Quantificação da Diversidade Biológica

Prática 3
Comparação de Índices de Diversidade com Amostras

Paulo Inácio Prado e Adriana Martini

7 de junho de 2010

1 Média do Índice de Simpson

Imagine que temos uma área de 50 ha de floresta, da qual queremos estimar a diversidade média por quadrícula de 20x20 metros. Para isto dividimos a área em quadrados de 20x20 m, dos quais tomamos um certo número ao acaso. Em cada quadrícula, identificamos as espécies de plantas, e contamos o número de indivíduos por espécie. Com estes dados, calculamos índices de diversidade por quadrícula, e a média destes valores. Usando as definições de amostragem estatística temos:

Unidades amostrais: a menor divisão de seu objeto de estudo que pode ser tomada de forma independente das demais. No caso são as quadrículas.

População estatística: conjunto de todas as unidades amostrais que podem ser tomadas. Aqui são todas as quadrículas da área.

Parâmetro: um atributo quantitativo da população que se pretende estimar com a amostra. No caso são as médias dos índices de diversidade por quadrícula na área de estudo.

Estimativa: valor estimado de um parâmetro, obtido da amostra. No caso a média dos índices na amostra de quadrículas.

Nosso objetivo com uma amostragem é fazer **inferências estatísticas** a respeito do parâmetro a partir das estimativas, como a precisão da estimativa, ou um teste de diferenças dos parâmetros de duas populações de onde foram tomadas amostras. Queremos, por exemplo, inferir se duas comunidades diferem quanto à média de um índice a partir de amostras tomadas

delas. Mas para isto temos que ter uma **amostra probabilística**, em que a probabilidade de cada unidade ser incluída na amostra é conhecida. A maneira mais simples de se fazer isto é uma amostragem ao acaso, pois aí determinamos que todas as quadrículas tem a mesma probabilidade de inclusão.

No caso de uma amostragem ao acaso, a média amostral é um estimador não enviesado da média paramétrica. Isto significa que, se repetíssemos a amostragem muitas vezes, e para cada uma delas calculássemos uma nova média amostral, os valores destas estimativas teriam, em média, o valor do parâmetro. Além disto, a distribuição destes valores em torno deste esperado deve seguir uma distribuição normal, indicando que amostras com estimativas muito diferentes do parâmetro devem ser menos prováveis.

Em aula vimos que isto é verdade para as médias de riquezas e de índice de Shannon. Neste exercício vamos verificar o mesmo para o índice de Simpson. Vamos usar os dados da parcela permanente de Barro Colorado [1], dividida em quadrículas de 20 x 20 m. Nesta parcela todas as árvores com DAP [2] igual ou superior a 5 cm foram mapeadas e identificadas. Os dados que usaremos são os índices de diversidade por quadrícula. Começamos importando o arquivo de dados [3] Para objeto um no R que chamaremos de `bci.20x20`:

```
> bci.20x20 <- read.csv2("bci20x20_indices.csv",row.names=1)
```

Este é um objeto do tipo "planilha", (*dataframe*), em que cada linha é uma quadrícula, e as colunas são os valores de riqueza, índice de Shannon (H') e índice de Simpson (D) das quadrículas. Verifique as primeiras e últimas linhas do *dataframe*:

```
> ## Primeiras linhas do dataframe
> head(bci.20x20)
```

```
     S        H         D
A1  22 2.996636 0.9444444
A10 23 2.983156 0.9408327
A11 24 3.040155 0.9455782
A12 21 2.832233 0.9259259
A13 19 2.754272 0.9217620
A14 20 2.755132 0.9159184
```

[1] `http://www.ctfs.si.edu/site/Barro+Colorado+Island/`

[2] Diâmetro à altura do peito, convencionado em 1,3 m acima do solo

[3] Este arquivo de dados, `bci20x20_indices.csv`, e os demais usados neste roteiro estão disponíveis na página da disciplina

```
> ## Ultimas linhas do dataframe
> tail(bci.20x20)
```

```
     S        H         D
Y5  31 3.246106 0.9533608
Y50 28 3.286463 0.9605142
Y6  30 3.214319 0.9511834
Y7  29 2.995579 0.9191883
Y8  29 3.243726 0.9551833
Y9  21 2.875591 0.9321330
```

O número de quadrículas é o número de linhas deste objeto de dados:

```
> nrow(bci.20x20)
```

```
[1] 1250
```

Obtemos a média e o desvio-padrão do índice de Simpson por quadrícula na parcela aplicando a função `mean` à terceira coluna do *dataframe*, que pode ser separada com o operador `$`:

```
> mean(bci.20x20$D)
> sd(bci.20x20$D)
```

Estes são os valores dos **parâmetros** média (μ) e desvio-padrão(σ) do índice de Simpson para esta área de 50 ha. Qual seria o valor da média se amostrássemos apenas algumas quadrículas? Podemos avaliar isto tomando uma amostra de 10 quadrículas com a função `sample` e calcular o valor médio de D desta amostra:

```
> set.seed(42)
> ##amostra de 10 valores
> am <- sample(bci.20x20$D,size=10)
> ## valores amostrados
> am
```

```
 [1] 0.9502041 0.9240889 0.9006544 0.9281046 0.8653061 0.9444444 0.9335938
 [8] 0.8965517 0.9506173 0.9276844
```

```
> ## media dos valores amostrados
> mean(am)
```

```
[1] 0.922125
```

A média estimada com a amostra não é exatamente o valor do parâmetro μ, pois nem todos os valores são incluídos na amostra. Como esta inclusão se dá ao acaso, cada amostragem pode ter um valor diferente, como vemos repetindo o mesmo procedimento mais duas vezes:

```
> ##nova amostragem
> am <- sample(bci.20x20$D,size=10)
> am
> mean(am)
> ## mais uma amostragem
> am <- sample(bci.20x20$D,size=10)
> am
> mean(am)
```

A média amostral varia a cada vez que repetimos o mesmo procedimento de amostragem, portanto ela é uma **variável aleatória**. Mas alguns valores podem ser mais prováveis do que outros. No caso de médias obtidas de amostras tomadas ao acaso, a teoria de amostragem prevê que são variáveis aleatórias Gaussianas (normais), cuja média corresponde ao valor da média paramétrica da população de valores de onde foi tomada a amostra, μ. O desvio-padrão desta distribuição normal corresponde a $\sigma/\sqrt{n}$, o desvio padrão da população de riquezas por quadrícula dividido pela raiz quadrada do tamanho da amostra.

Para verificar isto, usamos um *loop* com a função `for` para repetirmos a amostragem de 10 quadrículas 2000 vezes e guardamos as médias de D de cada uma destas amostras:

```
> ##vetor para guardar os resultados
> bci.20x20q10 <- c()
> ##loop para repetir a amostragem
> for(i in 1:2000){
+ bci.20x20q10[i] <- mean(sample(bci.20x20$D,size=10))
+ }
```

E então podemos avaliar a distribuição destes valores em relação ao valor do parâmetro e à normal teórica esperada com

```
> ##distribuição de densidade de valores
> plot(density(bci.20x20q10),xlab="Media de D",
+      main="Media de D por Quadricula",lwd=2)
> ## Valor da media parametrica
```

```
> abline(v=mean(bci.20x20$D),col="darkgrey",lwd=2,lty=2)
> ## Distribuicao normal esperada
> curve(dnorm(x,mean=mean(bci.20x20$D),sd=sd(bci.20x20$D)/sqrt(10)),
+       add=T,col="red",lty=2)
```

Agora vamos fazer as mesmas simulações para amostras de 25 e 50 quadrículas. Para cada tamanho amostral, repetimos a amostragem 2000 vezes, guardamos os valores das médias do Índice de Simpson (D) por quadrícula em cada amostragem:

```
> ##vetor para n=25
> bci.20x20q25 <- c()
> ##loop para repetir a amostragem
> for(i in 1:2000){
+   bci.20x20q25[i] <- mean(sample(bci.20x20$D,size=25))
+ }
> ##vetor para n=50
> bci.20x20q50 <- c()
> ##loop para repetir a amostragem
> for(i in 1:2000){
+ bci.20x20q50[i] <- mean(sample(bci.20x20$D,size=50))
+ }
```

e fazemos três gráficos de densidade dos valores, com as respectivas curvas normais teóricas:

```
> ## espaco para ate 4 graficos
> par(mfrow=c(2,2))
> ##Amostra de 10
> plot(density(bci.20x20q10),xlab="Media de D",
+      main="Media de D por Quadricula \n N=10",lwd=2,
+      xlim=range(c(bci.20x20q10,bci.20x20q25,bci.20x20q50)))
> abline(v=mean(bci.20x20$D),col="darkgrey",lwd=2,lty=2)
> curve(dnorm(x,mean=mean(bci.20x20$D),sd=sd(bci.20x20$D)/sqrt(10)),
+       add=T,col="red",lty=2)
> ##Amostra de 25
> plot(density(bci.20x20q25),xlab="Media de D",
+      main="Media de D por Quadricula \n N=25",lwd=2,
+      xlim=range(c(bci.20x20q10,bci.20x20q25,bci.20x20q50)))
> abline(v=mean(bci.20x20$D),col="darkgrey",lwd=2,lty=2)
> curve(dnorm(x,mean=mean(bci.20x20$D),sd=sd(bci.20x20$D)/sqrt(25)),
```

```
+       add=T,col="red",lty=2)
> ##Amostra de 50
> plot(density(bci.20x20q50),xlab="Media de D",
+      main="Media de D por Quadricula \n N=50",lwd=2,
+      xlim=range(c(bci.20x20q10,bci.20x20q25,bci.20x20q50)))
> abline(v=mean(bci.20x20$D),col="darkgrey",lwd=2,lty=2)
> curve(dnorm(x,mean=mean(bci.20x20$D),sd=sd(bci.20x20$D)/sqrt(50)),
+       add=T,col="red",lty=2)
> par(mfrow=c(1,1))
```

Pergunta 1: Qual o efeito do tamanho da amostra sobre

(a) a precisão das estimativas?

(b) a aproximação das estimativas a uma distribuição normal?

2 Simulação de Testes de Comparação de Médias de Índices

As propriedades das estimativas dos índices de diversidade médios que vimos acima nos permitem utilizar testes estatísticos usuais para comparar diferenças entre duas amostras. Nesta seção vamos avaliar a eficácia do teste t para comparar duas comunidades das quais temos estas estimativas.

2.1 Duas amostras de comunidades diferentes

Vamos comparar a riqueza média de espécies por quadrícula de 20x20 m nas parcelas de Barro Colorado e da Ilha do Cardoso. Começamos importando este segundo conjunto de dados para um objeto no R :

```
> peic.20x20 <- read.csv2("peic20x20_indices.csv")
```

As riquezas médias por quadrículas nestas duas áreas não são muito diferentes:

```
> ##S média/quadr BCI
> mean(bci.20x20$S)

[1] 23.6792

> ##S média/quadr PEIC
> mean(peic.20x20$S)
```

```
[1] 24.87109
```

Será que um teste t feito com amostras de 10 quadrículas detectaria esta diferença? Podemos simular uma amostra de cada local, e aplicar o teste:

```
> set.seed(42)
> ## Amostra de BCI, n=10
> bci.s <- sample(bci.20x20$S,size=10)
> ##Amostra de PEIC, n=10
> peic.s <- sample(peic.20x20$S,size=10)
> ## O teste t
> t.test(bci.s,peic.s)

        Welch Two Sample t-test

data:  bci.s and peic.s
t = 0.2502, df = 16.742, p-value = 0.8054
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.720485  4.720485
sample estimates:
mean of x mean of y
     24.1      23.6
```

A diferença entre as médias de riqueza nas duas amostras é ainda menor, e o teste indica que há uma probabilidade de $p = 0.81$ de que as amostras venham de comunidades com a mesma média, portanto não autoriza rejeitar esta hipótese. Portanto, nesta simulação o teste não teve força para detectar a diferença que existe na realidade. Podemos avaliar com que frequência isto acontece repetindo a amostragem e o teste muitas vezes, guardando o valor de p para cada um destes testes.

```
> ##vetor para os resultados
> ttest.q10 <- c()
> ##loop repete amostras e testes
> for (i in 1:2000){
+   s1 <- sample(bci.20x20$S,size=10)
+   s2 <- sample(peic.20x20$S,size=10)
+   ttest.q10[i] <- t.test(s1,s2)$p.value
+ }
```

E verificamos em quantos destes testes a hipótese nula (ausência de diferenças) foi refutada, ao nível de $p < 0.05$:

```
> ##Percentual de testes significativos
> sum(ttest.q10<0.05)/2000*100

[1] 7.9
```

Em menos de 10% das amostragens o teste foi capaz de detectar a diferença de riqueza por quadrícula, o que indica um poder pequeno neste caso. Isto depende da magnitude da diferença real e da variação da riqueza entre quadrículas, e da precisão da estimativa obtida com as amostras. Repita o mesmo procedimento para amostras de 50 quadrículas:

```
> ##Amostras N=50
> ttest.q50 <- c()
> for (i in 1:2000){
+   s1 <- sample(bci.20x20$S,size=50)
+   s2 <- sample(peic.20x20$S,size=50)
+   ttest.q50[i] <- t.test(s1,s2)$p.value
+ }
> ##% de testes significativos
> sum(ttest.q50<0.05/2000*100)
```

Pergunta 2: Qual o efeito do tamanho da amostra sobre o poder do teste? Proponha uma explicação.

2.2 Duas amostras da mesma comunidade

Na seção anterior avaliamos o poder do teste, que é a a probabilidade de detectar uma diferença quando ela existe. Não detectar esta diferença é chamado de **Erro tipo II**. O oposto, chamado **Erro Tipo I**, seria detectar uma diferença quando ela não existe. Isto também pode acontecer, pois duas amostras dificilmente terão o mesmo valor médio de diversidade, mesmo quando tomadas de uma mesma comunidade. Os testes estatísticos nos fornecem a probabilidade de estarmos cometendo este erro, que é a probabilidade da hipótese nula estar correta. Se esta probabilidade for baixa, podemos rejeitar a hipótese nula com pouca chance de estarmos errados. Em biologia, em geral usamos a convenção de que probabilidades abaixo de 0,05 são baixas o suficiente para rejeitar a hipótese nula.

Esta é uma **probabilidade de longo prazo** ou de repetição. Sua interpretação correta é que, se adotamos o nível de significância de 5% e repetimos a amostragem e o teste muitas vezes em situações em que a hipótese nula é correta, em cerca de 5% destas repetições a hipótese nula será rejeitada. Podemos simular esta situação tomando duas amostras de uma mesma comunidade e aplicando o teste t. Repetimos este procedimento muitas vezes e guardamos os valores de p de cada teste, para verificar quantos foram menores que 0,05. Faremos isto com as riquezas de espécies por quadrículas da parcela de BCI, para amostras de 10 e 50 quadrículas. Em cada caso, repetimos a amostragem e testes 2000 vezes:

```
> ##Vetores de resultados
> t.bci.n10 <- c()
> t.bci.n50 <- c()
> ## N=10
> for(i in 1:2000){
+   s1 <- sample(bci.20x20$S,size=10)
+   s2 <- sample(bci.20x20$S,size=10)
+   t.bci.n10[i] <- t.test(s1,s2)$p.value
+ }
> ##N=50
> for(i in 1:2000){
+   s1 <- sample(bci.20x20$S,size=50)
+   s2 <- sample(bci.20x20$S,size=50)
+   t.bci.n50[i] <- t.test(s1,s2)$p.value
+ }
```

E então calculamos ao percentual de testes significativos, ou seja, que tiveram $p < 0.05$:

```
> ##% de testes significativos
> ##N=10
> sum(t.bci.n10<0.05)/2000*100
> ##N=50
> sum(t.bci.n50<0.05)/2000*100
```

Pergunta 3: Qual o efeito do tamanho da amostra sobre a probabilidade de erro tipo II neste caso?

3 Comparação de Amostras com Perfis de Diversidade

Nosso último exercício é a comparação das parcelas BCI e PEIC com perfis de diversidade com o Programa PAST. Na planilha `PerfisDiversidade-PEIC_BCI.xls` vocês encontrarão os dados de abundâncias das espécies para três amostras diferentes. Na primeira coluna estão as abundâncias das espécies em uma área de 10 ha de floresta de restinga (`PEIC_10ha`). Na segunda coluna estão as abundâncias das espécies em uma área de 10 ha de floresta estacional tropical (`BCI_10ha`). Na terceira coluna estão as abundâncias das espécies em uma área de 50 ha da mesma floresta estacional tropical da segunda coluna citada acima (`BCI_50ha`).

Usando o programa PAST, calcule e analise os índices de diversidade das três amostras, a partir do Menu `Diversity:Diversity Indices`. Depois, produza o perfil de diversidade para as três áreas, a partir do `Menu Diversity:Diversity profiles`. Amplie o gráfico e analise cuidadosamente as três curvas produzidas.

ATENÇÃO: No programa PAST, o perfil de diversidade é calculado pela série de Rényi (1961) e não pela série de Hill (1973) e, por esse motivo, no eixo X do gráfico encontra-se o valor de α e não de a. Considerando que a relação entre as duas séries é $H_\alpha = ln(a)$, a interpretação dos perfis é basicamente a mesma que aquela apresentada na aula para a série de Hill.

Caso você queira fazer os perfis no `R` , use a função `renyi`, do pacote `vegan`, e os arquivos `csv` fornecidos:

```
> ## carregue o pacote
> library(vegan)
> ##Perfis
> ##Leitura dos dados
> bci.tot.10 <- read.csv2("bcitot10.csv")
> bci.tot.50 <- read.csv2("bcitot.csv")
> peic.tot <- read.csv2("peictot.csv")
> ##Calculo numeros de Hill
> perf.bci10 <- renyi(bci.tot.10,hill=T)
> perf.bci50 <- renyi(bci.tot.50,hill=T)
> perf.peic <- renyi(peic.tot,hill=T)
> ## Grafico
> plot(as.vector(perf.bci50),type="b",axes=F, xlab="a", ylab="N de Hill")
> axis(side=1,at=1:length(as.vector(perf.bci50)),labels=names(perf.bci50))
```

```
> axis(side=2)
> lines(perf.bci10,type="b",col="red")
> lines(perf.peic,type="b",col="blue")
> legend(x="topright",c("BCI, 50ha","BCI, 10ha","PEIC"),lty=1,col=c("black","red","bl
```

Pergunta 3: Descreva e explique as diferenças observadas entre as três amostras ao longo de todo o perfil de diversidade, dando maior ênfase às diferenças nos extremos do gradiente de α ou a, conforme a escala que tenha usado (Hill ou Renyi).

Referências

Hill, M. O., 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**:427–432.

Rényi, A., 1961. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, volume 1, pages 547–561.