,  size = 20, prob = 0.63)
```

Você pode plotar o gráfico da **função massa de distribuição** através do seguinte comando:

```
>plot(dbinom(seq(0,6, by =1), size = 6, prob = 0.3), type ="h",
xlab = "Número de girinos predados", ylab = "Probabilidade",
main = "Função massa de probabilidade")
```

O gráfico da **função de probabilidade acumulada** pode ser plotado com o seguinte comando:

```
>plot(pbinom(seq(0,6, by =1), size = 6, prob = 0.3),type ="h",
xlab = "Número de girinos predados", ylab = "Probabilidade",
main = "Função de probabilidade acumulada")
```

DISTRIBUIÇÃO POISSON

Na teoria da probabilidade e na estatística, a **distribuição de Poisson** é uma distribuição de probabilidade **discreta**. Expressa a probabilidade de uma série de eventos ocorrem em um período fixo de tempo, área, volume, quadrante, etc. Esta distribuição segue as mesmas premissas da distribuição binomial: i) as tentativas são **independentes**; ii) a variável aleatória é o número de eventos em cada amostra; e iii) a probabilidade é **constante** em cada intervalo.

A probabilidade de que existam exatamente k ocorrências (k sendo um número inteiro, não negativo, $k = 0, 1, 2, ...$) é:

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}\frac{e^{-\lambda}\lambda^k}{k!}$$

- e é base do logaritmo natural ($e = 2.71828...$),
- $k!$ é o fatorial de k,
- λ é um número real, igual ao número esperado de ocorrências que ocorrem num dado intervalo de tempo.

Se a $X \sim Pois(\lambda)$, isto é, X é uma variável aleatória com distribuição Poisson, então o valor esperado de X é

$$E[X] = \lambda$$

e a variância é

$$Var[X] = \lambda$$

Exemplo

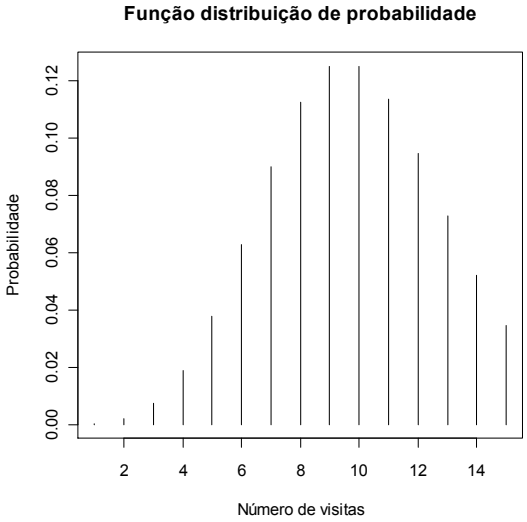
Suponha que um pesquisador registrou o número de visitas à flor de uma planta durante um período de 15 minutos. O número médio de borboletas que visitam no período de 15 minutos é 10 (λ). Determine a probabilidade de que cinco borboletas visitem a flor em 15 minutos. A probabilidade de uma borboleta visitar é a mesma para quaisquer dois períodos de tempo de igual comprimento. Trace um histograma dessa distribuição de probabilidade.

Solução

Admitindo que a visita ou não visita de uma borboleta em qualquer período de tempo é independente da visita ou não visita de uma segunda borboleta em qualquer outro período de tempo, fazemos λ = 10 e X = 5 na fórmula da distribuição poisson:

$$P(X = 5) == \frac{10^5 e^{-10}}{5!} = 0{,}0378$$

$$P(X = 5) == \frac{10^5 e^{-10}}{5!} = 0{,}0378$$



REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

Existem quatro funções que podem ser utilizadas para gerar os valores associados à distribuição poisson. Você pode obter uma lista completa das mesmas e as suas opções com o comando help:

```
>help(Poisson)
```

Quando você tem a média por unidade de tempo, área ou quadrante (λ) você pode utilizar o comando abaixo para descobrir a probabilidade para qualquer valor da variável X.

```
>dpois(x, lambda)
```

No caso do exemplo acima, para descobrirmos qual a probabilidade de que cinco borboletas visitem uma flor, precisamos digitar o seguinte comando:

```
>dpois (5, lambda = 10)
0.03783327
```

A probabilidade de que oito borboletas visitem uma flor é:

```
>dpois (8, lambda = 10)
0.1125
```

Função de probabilidade acumulativa - Para descobrir a probabilidade de valores menores ou iguais a X utilizamos o comando:

```
>ppois(x, lambda)
```

Para descobrirmos qual a probabilidade de duas ou menos visitas (1) à flor, precisamos digitar o seguinte comando:

```
>ppois (2, lambda = 10)
0.00276
```

A probabilidade de cinco ou menos visitas (1, 2, 3, 4) à flor é:

```
>ppois (5, lambda = 10)
0.06708
```

Inverso da função de probabilidade acumulativa - Um exemplo contrário ao comando anterior é quando você fornece um valor de probabilidade e o programa retorna o valor de X associado a ele. Para isso usa-se o seguinte comando:

```
>qpois (p, lambda)
```

Qual o valor de X (número de visitas) associado à probabilidade de 0.8?

```
>qpois (0.8, lambda = 10)
13
```

Qual o valor de X (número de visitas) associado a probabilidade de 0.1?

```
>qpois (0.1, lambda = 10)
6
```

Finalmente números aleatórios podem ser gerados de acordo com a distribuição Poisson com o seguinte comando:

```
>rpois (n, lambda)
```

Por exemplo, para gerar dez números aleatórios de uma distribuição Poisson com média (λ) 22.

```
>rbinom(10,  lambda = 22)
```

Você pode plotar o gráfico da **função massa de distribuição** através do seguinte comando:

```
>plot(dpois(seq(1,10, by =1), lambda = 10), type ="h",xlab = "Número de visitas", ylab = "Probabilidade", main = "Função massa de probabilidade")
```

O gráfico da **função de probabilidade acumulada** pode ser plotado com o seguinte comando:

```
>plot(ppois(seq(1,10, by =1), lambda = 10),type ="h", xlab = "Número visitas", ylab = "Probabilidade", main = "Função de probabilidade acumulada")
```

Podemos usar a distribuição de Poisson como uma aproximação da distribuição Binomial quando *n*, o número de tentativas, for grande e *p* ou *1 – p* for pequeno (eventos raros). Um bom princípio básico é usar a distribuição de Poisson quando ***n* ≥ 30 e *n.p* ou *n.(1-p)* < 5%.** Quando *n* for grande, pode consumir muito tempo em usar a distribuição binomial e tabelas para probabilidades binomiais, para valores muito pequenos de *p* podem não estar disponíveis. Se *n(1-p)* < 5, sucesso e fracasso deverão ser redefinidos de modo que *Np* < 5 para tornar a aproximação precisa.

```
>plot(dbinom(seq(1,50, by =1), size =50, prob = 0.09), type ="h", ylab = "Probabilidade",    main = "Distribuição Binomial")
```

```
>plot(dpois(seq(1,50, by =1), lambda = 50*0.09), type ="h", ylab = "Probabilidade", main = "Distribuição Poisson")
```

DISTRIBUIÇÃO NORMAL

A distribuição normal é uma das mais importantes distribuições com probabilidades contínuas. Conhecida também como Distribuição de Gauss ou Gaussiana. Esta distribuição é inteiramente descrita por parâmetros de **média (μ)** e **desvio padrão (σ)**, ou seja, conhecendo-se estes parâmetros consegue-se determinar qualquer probabilidade em uma distribuição Normal.

A importância da distribuição normal como um modelo de fenômenos quantitativos é devido em parte ao **Teorema do Limite Central**. O teorema afirma que "*toda soma de variáveis aleatórias independentes de média finita e variância limitada é aproximadamente Normal, desde que o número de termos da soma seja suficientemente grande*" (Fig. 7). Independentemente do tipo de distribuição da população, na medida em que o tamanho da amostra aumenta, a distribuição das médias amostrais tende a uma distribuição Normal.

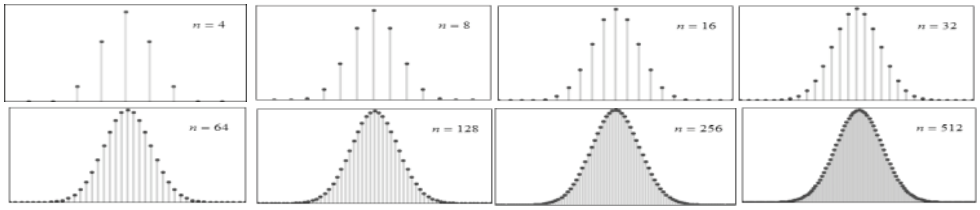


Figura 7. Gráficos demonstrando que mesmo com um grande número de variáveis aleatórias, as distribuições têm um padrão aproximadamente normal.

A **distribuição binomial** $B(n,p)$ é aproximadamente normal $N(np, np(1-p))$ para **grande *n*** e para ***p* não tão próximos de 0 ou 1**. Enquanto que a **distribuição Poisson** Pois(λ) é aproximadamente Normal $N(\lambda, \lambda)$ para **grandes valores de λ** .

A função de densidade de probabilidade da **distribuição normal** com média μ e variância σ^2 (de forma equivalente, desvio padrão σ) é assim definida,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Variáveis aleatórias com distribuição aproximadamente normal apresentam as seguintes propriedades:

- Metade (50%) está acima (e abaixo) da média
- Aproximadamente 68% está dentro de 1 desvio padrão da média

- Aproximadamente 95% está dentro de 2 desvios padrões da média
- Virtualmente todos os valores estão dentro de 3 desvios padrões da média

Na prática desejamos calcular probabilidades para diferentes valores de **μ** e **σ**. Para isso teríamos que realizar uma integral:

$$P(a < x < b) = \int_a^b \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} dx$$

Para facilitar, a variável **X** cuja distribuição é **N (μ , σ)** é transformada em uma forma padronizada **Z** com distribuição **N (0, 1)** (**distribuição Normal padrão**) cuja distribuição é tabelada. A quantidade **Z** é dada por :

$$X \sim N(\mu, \sigma) \quad \Rightarrow \quad Z = \frac{Y - \mu}{\sigma} \sim N(0,1)$$

Exemplo

Qual é a probabilidade de que um peixe capturado aleatoriamente tenha 20,15 cm ou mais, sabendo que a média da população é 17,1 cm e o desvio padrão é de 1,21 cm? Trace um histograma dessa distribuição de probabilidade.

Solução

$$Z_L = \frac{20.15 - 17.1}{1.21} = 2.52 \qquad Z_U = \infty$$

Para descobrir a probabilidade de se capturar um peixe maior que 20,15 cm, você precisa procurar pelo valor de Z = 2.52 em uma tabela de distribuição Z:

$P(X \geq 20.15) = P(Z \geq 2.52) = .0059 \ (\approx 1/170)$

Portanto, a probabilidade de se capturar um peixe aleatoriamente maior que 20,15 cm numa população com média 17,1 cm e desvio de 1,21 cm é de 0.006%.

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

Existem quatro funções que podem ser utilizadas para gerar os valores associados à distribuição Normal. Você pode obter uma lista completa das mesmas e as suas opções com o comando help:

```
>help(Normal)
```

Quando tem-se a média e o desvio padrão da população você pode utilizar o comando abaixo para descobrir a probabilidade para qualquer intervalo.

```
>pnorm(x, mean, sd, lower.tail = TRUE) ## Ficar atento para
      quando você quer medir intervalo acima da média ou abaixo
      dela. Quando for acima, você precisa substituir o TRUE
      por FALSE
```

No caso do exemplo acima, para descobrirmos qual a probabilidade de se capturar um peixe maior que 20,15 cm, precisamos digitar o seguinte comando:

```
>pnorm (20.15, mean = 17.1, sd = 1.21, lower.tail = FALSE)
0.0058567
```

Imagine que se tenha uma população com média 100 cm e um desvio padrão de 10 cm, para descobrir o intervalo associado com 95% de probabilidade você deve usar o seguinte comando:

```
>qnorm (0.95, mean = 100, sd = 10)
116.45
```

Para descobrir a probabilidade de se obter valores entre 80 e 120 cm, deve-se usar o seguinte comando:

```
>pnorm(120, mean=100, sd=10) - pnorm(80, mean=100, sd=10)
0.95449
```

Você pode plotar o gráfico da **função densidade de probabilidade** através do seguinte comando:

```
x = seq(70,130,length = 200)
y = dnorm(x, mean=100, sd=10)
plot(x, y, type="l", lwd=2, col="red", ylab =
"Probabilidade",main ="Função densidade de probabilidade")
```

O gráfico da **função de probabilidade acumulada** pode ser plotado com o seguinte comando:

```
x = seq(70,130,length = 200)
y = pnorm(x, mean=100, sd=10)
plot(x, y, type="l", lwd=2, col="red", ylab = "Probabilidade",main ="Função de probabilidade acumulada")
```

Exercícios

1) Uma aranha predadora que vive em flores polinizadas por pequenas mariposas consome em média cinco mariposas por hora. Qual a probabilidade da aranha predar duas mariposas em uma hora selecionada aleatoriamente?

2) Um pesquisador verificou que seis ovos de uma determinada ave são consumidos em média por hora em uma área de nidificação.
a) Qual é a probabilidade de que três ovos sejam predados?
b) Qual é a probabilidade de que três ou menos ovos sejam predados?

3) Um trabalho recente verificou que 1% dos fígados de cobaias submetidas ao tratamento com álcool apresentavam danos teciduais. Encontre a probabilidade de que mais de um fígado em uma amostra aleatória de 30 fígados apresente danos teciduais usando:

a) Distribuição Binomial
b) Distribuição Poisson

4) Uma nova técnica de amostragem registra dez indivíduos de lagartos por hora em uma área florestal. Encontre a probabilidade de que quatro ou menos indivíduos sejam registrados em uma hora aleatória.

5) Supondo que a probabilidade de um casal de ursos pandas ter filhotes albinos é de ¼. Se um casal produzir seis filhotes, qual é a probabilidade de que metade deles sejam albinos?

6) Se a probabilidade de um sapo capturar uma mosca em movimento é de 30%. Qual é a probabilidade de que em quatro tentativas ele capture no mínimo três moscas?

7) Um pesquisador extrai 15 amostras de DNA aleatoriamente de um banco de dados que produz 85% de amostras aceitáveis. Qual é a probabilidade de que dez amostras extraídas sejam aceitáveis?

8) Um população de crocodilos tem tamanho corporal médio de 400 cm e desvio padrão de 50 cm. Qual a probabilidade de capturarmos um crocodilo dessa população com tamanho entre 390 e 450 cm?

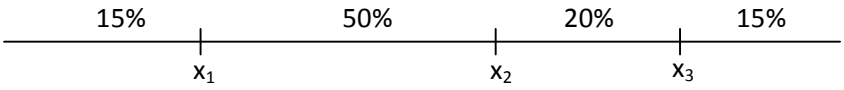
9) O comprimento do antebraço de uma espécie de morcego endêmica do Cerrado é de 4 cm com desvio padrão de 0,25 cm. A partir de qual comprimento os morcegos teriam os antebraços mais compridos nessa população?

10) Suponha que o tempo necessário para um leão consumir sua presa siga uma distribuição normal de média de 8 minutos e desvio padrão de 2 minutos.

(a) Qual é a probabilidade de que um leão consuma sua presa em menos de 5 minutos?
(b) E mais do que 9,5 minutos?
(c) E entre 7 e 10 minutos?

11) A distribuição dos pesos de coelhos criados em uma granja pode muito bem ser representada por uma distribuição Normal, com média 5 kg e desvio padrão 0,9 kg. Um pesquisador comprará 5000 coelhos e pretende classificá-los de acordo com o peso do seguinte modo: 15% dos mais leves como pequenos, os 50% seguintes como médios, os 20% seguintes como grandes e os 15% mais pesados como extras. Quais os limites de peso para cada classificação?

Classificação do pesquisador



Seja,
x_1 o valor do peso que separa os 15% mais leves dos demais,
x_2 o valor do peso que separa os 65% mais leves dos demais,
x_3 o valor do peso que separa os 85% mais leves dos demais.

Generalized Linear Models (GLM) – Modelos Lineares Generalizados

Muitos métodos estatísticos populares são baseados em modelos matemáticos que assumem que os dados seguem uma distribuição Normal, dentre eles a análise de variância e a

regressão múltipla. No entanto, em muitas situações a suposição de normalidade não é plausível. Conseqüentemente, o uso de métodos que assumem a normalidade pode ser insatisfatório e aumentam a probabilidade de cometermos erros inferenciais (erros do Tipo I e II). Nestes casos, outras alternativas que não pressupoem distribuição normal dos dados são atraentes e mais robustas.

Podemos usar modelos lineares generalizados (GLM) quando a variância não é constante, e/ou quando os erros não são normalmente distribuídos. Muitos tipos de dados têm erros não normais. No passado, as únicas maneiras capazes de lidar com esse problema eram a transformação da variável resposta ou a adoção de métodos não paramétricos. Em GLM, assumimos que cada resultado da variável dependente Y seja gerado a partir de uma variedade de diferentes tipos de distribuições que lidam com esse problema:

Poisson – úteis para dados de contagem

Binomial – úteis para dados com proporções

Gamma – úteis para dados mostrando um coeficiente constante de variância

Exponencial – úteis com dados de análises de sobrevivência

Existem muitas razões para usar GLMs, em vez de regressão linear. Dados de presença-ausência são (geralmente) codificados como 1 e 0, os dados proporcionais são sempre entre 0 e 100%, e os dados de contagem são sempre não-negativos. GLMs usados para 0-1 e dados proporcionais são normalmente baseados em distribuição binomial e para dados de contagem as distribuições de Poisson e binomial negativa são opções comuns.

A média, μ, da distribuição depende das variáveis independentes, X, e é calculada através de:

$$E(Y) = \mu = g^{-1}(X\beta)$$

onde E (Y) é o valor esperado de Y; Xβ é o preditor linear, uma combinação linear de parâmetros desconhecidos, β; g é a função de ligação.

GLM consiste em três etapas:

1. Uma hipótese sobre a distribuição da variável resposta Y_i. Isso também define a média e a variância de Y_i. (*e.x.*, Distribuição Poisson, Binomial, Gamma).

2. Especificação da parte sistemática. Esta é uma função das variáveis explicativas.

$$n_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{k1}$$

3. A relação entre o valor médio de Y_i e a parte sistemática. Esta é também chamada de ligação entre a média e a parte sistemática (Tabelas 2 e 3).

Tabela 2. Funções de ligações para GLM.

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

Tabela 3. Algumas das ligações mais comuns para GLM.

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0,1,\ldots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	0,1,2,...	μ_i
Gamma	Inverse	$(0,\infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0,\infty)$	$\phi\mu_i^3$

Likelihood

Os passos finais do processo de modelagem são constituídos pela estimativa dos parâmetros a partir dos dados e teste dos modelos uns contra os outros. Estimar os parâmetros dos modelos significa achar os parâmetros que fazem o modelo se ajustar melhor aos dados coletados. Nosso *goodness-of-fit* será baseado na probabilidade (*likelihood*) - a probabilidade de se encontrar nossos dados dado um modelo particular. Queremos a estimativa da máxima verossimilhança (*maximum likelihood estimate*) dos parâmetros – aqueles valores dos parâmetros que fazem os dados observados mais prováveis de terem acontecido. Uma vez que

as observações são independentes, a junção das probabilidades dos dados totais é o produto das probabilidades de cada observação individual. Por conveniência matemática, sempre maximizamos o logaritimo das probabilidades (*log-likelihood*) ao invés da probabilidade direto.

Likelihood Ratio Test

Os modelos GLM são ajustados aos dados pelo método de máxima verossimilhança, proporcionando não apenas estimativas dos coeficientes de regressão, mas também estimando erros padrões dos coeficientes. Nós podemos utilizar a *likelihood ratio test* (LRT) para escolher modelos em certas situações. A LRT compara dois modelos aninhados, testando se os parâmetros aninhados do modelo mais complexo diferem significativamente do valor nulo. Um modelo mais simples (com menos parâmetros) é aninhado em outro, mais complexo (com mais parâmetros), se o modelo complexo for reduzido para o mais simples pela retirada de um dos parâmetros. Em outras palavras, ele testa se há necessidade de se incluir um parâmetro extra no modelo para explicar os dados. O *residual deviance* para um GLM é $D_m = 2 (\log_e Ls - \log_e Lm)$, onde L*m* é a máxima verossimilhança sob o modelo em questão, e L*s* é a máxima verossimilhança sob um modelo saturado (modelo mais complexo) que dedica um paramêtro para cada observação e consequentemente ajusta os dados o mais próximo possível. O *residual deviance* é análogo à soma dos quadrados dos resíduos para um modelo linear. Em GLM para o qual o parâmetro de dispersão é fixado em 1 (binomial e Poisson), a razão da verossimilhança estatística do teste é a diferença dos *residual deviance* para os modelos aninhados. LRT apresenta uma distribuição de qui-quadrado com $k_1 - K_0$ graus de liberdade. Para GLM em que existe um parâmetro para estimar a dispersão (Gaussian, Quasi-poisson e Gamma), podemos comparar modelos aninhados por um teste F.

Akaike Information Criterion (AIC) - Critério de Informação de Akaike

O critério de Akaike é uma ferramenta para seleção de modelos, pois oferece uma medida relativa do *goodness-of-fit* (qualidade do ajuste) de um modelo estatístico. AIC não fornece um teste de um modelo no sentido usual de testar uma hipótese nula, ou seja, ele não pode dizer nada sobre o quão bem o modelo ajusta os dados em um sentido absoluto.

No caso geral, AIC é

$$AIC = 2K - 2\ln(L)$$

onde *k* é o número de parâmetros no modelo estatístico, e *L* é o valor maximizado da função *likelihood* para o modelo estimado. Dado um conjunto de modelos candidatos, o modelo preferido é aquele **com o valor mínimo de AIC**. O valor de AIC não só recompensa *goodness-of-fit*, mas inclui também uma penalização que é uma função crescente do número de parâmetros estimados. Esta penalidade desencoraja *overfitting* (aumentando o número de parâmetros livres no modelo melhora a qualidade do ajuste, independentemente do número de parâmetros livres no processo de geração de dados).

AIC_C é AIC com uma correção para amostras finitas:

$$AIC_c = AIC + \frac{2K\ (K+1)}{n\ -K-1}$$

onde *k* denota o número de parâmetros do modelo. Assim, AIC_C é AIC com uma maior penalização para os parâmetros extra.

Burnham & Anderson (2002) recomendam o uso do AIC_C, ao invés de AIC, se *n* for pequeno ou *k* é grande. Uma vez que o valor de AICc converge para AIC quando *n* se torna grande, AICc geralmente devem ser empregados independentemente do tamanho da amostra. Usar AIC, em vez de AIC_C, quando *n* não é muitas vezes maior do k^2 aumenta a probabilidade de seleção dos modelos que têm muitos parâmetros (*overfitting*).

Uma outra comparação entre os modelos pode ser baseada no cálculo do Peso do Akaike (*Akaike weigths* - Buckland et al. 1997). Se existem M modelos candidatos, então o peso para o modelo *i* é:

$$W_i = \frac{exp\ (\Delta/2)}{\exp\left(\frac{\Delta 1}{2}\right) + \exp\left(\frac{\Delta 2}{2}\right) + \cdots \exp\left(\frac{\Delta m}{2}\right)}$$

onde Δ é a diferença entre o valor do AIC entre modelo *i* e os modelos restantes. Os pesos do Akaike calculados desta forma são usados para medir a força da evidência em favor de cada um dos modelos, com um grande peso indicando alta evidência.

Dez orientações para Seleção de Modelo

1) Cada modelo deve representar uma hipótese (interessante) específica a ser testada.

2) Mantenha os sub-grupos de modelos candidatos curtos. É desaconselhável considerar tantos modelos quanto o número de dados que você tem.

3) Verificar a adequação do modelo: use o seu modelo global (modelo mais complexo) ou modelos subglobais para determinar se as hipóteses são válidas. Se nenhum dos modelos se ajustar aos dados, critérios de informação indicarão apenas o mais parcimonioso dos modelos mais pobres.

4) Evitar a dragagem de dados (e.g., procura de padrões após uma rodada inicial de análise).

5) Evite modelos *overfitted*.

6) Tenha cuidado com os valores faltantes (NA). Lembre-se de que valores faltantes somente para algumas variáveis alteram o tamanho do conjunto de dados e amostras dependendo de qual variável é incluída em um dado modelo. É sugirido remover casos omissos antes de iniciar a seleção de modelos.

7) Use a mesma variável resposta para todos os modelos candidatos. É inadequado executar alguns modelos com variável resposta transformados e outros com a variável não transformada. A solução é usar uma função de ligação diferente para alguns modelos (e.g., *identity* vs. *log link*).

8) Quando se trata de modelos com *overdispersion*, utilize o mesmo valor de *c-hat* para todos os modelos em um conjunto de modelos candidatos. Para modelos binomiais com *trials* > 1 ou com Poisson GLM, deve-se estimar o *c-hat* do modelo mais complexo (modelo global). Se *c hat* > 1, deve-se usar o mesmo valor para cada modelo do conjunto de modelos candidatos e inclui-lo na contagem dos parâmetros (K). Da mesma forma, para binomial negativa, você deve estimar o parâmetro de dispersão do modelo global e usar o mesmo valor em todos os modelos.

9) Burnham e Anderson (2002) recomendam evitar misturar a abordagem da teoria da informação e noções de significância (ou seja, os valores *P*). É melhor fornecer estimativas e uma medida de sua precisão (erro padrão, intervalos de confiança).

10) Determinar o ranking das modelos é apenas o primeiro passo. A soma do Peso Akaike é 1 para o modelo de todo o conjunto e pode ser interpretado como o peso das evidências em favor de um determinado modelo. Modelos com grandes valores do Peso Akaike têm forte apoio. Taxas de evidências, valores de importância, e intervalo de confianca para o melhor modelo são outras medidas que auxiliam na interpretação. Nos casos em que o melhor modelo do ranking tem um Peso Akaike > 0,9, pode-se inferir que este modelo é o mais parcimonioso. Quando muitos modelos são classificados por valores altos (ou seja, o delta (Q) AIC (c) < 2 ou 4), deve-se considerar a média dos parâmetors dos modelos de interesse que aparecem no topo. A média dos modelos consiste em fazer inferências com base no conjunto de modelos candidatos, em vez

de basear as conclusões em um único "melhor" modelo. É uma maneira elegante de fazer inferências com base nas informações contidas no conjunto inteiro de modelos.

Exemplos

A partir dos exemplos a seguir irei explicar os comandos básicos necessários para realizar as análises de GLM. É altamente recomendável que vocês recorram aos livros sugeridos no início desta apostila para um aprofundamento no assunto e para que possam realizar análises mais complexas.

Carregando pacotes necessários para as análises

```
>library(languageR)
>library(nlme)
>library(glmmML)
>library(lme4)
>library(AICcmodavg)
>library(bestglm)
>library(mgcv)
>library(MuMIn)
>library(pscl)
>library(MASS)
>library(bbmle)
>library(lattice)
>library(AED) ## Esse pacote tem deve ser baixado da página
#http://www.highstat.com/book2.htm
```

Primeiro Exemplo

```
>data(RoadKills)    ## Carregando dados - Os dados consistem do
número de mortes de anfíbios em uma rodovia em 52 sítios em
Portugal
```

Teoria: Ecologia de Paisagem
Variável dependente: Número de anfíbios mortos
Questão: Quais variáveis da paisagem melhor explicam a mortalidade de anfíbios?

```
>RK <- RoadKills        ## Renomeando para facilitar
```

Modelo Global

```
>M1 <- glm (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD + SQ.DWATCOUR + D.PARK,
      family = poisson, data=RK)
```

SELEÇÃO DO MELHOR MODELO

Akaike Information Criterion (AIC)

```
>step(M1) ## Esse comando faz a seleção automaticamente
```

Outra maneira de utilizar Akaike Information Criterion. É preciso construir os modelos de acordo com suas hipóteses ou retirando as variáveis que não apresentam um efeito significativo.

```
>M2 <- glm (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD + D.PARK, family =
      poisson, data=RK)
>M3 <- glm (TOT.N ~ MONT.S + SQ.POLIC + SQ.SHRUB + SQ.WATRES +
      L.WAT.C + SQ.LPROAD + D.PARK, family = poisson,
      data=RK)
>M4 <- glm (TOT.N ~ L.WAT.C + SQ.LPROAD + D.PARK, family =
poisson, data=RK)
```

Esse comando cria uma tabela colocando os modelos em ordem crescente de valores, ou seja, com o melhor modelo no topo. Ele apresenta o valor de delta que é a diferença entre o melhor modelo que recebe o valor de zero e os outros modelos.

WEIGHT = são usados para medir a força da evidência em favor de cada um dos modelos

```
>AIC <- ICtab (M1, M2, M3, M4, type = c("AIC"), weights = TRUE,
      delta = TRUE, sort = TRUE)
>AIC
```

Contudo, quando o número de amostras dividido pelo número de paramêtros for < 40 é recomendado utilizar um AIC corrigido (AIC_c) para pequenas amostras. Na verdade, como em

grandes amostras o valor de AIC_c tende ao valor de AIC sem correção, é recomendado sempre utilizar AIC_c.

```
>AICc <- ICtab(M1, M2, M3, M4, type = c("AICc"), weights = TRUE,
      delta = TRUE, sort = TRUE, nobs = 52)
>AICc
```

Terceira maneira de calcular AIC, AIC_c

Cria um vetor com lista de modelos:

```
>Modelos <- list()
>Modelos [[1]] <- glm(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC +
      SQ.SHRUB + SQ.WATRES + L.WAT.C + SQ.LPROAD + SQ.DWATCOUR +
      D.PARK, family = poisson (link = "log"), data=RK)
>Modelos [[2]] <- glm(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC +
      SQ.SHRUB + SQ.WATRES + L.WAT.C + SQ.LPROAD + D.PARK,
      family = poisson (link = "log"), data=RK)
>Modelos [[3]] <- glm(TOT.N ~ MONT.S + SQ.POLIC + SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD + D.PARK, family = poisson
      (link = "log"), data=RK)
>Modelos [[4]] <- glm(TOT.N ~ L.WAT.C + SQ.LPROAD + D.PARK,
family = poisson, data=RK)
```

Cria um vetor com nomes dos modelos

```
>(Modnames <- paste("Mod", 1:length(Modelos), sep=""))
```

Gera uma tabela com valores de AIC

```
>(res.table <- aictab(cand.set = Modelos, modnames = Modnames,
second.ord = FALSE))  ## FALSE: mostrar valores de AIC

>(res.table <- aictab(cand.set = Modelos, modnames = Modnames,
second.ord = TRUE)) ## TRUE: mostrar valores de AICc
```

TESTE DE HIPÓTESES - Likelihood ratio test (LRT)

DEVIANCE = RESIDUAL DEVIANCE = É 2 x a diferença entre o log likelihood do modelo que apresenta um ajuste perfeito (modelo saturado) e o modelo em questão. Quanto menor o *residual deviance*, melhor o modelo.

```
>drop1(M1,test = "Chi") # A diferença entre as deviance dos
      modelos apresenta uma distribuição chi- square com p1 - p2
      graus de liberdade

>DM1 <- glm(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD +         D.PARK, family =
      poisson, data = RK)

>drop1(DM1, test = "Chi")
```

```
Model:
TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB + SQ.WATRES + L.WAT.C +
    SQ.LPROAD + SQ.DWATCOUR + D.PARK
            Df Deviance     AIC    LRT   Pr(Chi)
<none>           270.23  529.62
OPEN.L       1   273.93  531.32   3.69 0.0546474 .
MONT.S       1   306.89  564.28  36.66 1.410e-09 ***
SQ.POLIC     1   285.53  542.92  15.30 9.181e-05 ***
SQ.SHRUB     1   298.31  555.70  28.08 1.167e-07 ***
SQ.WATRES    1   280.02  537.41   9.79 0.0017539 **
L.WAT.C      1   335.47  592.86  65.23 6.648e-16 ***
SQ.LPROAD    1   281.25  538.64  11.02 0.0009009 ***
SQ.DWATCOUR  1   272.50  529.89   2.27 0.1319862
D.PARK       1   838.09 1095.48 567.85 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Este resultado indica que podemos retirar a variável SQ.DWATCOUR, pois o modelo sem esta variável tem o mesmo poder de explicação do modelo com esta variável. Repita o processo até que nenhuma variável possa ser retirada do modelo.

OVERDISPERSION

Contudo a vida não é tão simples, antes de analisar os resultados e realizar as análises de seleção você precisa checar se os seus dados possuem *overdispersion*. A *overdispersion* significa que a variância é maior do que a média.

Como saber se os dados apresentam *overdispersion*?

```
>M1 <- glm (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD + SQ.DWATCOUR + D.PARK,
      family = poisson, data=RK)
>summary(M1)
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1071.44  on 51  degrees of freedom
Residual deviance:  270.23  on 42  degrees of freedom
AIC: 529.62

Number of Fisher Scoring iterations: 5
```

Veja que o resultado mostra que o parâmetro de dispersão para família Poisson tem que ser 1. Nesse caso o parâmetro de dispersão do seu modelo é 270,23/42 = 6,43. Desse modo, seu modelo apresenta *overdispersion* e você não pode continuar a análise considerando a família Poisson.

Existem duas alternativas: corrigir o Poisson com Quasi-Poisson ou usar a distribuição Binomial Negativa.

QUASI-POISSON

```
>M4 <- glm(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC+ SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD+ SQ.DWATCOUR + D.PARK,
      family = quasipoisson, data = RK)
>summary(M4)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.749e+00  3.814e-01   9.830 1.86e-12 ***
OPEN.L      -3.025e-03  3.847e-03  -0.786  0.43604
MONT.S       8.697e-02  3.309e-02   2.628  0.01194 *
SQ.POLIC    -1.787e-01  1.139e-01  -1.570  0.12400
SQ.SHRUB    -6.112e-01  2.863e-01  -2.135  0.03867 *
SQ.WATRES    2.243e-01  1.717e-01   1.306  0.19851
L.WAT.C      3.355e-01  1.005e-01   3.338  0.00177 **
SQ.LPROAD    4.517e-01  3.282e-01   1.376  0.17597
SQ.DWATCOUR  7.355e-03  1.188e-02   0.619  0.53910
D.PARK      -1.301e-04  1.445e-05  -9.004 2.33e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 5.928003)

    Null deviance: 1071.44  on 51  degrees of freedom
Residual deviance:  270.23  on 42  degrees of freedom
AIC: NA
```

Veja que o parâmetro de dispersão *f* é estimado em 5,93. Isto significa que todos os erros padrões foram multiplicados por 2,43 (a raiz quadrada de 5,93), e como resultado, a maioria dos parâmetros não são mais significativos. Não escreva na sua dissertação ou artigo que usou uma distribuição Quasi-Poisson. Quasi-Poisson não é uma distribuição. Basta dizer que você fez GLM com distribuição Poisson, detectou *overdispersion*, e corrigiu os erros padrões usando um modelo Quasi-GLM, onde a variância é dada por $f \times \mu$, onde μ é a média e *f* é o parâmetro de dispersão.

Seleção modelos em Quasi-Poisson

Quando inserirmos uma variável para a dispersão, os modelos não podem ser comparados por qui-quadrado. Eles são comparados por distribuição F.

```
>drop1(M4, test = "F")
```

```
Model:
TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB + SQ.WATRES + L.WAT.C +
    SQ.LPROAD + SQ.DWATCOUR + D.PARK
            Df Deviance F value    Pr(F)
<none>           270.23
OPEN.L       1   273.93  0.5739 0.452926
MONT.S       1   306.89  5.6970 0.021574 *
SQ.POLIC     1   285.53  2.3776 0.130585
SQ.SHRUB     1   298.31  4.3635 0.042814 *
SQ.WATRES    1   280.02  1.5217 0.224221
L.WAT.C      1   335.47 10.1389 0.002735 **
SQ.LPROAD    1   281.25  1.7129 0.197727
SQ.DWATCOUR  1   272.50  0.3526 0.555802
D.PARK       1   838.09 88.2569    7e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Repita o procedimento até que nenhuma variável possa ser retirada do modelo.

Modelo final selecionado

```
>M12 <- glm (TOT.N ~ D.PARK, family = quasipoisson, data = RK)
```

Gráfico com os dados ajustado para a curva Quasi-Poisson-Glm e intervalo de confiança de 95% (IC 95%).

```
>G <- predict (M12, newdata = RK, type = "link", se = TRUE)
>F <- exp(G$fit)
```

```
>FSEUP <- exp(G$fit + 1.96 * G$se.fit)
>FSELOW <- exp(G$fit - 1.96 * G$se.fit)
>plot(RK$D.PARK, RK$TOT.N, xlab = "Distance to park",
      ylab = "Número de anfíbios mortos")
>lines(RK$D.PARK, F, lty = 1, col = "red")
>lines(RK$D.PARK, FSEUP, lty = 2, col = "red")
>lines(RK$D.PARK, FSELOW, lty = 2, col = "red")
```

Em Quasi-Poisson não é possível calcular o valor de AIC. Por isso, é necessário calcular um valor de QUASI-AIC

```
>dd1 <- dredge (M4, rank = "QAICc", chat =
summary(M4)$dispersion)
>MQP1 <- get.models (dd1, 1:4)
model.avg(MQP1)
```

Os usuários devem ter em mente os riscos que correm usando tal "abordagem impensada" de avaliação de todos os modelos possíveis. Embora este procedimento seja útil em certos casos e justificado, ele pode resultar na escolha de um "melhor" modelo espúrio.

"Deixar o computador descobrir" é uma estratégia pobre e geralmente reflete o fato de que o pesquisador não se preocupou em pensar claramente sobre o problema de interesse e sua configuração científica (Burnham e Anderson, 2002).

Outra maneira de computar QAIC

```
>MQP <- list()
>MQP [[1]] <- glm (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC+ SQ.SHRUB
      + SQ.WATRES + L.WAT.C + SQ.LPROAD+ SQ.DWATCOUR + D.PARK,
      family = poisson, data = RK)
>MQP [[2]] <- glm (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC+ SQ.SHRUB
      + SQ.WATRES + L.WAT.C + SQ.LPROAD+ D.PARK, family =
      poisson, data = RK)
>MQP [[3]] <- glm (TOT.N ~ MONT.S + SQ.POLIC+ SQ.SHRUB +
      SQ.WATRES + L.WAT.C + SQ.LPROAD+ D.PARK, family =
      poisson, data = RK)
>MQP [[4]] <- glm (TOT.N ~ MONT.S + SQ.POLIC + SQ.SHRUB +
      L.WAT.C + SQ.LPROAD + D.PARK, family = poisson, data = RK)
```

```
>MQP [[5]] <- glm (TOT.N ~ MONT.S + SQ.POLIC+ SQ.SHRUB + L.WAT.C
      + D.PARK, family = poisson, data = RK)
>MQP [[6]] <- glm (TOT.N ~ MONT.S + SQ.POLIC+ L.WAT.C + D.PARK,
      family = poisson, data = RK)
>MQP [[7]] <- glm (TOT.N ~ MONT.S + L.WAT.C + D.PARK, family =
poisson, data = RK)
>MQP [[8]] <- glm (TOT.N ~ L.WAT.C + D.PARK, family = poisson,
data = RK)
>MQP [[9]] <- glm (TOT.N ~ D.PARK, family = poisson, data = RK)
```

Cria um vetor com nomes dos modelos:

```
>(Modnames <- paste ("MQP", 1:length(MQP), sep=""))
```

Overdispersion

```
>c_hat(MQP[[1]])
>c_hat(MQP[[2]])
>c_hat(MQP[[3]])
>c_hat(MQP[[4]])
>c_hat(MQP[[5]])
>c_hat(MQP[[6]])
>c_hat(MQP[[7]])
>c_hat(MQP[[8]])
>c_hat(MQP[[9]])
```

Gera uma tabela com valores de QAIC:

```
>(res.table <- aictab(cand.set = MQP, modnames = Modnames,
      second.ord = TRUE, c.hat = 5.92))
```

BINOMIAL NEGATIVA

odTest = Compara o log-likelihood do modelo de regressão binomial negativa com modelo de regressão Poisson.

```
>NB <- glm.nb(TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
SQ.WATRES + L.WAT.C + SQ.LPROAD + SQ.DWATCOUR + D.PARK,
link="log", data=RK)
>odTest(NB)
```

```
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  141.515 p-value = < 2.2e-16
```

O resultado mostra que a LRT entre Poisson e Binomial Negativa com uma diferença na deviance de 141.515 e com grau de liberdade 1 é $p < 0.0000$. Portanto, Binomial Negativa é melhor que Poisson.

Modelos de Binomial Negativa:

```
>NB1 <- glm.nb (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
     SQ.WATRES + L.WAT.C + SQ.LPROAD + SQ.DWATCOUR + D.PARK,
     link="log", data=RK)
>NB2 <- glm.nb (TOT.N ~ OPEN.L + MONT.S + SQ.POLIC + SQ.SHRUB +
     SQ.WATRES + L.WAT.C + SQ.LPROAD + D.PARK, link = "log",
     data = RK)
>NB3 <- glm.nb (TOT.N ~ OPEN.L + MONT.S + SQ.SHRUB + SQ.WATRES +
     L.WAT.C + SQ.LPROAD + D.PARK, link = "log", data = RK)
>NB4 <- glm.nb (TOT.N ~ OPEN.L + MONT.S + SQ.SHRUB + L.WAT.C +
     SQ.LPROAD + D.PARK, link = "log", data = RK)
>NB5 <- glm.nb (TOT.N ~ OPEN.L + MONT.S + L.WAT.C + SQ.LPROAD +
     D.PARK, link = "log", data = RK)
>NB6 <- glm.nb (TOT.N ~ OPEN.L + L.WAT.C + SQ.LPROAD +  D.PARK,
     link = "log", data = RK)
>NB7 <- glm.nb (TOT.N ~ OPEN.L + L.WAT.C + D.PARK, link = "log",
data = RK)
>NB8 <- glm.nb (TOT.N ~ OPEN.L + D.PARK, link = "log", data =
RK)
```

Seleção automática por AIC:

```
>AIC <- stepAIC(NB1)
>AIC
```

Seleção dos modelos por AIC_c:

```
>AICc <- ICtab (NB1, NB2, NB3, NB4, NB5, NB6, NB7, NB8, type =
      c("AICc"), weights = TRUE, delta = TRUE, sort = TRUE, nobs
      = 52)
>AICc
```

Likelihood Ratio Test (LRT)

```
>drop1(NB1,test="Chi")
```

Repita o procedimento até que nenhuma variável retirada apresente efeito siginificativo na comparação.

Para o modelo final, os autores justificaram a retirada de L.WAT.C porque seu valor estava muito próximo de 0.05.

Modelo Final:

```
>NB8 <- glm.nb(TOT.N ~ OPEN.L + D.PARK, link="log", data = RK)
>summary(NB8)
```

BINOMIAL NEGATIVA

```
>plot (NB8)
```

QUASI-POISSON

```
>mu <- predict (M12, type = "response")
>E <- RK$TOT.N - mu
>EP2 <- E / sqrt (7.630148 * mu)
>plot(x = mu, y = EP2, main = "Quasi-Poisson",
```

```
      ylab = "residuos",
      xlab = "predito")
abline(h = 0, v = 0)
```

Comparando os resíduos do modelo final da Binomial Negativa e Quasi-Poisson vemos que os resíduos da Binomial não apresentam um padrão, enquanto a Quasi-Poisson apresenta. Então, Binomial é melhor.

GLM BINOMIAL

Agora mostraremos um exemplo bem simples com dados de presença e ausência. GLM com dados binários ou proporção são também chamados de regressão logística.

```
>data(Boar)
>head(Boar)
```

Variável dependente: presença ou ausência de tuberculose.
Variável independente: Comprimento do javali (cabeça-tronco).

```
>B1 = glm ( Tb ~ LengthCT, family = binomial, data = Boar)
>summary(B1)
```

Likelihood Ratio Test:

```
>drop1 (B1, test="Chi")
```

```
Model:
Tb ~ LengthCT
         Df Deviance    AIC    LRT   Pr(Chi)
<none>        663.56 667.56
LengthCT  1   700.76 702.76 37.201 1.066e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Função para fazer o gráfico:

```
>MyData <- data.frame (LengthCT = seq
      (from = 46.5, to = 165, by = 1))
>Pred <- predict (B1, newdata = MyData,
      type = "response")
```

```
>Plot (x = Boar$LengthCT, y = Boar$Tb,
       xlab = "Comprimento",
      ylab = "Probabilidade de tuberculose")
>lines(MyData$LengthCT,Pred)
```

Segundo exemplo Binomial

```
>data(Tbdeer)
```

Variável dependente: proporção de infectados.
Variável independente: variáveis da paisagem.

Transforma a variável Fenced em vetor:

```
>Tbdeer$fFenced <- factor(Tbdeer$Fenced)
```

Transforma a variável dependente em proporção:

```
>Tbdeer$DeerPosProp <- Tbdeer$DeerPosCervi/
Tbdeer$DeerSampledCervi
```

Modelo Geral:

```
>Deer2 <- glm (DeerPosProp ~ OpenLand + ScrubLand +
      QuercusPlants + QuercusTrees + ReedDeerIndex + EstateSize
      + fFenced, family = binomial, weights =
      DeerSampledCervi,data = Tbdeer)
>summary(Deer2)
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.85  on 22  degrees of freedom
Residual deviance: 152.79  on 15  degrees of freedom
  (9 observations deleted due to missingness)
AIC: 227.87

Number of Fisher Scoring iterations: 4
```

Como na distribuição Poisson, quando trabalhamos com distribuição Binomial temos que verificar se existe *overdispersion* no modelo. Nesse caso, 152,79/15 = 10,18. A variância é maior que a média. Portanto, utilizamos um modelo corrigido por Quasi-Binomial.

QUASI-BINOMIAL

```
>Deer2 <- glm(DeerPosProp ~ OpenLand + ScrubLand + QuercusPlants
      + QuercusTrees + ReedDeerIndex + EstateSize + fFenced,
      family = quasibinomial, weights = DeerSampledCervi,data =
      Tbdeer)
```

Seleção do modelo por LRT

```
>drop1(Deer2,test="F")
```

Continue até que não seja permitido retirar mais nenhuma variável.

Modelo final:

```
>Deer8    <-    glm(DeerPosProp    ~    OpenLand,    family    =
      quasibinomial,weights = DeerSampledCervi,data = Tbdeer)
```

Função para fazer o gráfico:

```
>MyData <- data.frame(OpenLand = seq (from =
      min(Tbdeer$OpenLand),
      to = max(Tbdeer$OpenLand),by=0.01))
>P1 <- predict(Deer8, newdata = MyData, type = "link",
      se = TRUE)
>plot(MyData$OpenLand,exp(P1$fit)/(1+exp(P1$fit)),
     type="l",ylim=c(0,1),
     xlab="Porcentagem de área aberta",
     ylab="Probabilidade de infecção por E. cervi")
>lines(MyData$OpenLand,exp(P1$fit+1.96*P1$se.fit)/
       (1+exp(P1$fit+1.96*P1$se.fit)),lty=2)
>lines(MyData$OpenLand,exp(P1$fit-1.96*P1$se.fit)/
       (1+exp(P1$fit-1.96*P1$se.fit)),lty=2)
>points(Tbdeer$OpenLand,Tbdeer$DeerPosProp)
```

Este resultado sugere que quanto maior a porcentagem de área aberta menor a probabilidade de amostrar um veado com infecção por *E. cervi*.

Visualização dos resíduos:

```
>EP = resid(Deer8,type = "pearson")
>mu = predict(Deer8,type = "response")
>E = Tbdeer$DeerPosProp - mu
>plot(x = mu,y = EP, main="Pearson residuals")
>plot(Deer8)
```

Generalized Mixed Effects Models

São usados para modelos mais complexos com design em blocos, medidas repetidas, split plot e dados aninhados.

Aprensenta dois efeitos dentro da formúla do modelo:

EFEITO FIXO - depende somente da média – as variáveis independentes de interesse.
EFEITO ALEATÓRIO - depende somente da variância (não queremos medir o efeito, e.g. blocos).

Exemplo 1

```
>data(RIKZ)
```

Riqueza de animais marinhos bentônicos em nove praias, cada praia com cinco amostras.
NAP = altura da estação de amostral em relação ao nível da maré
PERGUNTA: Existe relação positiva entre a riqueza e a NAP?

Transforma praia em fator:

```
>RIKZ$fBeach <- factor(RIKZ$Beach)
```

Modelo

```
>Mlme1 <- lme (Richness ~ NAP, random = ~1 | fBeach, data=RIKZ)
summary (Mlme1)
```

```
Random effects:
 Formula: ~1 | fBeach
        (Intercept) Residual
StdDev:    2.944065  3.05977
```

Utilizando praia como efeito aleatório permite que cada praia tenha um intercepto diferente. Se o StdDev do efeito aleatório for zero, todos os interceptos ficam na linha predita. Veja o gráfico abaixo.

Função para fazer o gráfico:

```
>F0 <- fitted(Mlme1,level=0)
>F1 <- fitted(Mlme1,level=1)
>I <- order(RIKZ$NAP)
>NAPs <- sort(RIKZ$NAP)
>plot(NAPs,F0[I],lwd=4,type="l",ylim=c(0,22),  ylab="Riqueza  de
espécies",xlab="NAP")
for (i in 1:9){
   x1<-RIKZ$NAP[RIKZ$Beach==i]
   y1<-F1[RIKZ$Beach==i]
   K<-order(x1)
   lines(sort(x1),y1[K])
}
>text(RIKZ$NAP,RIKZ$Richness,RIKZ$Beach,cex=0.9)
```

Suponha que a relação entre riqueza de espécies e NAP é diferente em cada praia. Isto implica em que temos de incluir um interação entre NAP*Praia no modelo. Mas isso tem um custo muito alto elevando o modelo para 17 parâmetros. E não estamos interessados no efeito da praia. Contudo, não podemos ignorar uma possível variação entre praias e na interação NAP*Praias. Se fizermos isso, a variação sistemática vai aparecer nos resíduos, levando à inferências erradas. Podemos aplicar o Mixed Effects Model com intercepto e slope (inclinação) aleatórios.

```
>Mlme2 <- lme (Richness ~ NAP, random = ~ 1 + NAP | fBeach, data
= RIKZ)
>summary(Mlme2)
```

O valor 3,54 é a quantidade de variação no intercepto da população. O valor 1,71 é a variação no *slope* (inclinação) nas nove praias. A correlação mostra que praias com interceptos mais altos também tem inclinação negativa mais alta.

Veja o gráfico abaixo.

Função para fazer o gráfico:

```
>F0 <- fitted(Mlme2,level=0)
>F1 <- fitted(Mlme2,level=1)
>I <- order(RIKZ$NAP)
>NAPs <- sort(RIKZ$NAP)
>plot(NAPs,F0[I],lwd=4,type="l",ylim=c(0,22),  ylab="Riqueza  de
espécies",xlab="NAP")
for (i in 1:9){
   x1<-RIKZ$NAP[RIKZ$Beach==i]
   y1<-F1[RIKZ$Beach==i]
   K<-order(x1)
   lines(sort(x1),y1[K])
}
>text(RIKZ$NAP,RIKZ$Richness,RIKZ$Beach,cex=0.9)
```

Likelihood em Mixed Models

MAXIMUM LIKELIHOOD (ML) - escolhe os parâmetros tal que o valor de L é máximo. O problema é que ML ignora o fato que intercepto e *slope* são estimados no modelo.

RESTRICTED MAXIMUM LIKELIHOOD (REML) - corrige o grau de liberdade incluindo o intercepto e o *slope*.

Transformar algumas variáveis em fatores:

```
>RIKZ$fExp <- RIKZ$Exposure
>RIKZ$fExp[RIKZ$fExp==8]<- 10
>RIKZ$fExp <- factor(RIKZ$fExp,levels = c (10,11))
```

Modelos com ML e com REML:

```
>M0.ML <- lme (Richness ~ NAP, data = RIKZ, random = ~1| fBeach,
method = "ML")
>M0.REML <-lme (Richness ~ NAP, random = ~1|fBeach, data = RIKZ,
method = "REML")
>M1.ML <- lme (Richness ~ NAP + fExp, data = RIKZ, random = ~1|
fBeach, method = "ML")
>M1.REML <- lme (Richness ~ NAP + fExp, data = RIKZ, random =
      ~1| fBeach, method = "REML")
```

Tabela 4. Resultados para dois modelos usando ML (coluna da esquerda) e REML (coluna da direita). Números entre parênteses são erros padrões. O primeiro modelo (parte de cima da tabela) usa um intercepto e NAP como variável fixa e um intercepto aleatório. O segundo modelo (parte inferior da tabela) usa os mesmos termos, exceto que a variável nominal *exposure* é usada como uma variável fixa também.

Mixed model with NAP as fixed covariate and random intercept		
Parameter	Estimate using ML	Estimate using REML
Fixed intercept	6.58 (1.05)	6.58 (1.09)
Fixed slope NAP	–2.57 (0.49)	–2.56 (0.49)
Variance random intercept	7.50	8.66
Residual variance	9.11	9.36
AIC	249.82	247.48
BIC	257.05	254.52
Mixed model with NAP and exposure as fixed covariate and random intercept		
Fixed intercept	8.60 (0.96)	8.60 (1.05)
Fixed slope NAP	–2.60 (0.49)	–2.58 (0.48)
Fixed Exposure level	–4.53 (1.43)	–4.53 (1.57)
Variance random intercept	2.41	3.63
Residual variance	9.11	9.35
AIC	244.75	240.55
BIC	253.79	249.24

PROTOCOLO PARA *MIXED MODELS*

1 - Comece com um modelo onde o componente fixo contém todas as variáveis independentes e tantas interações possíveis.

2 - Ache a melhor estrutura para o modelo aleatório. Modelos com REML precisam ser comparados tanto para LRT como para AIC ou BIC;

3 - Depois de achar o modelo aleatório, temos que comparar os modelos fixos. Para isso temos que usar ML;

4 - Apresente o modelo final com REML;
PASSOS 1 e 2 - Selecionando efeito aleatório

```
>B1 <- gls(Richness ~ 1 + NAP * fExp, method = "REML", data =
RIKZ)
>B2 <- lme(Richness ~1 + NAP * fExp, data = RIKZ, random = ~1 |
fBeach, method = "REML")
>B3 <- lme(Richness ~ 1 + NAP * fExp,data = RIKZ, random = ~1
      NAP | fBeach, method = "REML")
```

Seleção de Modelos Aleatórios

```
AIC (B1, B2, B3)
```

ou

```
anova (B1, B2, B3)
```

PASSO 3 - Selecionando efeito fixo

```
>B2 <- lme (Richness ~ NAP * fExp, data = RIKZ, random = ~1 |
fBeach, method = "ML")
```

Fiquem atentos com valores de *P* próximos a 0,05.

```
>B3 <- lme (Richness ~ NAP + fExp, data = RIKZ, random = ~1 |
fBeach, method = "ML")
>B3a <- lme (Richness ~ NAP + fExp, data = RIKZ, random = ~1 |
fBeach, method = "ML")
>B3b <- lme (Richness ~ NAP + fExp, data = RIKZ, random = ~1 |
fBeach, method = "ML")
>AICc <- ICtab(B2, B3, B3a, B3b, type = c("AICc"), weights =
      TRUE, delta = TRUE, sort = TRUE, nobs = 45)
>AICc
```

PASSO 4 - Modelo Final com REML

```
>B2 <- lme (Richness ~ NAP + fExp, data = RIKZ, random = ~1 |
fBeach, method = "REML")
>plot(B2)
```

Exemplo Abelhas

Os dados são aninhados com múltiplas observações na mesma colméia. No total são 24 colméias com três medidas por colméia.

Mostrar comando VarIdent

```
>data(Bees)
```

Como variável dependente temos densidade de esporos medido em cada colméia. A variável independente *Infection* quantifica o grau de infecção, com valores 0, 1, 2 e 3. Embora *mixed effects modelling* podem lidar com um certo grau de dados desbalanceados, neste caso, é melhor converter a variável *Infection* em 0 (sem infecção) e 1 (infectado) porque existem poucas observações com valores 2 e 3.

Transformar a variável *Infection* em presença e ausência:

```
>Bees$Infection01 <- Bees$Infection
>Bees$Infection01[Bees$Infection01 > 0] <- 1
>Bees$fInfection01 <- factor(Bees$Infection01)
```

Transformar colméia em fator e logaritimizar esporos:

```
>Bees$fHive <- factor(Bees$Hive)
>Bees$LSpobee <- log10(Bees$Spobee + 1)
```

Plotar os dados por colméia:

```
>op <- par(mfrow = c(1, 2), mar = c(3, 4, 1, 1))
>dotchart(Bees$Spobee, groups = Bees$fHive)
>dotchart(Bees$LSpobee, groups = Bees$fHive)
```

```
>par(op)
```

Começaremos com uma regressão linear e plotaremos os resíduos por colmeia:

```
>M1 <- lm (LSpobee ~ fInfection01 * BeesN, data = Bees)
>E1 <- rstandard(M1)
>plot (E1 ~ Bees$fHive, ylab = "Resíduos", xlab = "Colméias")
>abline (0, 0)
```

Veja que algumas colméias apresentam os três resíduos acima do esperado, enquanto outras possuem três resíduos abaixo do esperado. Temos a opção de colocar colméia como *random effect*.

Vantagens

(1) requer um parâmetro extra (variância do intercepto), comparado com regressão linear que requer 23 parâmetros extras.
(2) Podemos fazer afirmações para colméias em geral não só para as 24 colméias do estudo.

Selecionando *random effect*

```
>M1 <- lme(LSpobee ~ fInfection01 * BeesN, random = ~ 1 | fHive,
     method = "REML", data = Bees)
>M2 <- lme(LSpobee ~ fInfection01 * BeesN, random = ~ 1 + BeesN
     | fHive, method = "REML", data = Bees)
>M3 <- lme (LSpobee ~ fInfection01 * BeesN, random = ~ 1 +
     fInfection01 | fHive, method = "REML", `  data = Bees)
>anova(M1,M2)
>anova(M1,M3)
```

Verificando o modelo selecionado:

```
>plot (M1, col = 1)
```

plota por infecção:

```
>boxplot (LSpobee ~ fInfection01, data = Bees, varwidth = TRUE)
```

Veja que há diferença na variação entre as categorias.

Inserimos um comando para dizer que as variâncias para infecção são diferentes.

varIdent = permite modelar diferentes variâncias para variáveis categóricas.

```
>M1 <- lme (LSpobee ~ fInfection01 * BeesN, random = ~ 1 |
     fHive, method = "REML", data = Bees)
>M4 <- lme (LSpobee ~ fInfection01 * BeesN, random = ~ 1 |
     fHive, method = "REML", data = Bees, weights = varIdent
     (form = ~ 1 | fInfection01))
>anova (M1,M4)
```

Selecionando estrutura fixa:

```
>M7full<- lme (LSpobee ~ fInfection01 * BeesN, random = ~
     1|fHive, weights = varIdent(form = ~ 1 | fInfection01),
     method = "ML", data = Bees)
>M7sub <- update(M7full, .~. -fInfection01 : BeesN )
>anova (M7full,M7sub)
>M8full <- lme (LSpobee ~ fInfection01 + BeesN, random = ~
     1|fHive, method = "ML", data = Bees, weights =
     varIdent(form =~ 1 | fInfection01))
>M8sub1 <- update (M8full, .~. -fInfection01 )
>M8sub2 <- update (M8full, .~. -BeesN )
>anova(M8full,M8sub1)
>anova(M8full,M8sub2)
>M9full<-lme(LSpobee ~ fInfection01, random = ~ 1|fHive,
     method="ML", data = Bees, weights = varIdent(form =~ 1 |
     fInfection01))
>M9sub1<-update(M9full, .~. -fInfection01 )
>anova(M9full,M9sub1)
```

Modelo final:

```
>Mfinal <- lme (LSpobee ~ fInfection01, random =~ 1|fHive, data
= Bees, weights = varIdent (form = ~ 1 | fInfection01),
method = "REML")
```

```
>plot(Mfinal)
```

Dados categóricos:

```
>data(ergoStool)
```

Esforço requerido por quatro diferentes mandíbulas para rasgar nove objetos diferentes.

```
>fm1Stool <- lme (effort ~ Type, data = ergoStool, random = ~ 1
| Subject)
>summary(fm1Stool)
```

Tentar exe,plicar os valores:

```
> (mean <- tapply(ergoStool$effort, ergoStool$Type, mean))
```

O primeiro parâmetro (intercepto) é a média da primeira categoria definida por ordem alfabética. Portanto, sempre que for comparar categorias, o intercepto será a categoria que começar com a menor letra do alfabeto.

O segundo parâmetro é a diferença entre o segundo parâmetro e o intercepto:

12.44 - 8.55 = 3.89

O terceiro parâmetro é a diferença entre o terceiro parâmetro e o intercepto:

10.77 - 8.55 = 2.22
9.22 - 8.55 = 0.66

As comparações podem ser alteradas de acordo com suas hipóteses. Comparações planejadas:

```
>contrasts(ergoStool$Type)<-cbind(c(3,-1,-1,-1),
c(0,2,-1,-1), c(0,0,-1,1))
>fm2Stool <- lme (effort ~ Type, data = ergoStool, random = ~ 1
| Subject)
>summary(fm2Stool)
```

Veja que o efeito totak de fm1Stool não muda quando alteramos os contrastes:

```
>anova(fm1Stool)
>anova(fm2Stool)
```

EXERCÍCIOS

EXERCÍCIO 1 – Carreguem os dados das corujas como demonstrado abaixo:

```
>library (AED)## O pacote AED tem que ser baixado da página
              ## http://www.highstat.com/book2.htm
>data(Owls)
```

Variável dependente = número de piados dos filhotes na ausência dos pais - *NegPerChick* - (Transforme em log essa variável).

Variáveis independentes = variáveis fixas [sexo dos pais, tratamento da alimentação (saciado e privado), hora de chegada dos pais] e variável aleatória (ninho)

Unidade amostral = ninho

Teoria: Ecologia Comportamental

Responda: Quais variáveis melhor explicam o comportamento de negociação dos filhotes de coruja?

EXERCÍCIO 2 – Carregue a planilha predador.csv

Variável dependente = presença ou ausência de predadores (larvas de odonata) em poças d'água com diferentes tamanhos onde foram amostrados girinos de *Pseudopaludicola falcipes.*

Variáveis independentes = tamanho das poças d'água

Unidade amostral = poça d'água

Teoria: Predação, Forrageio Ótimo

Responda: A probabilidade da presença de predadores está relacionada com o tamanho das poças d'água?

EXERCÍCIO 3 - Carregue os dados da planilha Solea.csv

Variável dependente = presença ou ausência do peixe *Solea solea* num estuário em Portugal.

Variáveis independentes = 11 variáveis preditoras

Unidade amostral = cada área de coleta ou ponto de coleta no estuário

Teoria: Ecologia de Paisagem

Responda: Quais variáveis melhor explicam a presença de *Solea solea* nos berçários de Portugal?

CURVA DE ACUMULAÇÃO DE ESPÉCIES

Curvas de acumulação de espécies, algumas vezes chamadas de **curva do coletor**, são representações gráficas que demonstram o número acumulado de espécies registradas (S) em função do esforço amostral (*n*). O esforço amostral pode ser o número de indivíduos coletados, ou uma medida tal como o número de amostras (e.g., quadrados) ou tempo amostral (e.g., meses). Colwell & Coddington (1994) sugeriram um método que consiste em montar várias curvas adicionando-se as amostras em uma ordem aleatória. Após construir várias curvas com este método, pode-se calcular uma curva do coletor média (baseada na riqueza média para cada número de amostra) e expressar a variação possível em torno dessa média. É importante frisar que esta variação não corresponde ao conceito estatístico de intervalo de confiança, já que é calculada por repetições das mesmas unidades amostrais (Santos 2003). Se as curvas de acumulação de espécies atingem um ponto em que o aumento do esforço de coleta não implica num aumento no número de espécies, isto significa que aproximadamente toda a riqueza da área foi amostrada (Fig. 8).

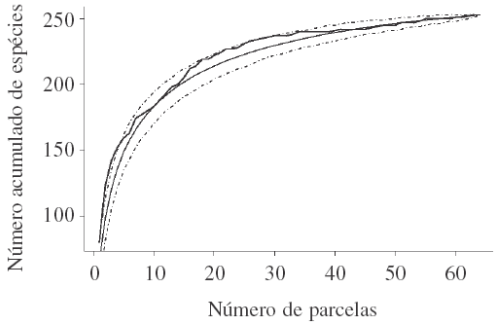


Figura 8. Exemplo de uma curva de acumulação de espécies.

RAREFAÇÃO

Esse método nos permite comparar o número de espécies entre comunidades quando o tamanho da amostra ou o número de indivíduos (abundância) 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 menor amostra ou com menos indivíduos (dentre todas amostras possíveis). Se considerarmos *n* indivíduos ($n < N$) para cada comunidade, quantas espécies iríamos registrar?

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

Onde:

E(S) = Número de espécies esperado
N = Número total de indivíduos na amostra
N_i = 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: i) as amostras a serem comparados devem ser consistentes do ponto de vista taxonômico, ou seja, todos os indivíduos devem pertencer ao mesmo grupo taxonômico; ii) as comparações devem ser realizadas somente entre amostras com as mesmas técnicas de coleta; iii) os tipos de hábitat onde as amostras são obtidas devem ser semelhantes; e iv) é um método para estimar a riqueza de espécies em uma amostra menor – **não pode ser usado para extrapolar e estimar riqueza**.

Exemplo:

Uma amostra de roedores tem quatro espécies e 42 indivíduos. A abundância de cada espécie foi 21, 16, 3, e 2 indivíduos. Desejamos calcular a riqueza de espécies esperada para amostras com 30 indivíduos.

$$E(s) = \left(1 - \frac{(42-21)/30}{42/30}\right) + \left(1 - \frac{(42-16)/30}{42/30}\right) + \left(1 - \frac{(42-3)/30}{42/30}\right) + \left(1 - \frac{(42-2)/30}{42/30}\right)$$

E(30) = 1 + 1 +0.981 + 0.923

E(30) = 3.9 espécies

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

Primeiramente carregue o pacote *vegan*:

```
>library(vegan)
```

O comando geral para realizar a análise de rarefação é:

```
>rarefy(x, sample, se = FALSE, MARG = 1)
```

Onde:

x = comunidade para a qual se deseja estimar a riqueza de espécies
sample = tamanho da sub-amostra (*n*)
se = desvio padrão
MARG = maneiras de visualizar o resultado – Utilizar número 2

Imagine que você tenha uma planilha aberta no R com o nome *rare.* Nesta planilha, existem três colunas referentes à três comunidades de roedores, e em cada linha a abundância de cada espécie (exemplo abaixo):

```
rare
roedore roedore roedore
s       s1      s2
21      16      10
16      15      10
3       13      10
2       31      10
0       1       10
0       1       10
0       1       10
0       1       0
```

Para obter-se o mesmo resultado do exercício anterior sem ter que realizar os cálculos manualmente, você precisa digitar o seguinte comando:

```
>rarefy(rare$roedores, sample = 30, MARG = 2)
>3.9
```

Para calcular a rarefação para diferentes valores de sub-amostras é precisa criar um comando com diversos tamanhos de amostras:

```
>amostras1 <-  c(seq(5, 40, by = 1))
>amostras2 <-  c(seq(5, 80, by = 1))
>amostras3 <- c(seq(5, 70, by = 1))
```

Rarefação para as três comunidades com vários valores de sub-amostras:

```
>roedor1 <- rarefy(rare$roedores, sample = amostras1, se = T,
MARGIN = 2)
>roedor2 <- rarefy(rare$roedores1, sample = amostras2, se = T,
MARGIN = 2)
>roedor3 <- rarefy(rare$roedores2, sample = amostras3, se = T,
MARGIN = 2)
```

Gráfico de rarefação para as três comunidades

```
>plot (amostras2, roedor2[1,], ylab = "Riqueza de espécies",xlab
= "No. de Individuos",ylim = c(1, 9), xlim = c(1,90), type= "n")
>text(30, 9, "Rarefação comunidade de roedores")
>lines (amostras1, roedor1[1, ], type = "b", col = "red", lwd =
1.7)
>lines (amostras2 + 0.2, roedor2[1, ], type = "b", col = "blue",
lwd = 1.7)
>lines (amostras3 + 0.4, roedor3[1, ], type = "b", col =
"black", lwd = 1.7)
>labs <- c ("Comunidade 1","Comunidade 2", "Comunidade 3")
>legend (locator(1), labs, lty = c(1,2,3), col = c("red",
"blue", "black") ,bty = "n")
>abline (h = 0, v = 40, col = "yellow")
```

ESTIMADORES DE RIQUEZA

Uma vez que determinar a riqueza total de espécies numa área é praticamente impossível, principalmente em regiões com alta diversidade 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). Nesta apostila serão considerados apenas os estimadores não paramétricos (que não são baseados nos parâmetros de um modelo de abundância das espécies), para outros estimadores veja Magurran (2004).

Chazdon et al. (1998) e Horter et al. (2006) definem quatro características para um bom estimador de riqueza:

i) Independência do tamanho da amostra (quantidade de esforço amostral realizado);

ii) Insensibilidade a diferentes padrões de distribuições (diferentes equitabilidades);

iii) Insensibilidade em relação à ordem das amostragens;

iv) Insensibilidade à heterogeneidade entre as amostras usadas entre estudos.

Tabela 5. Número de indivíduos registrados de cada espécie de anuros em 14 amostras no noroeste de São Paulo, Brasil. Será utilizado nos exemplos abaixo.

Espécies	AMOSTRAS														Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Delian	0	0	6	15	2	2	0	0	0	1	0	5	5	2	**38**
Dmelan	0	0	0	0	1	0	0	1	0	0	0	0	0	0	**2**
Dminu	0	2	1	15	8	2	0	1	2	2	0	4	0	2	**39**
Dnanu	4	0	3	15	2	2	0	7	0	2	0	3	2	2	**42**
Dmulle	0	0	0	3	12	0	2	0	0	0	0	0	0	0	**17**
Ebic	0	0	0	0	1	0	0	1	0	0	0	0	1	0	**3**
Esp	0	0	2	0	0	0	0	0	0	1	0	0	0	0	**3**
Enat	0	4	1	0	17	0	2	0	1	0	4	0	0	1	**30**
Halb	5	0	0	0	0	1	0	9	0	1	0	0	4	0	**20**
Hfab	0	0	0	0	0	0	0	4	0	0	0	0	0	0	**4**
Hran	14	0	0	5	0	1	0	0	0	2	0	0	8	0	**30**
Lchaq	0	0	0	0	11	0	3	0	0	0	0	0	0	0	**14**
Lfus	8	3	2	5	4	2	1	6	1	3	1	2	3	6	**47**
Llab	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**1**
Riqueza Total	**4**	**3**	**6**	**6**	**9**	**6**	**4**	**7**	**3**	**8**	**2**	**4**	**6**	**5**	

CHAO 1

Estimador simples do número absoluto de espécies em uma comunidade. É baseado no número de espécies raras dentro de uma amostra. Esse método **requer a abundância** das espécies.

$$Chao_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

onde:

S_{obs} = o número de espécies na comunidade

F_1 = número de espécies observadas com abundância de um indivíduo (espécies *singleton*)

F_2 = 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.

Exemplo:

Usando os dados da tabela 1 calcule o valor de Chao 1 para a comunidade:

Chao 1 = 14 + [(1^2)/(2*1)] = 14 + (1/2) = 14 + 0,5

Chao 1 = 14,5

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

Carregue os pacotes Vegan e BiodiversityR

```
>library(vegan)
>library(BiodiversityR)
```

Imagine que você tenha a mesma tabela acima salva no R com o nome "*est*". Após carregar essa tabela você pode obter o valor de Chao 1 através do seguinte comando:

```
>est <- read.table (estimadores, h = T)
>Chao1 <-estaccumR (est, permutations = 100)
>summary(Chao1, display = "chao")
```

Outra maneira de conseguir o mesmo valor:

```
>est1 <- colSums(est)## soma abundância de cada linha = abundância total por espécie
>Chao1 <- estimateR (est1)
>Chao1
```

CHAO 2

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 somente conhecer o número de espécies encontradas em somente uma amostra e o número de espécies encontradas exatamente em duas amostras. Essa variação ficou denominada Chao 2:

$$Chao_2 = S_{obs} + \frac{L^2}{2M}$$

onde:
L = número de espécies que ocorrem apenas em uma amostra (espécies *uniques*)
M = número de espécies que ocorrem em exatamente duas amostras (espécies *duplicates*)

O valor de Chao 2 é máximo quando todas as espécies menos uma são únicas (*singletons*). Neste caso, a riqueza estimada é aproximadamente o dobro da riqueza observada.

Collwel & Coddington (1994) encontraram que o valor de Chao 2 mostrou ser o estimador menos enviesado para amostras com tamanho pequeno.

Exemplo:

Usando os dados da tabela 1 calcule o valor de Chao 2 para a comunidade:

Chao 2 = 14 + [(2^2)/(2*3)] = 14 + (4/6) = 14 + 0.66
Chao 2 = 14.66

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

A função `poolaccum` do pacote *vegan* apresenta resultados mais completos com valores de riqueza de espécie estimado para cada amostra

```
>est <- read.table (estimadores, h = T)
>Chao2 <- poolaccum (est, permutations = 100)
>summary(Chao2, display = "chao")
```

Os comandos `specpool` e `diversityresult` são mais simples e diretos, pois apresentam somente o valor final estimado:

```
>Chao2 <- specpool(est)
>Chao2
>Chao2 <- diversityresult(est, index = "chao")
```

JACKKNIFE 1

Este estimador baseia-se no número de espécies que ocorrem em somente uma amostra (Q_1).

$$Jack_1 = S_{obs} + Q_1 \left(\frac{m-1}{m}\right)$$

Onde:

m = número de amostras

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

Exemplo:

Usando os dados da tabela 1 calcule o valor de Jaccknife 1 para a comunidade:

Jack 1 = 14 + 2 * [(14-1)/14] = 14 + 2 * (0.92) = 14 + 1.857

Jack 1 = 15.857

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

```
>est <- read.table(estimadores, h = T)
>Jackk1 <- poolaccum(est, permutations = 100)
>summary(Jackk1, display = "jack1")
```

Outra maneira de conseguir o mesmo valor:

```
>Jackk1 <- specpool(est)
>Jackk1
>Jackk1 <- diversityresult(est, index = "jack1")
```

JACKKNIFE 2

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.

$$Jack_2 = S_{obs} + \left[\frac{Q_1 (2m-3)}{m} - \frac{Q_2 (m-2)^2}{m(m-1)} \right]$$

Onde:

Q_1 = número de espécies registradas em apenas uma amostra

Q_2 = número de espécies registradas em exatamente duas amotras

m = número de amostras

Exemplo:

Usando os dados da tabela 1 calcule o valor de Jaccknife 2 para a comunidade:

Jack 2 = 14 + [2 *(((2*14)-3))/14))] – [3*((14-2)^2)/(14(14-1))] = 14 + 3,57 – 2,37

Jack 2 = 15.197

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

```
>est <-read.table(estimadores, h = T)
>Jackk2 <- poolaccum(est, permutations = 100)
>summary(Jackk2, display = "jack2")
```

Outra maneira de conseguir o mesmo valor:

```
>Jackk2 <- specpool(est)
>Jackk2
>Jackk 2 <- diversityresult(est, index = "jack2")
```

ACE (Abundance-based Coverage Estimator)

Este método trabalha com a abundância das espécies raras (abundância baixa). Entretanto, diferente dos estimadores anteriores, 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 (Santos 2003).

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

Onde:

$$\gamma^2_{ace} = max\left[\frac{S_{rare}}{C_{ace}}\frac{\sum_{i=i}^{10} i(i-1)F_1}{(N_{rare})(N_{rare}-1)} - 1\right]$$

$$C_{ace} = 1 + \frac{F_1}{N_{rare}}$$

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

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

REALIZANDO O EXERCÍCIO NO PROGRAMA R:

Comandos

```
>est <- read.table("estimadores.txt", h = T)
>ACE <- estaccumR(est, permutations = 100)
>summary(ACE, display = "ace")
```

Outra maneira de conseguir o mesmo valor:

```
>est1<-colSums(est) ## soma abundância de cada linha= abundância
                total por espécie
>ACE <- estimateR(est1)
>ACE
```

ICE (Incidence-based Coverage Estimator)

Este método trabalha com o número de espécies infreqüentes (que ocorrem em poucas unidades amostrais). Esse método permite ao pesquisador determinar os limites para os quais uma espécie seja considerada infreqüente. Em geral, são consideradas como tal espécies com incidência entre 1 e 10 indivíduos (Chazdon et al. 1998) ou 1 a 20 (Walther & Morand 1998). A riqueza estimada pode variar conforme se aumente ou diminua o limiar de incidência, e

infelizmente não existem critérios biológicos definidos para a escolha do melhor intervalo (Santos 2003).

$$ICE = S_{freq} + \frac{S_{\mathrm{inf}r}}{C_{ice}} + \frac{Q_1}{C_{ice}}\,\gamma_{ice}^2$$

onde:

$$\gamma_{ice}^2 = max\left[\frac{S_{\mathrm{inf}r}}{C_{ice}}\frac{m_{\mathrm{inf}r}}{(m_{\mathrm{inf}r-1})}\frac{\sum_{j=1}^{10} j(j-1)Q_1}{(N_{\mathrm{inf}r})^2} - 1\right]$$

$$C_{ice} = 1 + \frac{Q_1}{N_{\mathrm{inf}r}}$$

$$N_{inf\,r} = \sum_{j=1}^{10} jQ_j$$

REALIZANDO O EXERCÍCIO NO PROGRAMA R:

Comandos

```
>est <-read.table(estimadores, h = T)
>ICE <- poolaccum(est, permutations = 100)
>summary(ICE, display = "ice")
```

Outra maneira de conseguir o mesmo valor:

```
>ICE <- specpool(est)
>ICE
```

BOOTSTRAP

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.

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

Onde:

P_k = proporção do número de amostras em que cada espécie foi registrada
m = número de amostras

Exemplo:

Usando os dados da tabela 1 calcule o valor de bootstrap para a comunidade:

Bootstrap = 14 + [$(1- 8/14)^{14}$ +$(1- 2/14)^{14}$ +$(1- 10/14)^{14}$ +$(1- 10/14)^{14}$ +$(1- 3/14)^{14}$ +$(1- 3/14)^{14}$ +$(1- 2/14)^{14}$

+ $(1- 7/14)^{14}$ +$(1- 5/14)^{14}$ +$(1- 1/14)^{14}$ +$(1- 5/14)^{14}$ +$(1- 2/14)^{14}$ +$(1- 14/14)^{14}$ +$(1- 1/14)^{14}$]

Bootstrap = 14 + 1 ,127

Boostrap = 15,127

REALIZANDO O MESMO EXERCÍCIO NO PROGRAMA R:

Comandos

```
>est <-read.table(estimadores, h = T)
>BOOT <- poolaccum(est, permutations = 100)
>summary(BOOT, display = "boot")
```

Outra maneira de conseguir o mesmo valor:

```
>BOOT <- specpool (est)
>BOOT
>BOOT <- diversityresult (est, index = "boot")
```

EXERCÍCIOS

1) Utilize os dados da planilha rarefação – exercicios.csv que foi entregue no cd junto com a apostila.

a) Calcule a abundância total em cada uma das comunidades
b) Calcule a riqueza total em cada comunidade
c) Construa uma gráfico de rarefação comparando as quatro comunidades

2) Para esse exercício usaremos os dados disponíveis na pagina do Prof. Dr. Adriano Melo da Universidade Federal de Goiás.

Para carregar os dados vocês precisam digitar o comando abaixo:

```
japi <-
read.table('http://www.ecologia.ufrgs.br/~adrimelo/div/japi.txt'
, h=T)
```

a) Faça um gráfico com a curva do coletor e acumulação (rarefação) de espécies/amostra juntos no mesmo gráfico.

3) Utilizando a planilha "est.csv"

a) Faça um gráfico com o estimador de riqueza *bootstrap* e a riqueza observada

b) Faça um gráfico com o estimador de riqueza chao1 e a riqueza observada

c) Faça um gráfico com os estimadores jackknife 1 e 2 e a riqueza observada

ESTIMATES

O programa R tem grandes vantagens sobre outros programas estatísticos, por permitir realizar diversos tipos de análises, plotar gráficos, e alterar funções de acordo com suas necessidades (leia o início dessa apostila). No entanto, existe um programa gratuito, disponível na internet no endereço http://viceroy.eeb.uconn.edu/estimates voltado à analises com 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.

Aqui mostramos rapidamente como realizar as análises nesse programa.

1 – A planilha que você utilizará deve ser montada da seguinte maneira no Excel. A1 = nome da planilha; A2 = Número de espécies; B2 = Número de amostras. **NÃO coloque o nome das espécies.**

	A	B	C	D	E	F	G	H	I	J	K	L
1	"anuros"											
2	10	12										
3	2	2	2	2	2	2	2	2	2	2	2	2
4	2	2	2	2	2	2	2	2	2	2	2	2
5	2	2	2	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2	2	2
8	2	2	2	2	2	2	2	2	2	2	2	2
9	2	2	2	2	2	2	2	2	2	2	2	2
10	2	2	2	2	2	2	2	2	2	2	2	2
11	2	2	2	2	2	2	2	2	2	2	2	2
12	2	2	2	2	2	2	2	2	2	2	2	2

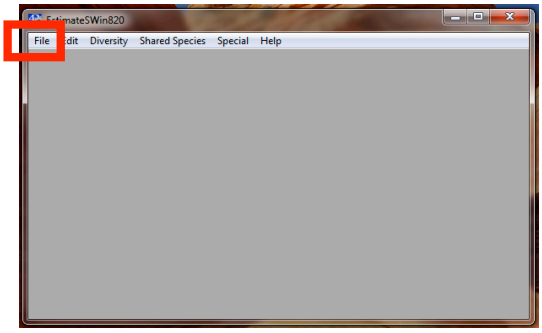
2 – Salve a planilha no formato **.txt – Texto separado por tabulação;**

3 – Depois de salvar a planilha no formato **Texto separado por tabulação**, abrir o programa Estimates;

4 - A tela abaixo deve aparecer;

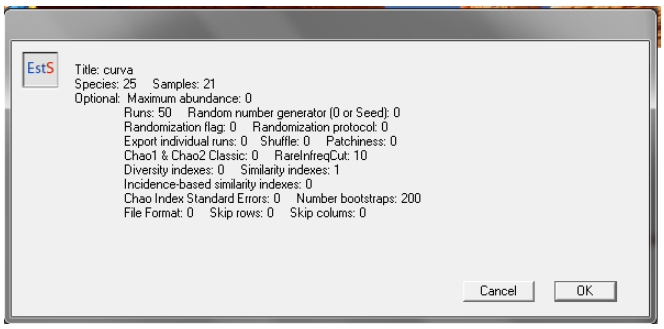
5 - Selecionar **FILE;**

6 – Selecionar a opção "**LOAD DATA INPUT FILE**" para carregar a planilha. Procure onde ela foi salva no seu computador;



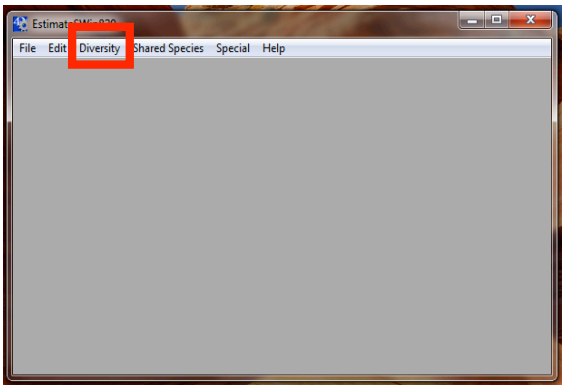
7 – Se o programa carregar a planilha corretamente, aparecerá a tela abaixo;

8 – Veja o número de espécies (Species) e amostras (Samples). Se estiver correto, clicar em **OK** nas telas que aparecerão;

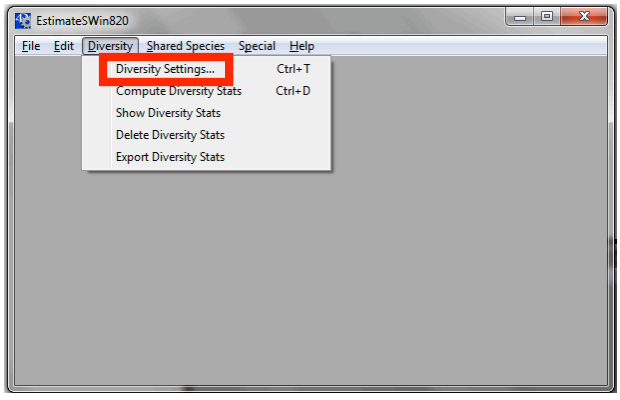


9 – Agora é necessário configurar o programa para realizar os testes;

10 – Clicar em "**DIVERSITY**", como demonstrado na tela abaixo;

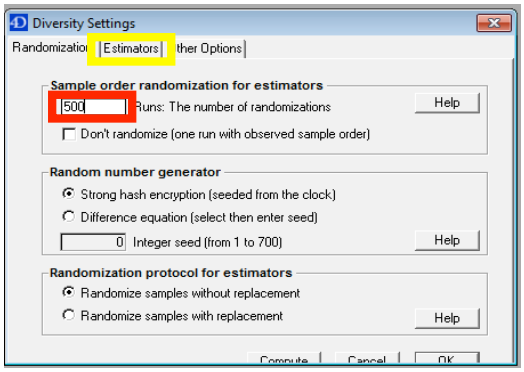


11 – Escolham a opção "**Diversity Settings**"



12 – Coloque **500** no lugar de **50** aleatorizações

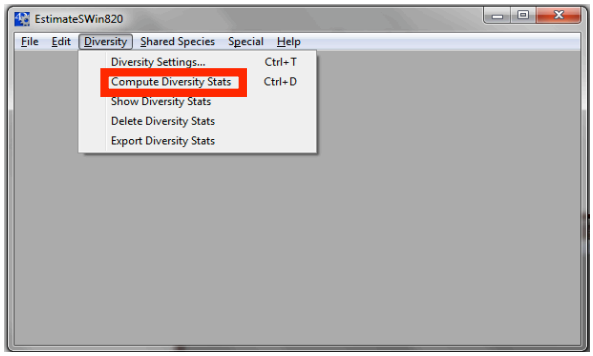
13 – Depois de colocarem 500 cliquem na **aba Estimators** (destacado em amarelo) e depois em **OK**;

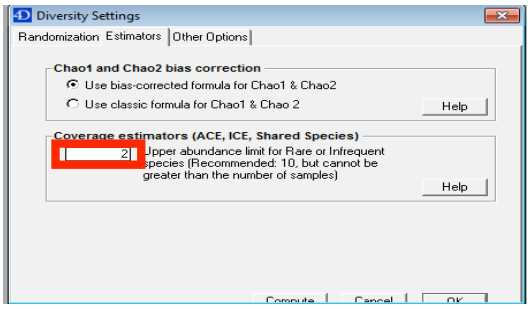


14 - Determine o número de espécies raras para o ACE e ICE. Esse número corresponde ao número de espécies que o programa irá considerar como espécies raras;

15 – Clicar em **OK;**

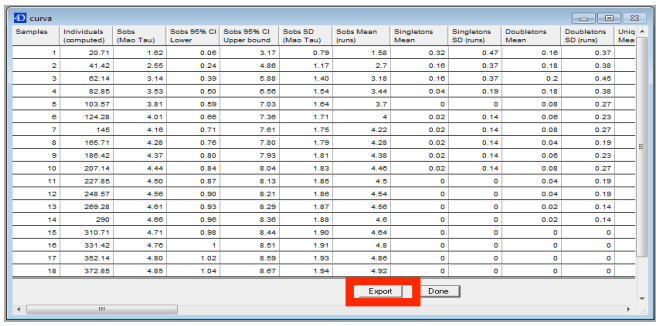
16 - Agora é só correr o teste. Clicar em **Compute Diversity Stats**;





17 – Aparecerá uma tela com os resultados do teste;

18 – Clicar em **Export** e salvar em algum lugar no seu computador, depois é só abrir com o Excel e fazer os gráficos no R;



ÍNDICES DE DIVERSIDADE E DIVERSIDADE BETA (β)

Índices de diversidade

Os índices de diversidade representam uma medida que combina a riqueza e abundância relativa (equitabilidade) das espécies de uma comunidade. O índice de Shannon (*H'*) é um dos mais utilizados na literatura para medir a diversidade de espécies. Este índice é derivado da teoria da informação e sua função foi derivada como:

$$H' = -\sum p_i \ln p_i$$

Onde p_i representa a proporção de indivíduos na *i*-nésima espécie em relação à abundância total na comunidade. Quanto maior o valor de *H'*, maior a diversidade da comunidade. Os valores de *H'* raramente ultrapassam 4, sendo que para que *H'* seja maior do que 5 a comunidade precisa ter mais de 10^5 espécies. Um dos problemas do índice de Shannon é que a diversidade é confundida pela riqueza de espécies e equitabilidade. Desse modo, tanto o número de espécies quanto o esforço amostral afetam o valor final do índice. Além disso, quando confrontamos valores de diversidade entre duas comunidades, por exemplo, *H'* = 2,71 e *H'* = 2,59, temos dificuldade para decidir se os valores são, de fato, diferentes.

Outro índice de diversidade muito usado por ecólogos é o índice de Simpson (*D*). Este índice mede a probabilidade de dois indivíduos coletados ao acaso pertencerem à mesma espécie através da fórmula:

$$D = \sum p_i^2$$

Onde p_i representa a proporção de indivíduos na *i*-nésima espécie em relação às abundância total na comunidade. Quanto maior o valor de *D*, menor a diversidade da comunidade. Alguns autores expressam a fórmula do índice de Simpson como 1 – *D* ou 1 / *D*. Este índice é considerado uma das medidas de diversidade mais robustas.

Apesar de existir um número impressionante de métricas para medir a diversidade biológica (Hulbert 1972, Magurran 2004), diversos autores desencorajam o uso dessas métricas para testar hipóteses ecológicas. Dentre os principais motivos destacamos: (1) ausência de uma base probabilística que nos permita assinalar valores de significância que, por sua vez, impede que façamos comparações biológicas entre duas comunidades; (2) todos os índices de diversidade são fortemente sensíveis ao número de indivíduos e de espécies; (3) problemas conceituais e de múltiplas definições que trazem pouco sentido biológico e dificultam a interpretação de padrões ecológicos. Dentre os autores que criticam a utilização de índices de diversidade na ecologia, se destacam pela clareza dos argumentos o trabalho marcante de

Hulbert (1971) e Gotelli & Graves (1996). Resumindo as idéias, a indefinição conceitual e técnica dos índices de diversidade sugerem que sua utilização seja abandonada (ou que sejam utilizados com rigor tremendo). Há quem se refira à "diversidade de espécies" como um "não-conceito" (Hulbert 1971). Como alternativa elegante, a utilização da riqueza de espécies e da abundância relativa como "métricas" distintas para medir a diversidade, bem como suas respostas às alterações ambientais, pode ser o melhor caminho para o desenvolvimento de bons estudos ecológicos.

Calculando os índices de diversidade no R

```
>library(vegan)

>mata.atlantica=read.table("mata.atlantica.txt", header=T)

>H=diversity(mata.atlantica, index="shannon")

>D=diversity(mata.atlantica, index="simpson")

>D.inv=diversity(mata.atlantica, index="invsimpson")

>riqueza=specnumber(mata.atlantica)

>diversidade.MA=cbind(riqueza, H, D, D.inv)

>diversidade.MA

>pairs(cbind(riqueza, H, D, D.inv), pch="+", col="black")
```

Praticando:

Exemplo 1: Bromélias geralmente acumulam água no fitotelmata e diversos grupos de artrópodes utilizam esses tanques para depositar ovos. Desse modo, as larvas aquáticas desses animais vivem imersas até atingirem a fase adulta. Uma bióloga coletou larvas em quatro espécies de bromélias-tanque (n=30 plantas de cada espécie) e dividiu cada bromélia em três grupos de tamanho: pequena (<100 ml de água acumulada; n=10/espécie), média (101 – 600 ml de água acumulada; n=10/espécie) e grande (> 601 ml de água acumulada; n=10/espécie). Utilize os arquivos "bromelias.txt" e "bromelia1.txt".

Pergunta 1: Qual espécie de bromélia possui maior diversidade de artrópodes aquáticos?

Pergunta 2: O volume de água afeta a diversidade de espécies de artrópodes aquáticos na Bromélia sp.1?

- Teoria: teoria da biogeografia de ilhas (volume de hábitat).

- Unidade amostral: bromélia

- Variável dependente: diversidade medida por algum índice de diversidade

- Variável independente: espécie de bromélia, volume (categorias pequena, média e grande)

Responda: Qual a espécie de bromélia com maior diversidade? O volume de água acumulada no fitotelmata aumenta a diversidade de artrópodes na Bromélia sp.1? Utilize as funções do R que aprendeu e calcule o índice de Shannon e Simpson.

Curvas de dominância ou Padrão de Distribuição da Abundância das Espécies (SADs)

Uma alternativa mais interessante para investigar concomitantemente a riqueza e a equitabilidade das espécies numa comunidade é a construção de curvas de dominância, conhecida na literatura ecológica por "Species Abundance Distributions" (SADs), curvas de dominância ou diagramas de abundância relativa. Essas curvas descrevem a abundância das espécies encontradas na comunidade (McGill et al. 2007). A maioria das comunidades é dominada por poucas espécies, um padrão conhecido como na literatura como "J" invertido. Uma maneira comum de representar graficamente as curvas de dominância é organizar as espécies em ordem decrescente de abundância no eixo *x* (i.e., da espécie mais abundante para a menos abundante) e o log da abundância de cada espécie no eixo *y* (Fig. 9a).

A representação desses diagramas evidencia as diferenças no padrão de equibilidade entre diferentes comunidades. Após o trabalho de Whittaker (1965), a utilização de diagramas de abundância relativa ganhou força, especialmente para ilustrar as modificações na flora ou na fauna durante a sucessão ecológica ou após um impacto ambiental. A informação mais básica que pode ser retirada dos diagramas está na inclinação das curvas; quanto maior a inclinação, maior a dominância da comunidade estudada (Fig. 9b). Além disso, quanto mais longa a curva, maior a riqueza de espécies da comunidade. Diversos trabalhos propuseram modelos teóricos para explicar os padrões de distribuição da abundância das espécies (Tokeshi 1999, Hubbel 2001, Magurran 2004, McGill et al. 2007). Alguns deles têm origem puramente estatística, como o modelo Log-normal, enquanto outros foram criados a partir de um arcabouço teórico (biológico) explícito, como os modelos *Broken-Stick* (nomeado "null" no pacote radfit do R), série geométrica ("preemption" no R), Zipf e Zipf-Mandelbrot.

A abundância esperada (LNa_r) segundo o modelo estatístico Log-normal para a espécie da ordem *r* é:

$$LNa_r = \exp(\log \mu + log\sigma N)$$

Onde *N* representa o desvio Normal e μ e σ são os coeficientes da fórmula. A abundância esperada (BSa_r) para a espécie na ordem (do inglês "rank") *r* para o modelo Broken-Stick é:

$$BSa_r = (J/S)\sum_{x=r}^{S}(1/x)$$

Onde *J* representa o número total de indivíduos na comunidade e *S* o número total de espécies. Para o modelo Série Geométrica, a abundância esperada (GSa_r) para a espécie da ordem *r* é:

$$GSa_r = J\alpha(1-\alpha)^{r-1}$$

Onde *J* representa o número total de indivíduos na comunidade e o coeficiente α é uma estimativa da taxa de decréscimo da abundância por ordem *r*. Para o modelo Zipf, a abundância esperada (Za_r) para a espécie da ordem *r* é:

$$Za_r = Jp_1 r^{\gamma}$$

Onde *J* representa o número total de indivíduos na comunidade, p_1 é a proporção ajustada da espécie mais abundante e γ é o coeficiente de decréscimo da abundância por ordem *r*. O modelo Zipf-Mandelbrot acrescenta um parâmetro na fórmula do Zipf para estimar a abundância (ZMa_r) da espécie da ordem *r*:

$$ZMa_r = Jc(r+\beta)^{\gamma}$$

Onde *J* representa o número total de indivíduos na comunidade, *c* e β são constantes de escala e γ é o coeficiente de decréscimo da abundância por ordem *r* (Wilson 1991).

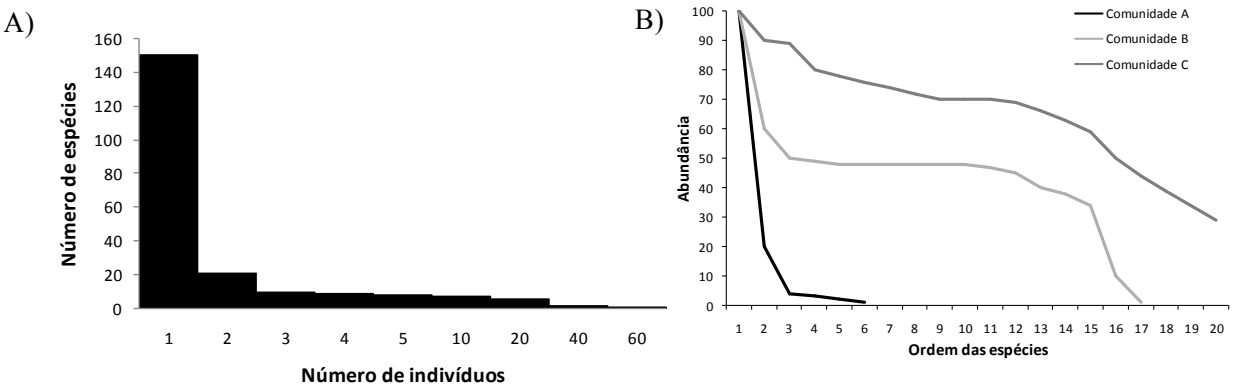


Figura 9. Duas representações comuns do padrão de distribuição da abundância das espécies. (A) Representação básica com o número de espécies com suas respectivas abundâncias organizadas em ordem decrescente. (B) Diagramas de abundância relativa (ou curvas de dominância) que podem ser utilizados para comparar o padrão de dominância entre diferentes comunidades.

Escolhendo o melhor modelo teórico no R

```
> library(vegan)
> rios=read.table("rios.txt", h=T)
> rios
> rad.rio1=radfit(rios[1,])
> rad.rio1
> plot(rad.rio1, xlab="Ordem das espécies", ylab="Abundância",
pch=19)
> rad.rio2=radfit(rios[2,])
> rad.rio2
> plot(rad.rio2, xlab="Ordem das espécies", ylab="Abundância",
pch=19)
> rad.rio3=radfit(rios[3,])
> rad.rio3
> plot(rad.rio3, xlab="Ordem das espécies", ylab="Abundância",
pch=19)
> par(mfrow=c(2, 2))
> plot(rad.rio1, main="Rio 1", xlab="Ordem das espécies",
ylab="Abundância", pch=19)
> plot(rad.rio2, main="Rio 2", xlab="Ordem das espécies",
ylab="Abundância", pch=19)
> plot(rad.rio3, main="Rio 3", xlab="Ordem das espécies",
ylab="Abundância", pch=19)
```

Praticando:

Exercício 1: A bióloga responsável pela Secretaria de Meio Ambiente do Município de Florianópolis/SC precisa determinar a qualidade da água das seis praias mais movimentadas da cidade. Este trabalho surgiu após reclamações de banhistas e de pescadores de algumas dessas praias. A bióloga mediu os níveis de colifórmes fecais e coletou peixes em vários pontos de cada praia. Um estagiário derrubou o computador da bióloga e perdeu todos os dados dessa pesquisa. Por sorte, a bióloga havia anotado todos os dados referentes aos peixes coletados nas praias. Porém, os dados sobre os níveis de colifórmes fecais só foram anotados em arquivo digital. Com recursos limitados, a bióloga não pôde refazer as análises da qualidade da água e precisa realizar uma avaliação indireta a partir dos dados de riqueza e abundância de peixes.

Teoria: Teoria do distúrbio + Distribuição da Abundância das Espécies (SADs)

Pergunta: Praias mais poluídas possuem padrão de distribuição da abundância da espécies mais equitativo?

Unidade amostral: Pontos de amostragem em cada praia

Variável dependente: Abundância relativa

Variável independente: Praia

Importe a planilha "peixes.floripa.txt" e indique a partir dos diagramas de abundância relativa qual a praia com melhor e pior qualidade da água. Informe os modelos teóricos que melhor explicam o padrão de distribuição de abundância de cada praia e faça um diagrama de abundância relativa para cada praia e uma figura contendo todos os diagramas na mesma janela.

Diversidade beta

Desde o início da ecologia, a identidade das espécies que constituem determinada comunidade (i.e., composição de espécies) tem gerado uma série de hipóteses importantes para o entendimento de como os organismos se distribuem no espaço e no tempo. Uma das principais perguntas sobre esse assunto é "O que torna comunidades de espécies mais ou menos similares em diferentes lugares e tempos?" (Vellend 2010). Após os influentes estudos do ecólogo Robert Whittaker (Whittaker 1960, 1972), o termo diversidade beta (i.e., variação na composição de espécies entre áreas) ganhou força na literatura ecológica. Nas duas últimas décadas, o número de trabalhos aumentou expressivamente com o desenvolvimento de novos métodos para medir a diversidade beta e de novos pacotes estatísticos. A grande quantidade de medidas, abordagens estatísticas, termos e interpretações para a diversidade beta aumentaram a confusão em relação às maneiras corretas de acessar e testar os padrões de modificação na composição de espécies (Tuomisto 2010a,b, Anderson et al. 2011). Nesta apostila utilizaremos um roteiro prático baseado em hipóteses sugerido recentemente por Anderson et al. (2011). Primeiro, é importante diferenciar dois tipos de conceito de diversidade beta, o conceito de substituição (*turnover*) e de variação. A substituição representa a modificação na composição de espécies de uma unidade amostral para a outra ao longo de um gradiente espacial, temporal ou ambiental. A substituição requer um gradiente que indique direção como, por exemplo, investigar a mudança na composição de espécies ao longo de um gradiente de profundidade em um lago (Fig. 10a). As principais questões testadas na análise de substituição são: (1) quantas novas espécies são encontradas ao longo de um gradiente e quantas delas foram inicialmente presentes e agora foram perdidas? (2) Qual a proporção de espécies encontradas em uma unidade amostral que não são compartilhadas com a próxima unidade do gradiente?

Por outro lado, a variação representa a modificação na composição de espécies entre um grupo de unidades amostrais (Fig. 10b). A variação é necessariamente não-direcional e representa a modificação das espécies dentro de uma extensão espacial ou temporal determinada, ou dentro de um mesmo fator (e.g., tipo de hábitat, fragmentos florestais). As principais questões testadas na análise de variação são: (1) podemos encontrar as mesmas espécies repetidamente entre diferentes unidades? (2) Qual a proporção esperada de espécies não compartilhadas entre todas as unidades amostrais?

Antes de usar os índices propostos nessa apostila, leia atentamente o artigo recentemente publicado na Ecology Letters (Anderson et al. 2011) para escolher corretamente o índice que responde a sua questão. Além disso, Koleff et al. (2003) e Legendre & Legendre (1998) são extremamente importantes para compreender a formulação e características de cada um dos índices de diversidade beta.

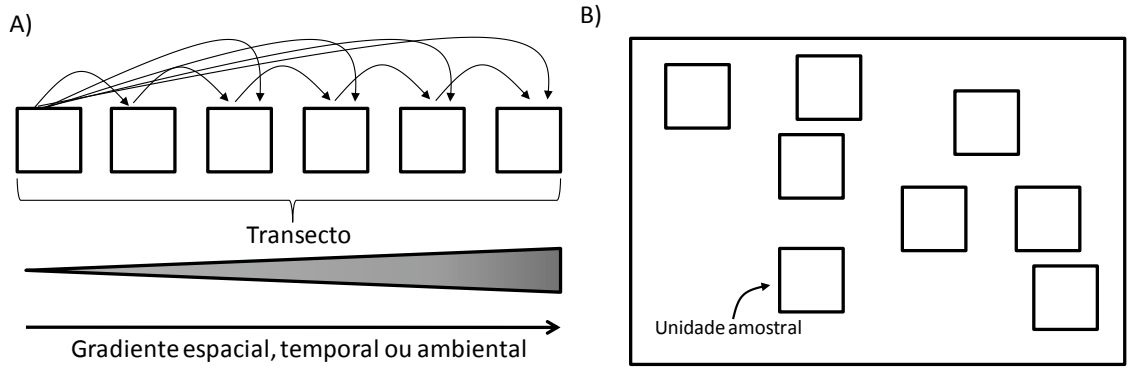


Figura 10. Diagrama esquemático dos dois tipos de diversidade beta: (A) substituição, mede taxa de modificação na composição de espécie em relação a um gradiente direcional; (B) variação, mede a diferença na composição de espécies entre grupos de unidades amostrais e é não-direcional (adaptado de Anderson et al. 2011).

Métricas para medir a diversidade beta

Um dos primeiros índices propostos para medir a diversidade beta é o Índice de Whittaker (β_w), que examina a taxa de diferenciação na diversidade alfa (riqueza local; α) entre duas ou mais comunidades em relação à diversidade gama (riqueza regional; γ). A fórmula foi proposta por Whittaker (1960) :

$\beta_w = \gamma / \bar{\alpha}$ = (b + c) / (2a + b + c)

Onde γ representa o total de espécies *S*, e α o valor médio da riqueza de uma amostra. O valor "a" representa o número de espécies compartilhadas, e os valores "b" e "c" o número de

espécies não compartilhadas entre duas comunidades. O foco dessa análise é na identidade da espécie e em quantas vezes a riqueza em uma região é maior do que o valor médio da riqueza na menor unidade amostral.

Índices binários (presença/ausência)

Os índices mais conhecidos e utilizados na ecologia são o índice de similaridade de Jaccard (*J*) e Sørensen (*S*). O inverso desses índices, i.e., o valor de dissimilaridade, são denomidados d_J e d_S. Para calcular cada um desses índices usamos as fórmulas:

$J = a / (a + b + c)$

$d_J = 1 - J$

$S = 2a / (2a + b + c)$

$d_S = 1 - S$

Onde "a" representa o número de espécies compartilhadas entre duas unidades amostrais *i* e *j*, "b" representa o número de espécies que ocorrem na comunidade *i*, mas não em *j*, e "c" representa o número de espécies que ocorrem na comunidade *j*, mas não em *i*. Os valores de *J* e *S* variam de 0 (comunidades sem nenhuma espécie compartilhada) a 1 (comunidades que compartilham todas as espécies, i.e., similaridade total). Os valores de dissimilaridade d_J e d_s variam de 0 (comunidades idênticas) a 1 (comunidades que não compartilham nenhuma espécies, i.e., dissimilaridade total). A diferença básica entre os índices *J* e *S* é que o segundo atribui maior peso à presença das espécies ("2a" na fórmula de *S*). Em teoria, uma espécie que ocorre em duas comunidades é mais importante do que uma espécie que não ocorre em nenhuma das duas comunidades (dupla ausência) (veja discussão em Anderson et al. 2011).

Índices quantitativos (abundância relativa)

Bray-Curtis

O índice de Bray-Curtis (BC_{ij}) é considerado um índice semi-métrico e utiliza a abundância das espécies em sua fórmula :

$$BC_{(x1,\ x2)} = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})}$$

Onde y_{1j} representa a abundância da espécie *j* na localidade *x1* e y_{2j} na localidade *x2*. Esse cálculo prossegue até a espécie *p*.

Medidas multivariadas

Uma medida de diversidade beta interessante para comparar *N* amostras é a dispersão em um espaço multivariado, com uma análise conhecida como teste de homogeneidade de dispersões multivariadas (Anderson 2006). Esta análise calcula o centróide (ou mediana especial) de um grupo específico (e.g., lagoa 1) e compara a dissimilaridade média das *n* observações individuais dentro desse grupo (e.g., abundância de cada espécie *p* na lagoa 1) utilizando uma medida apropriada de dissimilaridade (e.g., Bray-Curtis, Chao-Sørensen, Distância Euclideana, Jaccard, Sørensen). O cálculo do centróide para medidas que utilizam distância euclidiana é a média aritmética de cada variável. Porém, para calcular o centróide para índice de distância não-euclidianos (e.g., Jaccard) é necessário fazer uma análise de coordenadas principais (Anderson 2006). A hipótese nula desta análise é a de que a diversidade beta não é diferente entre as amostras de interesse. Para acessar a probabilidade de a hipótese nula ser verdadeira utiliza-se a estatística *F* de Levene comparando a distância média de cada observação ao centróide do seu grupo que, por sua vez, é definido por uma medida de dissimilaridade. Para gerar os valores do *P* são realizadas *n* permutações (e.g., 1000) (detalhes em Anderson 2006).

Calculando os índices de diversidade no R

1. Calculando o índice clássico de Whittaker (β_w):

```
> salinidade=read.table("salinidade.txt", header=T)
> salinidade
> diversidade.beta=betadiver(salinidade, "w")
> diversidade.beta
```

2. Calculando índices de Jaccard e Sørensen:

```
> jaccard=betadiver(salinidade, "j")
> sorensen=betadiver(salinidade, "sor")
> scores(jaccard)
> scores(sorensen)
```

3. Calculando os índices de Bray-Curtis e Morisita-Horn:

```
> library(vegan)
> data(mite)
> bray=vegdist(mite, "bray")
> bray
> morisita.horn=vegdist(mite, "horn")
> morisita.horn
# Testando hipóteses com as matrizes de
similaridade/dissimilaridade
> library(vegan)
> data(varespec)
> data(varechem)
> dist.species=vegdist(varespec, "bray")
> dist.chemical=vegdist(scale(varechem), "euclidean")
> associacao=mantel(dist.species, dist.chemical)
> associacao
```

4. Calculando os índices de Chao-Jaccard e Chao-Sørensen:

```
> CSoren.dist=ecol.dist(ilhas, chao.sorenson, type="dis")
> CSoren.simi=ecol.dist(ilhas, chao.sorenson, type="sim")
> CJaccar.dist=ecol.dist(ilhas, chao.jaccard, type="dis")
> CJaccar.simi=ecol.dist(ilhas, chao.jaccard, type="sim")
# se optar por calcular a similaridade entre duas localidades
use a seguinte função:
> IlhaA=ilhas[,1]
> IlhaB=ilhas[,2]
> CSoren.A.B=chao.sorenson(IlhaA, IlhaB)
> CJaccar.A.B=chao.jaccard(IlhaA, IlhaB)
> CSoren.A.B
> CJaccar.A.B
```

5. Calculando outros índices de similaridade com o pacote *fossil*:

```
> library (fossil)
> Comunidade.A <- c(1,0,4,3,5,0,0,7)
> Comunidade.B <- c(2,1,3,0,0,1,0,6)
> bray.curtis(Comunidade.A, Comunidade.B)
```

```
> jaccard(Comunidade.A, Comunidade.B)
> simpson(Comunidade.A, Comunidade.B)
> sorenson(Comunidade.A, Comunidade.B)
> morisita.horn(Comunidade.A, Comunidade.B)
```

6. Teste de homogeneidade de dispersões multivariadas:

```
> library(vegan)
> cafe=read.table("cafe.txt", header=T)
> tipo.matriz=factor(c(rep(1,16), rep(2,8)), labels =
c("com.mata","sem.mata"))
> dissimilaridade=vegdist(cafe, "bray")
> HDM=betadisper(dissimilaridade, tipo.matriz)
> valor.P=permutest(HDM, pairwise = F)
> plot(HDM)
```

Praticando:

Exercício 1: Baseado na teoria de que os organismos selecionam sua planta hospedeira considerando características fisiológicas e estruturais, um biólogo pretende testar se três clones (clones x1, x2, e x3) de uma planta X possuem composição de espécies de ácaros diferente. Ele coletou ácaros em 60 plantas (20 plantas de cada clone) em uma estação experimental que cultiva a planta X. Em cada planta, o biólogo coletou 10 folhas e identificou e quantificou todos os ácaros. Além disso, o biólogo mensurou o comprimento, largura e área foliar e a densidade de tricomas.

Pergunta: O clone afeta a composição de espécies de ácaros?

Teoria: Teoria do nicho (*species sorting*)

Unidade amostral: Folha

Variável dependente: Composição de espécies

Variável independente: comprimento, largura e área foliar, e a densidade de tricomas.

Importe a planilha “clone.col1.txt” e “clone.col2.txt”e verifique se os clones possuem composição semelhante ou diferente nas duas coletas hipotéticas. Após as análises, responda a pergunta do biólogo para cada coleta. Os resultados realmente permitem que a pergunta seja respondida? O que você pode interpretar com a coleta 1 e com a coleta 2?

Exercício 2: Uma atividade muito comum em países com megadiversiadade de aves tais como o Brasil é chamada de "birdwatching" (BW), que consiste no estudo e observação de aves a olho nu ou com binóculos. Turistas estrangeiros gastam milhões de dólares anualmente para observar aves em florestas tropicais. Em uma fazenda particular com 10000 ha de floresta amazônica, um biólogo comparou o impacto do BW na diversidade beta de aves. Ele comparou dez trilhas utilizadas para BW e dez trilhas bloqueadas para turismo e pesquisa. O biólogo acredita que o fluxo de turistas nas trilhas interfere no comportamento de forrageio de muitas espécies de aves e diminui a riqueza e diversidade beta em comparação com áreas sem esta atividade.

Pergunta: a diversidade beta é maior em áreas sem BW?

Teoria: Nicho, teoria do forregaio ótimo.

Unidade amostral: pontos de amostragem ao longo da trilha.

Variável dependente: diversidade beta.

Variável independente: tipo de trilha (indiretamente relacionado ao impacto do turismo).

- Importe a planilha "birdwatch.txt" e responda se o turismo (BW) afeta a diversidade beta de aves utilizando o teste de homogeneidade de dispersões multivariadas. Faça uma figura representando a dispersão multivariada das observações em relação ao centróide de cada grupo: trilha com turismo e trilha sem turismo. As dez primeiras linhas do arquivo "birdwatch.txt" representam trilhas bloqueadas a turistas e pesquisadores e as dez últimas linhas são trilhas utilizadas para BW.

INTRODUÇÃO À ESTATÍSTICA MULTIVARIADA

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*.

Devido à restrições de tempo, este módulo do curso terá um componente mais *informativo* que *formativo*. Procuraremos explicar a lógica por trás de cada teste, a sua aplicação em problemas comumente encontrados em estudos ecológicos, mas infelizmente não há tempo hábil para destrinchar detalhadamente como cada método funciona e o seu componente matemático.

Em geral, análises multivariadas têm três principais utilidades: encontrar a principal direção de variação dos dados, efetuar correlações 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*, não pode ser ignorada. Antes de iniciar a parte prática, gostaria de discutir alguns aspectos teóricos e filosóficos, grandemente baseada em James & McCulloch (1990).

A amostragem em campo deve ser adequada para o objetivo da análise. Se o objetivo do usuário for **estimar parâmetros**, a amostragem deve ser **aleatória** ou **estratificada**. Se o objetivo for a **detecção de padrões**, a amostragem deve ser **sistemática** (veja Hayek, 1994; Gotelli & Ellison, 2004; Sutherland, 2006; Greenwood & Robinson, 2006). Para estudos **experimentais**, deve haver sempre **aleatorização** (sorteio), ou seja, cada *unidade amostral* tem de ser independente da outra e ter a mesma chance de ser selecionada (veja Hurlbert, 1984). Este procedimento eliminaria qualquer fonte de confundimento e enviesamento da amostragem, por "dissolver" possíveis fatores que possam afetar a variável de interesse e que não foram medidos/considerados no estudo.

Além disso, ao desenhar o seu estudo, priorize ou a escala temporal ou a espacial. Sempre obtenha mais amostras que variáveis. Sempre que possível, evite perder dados (*missing values*, NA's), pois eles diminuem o poder do teste (mas veja Legendre & Legendre, 1998 para saber como lidar com NA's). Para avaliar a suficiência amostral, verifique se o mesmo padrão de classificação emerge com o aumento do número das amostras.

Por fim, 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, 1998). Por outro lado, a análise de ordenação é uma operação pela qual os objetos (ou descritores) são posicionados num espaço que contém menos dimensões que o conjunto de dados original; a posição dos objetos ou descritores em relação aos outros também podem ser usadas para agrupá-los.

Agrupamento

Análise de agrupamento hirerárquico (*cluster*)

A análise de agrupamento hierárquico é a mais utilizada em ecologia. No entanto, existem também outras análises não hierárquicas, como a K-means, que não serão abordadas neste curso. 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 apresentá-la 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.

Apesar da sua versatilidade, deve-se ressaltar que nem todos os problemas em ecologia são problemas de agrupamento. 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. Além disso, existem algumas críticas que merecem atenção: mesmo para um conjunto de dados aleatórios é possível encontrar 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. Normalmente, a análise de agrupamento tenta arranjar os objetos em grupos que são mutuamente excludentes, ou seja, o mesmo objeto não pode fazer parte de mais de um grupo. No entanto, existem algumas técnicas, chamadas de *fuzzy clustering*, que permitem uma gradação na classificação de objetos. Esta técnica não será abordada neste módulo, mas o leitor interessado é remetido à duas referências: Legendre & Legendre (1998) e Borcard et al. (2011).

Os passos para a análise de agrupamento são os seguintes:

1) A matriz deve conter os objetos a serem agrupados (p.ex. espécies) nas linhas e as variáveis (p.ex., locais de coleta ou medidas morfológicas) nas colunas. Primeiramente, se os dados forem de *abundância*, é mais correto realizar a **transformação** de Hellinger (Legendre & Gallagher, 2001). Se a matriz original contiver muitos valores

discrepantes (p.ex., uma espécie muito mais ou muito menos abundante que outras) é necessário **transformar** os dados usando Log (x+1)[1]. Se as variáveis forem medidas tomadas em diferentes escalas (metros, graus celcius etc), é necessário **padronizar** cada variável utilizando a seguinte fórmula:

$$Z = \frac{obs - média}{desvio}$$

Onde *obs* representa o valor da unidade amostral de interesse e os valores da média e do desvio padrão são calculados para cada variável.

2) Escolha do **método de agrupamento**

A escolha do método de agrupamento é crítico para a escolha de um coeficiente de associação. É importante compreender completamente as propriedades dos métodos de agrupamento para interpretar corretamente a estrutura ecológica que eles evidenciam (Legendre & Legendre, 1998). 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. Por motivos de espaço e tempo discutiremos somente os métodos hierárquicos, que são os mais comumente encontrados na literatura ecológica.

Métodos hierárquicos podem ser divididos naqueles que consideram o centróide ou a média 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. A matriz de similaridade ou distância é atualizada e reduzida de tamanho em cada etapa do agrupamento, por isso não exige tanto do computador (Legendre & Legendre, 1998).

[1] O uso do 1 é obrigatório pois Log de zero na base 10 não existe.

O método de Ward é baseado no critério de quadrados mínimos dos modelos lineares. 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. 2011).

3) Escolha dos **índices de similaridade** (coeficientes de distância ou de associação, ou índices de dissimilaridade).

Os índices de similaridade medem a distância entre dois objetos ou quantificam o quanto eles são parecidos. Lembre-se: as questões e hipóteses iniciais do estudo devem ser levadas em conta na escolha do índice (veja Anderson et al. 2011).

Índices binários assimétricos

Se os dados disponíveis foram de **presença-ausência** (binários), os índices recomendados são os de Jaccard e Sørensen. Os índices tradicionais de Jaccard e Sørensen são chamados de índices assimétricos, pois ao fazerem a comparação entre amostras não levam em conta duplas ausências. Essa característica é desejável ao analisar dados ecológicos porque o não encontro de duas espécies em duas localidades não é um indicativo de que duas localidades sejam similares, já que isto pode ter surgido por variação estocástica na amostragem, padrões de dispersão, etc. Além disso, as duplas-ausências não refletem necessariamente diferenças nas localidades (Legendre & Legendre, 1998; Anderson et al., 2011). Desta forma, somente serão considerados similares localidades que de fato compartilhem espécies.

Compare as fórmulas dos coeficientes de Jaccard e Sørensen (Pag. 89):

Como é possível perceber pelas fórmulas, o coeficiente de Sørensen dá um peso maior para as duplas presenças, pois elas são um indicativo mais forte de semelhança. No entanto, o índice de Sørensen é sensível à variações na riqueza entre as localidades.

Como uma alternativa, o índice de Simpson para similaridade múltipla entre comunidades foi proposto recentemente por Baselga et al. (2007) como uma modificação do índice de diversidade de Simpson. Este índice tem a vantagem de ser independente da riqueza e assim, consegue distinguir entre a substituição verdadeira e a simples perda de espécies. Isto é importante porque, como visto anteriormente, a diversidade beta pode ser causada por dois distintos fenômenos: aninhamento e substituição de espécies (*turnover*) que, por sua vez, são causados por processos ecológicos diferentes. Além disso, este índice leva em consideração a similaridade em toda comunidade e não par-a-par, como outros índices tradicionais (Baselga et

al. 2007). Se o leitor estiver interessado nesse assunto, existe outro índice de múltiplas comunidades proposto por Anne Chao (Chato et al. 2005, 2006; veja acima) que é implementado na função no programa SPADE da autora que usa tanto dados de incidência quanto de presença-ausência. Esta autora também propôs modificações nos índices clássicos de Jaccard e Sørensen para possibilitar a inclusão de dados de abundância. A implementação destes índices de Chao-Jaccard e Chao-Sørensen está disponível na função `chao.sorenson()` do pacote *fossil*.

Índices quantitativos assimétricos

Esses índices permitem a incorporação de dados de **abundância** nas análises. Os índices recomendados e os mais usados são os de Bray-Curtis, Gower2 (elimina duplas ausências, pode ser usado tanto para abundância quanto variáveis *dummy*) e Morisita-Horn. A grande vantagem deste último é a sua independência do tamanho amostral (Krebs, 1999).

Coeficientes de distância métricos

O principal coeficiente de distância usado em cologia é a distância euclidiana e suas demais variantes: distância euclidiana média, ponderada e padronizada. A distância euclidiana é recomendada nos casos em que as variáveis de estudo forem **contínuas**, **morfométricas** ou **descritores ambientais**.

Como avaliar a representatividade do dendrograma?

E como avaliamos se o dendrograma representa adequadamente a matriz de dados original? Existem basicamente duas formas: avaliar o coeficiente de correlação cofenética ou utilizar a distância de Gower (Borcard et al., 2011). A correlação cofenética é obtida simplesmente pela correlação de Pearson entre a matriz original de similaridade e a matriz cofenética. Esta é dada pela distância cofenética (distância onde dois objetos tornam-se membros de um mesmo grupo) entre todos os pares de objetos. Quanto maior a correlação, melhor a representatividade da análise. Normalmente, uma "regra de polegar" usada é somente admitir análises que produzam uma correlação maior que 0.8. Se o usuário não tem certeza de qual método de agrupamento ou coeficiente de distância usar, é possível (mas talvez não muito recomendado) realizar a análise com vários métodos e depois escolher o que produzir a maior

correlação utilizando um diagrama de Shepard (Borcard et al., 2011). Ainda, é possível utilizar a correlação de Kendall ou Spearman como alternativa para a de Pearson.

A distância de Gower é calculada como a soma dos quadrados da diferença entre as matrizes de distâncias cofenéticas e a original. O método de agrupamento que produzir a menor distância de Gower é aquele que fornece o melhor modelo de agrupamento para a matriz de distância. Mas observe que o método da correlação cofenética e a distância de Gower nem sempre concordam (Borcard et al., 2011).

Interpretação dos grupos: qual o nível de corte?

A análise de agrupamento é um procedimento heurístico e não um teste estatístico (Borcard et al., 2011). Portanto é necessário que o usuário interprete o resultado (dendrograma) à luz dos dados originais. Isto também enfatiza a necessidade de se escolher o método mais apropriado para o estudo, já que o resultado depende fortemente dos métodos. Existem várias formas propostas para escolher o nível de corte do dendrograma. É possível realizar uma inspeção visual e determinar quais agrupamentos fazem sentido, em relação ao conjunto de dados. Ainda, é possível utilizar matrizes modelos contruídas e depois compará-las com a original, posteriormente faz-se uma correlação entre essas matrizes para encontrar o nível de corte mais apropriado (Bini & Diniz-Filho, 1995). Outra "regra de polegar" normalmente usada é escolher o nível de corte como 50% de similaridade. Outra opção é adicionar valores de *bootstrap* aos nós do dendrograma e interpreter somente os nós co um valor alto, algo como 70%, de *bootstrap*. O livro Borcard et al. (2011, p. 65) traz mais alguns métodos para a escolha do nível de corte. Recomendamos ao leitor avaliá-los para determinar se algum se encaixa na proposta do seu estudo.

Outra alternativa para encontrar grupos em um dendrograma é oferecida pelo pacote *pvclust* (Suzuki & Shimodaira, 2005). Este pacote calcula automaticamente o valor de *P* para cada agrupamenteo formado. O pacote ainda emprega uma reamostragem em multiescala usando *bootstrap* que, por sua vez, utiliza tamanhos amostrais maiores e menores que a matriz original de dados, ao contrário da análise comum de bootstrap, na qual o tamanho amostral permanece constante e igual ao tamanho da matriz de dados (Shimodaira 2004). Assim, o valor de *P* é estimado pelo ajuste a uma curva teórica obtida de todos os tamanhos de amostragem, corrigindo assim para o enviesamento do tamanho amostral constante do bootstrap comum.

A seguir, faremos alguns exercícios que utilizarão o *pvclust* para selecionar os grupos do dendrograma.

Exercícios

1) No R existem dois pacotes que realizam a análise de agrupamento: a função `hclust()` do pacote *vegan* e o pacote *cluster*. Para começarmos a trabalhar, baixe e carregue o pacote *vegan*, depois carregue o arquivo de dados "mite" para o R da seguinte forma:

```
>library(vegan)
>data(mite)
```

a) Efetue a análise de agrupamento pela função `hclust()` utilizando o método UPGMA e o índice de Bray-Curtis. Lembre-se de dar nome ao objeto para poder plotar o dendrograma depois. Utilize a ajuda para encontrar como entrar com os argumentos da função.

b) Faça agora o dendrograma com outro índice de dissimilaridade e compare os resultados. São diferentes? No que eles influenciaríam a interpretação do resultado?

2) Agora vamos usar a abordagem proposta pelo *pvclust*. Primeiro instale o pacote e depois carregue-o. Em seguida, digite esta função no script do R:

```
dist <- function(x, ...){
          vegdist(x, ...)
          }
```

O *pvclust* é limitado porque só permite que usemos os indices de dissimilaridade da função `dist()`. Essa função faz com que possamos utilizar os índices da função `vegdist()` do pacote *vegan*. Se preferir, é possível usar os índices disponíveis na função `dsvdis()` do pacote *labdsv* substituindo-a na função acima. Importe o conjunto de dados "bocaina.txt" para o R e faça a análise utilizando o método UPGMA e o índice de Morisita-Horn. O *pvclust* agrupa os objetos que estão na coluna. Dese modo, se quisermos agrupar as espécies da comunidade devemos primeiro transpôr a matriz. Lembre-se de dar nome ao objeto para podermos plotar o dendrograma depois.

3) Calcule novamente o dendrograma usando o *pvclust* e o conjunto de dados dunedata$veg do pacote *ade4* utilizando o método UPGMA e a distância de Bray-Curtis.

IndVal

O objetivo desta análise é identificar **especies indicadoras** de grupos pré-estabelecidos. 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). As vantagens desta análise é que ela é baseada na abundância das espécies dentro do grupo e mede a associação entre as espécies e os grupos. A análise originalmente proposta por Dufrêne & Legendre (1997) parecia um pouco circular, já que a classificação das localidades para a formação dos grupos é feita a partir de dados das espécies, então as espécies indicadoras já seriam aquelas que foram usadas pra formação dos grupos. Uma forma de contornar essa circularidade seria utilizar alguma informação independente para a formação dos grupos como, por exemplo, algum descritor ambiental. Algumas melhorias foram realizadas na análise original e estão disponíveis em De Cáceres & Legendre (2009), incluindo um novo pacote chamado *indicspecies* disponível na página pessoal do autor (http://sites.google.com/site/miqueldecaceres/software).

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.

Espécies indicadoras podem mostrar características particulares de um determinado grupo, podendo inferir, por exemplo, situações de eutrofização de ambiente aquático. Por exemplo, algumas espécies quando muito abundantes em determinado local podem indicar que o ambiente está poluído. **A espécie indicadora é definida como a mais característica de um determinado grupo**.

A análise procede da seguinte forma:

1º Uma matriz de distância é construída e as unidades amostrais são classificadas com alguma análise de agrupamento, hierárquico ou não;

2º A variável ambiental para a qual se deseja classificar os grupos é inserida;

3º As espécies indicadoreas de cada grupo são formadas através do cálculo da especificidade e fidelidade, obtendo-se o valor de IndVal para cada espécie;

4º Por fim, o conjunto de dados originais é comparado para ver se análise faz sentido.

O índice é calculado seguindo a fórmula abaixo para cada espécie:

$IndVal_{ij} = A_{ij} * B_{ij} * 100$,

onde A_{ij} é a especificidade da espécie *i*, que é dada pela abundância média dessa espécie no grupo *j* dividiva pela soma das abundâncias médias da espécie *i* em todos os grupos. B_{ij} é a fidelidade da espécie, que é dada pelo número de locais do grupo *j* onde a espécie *i* ocorre dividido pelo número de locais do grupo *j*.

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.

Na interpretação do resultado, uma espécie pode ser **indicadora perfeita**, quando ocorre em somente um grupo restrito de locais que têm uma dada característica e também ocorre em todos locais daquele grupo, ou seja, ela tem uma alta fidelidade e especificidade. Uma espécie pode ser ainda **indicadora assimétrica** quando a mesma não tem alta fidelidade, mas alta especificidade. Ao contrário, uma espécie **indicadora simétrica** tem alta fidelidade, mas baixa especificidade.

Exemplo

```
>install.packages("labdsv")
>library(labdsv)
>mam.cerrado=read.table(file.choose(), h=T)
>?indval
>fitofis=c(rep(1,4), rep(2,4), rep(3,4), rep(4,4), rep(5,4))
>resultado=indval(mam.cerrado, fitofis)
>summary(resultado)#para apresentar uma tabela dos resultados
>resultado$maxcls
>resultado$indcls
>resultado$pval
>tab.resultado=cbind(resultado$maxcls,resultado$indcls,resultado
$pval)
>colnames(tab.resultado)<-c("maxgrp", "ind. value","P")
>tab.resultado
```

Exercícios

1) Importe o conjunto de dados "indvalR.txt". Nestes dados, as espécies de cladóceros estão nas colunas e as unidades amostrais (lagoas) nas linhas, existe também informação sobre a turbidez (variável contínua) da água, para o qual iremos tentar encontrar espécies indicadoras de cada faixa. Esta coluna deve ser selecionada para compor os grupos.

2) Importe conjunto de dados "exemploIndval.txt". Neste conjunto, as espécies de anfíbios anuros estão nas colunas e os locais de reprodução estão nas linhas. O arquivo "gruposIndval.txt" classifica os locais de acordo com o nível de poluição. Calcule o IndVal para cada espécie e descubra se existe alguma espécie que pode ser indicativa de locais poluídos.

Comparação de médias entre grupos

Análise de Similaridade (ANOSIM)

A análise de similaridade (ANOSIM, ANalysis Of SIMilarity) é um tipo particular de análise de variância multivariada (MANOVA, Multivariate ANalysis Of VAriance) para comparação de médias, mas que não requer que os dados tenham distribuição normal multivariada e homogeneidade de variância. Esta análise testa se a similaridade é menor *dentro* do que *entre* grupos definidos numa matriz. Por exemplo, quando temos dois ambientes muito distintos (p.ex., um conjunto de riachos poluídos e outro saudável) e queremos avaliar se abundância de espécies é diferente entre estes dois tipos de ambientes. O teste ranqueia as similaridades dando o ranque de 1 para a maior similaridade entre um par de objetos (McCune & Grace, 2002). A estatística do teste, R, varia de -1 a 1, quanto mais positivo for o valor, maior a diferença entre os grupos. A estatística R é dada por:

$$R = \frac{(\bar{r}_b - \bar{r}_w)}{(M/2)}$$

onde r_b é a similaridade ranqueada entre grupos; r_w é a similaridade ranqueada dentro do grupo; M=n(n-1)/2; n=número de total de unidades amostrais. O ANOSIM também pode ser utilizado com dados de incidência para avaliar se a composição de espécies difere entre locais.

A MANOVA é raramente utilizada para analisar dados ecológicos de campo, devido às restrições mencionadas acima (McCune & Grace, 2002). Logo, não a incluímos neste curso. Por outro lado, a MANOVA, ou a sua variação PERMANOVA, é comumente utilizada para analisar dados de experimentos cujo desenho se encaixa nas premissas do teste (McCune & Grace, 2002). O ANOSIM é muito robusto quando temos somente dois grupos para os quais

queremos comparar a diferença. Quando temos mais de dois grupos, o procedimento mais recomendado é o MRPP, que veremos a seguir.

Procedimento de permutaçao multi-resposta (MRPP)

O MRPP é um procedimento não-paramétrico muito similar ao ANOSIM, diferindo somente na estatística do teste. Além disso, o MRPP é usualmente utilizado quando há mais de dois grupos para os quais se deseja testar se há diferença (McCune & Grace, 2002; p.188), enquanto o ANOSIM é mais recomendado quando se tem dois grupos.

Os procedimentos do teste incluem o cálculo de uma estatística δ, que é dada por:

$$\delta = \sum_{i=1}^{g} C_i x_i$$

onde g é o número de grupos, e C um peso que depende do número de ítens nos grupos. Existem vários métodos para atribuir peso, o mais usado e recomendado é $C_i=n_i/N$; onde n é o número de itens no grupo i e N é o número total de itens. São calculados dois valores de δ, um observado e outro simulado, que re-ordena as unidades amostrais dentro dos grupos. Posteriormenre, o valor de δ entra no cálculo da estatística do teste, R, que é dada por:

$$R = 1 - (\frac{\delta \text{ observado}}{\delta \text{ esperado}})$$

O valor de R mede o tamanho do efeito e é então independente do tamanho amostral. O R do MRPP funciona de maneira *oposta* ao R do ANOSIM: quanto maior o seu valor, menor a diferença entre os grupos (McCune & Grace, 2002; p.191).

Exemplo

```
>library(vegan)
>bocaina
>?anosim
>vec.bocaina=factor(c(rep(1, 7), rep(2,7)),
labels=c("Temporárias", "Permanentes"))
>bocaina.pad=decostand(bocaina, "pa")
>anosim(bocaina.pad, vec.bocaina)
>plot(anosim)
```

Teoria: Teoria de história de vida
Hipótese: As poças temporárias e permanentes terão similaridades diferentes
Unidade amostral: espécies
Amostras: Poças

Exercício

1) Na perspectiva de metacomunidades (Leibold et al., 2004), a dispersão dos organismos tem um papel proeminente para entender como as espécies estão distribuídas na natureza. Com o objetivo de testar se a dispersão influencia a composição de espécies de cladóceros e copépodos, e portanto a estrutura da metacomunidade, um pesquisador selecionou dois conjuntos de lagos: em um deles todos os lagos são isolados e no outro os lagos são conectados. Importe para o R o conjunto de dados "lagos.txt" e responda a pergunta se o fato de os lagos estarem conectados ou não influencia a composição de espécies desses microcrustáceos.

2) Refaça o mesmo teste para encontrar se a abundância relativa é diferente entre os lagos.
Explore os resultados com as funções `summary()`, `plot()`, `names()`.

3) Importe o conjunto de dados "anosim.txt" para o R. Este conjunto consiste de um levantamento de artrópodos de serrapilheira coletados em uma região de mata ombrófila densa (cinco primeiras unidades amostrais) e uma região de mata ombrófila mista (demais unidades amostrais). Faça um teste para calcular se a abundância dos artrópodes é diferente entre esses dois grupos de unidades amostrais.

4) Importe o conjunto de dados "mrpp.txt" para o R e responda se a composição de espécies vegetais é diferente entre as fitofisionomias de cerrado.

Ordenação irrestrita

Análise de Componentes Principais (PCA)

Ao contrário de análises de agrupamento (ou classificação), análises de ordenação não buscam por uma descontinuidade nos dados, mas sim analisar como os objetos se distribuem ao longo de gradientes. A ordenação representa uma situação mais próxima da prática em estudos ecológicos. A análise de componentes principais (PCA) é principalmente usada para reduzir a dimensionalidade dos dados, e também verificar como as amostras se relacionam, ou seja, o

quão semelhantes são segundo as variáveis utilizadas. O resultado prático é produzir um diagrama de ordenação que sintetize os dados, no qual os objetos mais próximos são mais semelhantes. Além disso, o método matemático procura maximizar a variância entre os objetos. Diferentemente de outras análises de ordenação, só é possível utilizar a distância euclidiana como coeficiente de similaridade na PCA. Logo, é mais recomendado usá-la para analisar variáveis ambientais ou medidas morfológicas.

A PCA tem como principais *vantagens*: retirar a multicolinearidade das variáveis, pois permite transformar um conjunto de variáveis originais intercorrelacionadas em um novo conjunto de variáveis não correlacionadas (componentes principais). Para visualizar o correlograma dos dados, utilize a função `cor()` e digite a matriz de dados como argumento. Além disso, reduz muitas variáveis a eixos que representam algumas variáveis, sendo estes eixos perpendiculares (ortogonais) explicando a variação dos dados de forma decrescente e independente.

As *desvantagens* são: a sensibilidade a *outliers,* não recomendada quando se tem duplas ausências (muitos zeros na matriz) e dados ausentes. A PCA também não é recomendada quando se tem mais variáveis do que unidades amostrais.

Conceitos importantes

Combinações lineares: equação que agrupa as diferentes variáveis, como em uma regressão múltipla.

Componentes principais: são as combinações lineares das variáveis, eixos ortogonais (independentes) que resumem (explicam) a variação dos objetos, e como tal podem ser consideradas como "novas" variáveis e usadas em análises posteriores. O número de componentes principais é igual ao número de variáveis. O primeiro componente principal resume a maior variação dos dados, o segundo, a segunda direção de maior variação dos dados e asim por diante.

Autovalores (*eigenvalues*): esses valores representam a variância dos componentes principais e traz a porcentagem de explicação de cada eixo. O número de autovalores é o mesmo do número de variáveis. Os autovalores serão maiores para aquelas variáveis que forem mais importantes na formação do eixo.

Autovetores (*eigenvectors*): o mesmo que ***Loading***, ou seja, coeficientes de combinação linear. Os autovetores são os eixos principais de dispersão da matriz e medem a importância de uma

variável em cada eixo. Desse modo, representam o peso de uma variável para a construção de um eixo e variam de -1 a 1 (correlação de Pearson);

Centróide: média ponderada de um conjunto multivariado, a menor distância média de todos os objetos num espaço multivariado;

Escores (Z_1, Z_2, Z_n): posição das unidades amostrais ao longo de um eixo de ordenação, pode se referir tanto à unidades mostrais quanto à variáveis. Escores são fornecidos pela substituição dos valores assumidos pelas variáveis originais nas combinações lineares. São utilizados para ordenar as unidades amostrais em um diagrama uni, bi ou tridimensional.

Inércia: a soma de todas as correlações das variáveis com elas mesmas, mede a quantidade de variância total que é explicada por um eixo.

Loadings (coeficiente de estrutura): correlação de Pearson entre os escores e as variáveis.

O procedimento da análise é o seguinte: uma matriz de similaridade é extraída de uma matriz de dados quantitativos utilizando a distância euclidiana. Se os dados estiverem em escalas diferentes, lembre-se de padronizá-los primeiro, ou usar a matriz de correlação ao invés da matriz de covariância. Os autovalores são então extraídos da matriz de similaridade para o cálculo dos autovetores, e então os componentes principais são calculados. A matriz de escores é extraída a partir da matriz de autovetores.

Um passo importante é selecionar quais são os eixos que foram os mais importantes, ou seja, aqueles que resumem a maior quantidade de variação dos dados. Para isso existem vários métodos (veja Jackson, 1993 e Peres-Neto et al. 2005): O critério de Kaiser-Guttman sugere calcular a média de todos os autovalores e interpretar somente aqueles cujo os autovalores sejam maiores que a média. Uma "regra de polegar" sugere escolher todos os componentes principais até atingir 75% de explicação. Outra opção é realizar um screen-plot que plota os componentes principais no eixo x e os autovalores no eixo y, os componentes com menor explicação tendem a estar numa linha reta; logo deve-se interpretar somente os componentes principais que não estão nesta reta. O critério da esferidade de Bartlett sugere que os componentes principais sejam selecionados até que as duas últimas medidas de explicação formem uma esfera. Finalmente, o método de *Broken Stick* sugere considerar somente os eixos maiores que o valor predito pelo modelo de Broken Stick. Este é o critério mais utilizado por ser um método estatísico e não heurístico, por isso vamos utilizá-lo no exemplo desta seção.

A PCA produz melhores resultados quando as variáveis possuem uma forte estrutura de correlação entre si (ou seja, qundo as variáveis são redundantes) e ao fazer esta análise, deseja-

se justamente eliminar a correlação entre as variáveis, produzindo assim novas variáveis que não correlacionadas. Além disso, a PCA também é muito sensível a valores discrepantes e *outliers*. Se a porcentagem de explicação dos eixos for muito similar entre si indica que não há uma associação entre as variáveis, i.e., não há uma estrutura clara nos dados.

Como perceber se a PCA foi a análise adequada? Aqui não existe um número mágico como o coeficiente de correlação cofenético. Então, um critério que se utiliza nestes casos é (dependendo do conjunto de dados analisado) utilizar a análise somente se os dois, ou no máximo, os três primeiros eixos explicarem em torno de 70% da variação dos dados. Se isso não acontecer, deve-se considerar outras análises, como veremos a seguir. Caso contrário, se considerarmos quatro ou cinco eixos, a interpretação pode ficar complicada. Um exemplo de interpretação de um biplot de PCA pode ser encontrado nas páginas 125-126 de Borcard et al. (2011).

Exercícios

1) Carregue o pacote MASS que já instalado no R. Ative o pacote de dados Crabs, `data(crabs)`. Este conjunto traz medidas morfológicas de dois morfo-tipos da espécie de caranguejo *Leptograpsus variegatus* coletada em Fremantle, Austrália. Calcule uma PCA e veja se existe uma semelhança morfológica entre os dois morfo-tipos. Lembre-se de dar nome ao objeto e use a função `biplot.rda()` para plotar o resultado do teste, utilize o argumento `scaling=1` e `scaling=2`. Dica: a projeção de um objeto perpendicular à seta do descritor fornece a posição aproximada do objeto ao longo desse descritor. A distância dos objetos no espaço cartesiano reflete a distância euclidiana entre eles.

2) Importe o arquivo "DoubsEnv.csv" para o R. Este conjunto fornece os descriores ambientais em 30 locais do rio Doubs, próximo à fronteira França–Suiça e consiste de 11 variáveis ambientais relacionada à hidrologia, geomorfologia e química do rio. Calcule uma PCA com a função `rda()` do pacote *vegan*. Para ver como entrar com os argumentos na função, digite `?rda`, utilize o argumento `scale=T` para padronizar as variáveis. Para ver quais eixos reter para plotar e interpretar, carregue e utilize a função `evplot()` escrita por Bocard et al. (2011) disponível no arquivo "evplot.R". O argumento da função deve ser os autovalores, portanto extraia-os utilizando `objeto1=objeto$CA$eig`.

Análise de Coordenadas Principais (PCoA)

A Análise de Coordenadas Principais é muito semelhante à PCA, diferindo somente pelo fato de que com ela é possível usar qualquer coeficiente de similaridade, e não só a distância euclidiana, como na PCA. Daí advém uma de suas grandes vantagens: é possível realizar a análise se só a matriz de similaridade estiver disponível. Além disso, a PCoA é adequada quando o número variáveis é maior que o número de amostras, ao contrário da PCA e também é robusta para valores ausentes, duplas ausências ou mesmo dados de incidência (variáveis *dummy*). É bastante útil para se analisar variações sazonais e gradientes de diversidade ou mesmo quando existem poucas unidades amostrais. No entanto, não informa quais variáveis influenciam a distribuição dados objetos e também não fornece a relação entre as variáveis e os eixos principais, somente as unidades amostrais. Outra desvantagem do método é a impossibilidade de interpretar os eixos com base na projeção dos descritores num 'continuum', ou em subconjuntos.

Os procedimentos para a análise são muito semelhantes à PCA, a única diferença é que a matriz de similaridade original passa por uma transformação denominada centralização dupla. Este procedimento é usado para manter a relação euclidiana entre as unidades amostrais. A PCoA produz *n-1* eixos, quando o número de unidades amostrais é igual ou maior que o número de variáveis.

Uma maneira de perceber se a análise foi adequada é verificar se foram produzidos autovalores negativos e altos, se sim, a matriz de distância que está sendo usada pode não ser adequada para a ordenação, pois a representação cartesiana pode estar distorcida. Para corrigir isso existem alguns métodos implementados na função `pcoa()`, do pacote *ape*. Na PCoA também os próprios autovetores são os escores, que podem então ser utilizados para ordenar as unidades amostrais.

Exercício

1) Importe o conjunto de dados "bocaina_temporal.txt" para o R. Este conjunto de dados consiste da abundâncias das espécies (nas linhas) de girinos que ocorreram em 13 poças durante 11 meses (colunas) no PARNA Serra da Bocaina. Faça uma PCoA utilizando o coeficiente de Bray-Curtis com a função `pcoa()` do pacote *ape* para descobrir se as espécies podem ser agrupadas de acordo com um padrão de ocorrência temporal. Construa o biplot com a função `biplot.pcoa()`.

Escalonamento multidimensional não-métrico (nMDS)

Este método é muito parecido com o anterior. Assim como a PCoA, o nMDS também permite utilizar qualquer coeficiente de distância para construir a matriz de similaridade e também aceita valores ausentes e duplas ausências. Mas, diferentemente da PCoA, o nMDS é uma técnica iterativa que visa minimizar o STRESS (STandard REsiduals Sum of Squares), uma medida do quanto as posições de objetos em uma configuração tridimensional desviam-se das distancias originais ou similaridades após o escalonamento. A análise procede pela atribuição de escores aleatórios aos eixos de ordenação escolhidos pelo usuário. Posteriormente, uma matriz de distância é calculada entre as unidades amostrais. Essa matriz é então correlacionada com a matriz de distância construída a partir dos dados originais. Os escores dos eixos de ordenação são aleatorizados até que a correlação entre a matriz de distância obtida com a aleatorização dos escores e a matriz de distância dos dados originais seja a maior possível e o valor de STRESS é então calculado. Este valor varia de 0 até 1, um bom ajuste é produzido quando o STRESS se aproxima de 0. Logo, o STRESS pode ser utilizado como uma medida do quão adequada a análise é. Uma "regra de polegar" (Clarke, 1993) sugere que:

- Stress <0.05 representação excelente;
- Stress <0.1 boa ordenação. Improvável de produzir algo melhor aumentando-se as dimensões do diagrama de Shepard;
- Stress <0.2 ordenação razoável. Não é possível discutir detalhes minusciosos, mas o aumento das dimensões do diagrama Shepar pode melhorar a representação;
- Stress >0.2 ordenação inviável e a interpretação pode ficar comprometida. Com valores de stress entre 0.35 e 0.4 as amostras estão posicionadas aleatoriamente, mantendo pouca ou nenhuma relação com a similariadde original.

Ao contrário da PCA e da PCoA, o nMDS permite escolher o número de eixos que se deseja produzir previamente à análise. Outras variantes do nMDS foram propostas, como o Hybrid MDS, que permite combinar coeficientes métricos e não métricos, mas não foram muito populares e não está disponível no R. A análise leva em conta o ranque das distâncias, e portanto não assume a linearidade entre as amostras, uma característica desejável quando se analisa dados de comunidades de espécies. No entando, essa característica não exclui a necessidade de se transformar os dados, se for preciso. As principais desvantagens do nMDS são: a análise não fornece a porcentagem de explicação de cada eixo, já que o número de eixos é escolhido previamente pelo usuário. Lembre-se de que na PCoA e PCA os eixos escolhidos são aqueles que produzem os maiores autovalores. O usuário deve fornecer o valor de STRESS, o coeficiente de distância utilizado e finalmente, se foi feita alguma transformação nos dados

previamente. Como o nMDS é uma técnica iterativa, é possível realizar a análise várias vezes como um procedimento para diminuir o valor de STRESS.

Exercício

1) Utilize a função `metaMDS()` do pacote vegan para ordenar os dados do arquivo "DoubsSpe.csv". Este conjunto de dados consiste da abundância de peixes coletados em vários trechos do rio Doubs, próximo à fronteira França-Suiça, utilize a distância de Bray-Curtis primeiramente e depois escolha um outro índice que também incorpore abundância e plote o resultado. Os resultados forram muito diferentes?

Ordenação restrita

Análise de Correspondência Canônica (CCA) e Análise de Redundância (RDA)

As duas principais análises de ordenação restritas (*constrained ordination*) utilizadas em ecologia são a Análise de Correspondência Canônica (CCA) e a Análise de Redundância (RDA). Estas duas análises são os equivalentes restritos da Análise de Correspondência (CA) (não abordada no curso) e da PCA, respectivamente. O principal objetivo destas análises é identificar a influência de variáveis ambientais sobre os padrões de composição e abundância das espécies numa comunidade. Estas análises são particularmente úteis para analisar a distribuição de espécies ao longo de gradientes ambientais, por isso são chamadas de "análises direta de gradientes" (*direct gradient analysis*).

A CCA avalia a estrutura de correlação dentro de um conjunto de dados (e.g., matriz de abundância de espécies) e entre a matriz de espécies e a matriz ambiental. Estas análises são chamadas de restritas por que restrigem a ordenação dos objetos de uma matriz por uma regressão linear múltipla de uma segunda matriz. Em termos práticos, se o usuário está interessado em saber o quanto da estrutura da comunidade pode estar relacionada a descritores ambientais e se se espera que as espécies respondam de forma unimodal a estes gradientes, então a análise de escolha é a CCA. Similarmente, a RDA também busca encontrar o quanto da composição e abundância das espécies na comunidade estão relacionadas com descritores ambientais, mas assume que existe uma resposta linear das espécies aos gradientes ambientais. Enquanto o presuposto da CCA parece ser mais ecologicamente plausível, os dados do usuário podem ser apropriados para uma RDA se a amostragem não compreender todo o gradiente ambiental. Por outro lado, a CA pode ser mais apropriada se o gradiente que influencia a

distribuição de espécies não tiver sido medido. Uma análise recentemente proposta permite analisar dados nos quais as espécies apresentem respostas mistas aos gradientes. O OMI (sigla para Outlying Mean Index, Dolédec et al., 2000) está disponível na função `niche()` do pacote *ade4*.

A CCA maximiza a separação dos nichos das espécies. Assim, as respostas das espécies diante do gradiente ambiental assumiriam a forma de curvas unimodais. Muitas variáveis ambientais podem ser utilizadas com o objetivo de explicar a distribuição das espécies, resultando em nichos p-dimensionais, no entanto a análise perde poder à medida que a matriz ambiental contiver mais e mais descritores do que unidades amostrais. A matriz de espécies pode conter somente dados de incidência. A RDA é conceitualmente equivalente a uma regressão linear múltipla multivariada, seguida de uma PCA baseada nos valores ajustados.

Diferentemente de outras análises, como PCA, PCoA e nMDS, todas as análises de correspondência, incluido a CCA, não calculam uma matriz de distância. Ao contrário, são baseadas nas distâncias de χ^2 onde as amostras são ponderadas de acordo com o total, fazendo com que haja uma distinção exagerada em amostras com muitas espécies raras. Por esse motivo, o uso da CCA deve ser restrito à situações onde as espécies raras foram adequadamente amostradas e são consideradas indicadores de características do ecosistema, do contrário, considere retirar espécies raras previamente à análise (Bocard et al., 2011, p.198-9).

O resultado prático destas duas análises, CCA e RDA, é um biplot no qual as variáveis ambientais são plotadas como setas e as espécies como pontos. Quanto menor o ângulo da seta em relação a um eixo, maior será a correlação daquela variável com o eixo. Geralmente em uma análise de ordenação, os números que estão plotados nos eixos são os autovalores. Também é pouco comum plotar a correlação nos outros eixos. Se essa informação estiver disponível, o usuário pode projetar a ponta da seta representando a variável no eixo da correlação para encontrar a correlação da variável com o eixo. O usuário pode saber a posição de uma amostra no eixo simplesmente projetando perpencidularmente a amostra no eixo. De forma similar, uma amostra pode ser projetada numa seta para saber em qual posição da variável uma amostra se encontra. No caso da CCA, ao projetar a espécie na seta da variável o usuário encontra o ótimo da espécie ao longo daquele gradiente. Quanto maior a seta, mais importante é a variável para explicar a distribuição das espécies. As espécies que estiverem no "quadrante" para o qual a seta aponta estão positivamente correlacionadas com variável. Ao contrário, as espécies que estiverem no "quadrante" oposto, estão negativamente correlacionadas com a variável. Mais detalhes de interpretação do gráfico produzido pela análise podem ser encontradas em Legendre & Legendre (1998; p. 586–587), Zurr et al. (2007; p. 240-2) e Bocard et al. (2011; p.166-7).

Se não temos uma hipótese a ser testada ou estamos particularmente interessados em descrever um padrão, um problema que pode surgir é que a grande quantidade de variáveis plotadas pode dificultar ou até mesmo confundir a interpretação dos dados. Para contornar essa questão, várias técnicas foram desenvolvidas, uma delas é a seleção "forward" de variáveis. Neste procedimento, somente as variáveis que forem significativas após uma aleatorização dos dados entram no modelo. No entando, um estudo recente (Blanchet et al., 2008) demonstrou que este procedimento pode levar à conslusões equivocadas. Portanto, as opções que temos são: avaliar a estrutura de correlação entre as variáveis e plotar somente as que não forem correlacionadas, ou delinear o estudo previamente à coleta das variáveis para diminuir a quantidade de informação a ser adicionada ao modelo.

Como decidimos qual análise usar: respostas lineares ou unimodais?

Muitos pacotes estatísticos disponíveis comercialmente, e.g., CANOCO, implementam um teste de aleatorização de Monte Carlo para avaliar a significância dos autovalores dos eixos canônicos baseado na estatística F (veja fórmula 6.2 em Bocard et al., 2011). Este teste avalia se as espécies exibem uma resposta linear ou unimodal aos gradientes ambientais, e portanto é crítico para a escolha correta do teste. No R este procedimento é implementado pela função genérica `anova()`, com os argumentos `by="axis"`, que indica que todos os eixos serão testados e `step=999` que indica o número de repetições do procedimento de aleatorização. Este analisa testa a significâncias dos eixos.

Exercícios

1) Calcule uma RDA com os dados "DoubsEnv.csv" e "DoubsSpe.csv", verifique se a análise foi aproprida e interprete o biplot.

2) Carregue os dados "mite.env" e "mite" e calcule uma CCA com esses dados, verifique se a análise foi aproprida e interprete o biplot.

RDA e CCA parcial

Como mostrado acima, RDA e CCA compõem um conjunto de análises chamadas análises canônicas assimétricas, que permitem a comparação de duas ou mais tabelas de dados. São chamadas análises assimétricas por que o conjunto de dados não têm a mesma função. O

exemplo mais famoso é a comparação de uma tabela de composição de espécies com uma segunda tabela de descritores ambientais (i.e., análise direta de gradientes). A ideia básica da RDA é "limitar" a matriz Y de composição de espécies a uma combinação linear com as variáveis ambientais. Em resumo, a RDA pode ser considerada uma regressão múltipla com todas as espécies sendo testadas simultaneamente (ter Braak & Smilauer 2002). Tanto a RDA parcial quanto a CCA parcial (daqui em diante RDA_p e CCA_p) têm a mesma lógica da RDA e CCA, porém as parciais utilizam uma terceira matriz no cálculo. A RDAp e CCAp possuem dois grupos de variáveis explanatórias: uma matriz X com as variáveis explanatórias que serão utilizadas no modelo, e uma matriz W com as covariáveis (e.g., variação espacial ou temporal); o efeito das covariáveis em Y (geralmente matriz de composição de espécies) é controlado na análise. Em geral, a matriz W contém variáveis cujos efeitos sobre a matriz Y são conhecidos. Por exemplo, coletas realizadas em tempos diferentes (e.g., dia, semana, mês) podem ser consideradas como covariáveis e, dessse modo, devem ser controladas com RDA_p ou CCA_p. Para analisar a relação da matriz Y com a matriz X na presença da covariável W é necessário: (i) calcular os resíduos de Y sobre W (chamados de $Y_{res|w}$) e os resíduos de X sobre W (chamados $X_{res|w}$); (ii) calcular a RDA (ou CCA) entre $Y_{res|w}$ e $X_{res|w}$ ou entre Y e $X_{res|w}$. Para testar a significância das análises RDA_p ou CCA_p são utilizados métodos de permutação. É importante notar que uma hipótese nula pode ser formulada sobre a relação entre X e Y. A partir dessa hipótese nula e dos testes de permutação, valores de probabilidade são acessados por meio de aleatorizações (veja detalhes metodológicos em Legendre & Legendre 1998; Bocard et al. 2011). Para calcular a força da relação entre Y e $X_{res|w}$ (R^2 canônico) usa-se a seguinte fórmula:

$$R^2_{Y|X_{res|w}} = \frac{SS\ (Y_{fit})}{SS\ (Y)}$$

Onde SS (Y_{fit}) representa a soma dos quadrados dos valores ajustados de Y, e SS(Y) a soma dos quadrados dos valores observados de Y. Para calcular a soma dos quadrados, o cálculo mais apropriado é: SS (Y_{fit}) = SS ($Y_{fit|(X+W)}$) − SS($Y_{fit|W}$), e SS (Y_{res}) = SS (Y) − SS ($Y_{fit|(X+W)}$). A soma de (X + W) representa a concatenação de X e W na mesma matriz. Y_{fit} é representado como uma regressão múltipla de Y contra X, ou seja, os valores ajustados de Y conforme fórmula da regressão, $Y_{fit}=X[X'X]^{-1}X'Y$.

Cuidado! No caso de interação entre a variável temporal e as variáveis ambientais ou espaciais, abordagens adicionais são necessárias para validar o modelo (mais detalhes em Legendre & Legendre 1998).

Na função `rda()` do *vegan*, a variação em Y explicada pelas variáveis ambientais é denominada "constrained variance" e a variação não-explicada (residual) é chamada "unconstrained variance".

Praticando:

Exemplo 1: Uma pesquisadora pretende testar como a composição de espécies de ácaros (matriz Y) varia na espécie de planta *Tibouchina granulosa* (Melastomataceae) na Serra do Mar. Para cada planta, ela anotou as seguintes variáveis: espessura da folha (esfl), área foliar (arfl) e densidade de tricomas (dtri). A pesquisadora tinha conhecimento de que a quantidade de água no substrato (quag), o tipo de solo (tiso) e a densidade da planta competidora *Tibouchina clavatium* (dens.tc) afetavam características estruturais da planta *T. granulosa*. Por isso, ela coletou esses dados para utilizar como covariáveis na análise.

- **Principal teoria**: Teoria do nicho

- **Pergunta**: a estrutura foliar de *T. granulosa* determina a composição de espécies de ácaros?

- **Unidade amostral**: planta.

- **Variável dependente**: composição de espécies.

- **Variável independente**: planta, variáveis ambientais (i.e., comprimento, largura, espessura e área foliar, densidade de tricomas).

- **Covariáveis**: quantidade de água no substrato e tipo de solo.

Exemplo 2: Um pesquisador pretende comparar a comunidade de ácaros associados à seringueiras em diversas regiões do Brasil. A principal questão é investigar se a composição de espécies de ácaros é influenciada por características ambientais (estrutura da planta hospedeira) e espaciais (oito localidades nos seguintes estados: AM, BA, ES, MS, MT, PA, SP). O pesquisador dividiu as características ambientais em duas escalas: uma ao nível da planta (densidade de tricomas, espessura foliar) e outra ao nível bioquímico (teor de nitrogênio, enxofre, proteínas e açúcares solúveis) e anotou as coordenadas geográficas dos pontos de coleta de cada planta.

- **Principais teorias**: Teoria do nicho e teoria neutra

- **Pergunta**: qual a importância relativa das características ambientais e espaciais na determinação da composição de espécies de ácaros associados à seringueira?

- **Unidade amostral**: planta.

- **Variável dependente**: composição de espécies.

- **Variável independente**: planta, variáveis ambientais e espaciais.

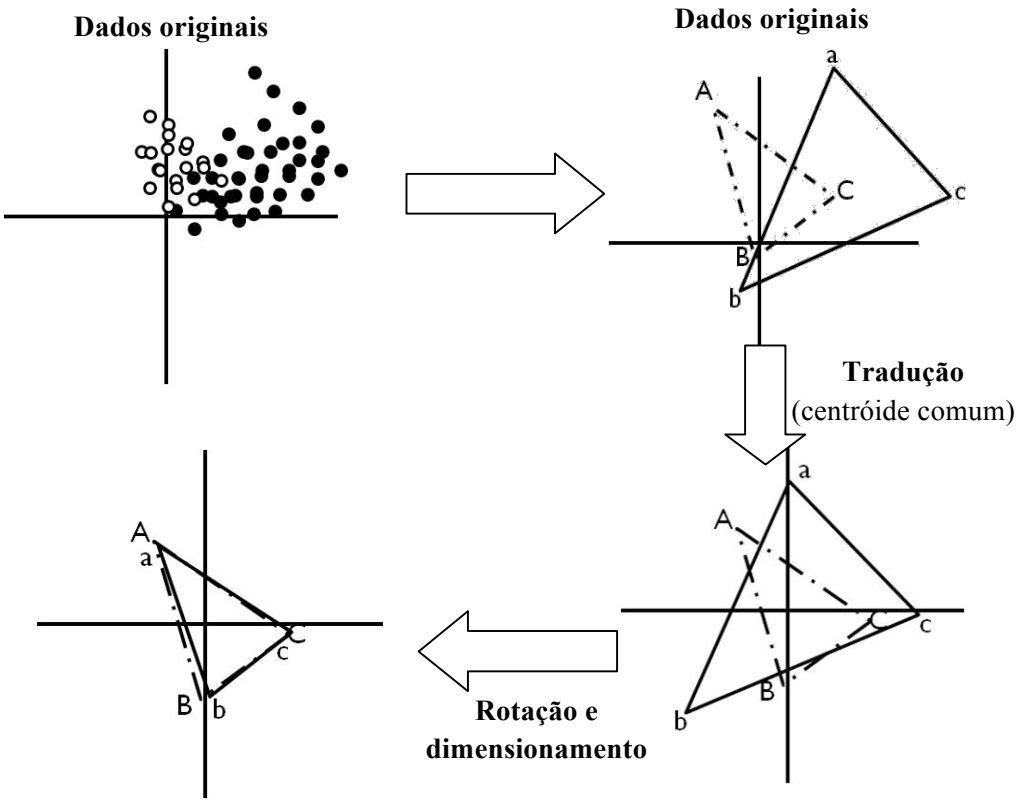
Análise de Procrustes

A análise de Procrustes é um método que compara dois grupos de dados. Esta análise mede o grau de concordância entre duas matrizes. Em outras palavras, o método combina pontos correspondentes (chamados "marcos") que são representados pela ordenação de espécies e características ambientais (quando aplicados à ecologia de comunidades) amostrados nas mesmas unidades amostrais. O objetivo da análise é de minimizar os desvios da soma de quadrados, o que define a estatística do teste (m^2) por meio da tradução (combina os dados de maneira que possuam o mesmo centróide), rotação e dilatação (dimensionamento dos dados) de um conjunto de dados para que seja "combinável" com a configuração "alvo" (*target matrix* ABC; veja esquema abaixo). Desse modo, quanto menor o valor dos resíduos, maior a concordância entre o conjunto de dados. Para testar a significância do valor de m^2 observado, são realizadas várias aleatorizações (definidas pelo usuário) com os dados originais para gerar n valores de m^2. Esta aleatorização é conhecida como PROtest na literatura. Os valores de m^2 e de P são definidos por:

$$m^2 = 1 - (\text{TraceW})^2$$

$$P = 1 + m^2_{small} / 1 + n$$

Para obter a matriz W é necessário decompor a matriz $Y_{(nxp)}$ em duas matrizes ortogonais $V_{(nxp)}$ e $U'_{(pxp)}$, e na matriz diagonal W. Para o cálculo do m^2, TraceW representa a soma dos elementos da diagonal principal (ou traço) da matriz W. A demonstração matemática dessa função não está no escopo dessa apostila. Para mais detalhes consulte Legendre & Legendre (1998). Para testar a significância do valor observado (m^2_{obs}), m^2_{small} indica o número de valores de m^2 simulados que são menores ou iguais ao m^2_{obs}, e n representa o número de aleatorizações. Por exemplo, se 12 valores encontrados na aleatorização (n = 9999 aleatorizações) são menores ou iguais ao m^2_{obs} observado, a probabilidade de que a hipótese nula seja verdadeira (ou seja, os dados não são concordantes) é $P = (1 + 12) / (1 + 9999) = 0{,}0013$.



Praticando:

Exemplo 1: Um pesquisador pretende testar se peixes e macro-invetebrados aquáticos têm respostas concordantes em relação aos lagos que ocorrem na região de Linhares, ES. Um dos objetivos desse pesquisador foi usar espécies-chave para reduzir o custo de se coletar vários táxons em uma mesma região. Em teoria, se espécies de táxons distintos respondem da mesma maneira em relação à diversas localidades (i.e., respostas concordantes), a resposta de um grupo taxonômico pode ser extrapolada para grupos concordantes. Cada lago (n = 25) foi dividido previamente em 30 parcelas "imaginárias" (selecionadas com imagens aéreas dos lagos). Foram sorteadas 5 parcelas/lago para fazer a coleta de peixes e macro-invertebrados com os métodos apropriados.

- **Principal teoria**: Teoria do nicho (baseando-se nas idéias de concordância de comunidades; *Community concordance* em inglês). Em um contexto de metacomunidades é importante conhecer a perspectiva de *species sorting*.

- **Pergunta**: peixes e macro-invertebrados possuem distribuição concordante em lagos da região de Linhares?

- Unidade amostral: parcela.

- Variável dependente: composição de espécies.

- Variável independente: lago.

Exercício 1:

O biólogo responsável pela gestão de uma RPPN (Reserva Particular do Patrimônio Natural) deseja utilizar um grupo indicador de qualidade ambiental. O proprietário da RPPN precisa reduzir os custos necessários para amostrar artrópodes e vertebrados e requisitou ao biólogo que optasse por um dos grupos. O biólogo tem dois problemas para resolver: o primeiro é que artrópodes e vertebrados podem responder de maneira diferente à qualidade ambiental, o segundo é qual dos grupos deveria escolher para trabalhar. Para resolver o primeiro problema, faça uma análise Procrustes e indique para o biólogo se as comunidades são concordantes ou não. O biólogo recuperou dados de coleta de artrópodes (artropodes.txt) e vertebrados (vertebrados.txt) em 50 pontos localizados em ambientes da RPPN. Os pontos foram definidos de acordo com diferentes tipos de solo e vegetação.

LEITURA RECOMENDADA

As maioria das referências (artigos e livros) citadas nesta apostila se encontram no CD entregue na primeira aula. Abaixo seguem uma lista de referências, algumas com comentários, cuja leitura recomendamos.

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology, 26: 32–46.

*Artigo da PERMANOVA

Anderson, M.J. et al. 2011. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. Ecology Letters 14: 19-28.

Baselga, A., Jimenez-Valverde, A. & Niccolini, G. 2007. A multiple-site similarity measure independent of richness. Biology letters 3:642-645.

*Descreve e implementa o índice de similaridade de Simpson

Bini, L. M. & Diniz-Filho, J.A.F. 1995. Spectral decompositions in cluster analysis with applications to limnological data. Acta Limnologica Brasiliensia 7: 35-40.

Blanchet, F. G., Legendre, P. & Borcard, D. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

*Artigo mostrando que o método "forward selection" para selecionar variáveis numa CCA não é a melhor opção.

Burnham, K.P. & Anderson, D.R. 2010. Model selection and multimodel inference: A pratical information-theoretic approach. Berlin, Springer.

Bocard, D. et al. 2011. Numerical ecology with R. Berlin: Springer.

**Escrito por autores de ponta em análises multivariadas, traz a implementação de testes abordados no livro de 1998 em R.

Chao A, Chazdon RL, Colwell RK, Shen T-J. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters 8:148– 159.

Chao A, Chazdon RL, Colwell RK, Shen T-J. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62:361–371.

Clarke, K. R. (1993). Non-parametric multivariate analysis of changes in community structure. Australian Journal of Ecology 18, 117-143.

**Artigo que descreve o ANOSIM e uma ótima referência para o nMDS também.

Clarke, K.R. & Warwick, R.M. 2000. Change in Marine Communities: An Approach to Statistical Analysis and Interpretation. 2nd eds. Plymouth Marine Laboratory & PRIMER-E: Plymouth.

*Manual do software Primer que traz também um pouco de teoria dos testes.

Cook, D. & Swayne, D.F. 2007. Graphics for data analysis interactive and dynamics with R and GGobi. Berlin: Springer.

*Este livro traz a implementação das funcionalidades do pacote ggobi, mais informações em: http://www.ggobi.org/.

Crawley, M.J. 2007. The R book. Nova York: Wiley.

*Livro que vai do básico ao avançado, tem informações sobre linguagem R, estatística univariada, multivariada e modelagem. Relativamente fácil de compreender. Cap. 5 e 27 traz funções para criação e manipulção de gráficos passo-a-passo

De Cáceres, M. & Legendre, P. 2009. Associations between species and groups of sites: indices and statistical inference. Ecology 90(12): 3566-3574.

*Artigo que expande o IndVal propondo variantes do índice.

Dolédec, S.; Chessel, D. & Gimaret-Carpentier, C. 2000. Niche separation in community analysis: a new method. Ecology 81(10): 2914–2927.

Dufrene, M. & Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67(3):345-366

*Artigo que propõe o IndVal

Ford ED. 2000. Scientific method for ecological research: Cambridge Univ Press.

Godfrey-Smith P. 2003. Theory and reality: An introduction to the philosophy of science: University of Chicago Press.

Gotelli N.J. & Ellison A.M. 2004. A primer of ecological statistics. Sunderland: Sinauer.

* O cap. 7 deste livro trás um apanhado geral sobre desenhos amostrais voltados

para experimentação e os dois últimos capítulos são uma introdução à estatística multivariada.

Greenwood, J. J. D. & Robinson, R. A. 2006. Principles of sampling. In: Sutherland, W. J. (ed.) Ecological Census Techniques, a handbook. 2 Ed. Cambridge: Cambridge University Press.

* Excelente abordagem sobre métodos de amostragem para pesquisas de campo.

Hayek, L-A. C. 1994. Research design for quantitative amphibian studies. In: Heyer, W.R. et al. (eds.) Measuring and monitoring biological diversity, standard methods for amphibians. Washington: Smithsonian Books.

Hurlbert SH. 1984. Pseudoreplication and the Design of Ecological Field Experiments. Ecological Monographs 54:187-211.

* Artigo clássico sobre amostragem e desenho experimental, além de uma leitura agradável.

Hurlbert, S.H. 1971. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology 52(4):577-586.

Husson, F.; Lê, S. & Pagès, J. 2011. Exploratory Multivariate Analysis by Example Using R. CRC Press.

*Traz alguns exemplos de ecologia.

Jackson D.A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74:2204-2214.

James, F.C. & McCulloch, C. E. 1990. Multivariate analysis in ecology and systematics: Panacea or pandora's box? Annual Review of Ecology and Systematics21:129-66.

*texto crítico que deve de ser lido por todo usuário de análises multivariadas. Bom também para escolher a análise correta.

Krebs, C. J. 1999. Ecological Methodology. 2 ed. Menlo-Park: Benjamin-Cummings.

*Texto bom para descrições e exemplos de coeficientes de similaridade e índices de diversidade, mas desatualizado infelizmente.

Legendre, P. & Legendre, L. 1998. Numerical ecology. 2 ed. inglesa. Elsevier.

**Este é o manual teórico essencial e leitura obrigatória para qualquer análise multivariada.

Magurran A.E. 2004. Measuring biological diversity. Oxford: Blackwell publishing.

McCune, B. & Grace, J. B. 2002. Analysis of Ecological Communities. MjM Software Design, Oregon: Gleneden Beach.

*Este é o manual que acompanha o programa PC-ORD, mas também traz um conteúdo teório bastante útil.

McGill BJ, et al. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology Letters 10:995–1015.

Murrell, P. 2006. R graphics. Boca Raton: Chapman & Hall/CRC.

Oksanen, J. 2011. Constrained Ordination: Tutorial with R and vegan. Disponível em: http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio2.pdf

The Ordination web page http://ordination.okstate.edu/

*página com vários recursos para auxiliar na execução de análises de ordenação, exemplos de planílha para entrada de dados em programas e um glossário termos em análise de ordenação podem parecer complicados no início e de fácil confusão.

Owen, W. J. The R Guide disponível em http://www.mathcs.richmond.edu/~wowen/TheRGuide.pdf.

* Este é um manual pequeno (49 páginas) fácil de entender para iniciantes não só no R mas também em computação. Uma boa pedida como texto inicial.

Palmer, M. W. 1993. Putting things in even better order: The advantages of canonical correspondence analysis. Ecology 74,2215-2230.

*Revisão sobre CCA

Paradis, E. 2005. R for beginners. Disponível em http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

* Este manual dá algumas noções iniciais de como lidar com objetos e gráficos no R, além de rudimentos de programação e análises estatísticas elementares.

Peres-Neto PR, Jackson DA, Somers KM. 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Computational Statistics &Data Analysis 49:974-997.

Pillar VDP. 1999. How sharp are classifications? Ecology 80:2508-2516.

R Labs for Vegetation Ecologists<http://ecology.msu.montana.edu/labdsv/R/labs/>

*Esta página traz uma introdução à análise de dados em R para ecólogos de comunidade.

Santos, A.J. 2003. Estimativas de riqueza em espécies. In: Cullen Jr., L. et al. (Org.). Métodos de estudo em biologia da conservação e manejo da vida silvestre. Curitiba: Ed. UFPR e Fundação O Boticário de Proteção à Natureza, p. 19-41.

Sarkar, D. 2008. Lattice, multivariate data visualization with R. Berlin: Springer.

Statistica electronic textbook <http://www.statsoft.com/textbook/>

*Esta é uma página que contém um livro-texto preparado pelos criadores do Statistica

Sutherland, W. J. 2006. Planning a research programme. In: Sutherland, W. J. (ed.)Ecological Census Techniques, a handbook. 2 Ed. Cambridge: Cambridge University Press.

*Boa leitura para "treinar" o raciocínio e planejar o trabalho de campo.

Ter Braak, C. J. F. (1986) Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167-1179.

*Artigo que propôs a CCA

ter-Braak CJE, M.Verdonschot PE. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology Aquatic Sciences 57(3):254-289.

Venables, W. N. & Ripley, B.D. 2000. S programming. Springer.

*Leitura avançada sobre programação em linguagem S, similar à R. O Cap. 12 deste manual contém mais detalhes de como criar e manipular gráficos

Venables, W. N. & Ripley, B.D. 2002. Modern applied statistics with S. 4.ed. Springer.

*Um livro para usuários avançados mas que traz muita informação sobre testes e um pouco de programação. Boa leitura para quem desejar se aventurar no R.

Venables, W. N. & Smith, D. M. 2010. An introduction to R. Disponível em http://brieger.esalq.usp.br/CRAN/doc/manuals/R-intro.pdf

* Este é o manual oficial do R development core team atualizado a cada versão lançada do R. Contém mais detalhes de como criar e manipular objetos no R, assim como as classes de objetos, gráficos, importação e exportação de dados, além de rudimentos de programação e análises estatísticas básicas, mas de difícil leitura.

Verzani, J. Simple R. Disponível em http://www.math.csi.cuny.edu/Statistics/R/simpleR/printable/simpleR.pdf

* Outro manual simples e de fácil consulta, bom como texto introdutório.

Wickham, H. 2009. ggplot2, Elegant graphics for data analysis. Berlin: Springer.

WolframathWorld<http://mathworld.wolfram.com/>

Zuur, A. F.; Ieno, E.N. & Meesters, E. H.W.G. 2009. A Beginner's Guide to R. Berlim: Springer.

* Este é um livro da série use R! da Springer de grande valia para os iniciantes, pois consegue atingir o equilíbrio entre detalhamento e volume de informação.

Zuur, A. et al. 2007. Analysing ecological data. Berlin: Springer.

*Capítulos 11-15 trazem implementação de análises multivariadas em R com exemplos de ecologia.