Week 2 Exercise: Local diversity

Z620: Quantitative Biodiversity, Indiana University

November 8, 2014

Overview

In this exercise, we will use R and RStudio to explore aspects of local or site-specific diversity, also known as alpha (α) diversity. We will use the tropical forest dataset from Barro-Colorado Island (BCI), a 1,560-hectare island in the center of the Panama Canal that is administered by the Smithsonian. After setting up our working environment, we will explore the BCI dataset and ways to quantify and examine the α-diversity within individual plots of tropical trees.

RETRIEVE AND SET YOUR WORKING DIRECTORY

```
## [1] "/Users/lisalocey/GitHub/Quantitative_Biodiversity/Assignments/Week2"
```

INSTALL PACKAGES

Ecologists have developed many packages for conducting quantitative analyses in R. You can use the 'help' function to learn more about package installation and add-ons. In this class, we will often use the package called 'vegan', which contains tools commonly used in ecological research, including analyses of (α)-diversity. To install vegan into your R environment type: install.packages("vegan") into your RStudio console.

```r
require("vegan") #You can also type: library("vegan"). The difference being that "require"
```

```
## Loading required package: vegan
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.0-10
```

```r
#loads additional packages (dependencies) that vegan requires for some analyses.
```

LOADING BCI FROM VEGAN

Conveniently, the vegan library contains a version of the BCI dataset, which is obtained by censusing the trees of 50 1-hectare plots every several years. More information on the BCI forest dataset can be found here: http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/.

```r
data(BCI) # Loading the BCI dataset
```

EXPLORING THE SITE-BY-SPECIES MATRIX

In vegan, the BCI data are organized into a **site-by-species matrix**, that is, a data table where the abundances of all 225 tree species (columns) are given for each of the 50 sites (rows). Consequently, this table should have 50 rows (sits) and 225 columns (species). Let's verify this using the dimension or "dim" function:

```
dim(BCI)   # Prints the dimensions of the site (row) by species (column) matrix
```

```
## [1]  50 225
```

The **site-by-species matrix** is one of the most basic data structures used by ecologists. It contains the abundances (or presence and absence) of all species found among a given number of sites, quadrats, transects, etc. With the BCI site-by-species matrix loaded, let's print the abundances of 4 species found in the first 8 sites.

```
BCI[1:8, 14:17] # print abundances for sites (rows) 1 to 8 and for species (columns) 14 to 17
```

```
##    Andira.inermis Annona.spraguei Apeiba.aspera Apeiba.tibourbou
## 1              0               1            13                2
## 2              0               0            12                0
## 3              0               1             6                1
## 4              0               0             3                1
## 5              1               0             4                0
## 6              1               0            10                0
## 7              0               0             5                0
## 8              0               1             4                1
```

Here, we can see that Cabbage Bark (*Andira inermis*) is absent from six of the eight sites and is only found as a single individual in two sites. On ther other hand, *Apeiba aspera* (locally known as Monkey Comb) is found at all eight sites and is relatively much more abundant than any of the other three species that we selected.

EXPLORING LOCAL DIVERSITY

So far, you have been introduced to an important data structure, the site-by-species matrix. You have learned to print its dimensions and to print the abundances of particular species (columns) found in specific sites (rows). Now, let's focus on a specific site by selecting the first row and assigning it to a variable called "Site1":

```
Site1 <- BCI[1,] # assign the first row (site) to the variable Site1
dim(Site1) # print the dimensions of Site1 to the screen
```

```
## [1]    1 225
```

You can see that Site1 has one row (it is only 1 site) and 225 columns (225 potentially present species). Remember that the BCI site-by-species matrix also accounts for absences, i.e., with zeros. Let's find out how many species actually occupied Site1 when it was censused:

```
S <- specnumber(Site1) # Find the number of species in Site1 and assign it to a varible S1
cat('There are', S, 'species at Site1') # "cat()" concatenates strings and numbers
```

```
## There are 93 species at Site1
```

Now we know that the recorded species richness (usually denoted as S) of Site1 is 93. Local species richness is the most basic aspect of diversity and is, in fact, what we usually refer to as α-diversity. Beyond knowing the number of species, we are often interested in knowing how many individuals were found among them. That is, the number of individuals recorded at Site1.

```
N <- sum(Site1) # find the number of individuals at Site1 and assign it to a variable N1
cat('There are', N, 'individuals among', S, 'species at Site1') # Mario, feel free to edit this as you
```

```
## There are 448 individuals among 93 species at Site1
```

Now we know the recorded species richness (S) and recorded total abundance (usually denoted as N) of Site1 in the BCI site-by-species matrix. N and S are two of the most common pieces of data we obtain when sampling ecological communities. Finally, we are beginning to quantify our data!

The species abundance distribution

A next natural step in exploring aspects of local diversity is to ask how abundance varies among species. One of the most common ways to visually explore this relationship is to rank the species from greatest to least, that is, as a rank-abundance curve (RAC). Let's plot the RAC for Site1.

Let's begin by removing the zeros from Site1. This removes the species that, while in the BCI site-by-species matrix, were not found at Site1.

```
Site1 <- Site1[ !Site1 %in% c(0) ] # Remove all zeros from Site1 and reassign the new vector
```

Now, we can plot the RAC, accounting only for the species that were recorded at Site1.

```
Site1 <- Site1[order(Site1, decreasing=TRUE)] # Rank the taxa by abundance
RAC <- as.vector(Site1, mode='numeric') # Convert Site1 from a data frame to a vector (y-values)
ranks <- as.vector(seq(1, S), mode='numeric') # Create a sequence of S ranks (x-values)
plot(ranks, log(RAC), type='p', xlab="Rank", ylab="log(abundance)") # plot the RAC
```
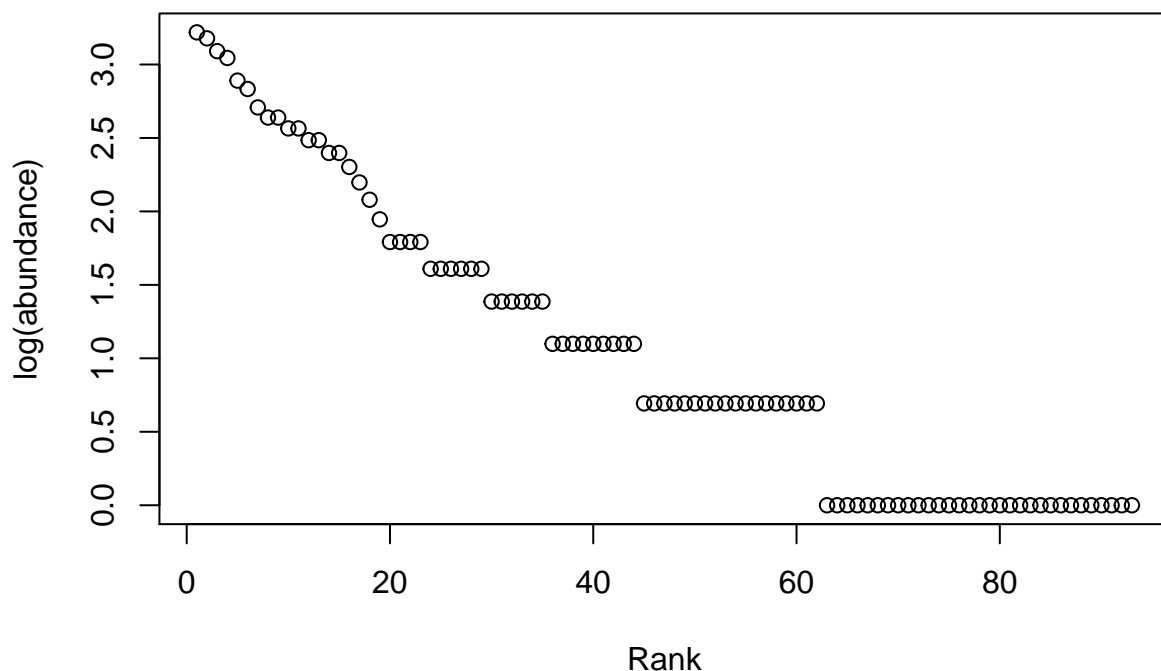
Looking at the RAC for Site1, we can see that abundance is distributed unevenly among species, even when abundances are log-transformed. In fact, few species have more than 10 individuals and most species have less than 5. This sort of uneven distribution of abundance among species is one of the most ubiquitous patterns in ecology and has provoked a long history of study, theories, and explanations (McGill et al. 2007).

You may wonder whether all sites (rows) in the BCI site-by-species matrix also have highly uneven RACs. But, before moving on to other sites, let's explore aspects of local diversity at Site1 more deeply, that is, by quantifying species diversity in ways other than richness, and by quantifying evenness, dominance, and rarity among the species recorded at Site1.

Species Evenness

Species evenness is generally defined as the similarity in abundance among species. Take for instance, the RAC for Site1. While there are many species with similar but low abundances, the few species with relatively high abundances actually drives the pattern to be uneven and, hence, the slope of the RAC to be steep. In fact, one way of quantifying evenness is simply to find the slope of regression through the points in RAC. This is referred to as the NHC evenness index (Nee et al. 1992).

```
fit  <- lm(log(RAC) ~ ranks)
summary(fit)
```

```
##
## Call:
## lm(formula = log(RAC) ~ ranks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48067 -0.20508 -0.07006  0.21552  0.57713
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.702031   0.056952   47.44   <2e-16 ***
## ranks       -0.035260   0.001052  -33.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2724 on 91 degrees of freedom
## Multiple R-squared:  0.925,  Adjusted R-squared:  0.9242
## F-statistic:  1123 on 1 and 91 DF,  p-value: < 2.2e-16
```

```
plot(ranks, log(RAC), type='p', xlab="Rank", ylab="log(abundance)") # plot the log-transformed RAC
abline(fit, col = 'red')
```
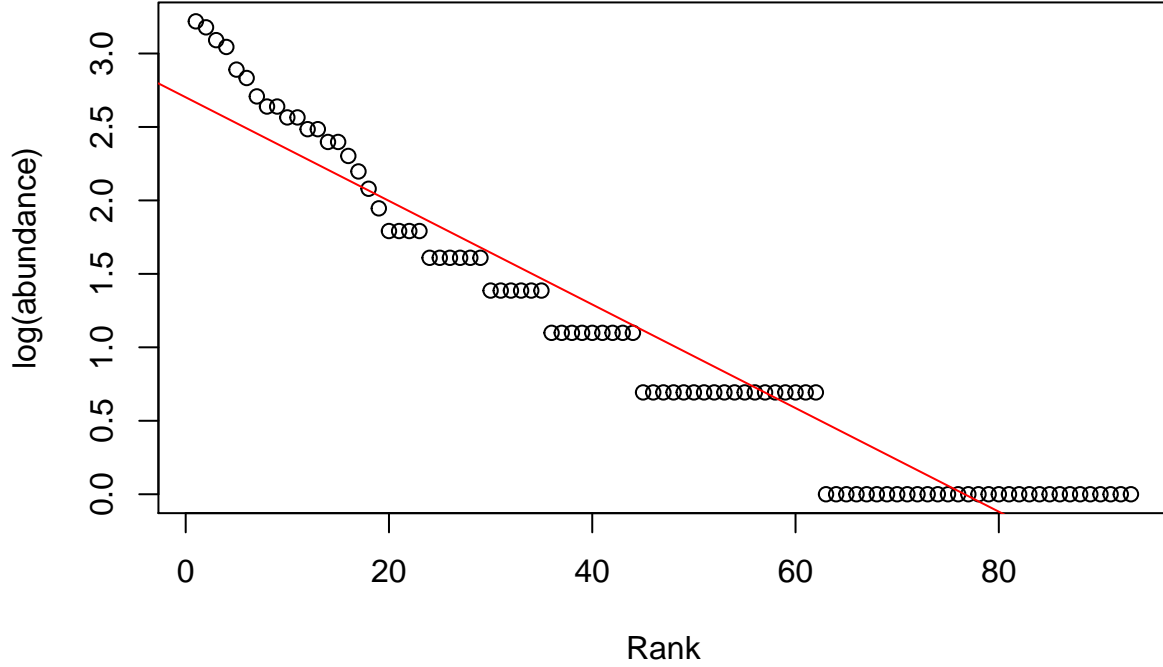
We can see that the slope is nearly -0.035 with a good overall fit. Note, that the slope of this relationship must be either 0.0 or negative because the RAC is constrained to between perfect evenness (slope of 0) and infinite unevenness (negative infinity). Also, notice that while a slope of 0 implies a perfectly even distribution of abundance among spcies, a slope of -0.035 really doesn't communicate much. Likewise, the number of ranks (i.e. species) influences the slope by 'stretching' out the relationship. Imagine what the slope would be for $S = 100{,}000$ ranks!

Desired properties of evenness

Shortcomings of the NHC index reveal important considerations when quantifying species evenness. First, we want to be able to intuit the values of our metric. Second, we want our metric to be bounded between a minimum (e.g. 0.0) and a maximum (e.g. 1.0). Third, we do not want evenness among species to simply reflect the number of species (i.e. S); otherwise, comparing evenness among sites with different numbers of species really becomes a comparison of species richness; not evenness *per se*. We also don't want species evenness to simply reflect differences among the most abundant or among the most rare species. In short, we desire an intuitive, bounded, and largely unbiased metric of species evenness.

Seveal well-known metrics of evenness have been derived (see Smith and Wilson 1996) that range between 0 (no evenness) and 1 (perfect evenness), are largely independent of the number of species (i.e. S), and are more or less unbiased towards either very abundant species or very rare species. Here, we will introduce four: Simpson's evenness, Smith and Wilson's evenness index, Heip's evenness index, and Shannon's evenness.

Shannon's (or Pielou's) evenness

Shannon's evenness metric is derived from Shannon's information entropy, a measure of uncertainty, and is standardized by S to remove the bias of S on the statistic. Shannon's evenness is calculated as: $J' = \frac{H'}{ln(S)}$,

Simpson's evenness

Simpson's evenness is calculated as $E_{\frac{1}{D}} = \frac{1}{S} \sum \frac{N(N-1)}{n_i(n_i-1)}$, where S is species richness, N is total abundance, i is the ith species. Let's use the vegan package to calculate Simpson's evenness for Site1:

Dominance and Rarity

Ken hasn't touched from here on down

Calculate Shannon index. 'Margin = 1' means diversity is calculated row-wise; 'Margin = 2' means diversity is calcualted column-wise.With base = exp(1), we are esimating Shannon's index using the natural logrithm of each taxon's relative abundance using the equation; $H' = -\sum(p_i \times ln(p_i))$
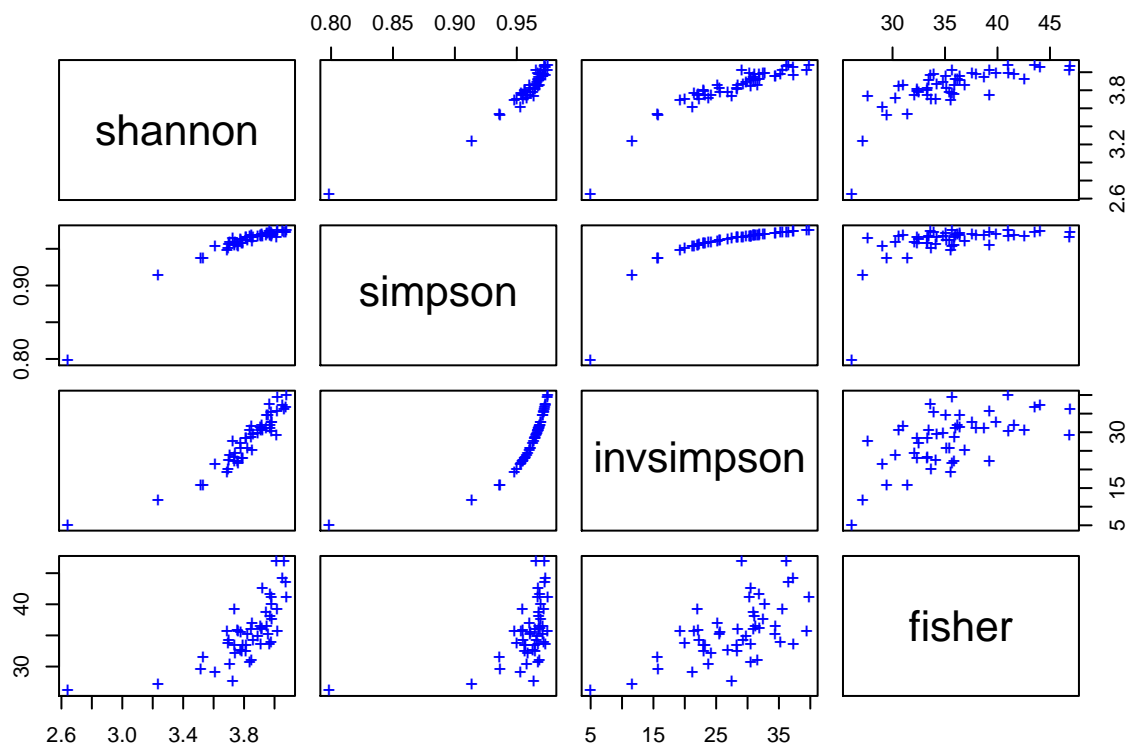
```
shannon <- diversity(BCI, index="shannon", MARGIN = 1, base=exp(1))
```

Can also calculate other diversity metrics

```
simpson <- diversity(BCI, "simpson")
invsimpson <- diversity(BCI, "inv")
fisher <- fisher.alpha(BCI)
```

Let's plot **Pairs is not a normal plotting tool, introduce first**

```
pairs(cbind(shannon, simpson, invsimpson, fisher), pch="+", col="blue")
```
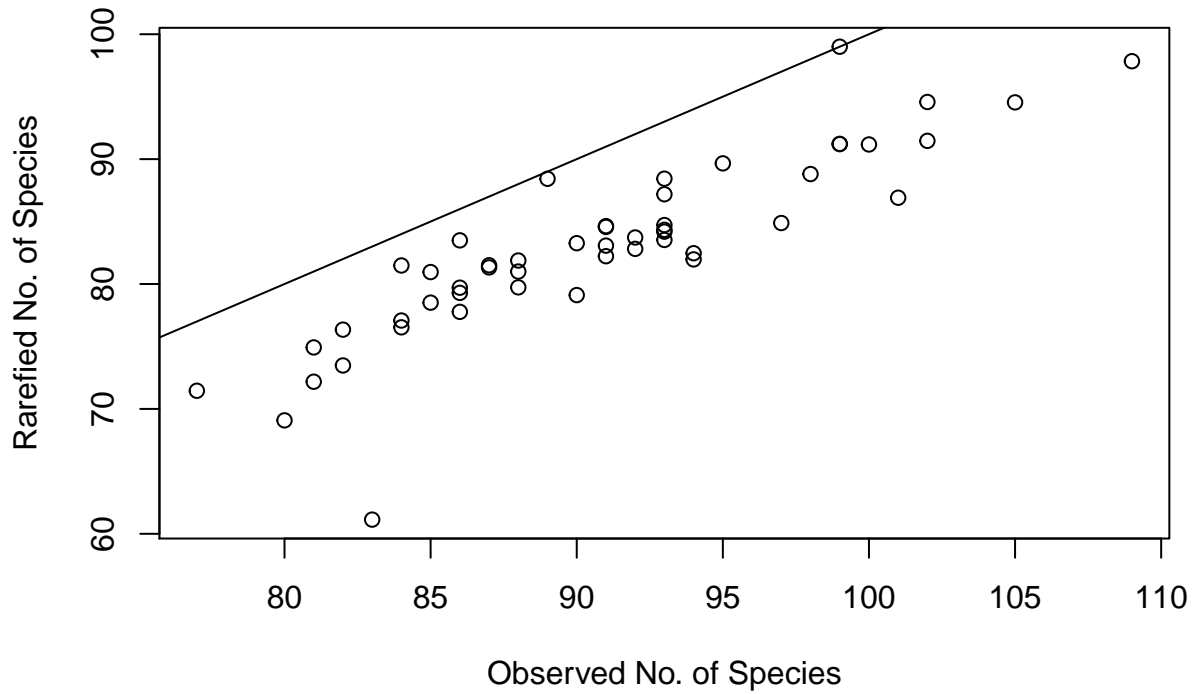
Species richness
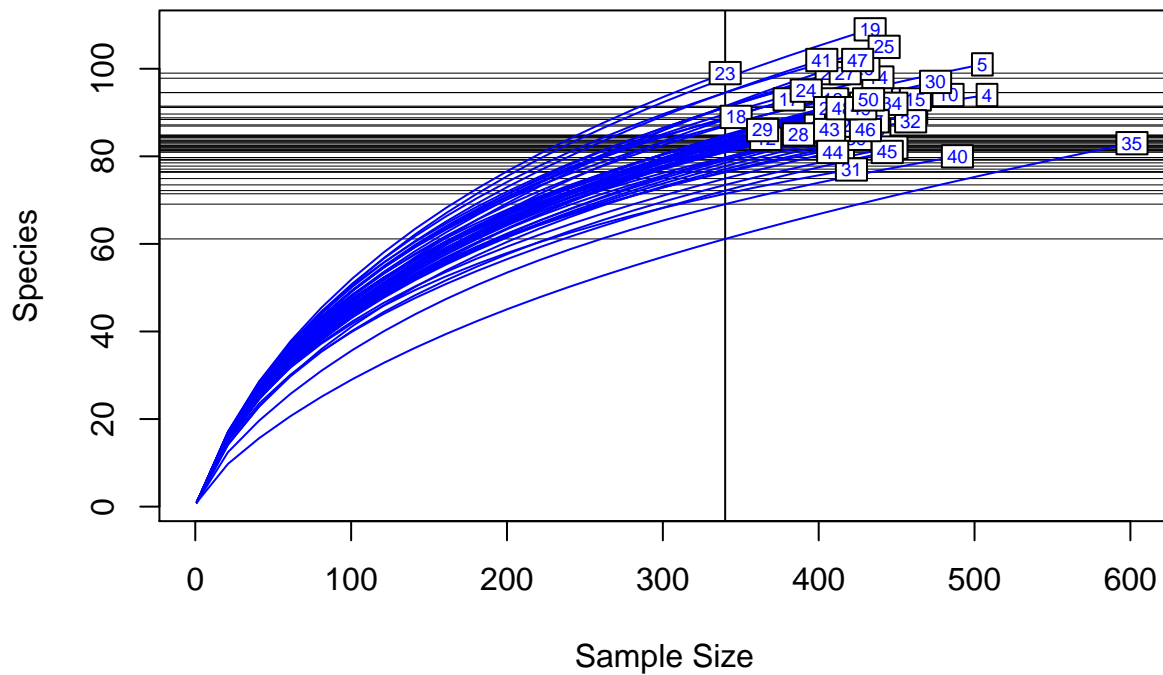
```
S <- specnumber(BCI)
```

Rarefaction

```
(raremax <- min(rowSums(BCI)))
```

```
## [1] 340
```

```
Srare <- rarefy(BCI, raremax)
plot(S, Srare, xlab = "Observed No. of Species", ylab = "Rarefied No. of Species")
abline(0, 1)
```



```
rarecurve(BCI, step = 20, sample = raremax, col = "blue", cex = 0.6)
```

Example: http://www.jennajacobs.org/R/rarefaction.html

Calculating relative abundances

```
BCI_t <- t(BCI)

BCIrel <- BCI_t
for(i in 1:ncol(BCI_t)){
  BCIrel[,i] = BCI_t[,i] / sum(BCI_t[,i])
  }
```

What's one way to test that this worked?

```
colSums(BCIrel)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

HOMEWORK

1) Recreate this exercise
2) Redo the regression and ANOVA with log10-transformed zooplankton biomas data (often, log-transformations help with meeting assumptions of normality and equal variance)
3) use knitr to create a pdf, push it to GitHub, and create a pull request