Week 2 Exercise: Local Diversity

Z620: Quantitative Biodiversity, Indiana University

January 23, 2015

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). BCI is a 1,560-hectare island in the center of the Panama Canal that is administered by the Smithsonian Trophcal Research Institution. We will explore the BCI dataset and ways to quantify and examine the α-diversity within individual plots of tropical trees.

Initial Setup

Retrieve and Set Your Working Directory

```
getwd() # prints the working directory
setwd("~/GitHub/QuantitativeBiodiversity/Assignments/Week2")
```

Install Packages

Recall, ecologists have developed many packages for conducting quantitative analyses in R. In this excercise, we will rely heavily on 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 the RStudio console. Alternatively, we will commonly use the following line of code. This code tries to load a packages (via `require()`), but if not found it will install and then load.

```
require("vegan")||install.packages("vegan");require("vegan")
```

```
## Loading required package: vegan
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.2-0
```

Loading BCI Data 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/. <– **Jay, we cannot make everything a link if we are planning on providing handouts** Or, `help(BCI)` for specific details about the BCI dataset in 'vegan'. Let's load the data:

```
data(BCI)
```

Exploring the Site-By-Species Matrix

In vegan, the BCI data are organized into a **site-by-species matrix**, that is, a data frame where the abundances of all 225 tree species (columns) are given for each of the 50 sites (rows). Consequently, this data frame should have 50 rows (sites) 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) data frame
```

```
## [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.glabra 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 indexed in the data frame.

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
# the prefered notation for names is all lowercase.
# I know that I don't always do this, but a good habit to learn.
# Accepted identifiers like "S" would be an exception {MEM}
dim(Site1) # print the dimensions of Site1 to the screen
```

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

You can see that Site1 consists of 225 potentially present species (columns). Remember that **<–However? {MEM}** 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 S
cat('There are', S, 'species at Site1') # "cat()" concatenates strings and numbers --> Is this necessary
```

```
## 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** –> needs to be clearer {JTL}]. 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')
```

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

```
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! And from this, we can gain insight into some other important attributes of biodiversity.

The Species Abundance Distribution

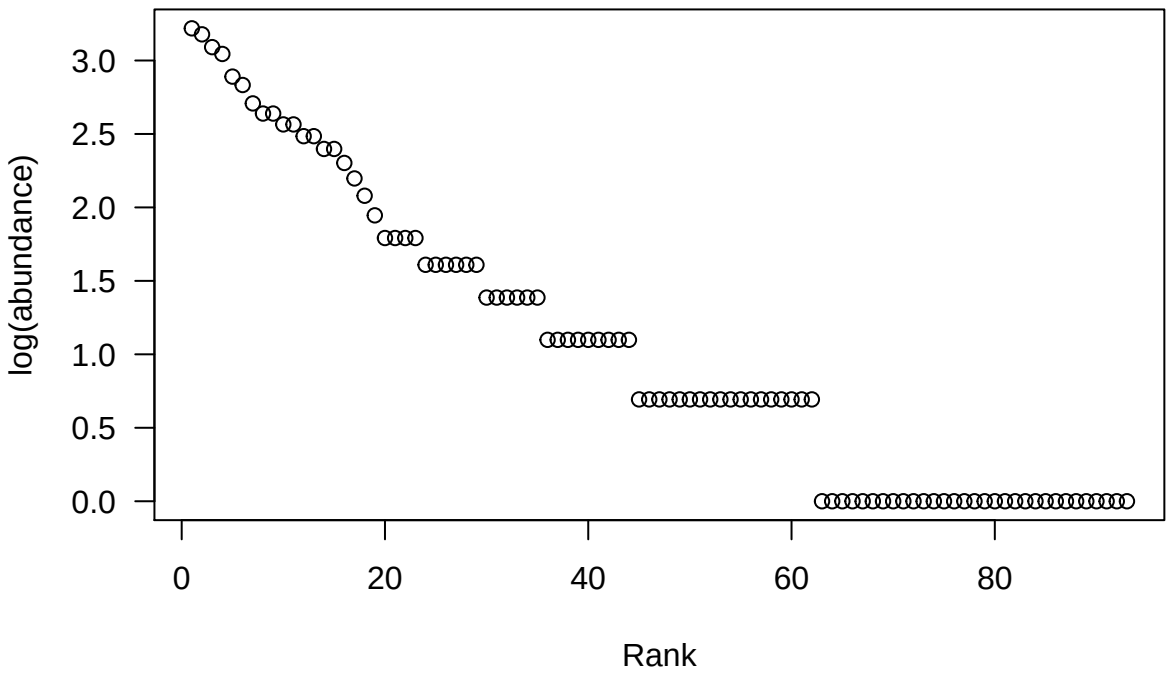
A next natural step in exploring aspects of local diversity is to ask how abundance varies among species. One of the most common ways to visually explore this relationship is to rank the species from greatest to least, that is, as a rank-abundance curve (RAC). Let's plot the RAC for Site1.

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 **Should w
Site1b <- subset(Site1, select = Site1 > 0) # Better use of R commands {MEM}
```

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 into a vector
ranks <- as.vector(seq(1, S), mode = 'numeric') # Create a sequence of S ranks
plot(ranks, log(RAC), type = 'p', las = 1,      # Plot the RAC
     xlab = "Rank", ylab = "log(abundance)")    # What do the commands mean?
```



**In the above code you are using mode="numeric" which is acutally something I would usually ignore. But don't you want this to actually be an integer anyways? Just a comment. However, if not needed we could just remove. R is good about not needing the user to spell out everything like this {MEM}*

Looking at the RAC for Site1, we can see that abundance is distributed unevenly among species, even when abundances are **prefer natural log or log10** log-transformed. In fact, few species have more than 10 **use log10?** 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 we try to answer this question, 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.

—JTL Stopped here

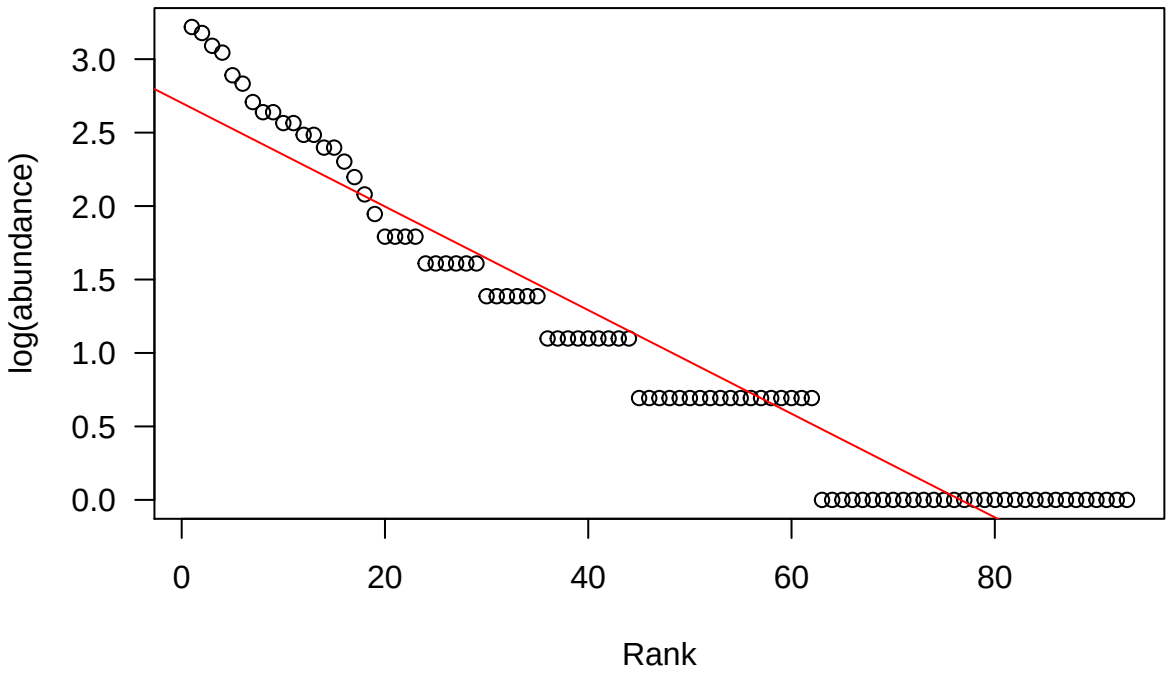
Species Evenness

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)                # Simple Linear Regression
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', las = 1,
     xlab = "Rank", ylab = "log(abundance)")
abline(fit, col = 'red')                    # Add Regression Line
```



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.

1. We want to be able to intuit the values of our metric.
2. We want our metric to be bounded between a minimum (e.g. 0.0) and a maximum (e.g. 1.0).
3. 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 three: Simpson's evenness, Smith and Wilson's evenness index, and Shannon's evenness. **Not the order you present them; also, that is a long sentence {MEM}**

Shannon's (or Pielou's) evenness, J'

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 (J') is calculated as: $J' = \frac{H'}{ln(S)}$, where H' is Shannon's information entropy and where p_i is the relative abundance of the ith species. Shannon's evenness was derived by Pielou (1969, 1975) and is sometimes referred to as Pielou's evenness.

The vegan package does not have functions to calculate Shannon's evenness but does have a function to calculate Shannon's entropy, sometimes referred to as Shannon's diversity, the Shannon-Weiner diversity index, among others.

```
H <- diversity(Site1, index="shannon") # find Shannon's entropy (H')
J <- H/log(S) # find Shannon's evenness (J')
J
```

```
##         1
## 0.8865579
```

You can see that Shannon's evenness is quite high for Site1. Yet, this doesn't intuitively reflect what we saw in the RAC. In fact, because of the way Shannon's is calculated, values for ecological communities are commonly greater than 0.5. Consequently, an intermediate value of J' does not necessarily mean an intermediate evenness. Compare this to another evenness metric, i.e., Simpson's evenness.

Simpson's evenness, $E_{1/D}$

Simpson's evenness metric essentially reflects the sample variance of the SAD, and is calculated as $E_{1/D} = \frac{1}{S} \sum \frac{N(N-1)}{n_i(n_i-1)}$, where S is species richness, N is total abundance, i is the ith species.

Once, again vegan has no function for Simpson's evenness but does calculate Simpson's diversity, i.e., the probability that the next sampled individual belongs to a different species:

```
D <- diversity(Site1, "simp") # find Simpson's diversity (D)
E <- (1/D)/S # find Simpson's evenness (E)
E
```

```
##          1
## 0.01103259
```

Here, we can see that Simpson's evenness reveals that species evenness for Site1 is quite low, i.e., > 0.1. However, Simpson's evenness has also been criticized for being biased towards the most abundant taxa, i.e., revealing differences in abundances among the first few ranks and is less influenced by the many similarly abundant but rarer species. Let's examine the value of evenness for Site1 using a different and purportedly less-biased evenness metric, E_{var}.

Smith and Wilson's evenness index, E_{var}

Smith and Wilson (1996) reviewed metrics of species evenness. Some of their conclusions about the desired characteristics of species evenness are reflected in this assignment. They derived a robust measure of species evenness called (E_{var}). This metric represents the sample variance of logarithmic abundances. Abundances are log-transformed to decrease bias towards very rare or very abundant species. E_{var} is independent of S and is standardized to take values between 0 (no evenness) and 1 (perfect evenness). The metric is calculated as: $E_{var} = 1 - \frac{2}{\pi \cdot \arctan(\sum_{i=1}^{i=S} ln(n_i) - \sum_{j=1}^{j=S} ln(n_j)/S)}$.

While considerably more involved to calcualte, E_{var} really reduces to finding the sample variance of the log-transformed species abundances and then standardizing it to a value between 0 and 1.

```
X <- var(log(RAC))
Evar <- 1 - (2/pi)*atan(X) # these operations make the value of Evar range between 0 and 1
Evar # print Evar for Site1
```

```
## [1] 0.5067211
```

We can confirm this with a more explicit R chunk:

```
P <- log(RAC) # log-tranform the abundances of the RAC and assign them to a vector P
AvgAb <- mean(P) # find the average of the log abundances
X <- 0 # declare a scalar variable X
Evar <- 0 # declare a scalar varible Evar

for (x in P) {
  X = X + (x - AvgAb)^2 / (S - 1)
}

Evar = 1 - (2/pi)*atan(X) # these operations make the value of Evar range between 0 and 1
Evar # print Evar for Site1
```

```
##         1
## 0.5067211
```

As you can, the value of E_{var} suggests an intermediate value of species evenness. Likewise, you can see that different evenness metrics, even when bound between 0 and 1 still return very different values. This is not because each is inaccurate but because each metric emphasizes different aspects of the SAD (or RAC). Where Simpson's evenness focuses more greatly on dominant species, E_{var} is nearly as influenced by the abundance of dominant species as that of rare ones. In effect, the metric of evenness one used, even if only using NHC, depends on the aspects of similarity in species abundances one prefers to emphasize.

Species Diversity

You saw in the derivation of Shannon's evenness and Simpson's evenness that we called the R diversity function, and then specified either "shannon" or "simp". In effect, in estimating Shannon's or Simpson's evenness we first estimated Shannon's diversity (H') and Simpson's diversity (D)! Species diversity is commonly defined as the relationship between species richness and species evenness, often of a form similar to $Diversity = \frac{Evenness}{Richness}$. Can you see this in looking back at how we calculated Shannon's and Simpson's evenness? Here, we will estimate popular indices of diversity (Shannon's, Simpson's, Fisher's alpha) using our own derivation and then check this against vegan's estimates.

Shannon's diversity (or entropy)

Shannon's diversity metric is really just Shannon's information entropy, a measure of uncertainty. This metric is calculated as $H' = -\sum p_i ln(p_i)$. Let's calculate Shannon's diversity for Site1 and then compare it to vegan's estimate:

```
H <- 0
for (sp in RAC){
  p = sp / sum(RAC)
  H = H - p*log(p)
}
H
```

```
## [1] 4.018412
```

```
diversity(RAC, index="shannon")
```

```
## [1] 4.018412
```

Are they the same? {MEM}

Simpson's diversity (or dominance)

Simpsons diversity is a straightforward metric and is calculated as $D = \sum p_i^2$ where p_i is the proportion of individuals found in the ith species. Simpson's index is often expressed as $1/D$ or 1-D, so that diversity naturally increases with $1/D$. Let's calculate Simpson's diversity for Site1 and then compare it to vegan's estimate:

```
D <- 0.0
N <- sum(RAC)
for (ni in RAC){
  D = D + (ni*ni)/(N*N)
}
```

```
invD <- 1/D # using the 1/D variation
invD
```

```
## [1] 39.41555
```

```
invD <- diversity(RAC, "inv") # using vegan
invD
```

```
## [1] 39.41555
```

```
D <- 1 - D # using the 1 - D variation
D
```

```
## [1] 0.9746293
```

```
D <- diversity(RAC, "simp") # using vegan
D
```

```
## [1] 0.9746293
```

Fisher's α

R.A. Fisher (1943) derived one of the first and most successful models for how abundance varies among species, i.e., the log-series distribution. This model has only a single fitted parameter, i.e., α, which is implicitly defined by the log-series model: $S = \alpha * ln(1 + n/\alpha)$. However, this does not tell us exactly how to estimate α, and this is because α is a fitted parameter. Later, we will explore the log-series. For now, let's concentrate on the Fisher's α, which has often been used as a diversity metric and is the root of 'alpha' diversity. Fisher's α, according to the authors of the vegan package is asymptotically similar to inverse Simpson's. Let's compare using Site1.

```
invD <- diversity(RAC, "inv")
invD
```

```
## [1] 39.41555
```

```
Fisher <- fisher.alpha(RAC)
Fisher
```

```
## [1] 35.67297
```

As we can see, the two measurements are somewhat similar. They would converge if our community was much greater in total abundance and richness. However, discussion of Fisher's α introduces a new concept, that is, of estimating diversity instead of just calculating a diversity metric. The difference being that an estimate of diversity implicitly or explicitly accounts for samplign error, that is, the fact that when samplign most ecological communities that we are not observing every single individual.

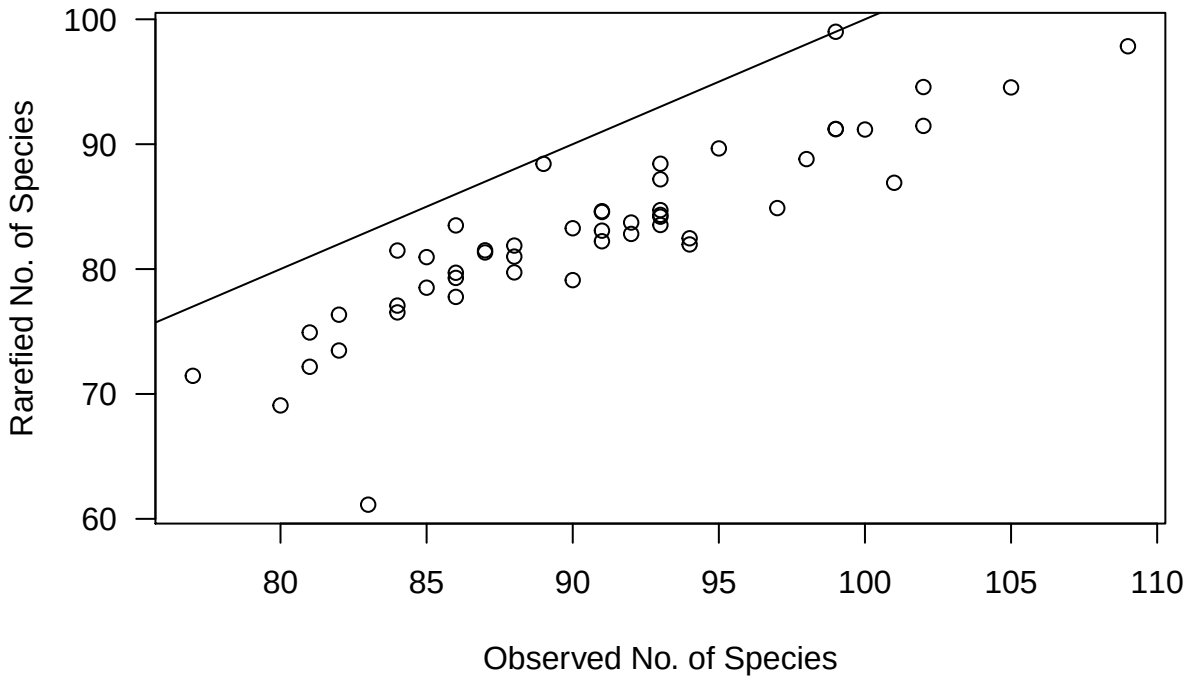
Estimating Richness and Diversity

Rarefaction

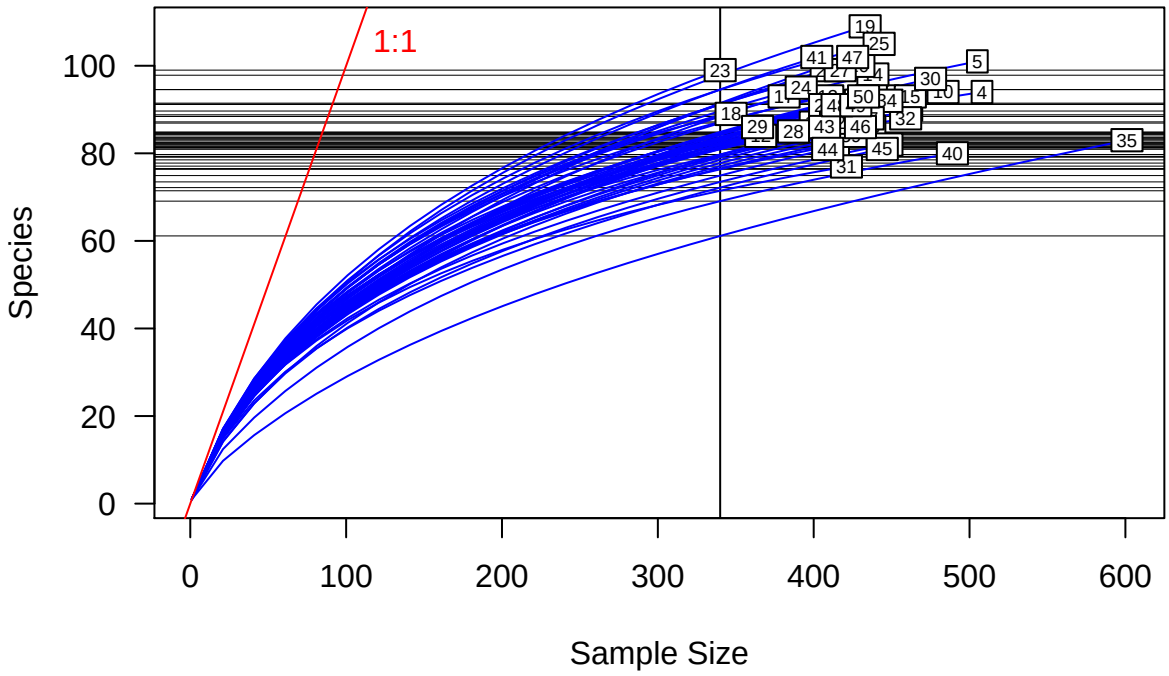
```
S <- specnumber(BCI) # <- You needed to add this so that you had S for each site {MEM}
(raremax <- min(rowSums(BCI))) # <-- Why the parentheses??? {MEM}
```

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

```
Srare <- rarefy(BCI, raremax)
plot(S, Srare, las = 1,
     xlab = "Observed No. of Species", ylab = "Rarefied No. of Species")
abline(0, 1)                  # Addes line with interercep = 0, slope = 1
```




```
rarecurve(BCI, step = 20, sample = raremax, col = "blue", cex = 0.6, las=1)
abline(0, 1, col = 'red')
text(105, 105, "1:1", pos = 4, col = 'red')
```



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

Other Diversity Estimators

Homework

1) As stated by Magurran (2004) the $D = \sum p_i^2$ derivation of Simpson's D really only applies to communities of infinite size. For anything but an infinitely large community, Simpson's index is calculated as $D = \sum \frac{n_i(n_i-1)}{N(N-1)}$. Calculate Simpson's D, 1 - D, and Simpson's inverse (i.e. 1/D) for Site1.

2) ...

3) use knitr to create a pdf, push it to GitHub, and create a pull request