Local Diversity (alpha)

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())
getwd()
setwd("~/GitHub/QuantitativeBiodiversity/Assignments/Alpha")
```

Install Packages

In this excercise, we will rely heavily on a **contributed R package** called `vegan`. `Vegan` 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. If you are running `vegan` for the first time, you will need to install it with the `install.packages` function. Otherwise, we recommend you just use the `require` function to load `vegan` and its dependencies.

```
#install.packages("vegan")
require("vegan")
```

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` package 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:

```
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 row in the site-by-species matrix corresponds to a site, while each column corresponds to a species. 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:

```
dim(BCI)
```

```
## [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.

```
BCI[1:8, 14:17]
```

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 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). In this part of the exercise, you will see that there are several ways to estimate richness, which 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, ]
dim(site1)
```

```
## [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.

```
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()` also calculates observed richness.

Question 2: Does `specnumber()` (from `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) where $C = 1 - \frac{n_1}{N}$, where 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.

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: Answer the following questions about the coverage:

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?
d. What would we conclude from Good's Coverage if n_1 equaled N?

Answer 4a:
Answer 4b:
Answer 4c:
Answer 4d:

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.

In the following R chunk, we load this microbial dataset as a matrix and identify a sample (i.e., site) for subsequent analysis. Notice the commands we are using to load this data. We recommend that you explore the various ways to load data into R before working with your own data.

```
soilbac <- read.table("data/soilbac.txt", sep = "\t", header = TRUE, row.names = 1)
soilbac <- as.data.frame(t(soilbac))
soilbac1 <- soilbac[1,]
```

Question 5: Answer the following questions about the soil bacterial dataset.

a. How many sequences did we recover from the sample `soilbac1`, i.e. N?
b. What is the observed richness 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 commonly used richness estimators that were developed by Anne Chao. Both estimators fall into the category of being non-parametric, which means that they are not based on an underlying distribution (e.g. Normal distribution, Student's t-distribution) and hence, make few statistical assumptions.

Chao1 is an ***abundance-based estimator*** and is 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). Chao1 tends to be used for analysis of samples with low-abundance taxa. Becasue it requires singleton and doubleton data to calculate, Chao1 cannot be used on with a site x species matrix where abundances have been relativized.

Let's write a function for Chao1:

```
S.chao1 <- function(x = ""){
  S.obs(x) + (sum(x == 1)^2) / (2 * sum(x == 2))
  }
```

In contrast, **Chao2** is ***incidence-based estimator*** that uses presence-absence data for examining richness across multiple sites. 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. Because it is an incidence-based estimator, Chao2 is commonly used when estimating richness with clonal organisms and/or percent cover data (e.g., corals)

Let's write a function for Chao2. In thie function, the first argument is site as either row number or row title (e.g., 1 or "T1_1") and the second argument is the site x species matrix (e.g., soilbac):

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

Notice that this function is a bit more complicated than previous functions we have written. Try to read through each step and figure out what is happening.

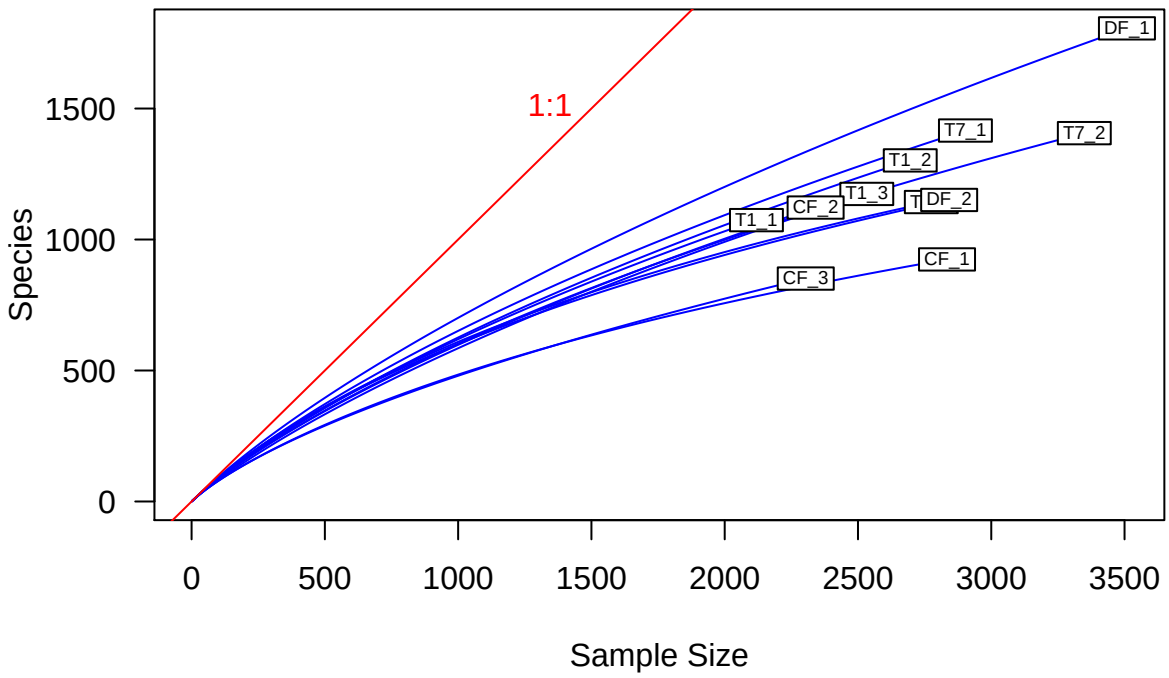
Rarefaction

Often researchers want to compare the richness of two or more 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 rarely met. 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.

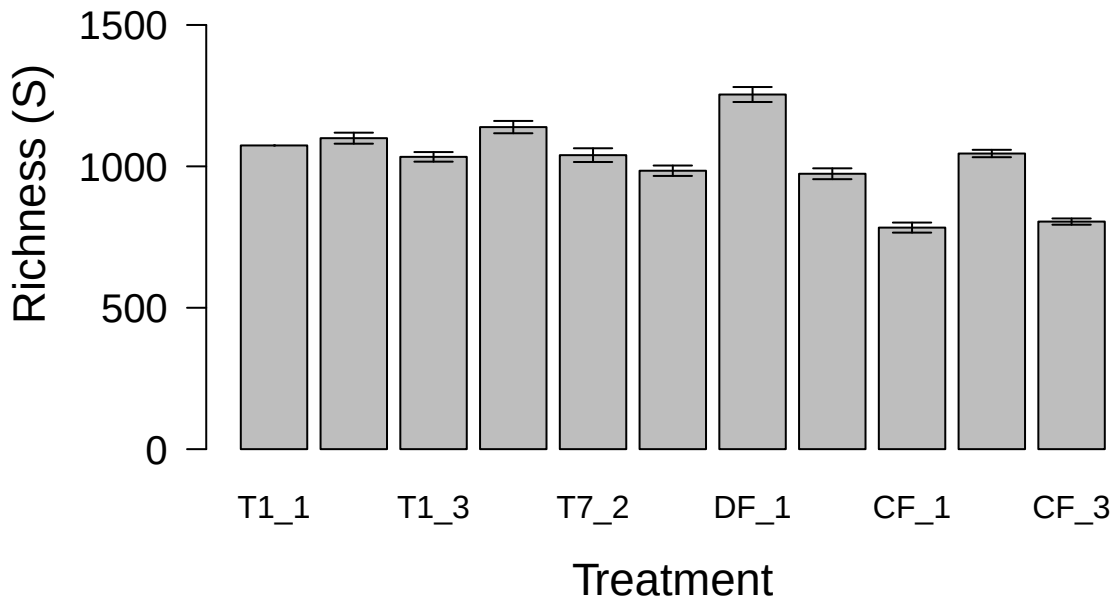
In the following section, we will calculate observed richness for soil bacteria collected from 11 of the KBS sites, where T1 = agriculture, T7 = grassland, DF = deciduous forest, and CF = coniferous forest. Then, we will identify the sample with the fewest sequences (i.e., N) and use `vegan` to rarefy to that sampling effort. Last we will use commands from `vegan` to construct rarefaction curves for the different samples.

```
soilbac.S <- S.obs(soilbac)
min.N <- min(rowSums(soilbac))
S.rarefy <- rarefy(x = soilbac, sample = min.N, se = TRUE)
rarecurve(x = soilbac, step = 20, col = "blue", cex = 0.6, las=1)
abline(0, 1, col = 'red')
text(1500, 1500, "1:1", pos = 2, col = 'red')
```



We can use the information from the `rarefy` function to create a barplot that compares each site. Remember, we can calculate 95% confidence intervals using $95\%CI = \bar{x} \pm SEM \times 1.96$.

```
opar <- par(no.readonly = TRUE)
par(mar=c(5.1, 6.1, 4.1, 2.1))
S.plot <- barplot(S.rarefy[1, ], xlab = "Treatment", ylab = NULL,
                  ylim =c(0, round(max(soilbac.S), digits = 0)),
                  pch = 15, las = 1, cex = 1, cex.lab = 1.4, cex.axis = 1.25)
arrows(x0 = S.plot, y0 = S.rarefy[1, ], y1 = S.rarefy[1, ] - (S.rarefy[2, ] * 1.96),
       angle = 90, length=0.1, lwd = 1)
arrows(x0 = S.plot, y0 = S.rarefy[1, ], y1 = S.rarefy[1, ] + (S.rarefy[2, ] * 1.96),
       angle = 90, length=0.1, lwd = 1)
title(ylab = "Richness (S)", line = 4, cex.lab = 1.4)
```




```
par(opar)
```

Notice that we did a few things differently here. Why did we have to plot the y-axis label manually? What did the `par(mar=c())` function do?

5) Species evenness

There is more to α-diversity than just the number of species in a sample. 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, and biological invasions relate to how abundances vary among taxa.

Visualizing Evenness: The Rank Abundance Curve (RAC)

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

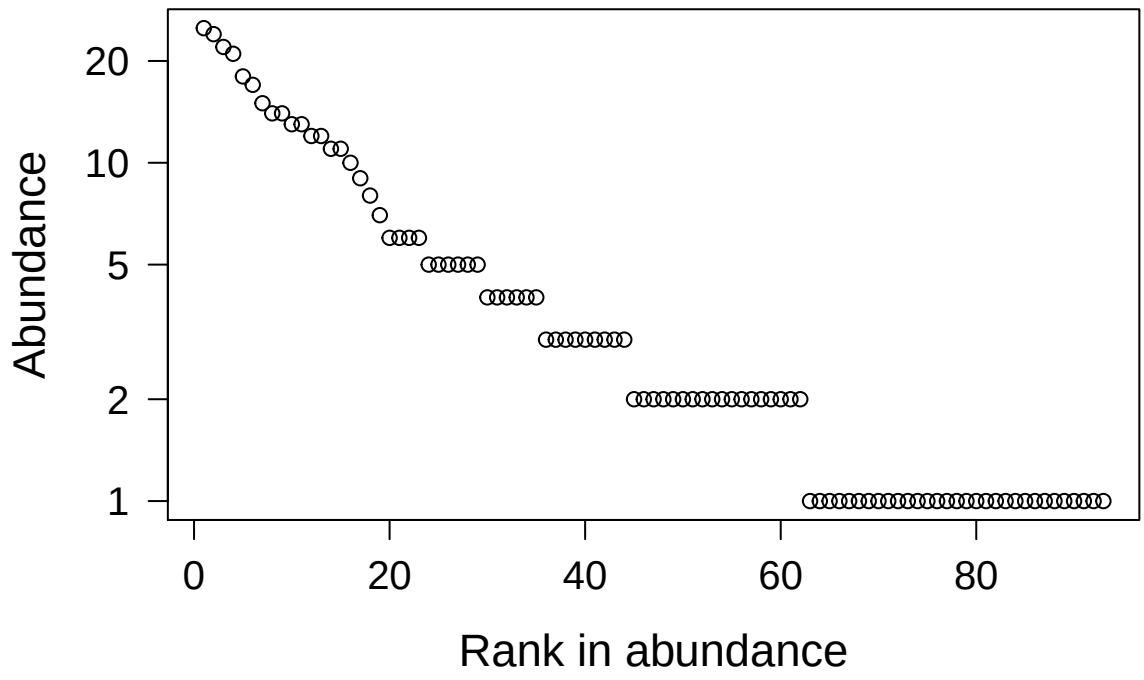
Let's write a function to construct an RAC. First, we will remove species that have zero abundances. Then, we will order the vector from greatest (most abundant) to least (least abundant).

```
RAC <- function(x = ""){
  x = as.vector(x)
```

```
  x.ab = x[x > 0]
  x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
  return(x.ab.ranked)
  }
```

Now, let's examine the RAC for `site1` of the BCI data set. We will do this by creating a sequence of ranks and plotting the RAC with natural-log transformed abundances.

```
rac <- RAC(x = site1)
ranks <- as.vector(seq(1, length(rac)))
opar <- par(no.readonly = TRUE)
par(mar=c(5.1, 5.1, 4.1, 2.1))
plot(ranks, log(rac), type = 'p', axes=F,
     xlab = "Rank in abundance", ylab = "Abundance",
     las = 1, cex.lab = 1.4, cex.axis = 1.25)
box()
axis(side = 1, labels=T, cex.axis = 1.25)
axis(side = 2, las = 1, cex.axis = 1.25,
     labels=c(1, 2, 5, 10, 20), at=log(c(1, 2, 5, 10, 20)))
```




```
par <- opar
```

Question 8: What effect does log-transforming the abundance data have on how we interpret evenness in an RAC?

Answer 8:

It is clear from looking at the RAC for `site1` that the abundance among species is unequally distributed. 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) (http://www.ncbi.nlm.nih.gov/pubmed/17845298).

Now that we have visualized unevennes, it is time to quantify it. Based on decades work, researchers have identified desirable features of an evenness metric. One of these features is that the values generated by the metric should be relatively easy to inuit. As such, useful metrics are often bound betwen a minimum evenness value of 0.0 and a maximum evenness value of 1.0. Another important feature is that evenness values should independent of S; 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 ($E_{1/D}$) and Smith and Wilson's evenness index (E_{var}).

Simpson's evenness ($E_{1/D}$)

Not sure how to handle here, but we haven't defined the SAD yet, so this explanation is not very useful; some revising probably needed

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 does not have a function for Simpson's evenness. But we can derive Simpson's evenness from Simpson's diversity, which estimates the probability that the next sampled individual belongs to a different species. In the following R chunk we will estimate Simpsons diversity (D) using the `diversity` function in `vegan` and observed richness.

```
SimpE <- function(x){
  x = as.data.frame(x)
  D <- diversity(x, "inv")
  S <- S.obs(x)
  E <- (D)/S
  return(E)
  }
```

Now, let's calculate Simpson's evenness using our previous RAC from `site 1` in the BCI site-by-species matrix.

```
SimpE(site1)
```

```
##         1
## 0.4238232
```

We can see that Simpson's evenness for `site1` is moderately even (~0.42). However, Simpson's evenness has been criticized for being biased towards the most abundant taxa. The value of the metric is suppodely sensitive to 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 using elementary trigonometry. Specifically, 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. In the end, an E_{var} function can be written as follows:

```
Evar <- function(x){
  x <- as.vector(x[x > 0])
  1 - (2/pi)*atan(var(log(x)))
  }
```

Now let's use the E_{var} function to estimate evenness for `site1` of the BCI site-by-species matrix.

```
Evar(rac)
```

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

Question 9: Compare estimates of evenness for `site1` of BCI using $E_{1/D}$ and E_{var}. Do they agree?

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, the interaction between richness and evenness is important. Here, we will estimate popular indices of diversity, which explicitly incorporate richness and evenness. We will write our own diversity functions and compare them agains the functions in `vegan`.

Shannon's diversity (a.k.a Shannon's entropy)

Shannon's diversity metric is derived from Shannon's information entropy, and is essentially 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 `site1` in the BCI site-by-species matrix and then compare it to the `vegan` estimate:

```
H <- function(x = ""){
  H = 0
  for (n_i in x){
    p = n_i / sum(x)
    H = H - p*log(p)
```

```
  }
  return(H)
}
```

Now we will use `vegan` to estimate Shannon's index:

```
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 index naturally increases with diversity. Let's calculate Simpson's diversity for `site1` and then compare it to the `vegan` estimate:

```
D <- function(x = ""){
  D = 0
  N = sum(x)
  for (n_i in x){
    D = D + (n_i^2)/(N^2)
  }
  return(D)
}
```

And now let's express Simpsons's diversity as $1/D$ (invD) and 1-D:

```
D.inv <- 1/D(rac)
D.min <- 1-D(rac)
```

Finally, we can use `vegan` to estimate Simpsons's index:

```
diversity(rac, "simp")
diversity(rac, "inv")
```

7) Moving beyond univariate metrics of α-diversity

Just as α-diversity consists of more than just the number of species, it also consists of and evenness. Looking at one vs. the other doesn't give a complete picture of α-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, including the sum of the abundances (N), the observed species richness (S), as well as the information needed to estimate richness and to calculate evenness and diversity.

Question 10: What might the shape of a rank-abundance distribution tell us about the forces, processes, and/or mechanisms influencing the structure of our system, e.g., an ecological community?

Answer 10:

The uneven shape of the RAC is one of the most intensively studied patterns in ecology, and underpins all or most ecological theories of biodiversity. Predicting the form of the RAC is the first test that any biodiversity theory must pass and there are no less than 20 models that have attempted to explain the uneven form of the RAC across ecological systems. These models attempt to predict the form of the RAC according to mechanisms and processes that are believed to be important to the assembly and structure of ecological systems.

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
```

As you can see, vegan fits five models (Null, Preemption, Lognormal, Zipf, and Mandelbrot) using the rank-abundance curve, which vegan refers to as the rank-abundance distribution (RAD). Before explaining what these models represent, let's run through vegan's output.

Next to "RAD models", we see "family poisson", which tells us that by default, vegan assumes Poisson distributed error terms. Below this, we see that vegan returns the number of species and the number of individuals for the empirical RAC. Next, we see a table of information, the first columns of which are par1, par2, and par3. These columns pertain to model parameters and reveal that the different models use different numbers of parameters; the null model uses none. Next, we see a column for Deviance, which is a quality of fit statistic based on the idea of residual sum of squares. After Deviance, we see columns for AIC and BIC, which are the estimated Akaike Information Criterion and the Bayesian Information Criterion, respectively. The AIC and BIC values both relate to the relative quality of a statistical model for a given set of data. That is, AIC and BIC estimate the quality of each model, relative to the other models. AIC and BIC are known as "penalized likelihood criteria" because they also penalize models for the number of parameters the models use.

Question 11: Why is it important to account for the number of parameters a model uses when judging how well it explains a given set of data?

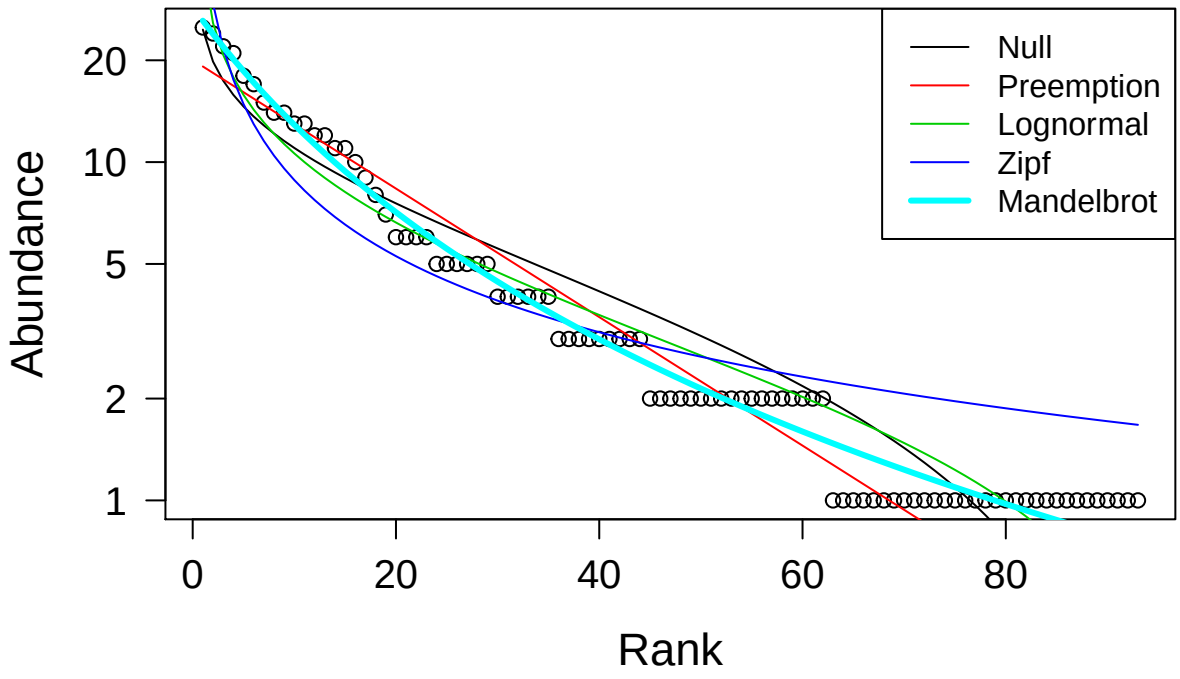
Answer 11:

Whether using the AIC or BIC value, we interpret a lower value as a better quality model. Whether looking at the AIC or BIC values, we can see that the Mandelbrot distribution seems to provide the best fit to our

empirical RAC. We can also see that the Zipf distribution appears to provide the worst fit. This result is also supported by the Deviance values, which are highest for the Zipf and lowest for the Mandelbrot.

Let's visualize our results by plotting the empirical RAC and the predicted RAC for each model:

```
plot(RACresults, las=1, cex.lab = 1.4, cex.axis = 1.25)
```



Visual inspection confirms our statistical results, i.e., the Mandelbrot model appears more like the empirical data (black circles) than the other models, while the Zipf under predict abundant species (i.e. rank) and over predicts the rare species.

But, what should we conclude from this curve fitting exercise? That is, what do these models represent?

Interpreting vegan's RAC models:

Null: A **broken stick model** (Pielou 1975) where the expected abundance of a species at rank r is a(r) = N/S * sum(from x=r to S) * 1/x. Where N is the total number of individuals and S is the total number of species. This gives a constraint-based Null model where the N individuals are randomly distributed among S species, and there are no fitted parameters. Null models often reveal that realistic patterns can be expected from random sampling, and have been extremely useful in ecology and evolution.

Preemption: The **niche preemption model**, a.k.a. geometric series or Motomura model. Envisioned an environment occupied by a single species. Now, envision that a second species colonizes the environment and takes some portion of resources equal to α. Then, envision that a third species colonizes the environment and takes a portion of resources equal to α away from the second species. Imagine this process continuing

until N is zero. The only estimated parameter is the preemption coefficient α which gives the decay rate of abundance per rank. The expected abundance (a) of species at rank r is $a(r) = N * \alpha * (1 - \alpha)^(r - 1)$.

Question 12: a.) What does the preemption model assume about the relationship between total abundance (N) and total resources that can be preempted? b.) Why does the niche preemption model look like a straight line in the RAD plot?

Answer 12a:
Answer 12b:

Lognormal: Many statistical models assume that the distribution of values (e.g. abundances) are normally distributed, e.g., they conform to the shape of a symmetrical bell curve, more precisely known as the Gaussian distribution. In contrast, the log-Normal model assumes that the logarithmic abundances are normally distributed. The expected abundance of a species at rank r is then: a[r] = exp(log(mu) + log(sigma) * N), where N is a Normal deviate. A Normal deviate is simply the number of standard deviations a score is from the mean of its population. The log-normal model was introduced into ecology by Frank Preston in 1948 and is one of the most widely successful species abundance models.

Zipf: The Zipf model is based on Zipfs Law, an well-known observation that many types of ranked data are fit by a simple scaling law (aka power law). In short, the abundance of a species in the RAC is inversely proportional to its rank. The expected abundance (a) of species at rank r is: $a(r) = N * p1 * r^\gamma$, where $p1$ is the fitted proportion of the most abundant species, and γ is a decay coefficient.

Mandelbrot: Shortened name for the Zipf–Mandelbrot model, a generalization of the Zipf model made by the mathematician and father of fractal geometry, Benoit Mandelbrot. This model adds one parameter (β) to the Zipf model. The expected abundance of a species (a) at rank r is $a(r) = N * c * (r + \beta)^\gamma$. Here, the p1 parameter of the Zipf model changes into a meaningless scaling constant c.

Homework

1) As stated by Magurran (2004) the $D = \sum p_i^2$ derivation of Simpson's Diversity only applies to communities of infinite size. For anything but an infinitely large community, Simpson's Diversity 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 site 1 of the BCI site-by-species matrix.

2) Along with the rank-abundance curve (or rank-abundance distribution), another way to visualize the distribution of abundance among species is with a histogram (aka frequency distribution) that shows that frequency of different abundance classes. For example, in a given sample, there may be ten species represented by a single individual, 8 species with two individuals, 4 species with three individuals, and so on. In fact, the rank-abundance curve and the frequency distribution are the two most common ways to visualize the species abundance distribution (SAD) and to test species abundance models and biodiversity theories. To address this homework question, use the R function **hist()** to plot the frequency distribution for site 1 of the BCI site-by-species matrix, and describe the general pattern you see.

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