Week 2 Exercise: Local Diversity

Z620: Quantitative Biodiversity, Indiana University

January 23, 2015

Overview

In this exercise, we will explore aspects of local or site-specific diversity, also known as alpha (α) diversity. After introducting one of the primary ecological data structures – the **site-by-species matrix** – we will quantify two of the fundamental components of (α) diversity: **richness** and **evenness**. From there, we will then discusss ways to integrate richness and evenness, which will include univariate metrics of "diversity" along with an investigation of **species abundance distributions (SAD)**.

1.) Setup

Retrieve and Set Your Working Directory

```
rm(list=ls()) # removes all variables from your workspace
getwd() # prints the working directory
setwd("~/GitHub/QuantitativeBiodiversity/Assignments/Week2") # sets the working directory
```

Install Packages

In this excercise, we will rely heavily on a **contributed R package** called `vegan`, which contains tools commonly used in ecological research, including analyses of α-diversity. Jari Oksanen has created an excellent tutorial that provides an overview of the `vegan` package: http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

Let's install the `vegan` package and its dependencies:

```
#install.packages("vegan") # use this if running for the first time
require("vegan") # otherwise, use this command
```

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

2.) Loading Data

We will start by using 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 Tropical Research Institution (http://www.stri.si.edu/english/research/facilities/terrestrial/barro_colorado/). Conveniently, the `vegan` library contains a version of the BCI dataset. The dataset is a census of all trees of at least 10cm in diameter at breast height (DBH) that occur in 50 one-hectare plots. More information on the BCI forest dataset can be found elswhere (http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/). You can also learn more about the BCI dataset associated with the 'vegan' package by typing `help(BCI)` in the command line. Let's load the BCI data using the `data` function:

```r
data(BCI)
```

3.) Exploring The Site-By-Species Matrix

The **site-by-species matrix** is a primary ecological data structure that contains abundances, relative abundances, or the presence of species (or other taxonomic units) observed at different sampling sites.

	Species1	**Species2**	**Species3**	**Species4**	...
Site1	90	76	1	0	
Site3	86	47	0	123	
Site3	23	89	0	46	
...					

Each site corresponds to a row in the site-by-species matrix, while each species corresponds to a column. Throughout the course, we will draw inferences about aspects of diversity from the site-by-species matrix.

The version of the BCI site-by-species matrix provided in `vegan` contains the abundances of 225 tree species (i.e. 225 columns) for each of 50 sites (i.e. 50 rows). While it is conventional to refer to a matrix or table by its rows and then by its columns, each row could just as easily be a species and each column could be a site, i.e., a species by site matrix. Some programs that you encounter will prefer data to be in one format or the other. In such case, you can use the transpose function (`t()`) to modify the site-by-species matrix.

Let's verify the structure of the BCI site-by-species matrix using the `dim` function, which was introduced last week:

```r
dim(BCI)  # Prints the dimensions of the site (row) by species (column) data frame
```

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

With the BCI site-by-species matrix now loaded, let's print the abundances of four species found in the first eight sites using the indexing tools that were introduced last week.

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

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 (sites 5 and 6). 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.

Question 1: Makes some additional observations about the occurrence and abundance of species in the samples you've indexed.

> ***Answer 1***:

4) Species Richness

Species richness (S) is simply the number of species in a system or the number of species observed in a sample. Species richness is the most basic aspect of diversity. In fact, it is usually what most people are referring to when they talk about α-diversity. Calculating species richness for a sample is often straightforward, i.e., count the number of species present in a sample. However, estimating species richness for a community from an incomplete sample requires assumptions about the nature of the sampling effort (e.g. biases, coverage). Consequently, estimating richness is an area of active research and, as you'll see, there are several ways to estimate richness that attempt to account for the number of species that were not detected.

Observed Richness

The simplest way to calcuate species richness is to just add up the number of species that were detected in a sample of a site. Let's calculate species richness using this approach for one of the BCI sites. Using some of the R basics from last week, let's: 1. assign the first row to a variable called "site1", and 2. check the dimensions of this new vector

```
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
```

While there are 225 species in the BCI dataset, and hence, 225 columns in the BCI site-by-species matrix, each of the 225 species was not likely present at each site. Consequently, it is important to note that site-by-species matrices account for absences (i.e., zero occurences). Let's write a function that calculates observed species richness of a site, i.e. a row in the site-by-species matrix, but ignore the 0's, i.e. absences.

```
S.obs <- function(x){
  rowSums(x>0)*1
  }
```

The writing and use of functions is central to programming. The basic concept behind functions is to define a piece of code that operates on one or more variables. So, instead of repeatedly rewriting code to calculate richness, we define a function called `S.obs` that calculates richness of a given vector (i.e., `x`). We can then call the function by typing `S.obs()` and placing our vector within the parentheses. There is also a function in the `vegan` package called `specnumber` that also calculate observed richness.

Question 2: Does `vegan` return the same value for observed richness of `site1` as our function `S.obs`?

> ***Answer 2***:

But How Well Did You Sample Your Site?

Accurate estimates of richness are influenced by sampling effort and biases. Even when the sampling effort is un-biased, the more individuals that are censused, the more likely you are to encounter new species. One index that provides an estimate of how well a site is sampled is known as **Good's Coverage (C)**: C = $1 - \frac{n_1}{N}$ Here, n_1 is the number of singleton species (species only detected once) and N is the total number of individuals in the sample. Examining the equation for Good's C reveals that the fraction is simply the portion of N represented by singleton species. Subtracting this from 1 give the portion of N belonging to species sampled more than once. By this formulation, Good's C assumes that no species in the sampled environment can actually be represented by a single individual, i.e., singletons reflect a sampling artifact and Good's C reveals the magnitude of the sampling artifact.

Question 3: What would we conclude if n_1 equaled N?

> ***Answer 3***:

Let's write a function and estimate Good's Coverage for site1 of BCI:

```
C <- function(x){
  1 - (sum(x==1)/rowSums(x))
  }
```

Question 4: a.) Have the researchers at BCI done a good job of sampling `site1`? b.) What is the range of values that can be generated by Good's Coverage? c.) What portion of taxa in `site1` were represented as singletons?

> ***Answer 4a***:
> ***Answer 4b***:
> ***Answer 4c***:

Estimating Richness

There are few systems on earth that have been better surveyed than the trees at BCI. For most ecological systems, sample size is much smaller than **N** and many taxa can easily go undetected. To address this question, we are going to introduce a new data set. This dataset is derived from bacterial 16S rRNA gene sequences, which were collected from multiple plots at the KBS Long-Term Ecological Site(http://lter.kbs.msu.edu/). Even though we obtained a fair number of sequences from each sample, soil bacteria are abundant and thought to be some of the most divese communities on earth.

Let's load this microbial dataset and print the abundances of operational taxonomic units (OTU's) recovered from the first plot.

```
soilbac <- read.table("data/soilbac.txt", sep = "\t", header = TRUE) # read in the data
soilbac <- as.matrix(t(soilbac), mode='numeric') # transform the data to a matrix and transpose, using
soilbac <- matrix(soilbac, ncol= ncol(soilbac), nrow=nrow(soilbac), dimnames = NULL) # remove row and c
soilbac1 <- soilbac[2,] # row 1 is a sample taken from an agricultural site
```

Question 5: a.) How many sequences did we recover from the sample `soilbac1`, i.e. the row total? b.) What is the observed richeness of `soilbac1`? c.) How does coverage compare between the BCI sample (`site1`) and the KBS sample (`soilbac1`)?

> ***Answer 5a***:
> ***Answer 5b***:
> ***Answer 5c***:

There are number of statistical techniques developed to estimate richness based on the coverge of species in a sample (e.g., Hughes et al. 2001 – http://aem.asm.org/content/67/10/4399). Here, we will highlight two of the most common richness estimators (i.e. Chao1 and Chao2) and reveal how richness can be estimated from species abundances within a single sample (i.e. via Chao1) or from species occurrences (or incidence or presence/absence) across multiple samples (i.e. via Chao2). These richness estimators are named after the person who derived them, i.e. Professor Anne Chao, and both are non-parametric. By non-parametric we mean they are not based on an underlying distribution (e.g. Normal distribution, Student's t-distribution) and hence, make few statistical assumptions about the underlying nature of the true 'population'.

Chao1 is abundance-based and useful for examining richness of a single site. It is calculated using observed richness (`S.obs`), the observed number of **singletons** (species with an observed abundance of 1), and the observed number of **doubletons** (species with observed abundance of 2). Consequently, this estimator cannot be used on relative abundance and tends to be used in communities with many low-abundance taxa. Let's write a function for Chao1:

```
S.chao <- function(x){
  S.obs(x) + (sum(x==1)^2)/(2*sum(x==2)) ### should we use which() and length() here?
  }
```

Note, estimated richness by definition must be greater than observed richness.

Chao2 is incidence-based and useful for examining richness across multiple sites. Like Chao1, Chao2 is calculated using observed richness (`S.obs`). However, in Chao2, **singletons** and **doubletons** refer to species observed once and twice, respectively, across sites or samples.

While Chao2 estimator does not account for species abundances (only incidence, i.e., presence/absence) but it does preserve some of the spatial nature of the data by recognizing that species could have been present at different sites. Let's write a function for Chao2:

```
S.chao2 <- function(site = " ", community = " "){
  site.obs = community[site, ]
  x = site.obs[site.obs > 0]
  S.obs = length(x)
  community.pa <- (community>0)*1
  Q1 = length(which(colSums(community.pa) == 1))
  Q2 = length(which(colSums(community.pa) == 2))
  S.chao2 = S.obs + (Q1^2)/(2*Q2)
  return(S.chao2)
  }
```

Note, estimated richness by definition must be greater than observed richness.

Question 6: TENTATIVE QUESTIONS: How do estimates from Chao1 and Chao2 differ for BCI?

> ***Answer 6***:

Rarefaction

Often researchers want to compare the richness of two or more samples, as opposed to Chao2 which predicts a single richness value for a set of samples. Ideally, we would sample all of our sites with equal effort, all sites would have equal densities, and all individuals would have equal probability of detection. However, these conditions are often not met in real systems and in most studies. Likewise, sites with greater abundance generally have greater species richness, which means that differences in richness may largely be due to differences in sample size.

A common way to reduce the bias associated with different N and to compare sites with both different N and different S, is to **rarify** samples down to a "lowest common denominator". For example, if there are 2 sites, one with N=100 and one with N=50, we could randomly sample (without replacement) 50 individuals from the site with greater N. Generating many random samples (each of 50 individuals) from the larger site will allow us to calculate the mean (i.e. expected S) and standard error. Likewise, we'll be able to tell whether the S of the smaller sample falls within the confidence-intervals for expected S of the larger sample and hence, whether the difference in S between the two sites is simply due to difference in N.

Let's look at the richness of soil bacteria for all 11 of the KBS sites, where T1=agriculture, T7=grassland, DF=deciduous forest, and CF=coniferous forest.

```
soilbac.S <- S.obs(soilbac)           # observed richness
min.N <- min(rowSums(soilbac))        # sample with fewest sequences
#S.rarefy <- rarefy(soilbac, min.N, se = TRUE)    # richness when rarefied to sample with fewest sequen
#rarecurve(soilbac, step = 20, sample = min.N, col = "blue", cex = 0.6, las=1) # vertical line = min.N;
#abline(0, 1, col = 'red')             # 1:1 line for reference
#text(1500, 1500, "1:1", pos = 2, col = 'red')
```

Question 6: How would our interpretation of diversity be affected if we used observed vs. estimated richness in our rarefaction analysis?

Answer 6:

5) Species evenness

There is more to (α) diversity than just the number of species, whether observed or estimated. Specifically, it is important to consider how abundance varies among species, that is, **species evenness**. Many important biodiversity issues such as species coexistence, community stability, the detection of rare taxa, the observation of compositional change and turnover due to disturbance, invasion, and random population fluctuations relate to abundances vary among taxa. In fact, the foreword of the "Biological Diversity: frontiers in measurement and assessment" as well as Chapter 1 and several other chapters are often concerned with quantifying how abundance varies among taxa.

Question 7: a.) In regards to any system your prefer, how would you interpret a highly even distribution of abundance (high evenness) versus a highly uneven one (low evenness)? b.) How might this relate to your ability to adequately sample the system?

Answer 7a: ***Answer 7b***:

Visualizing Evenness: The Rank Abundance Curve (RAC)

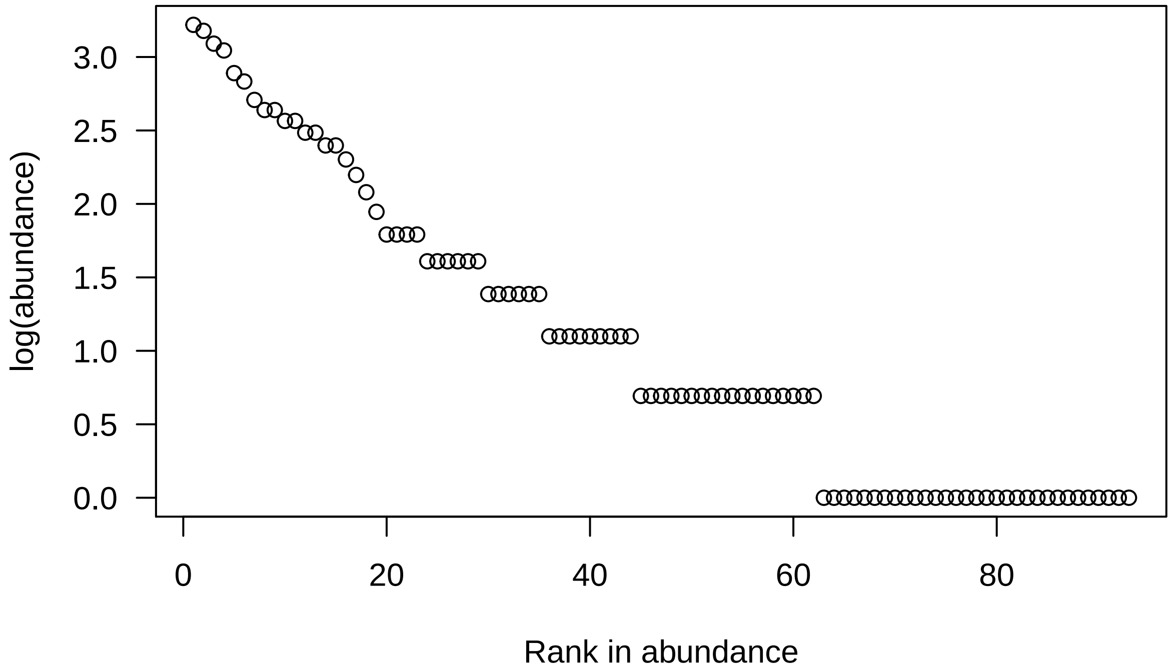
One of the most common ways to visualize evenness is to rank the species from most abundant to least abundant without respect to species labels (and hence no worries about 'ties' in abundance). This **rank-abundance curve (RAC)** is also referred to as a rank-abundance distribution and also as a Whittaker plot.

Let's code a function to construct an RAC.
First, we will remove species that have zero abundances. Then, we order the vector from greatest (most abundant) to least (least abundant).

```
RAC <- function(x){
  x.ab <- x[x > 0] # Remove all zeros from Site1 and reassign the new vector
  x.ab.ranked <- x.ab[order(x.ab, decreasing = TRUE)]  # Use `order` function and "decreasing" argument
  }
```

Let's examine the RAC for `site1` of the BCI data set.

```
rac = RAC(site1)
ranks <- as.vector(seq(1, length(rac)))          # Create a sequence of S ranks
plot(ranks, log(rac), type = 'p', las = 1,      # Transform abundances using the natural logarithm and p
xlab = "Rank in abundance", ylab = "log(abundance)")        # "type = 'p'" = points; "las = 1" horizont
```



Question 8: What effect does log-transforming the abundance data have on how we examine differences among species?

> ***Answer 8***:

It is clear from looking at the RAC for `site1` that abundance among species is unequally distributed, even when log-transformed. This is true even when abundances are log-transformed (note the plotting options above), which deemphasizes the influence of the few most abundant species. 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.

Now that we have visualized unevennes, it is time to quantify it.
This leads us to the desired features of evenness metrics. While there are at least 10, we will concentrate on a few of the most important. First, we want to be able to intuit the values of our metric; the reason why is intuitive. Second, we want our metric to be bound between a minimum evenness (i.e. 0.0) and a maximum evenness (i.e. 1.0). Third, we want evenness to be independent of S because we don't want evenness to simply be a reflection of richness. Here, we will introduce two metrics of evenness that meet the above criteria: Simpson's evenness and Smith and Wilson's evenness index

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.

The `vegan` package has no function for Simpson's evenness. But we can derive Simpson's evenness from Simpson's diversity, i.e., the probability that the next sampled individual belongs to a different species:

```
SimpE <- function(x){
  D <- diversity(x, "simp") # Simpson's diversity (D) via `vegan`
  S <- specnumber(x) ## would like to stick with our S.obs funciton, but doesn't seem to work..strange.
  E <- (1/D)/S # Simpson's evenness (E)
  return(E)
  }
```

Let’s calculate Simpson’s evenness using our previous RAC from site 1 in the BCI site-by-species matrix.

```
SimpE(rac)
```

```
## [1] 0.01103259
```

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 been criticized for being biased towards the most abundant taxa, i.e., the value of the metric can easily be driven by differences in the few most abundant species. Let’s examine the value of evenness for `site1` using a different evenness metric that is less biased in this way, i.e., E_{var}.

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

After reviewing existing metrics, Smith and Wilson (1996) derived a more robust measure of evenness, which they called (E_{var}). This metric is standardized to take values between 0 (no evenness) and 1 (perfect evenness). Abundances are transformed to their natural logarithms to decrease bias towards the most abundant species, that is, the potential for a metric’s value to be influenced more by large numbers than small ones. E_{var}, like all desireable measures of evennness, is independent of richness S. 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 seemingly more involved to calcualte, E_{var} simply reduces to finding the sample variance of the log-transformed abundances and then standardizing it to take values between 0 and 1 by using elementary trigonometry. E_{var} uses the arctangent, which varies between -$\pi/2$ and $\pi/2$ without being periodic like sine waves. Multiplying the arctangent by 2/pi forces the result to take values between 0 and 1. Subtracting this from one allows low evenness to be associated with values near 0 and high evenness to be associated with values near 1. We can confirm all this with a very short and simply R chunk:

```
Evar <- function(x){
  1 - (2/pi)*atan(var(log(x)))
  }
```

Let’s calculate E_{var} for site 1 of the BCI site-by-species matrix.

```
Evar(rac)
```

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

As you can see, the value of E_{var} suggests an intermediate value of species evenness for site1. Likewise, you can see that different evenness metrics, even when bound between 0 and 1 still return very different values.

Question 9: Does this disagreement reveal that each metric is inaccurate, or something else? Explain.

> ***Answer 9***:

6) Integrating richness and evenness: "diversity metrics"

So far, we have introduced two primary aspects of diversity, i.e., richness and evenness. While we often examine each of these independently, e.g., estimating E_{var} or conducting rarefaction, the interaction of richness and evenness is also important. In fact, richness and evenness are the two primary components of metrics of diversity, which are commonly derived as a relationship of a form similar to $Evenness = \frac{Diversity}{Richness}$. For example, in estimating Simpson's evenness we first estimated Simpson's diversity (D), which is usually of the form 1-D or 1/D so that lower values have lower diversity. We then divided 1/D by S, i.e. richness, to obtain a measure of evenness. Here, we will estimate popular indices of diversity 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 used across the natural sciences and is calculated as $H' = -\sum p_i ln(p_i)$. Let's calculate Shannon's diversity for the RAC of site 1 in the BCI site-by-species matrix 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
```

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's calculation for the inverse of Simpson's diversity
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, Because α is a fitted parameter, it is less straightforward to estimate and we will not attempt to code a function for it, here. Fisher's α has often been used as a diversity metric and is the root of 'alpha' diversity and, according to the authors of the vegan package, it is asymptotically similar to inverse Simpson's. Let's do this comparison using the RAC from site 1 of the BCI site-by-species matrix.

```
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.

7) Integrating richness and evenness: Species Abundance Distributions (SAD)

Just as alpha-diversity consists of more than just the number of species, it also con and evenness. Looking at one vs. the other doesn't give a complete picture of alpha diversity. Diversity metric attempt to do this, but have some limitations. The SAD is perhaps a better and more "integrative" way of handling the data that you get from a sample.

Species abundance models

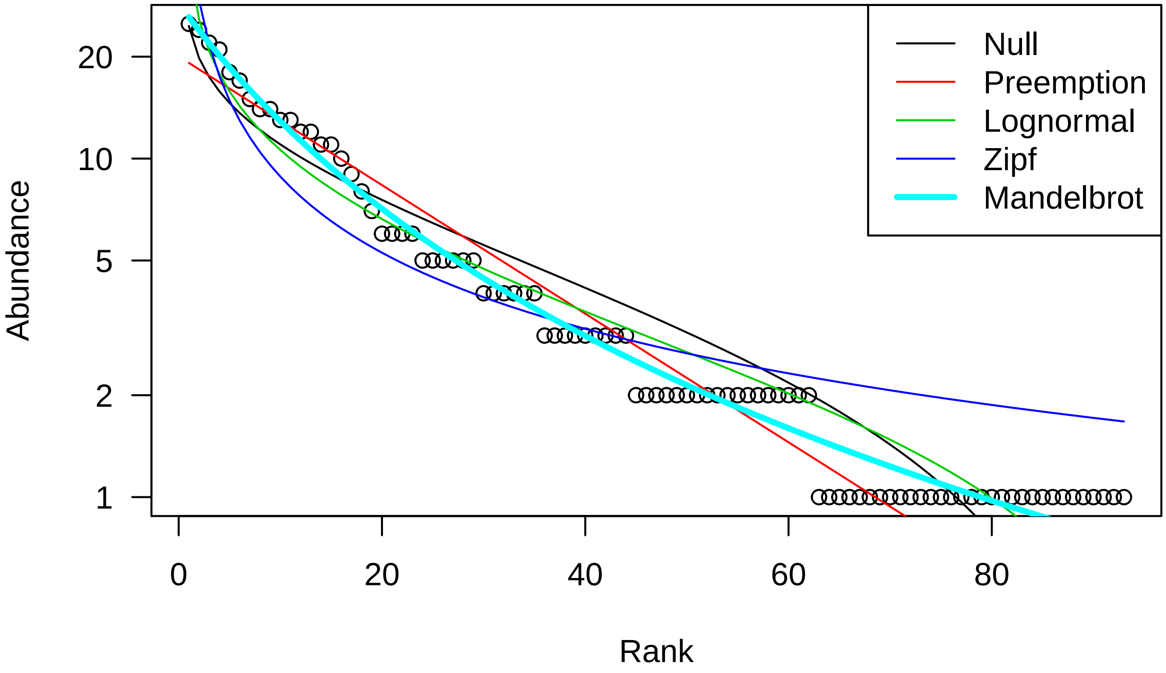
Recall that the species rank-abundance curve RAC is simply the abundance of species ranked from most-to-least abundant. The RAC is a simple data structure that is both a vector of abundances and a row in the site-by-species matrix (minus the zeros, i.e., absences). Despite it's simplicity the RAC contains a lot of information. For example, the sum of the abundances in the RAC is commonly referred to as the observed **total abundance** and denoted as **N** while the length of the RAC (the number of ranks) is simply the observed **species richness**, denoted as **S**. The RAC also contains the information needed to estimate richness and calculate evenness and diversity. Moreover, the RAC represents a type of species-abundance distribution (SAD), the form of which has been intensively studied and modeled. Species abundance models attempt to predict the form of the SAD (e.g. as an RAC) according to mechanisms and processes that underpin general ecological and biodiversity theory.

Let's use the radfit() function in the vegan package to fit the predictions of various species abundance models to the RAC of site1 in BCI

```
RACresults <- radfit(rac)
RACresults
```

```
##
## RAD models, family poisson
## No. of species 93, total abundance 448
##
##             par1      par2     par3    Deviance AIC      BIC
## Null                                    39.5261 315.4362 315.4362
## Preemption  0.042797                    21.8939 299.8041 302.3367
## Lognormal   1.0687    1.0186            25.1528 305.0629 310.1281
## Zipf        0.11033  -0.74705           61.0465 340.9567 346.0219
## Mandelbrot  100.52   -2.312    24.084    4.2271 286.1372 293.7350
```

```
plot(RACresults, las=1)
```



which is implicitly defined by the log-series model: $S = \alpha * ln(1 + n/\alpha)$.

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