Assignment: Local (alpha) Diversity

Erik Parker; Z620: Quantitative Biodiversity, Indiana University

23 January, 2017

OVERVIEW

In this exercise, we will explore aspects of local or site-specific diversity, also known as alpha (α) diversity. First we will quantify two of the fundamental components of (α) diversity: **richness** and **evenness**. From there, we will then discuss ways to integrate richness and evenness, which will include univariate metrics of diversity along with an investigation of the **species abundance distribution (SAD)**.

Directions:

1. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the exercise as possible during class; what you do not complete in class will need to be done on your own outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this exercise document; they also correspond to the handout. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">".
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For homework, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, please submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file `alpha_assignment.Rmd` and the PDF output of `Knitr` (`alpha_assignment.pdf`).

1) R SETUP

In the R code chunk below, please provide the code to: 1) Clear your R environment, 2) Print your current working directory, 3) Set your working directory to your `/Week2-Alpha` folder, and 4) Load the `vegan` R package (be sure to install if needed).

```
rm(list=ls())
getwd()
```

```
## [1] "/var/host/media/removable/USB Drive/GitHub/QB2017_Parker/Week2-Alpha"
```

```
setwd("/media/removable/USB Drive/GitHub/QB2017_Parker/Week2-Alpha/")
#install.packages("vegan")
require(vegan)
```

```
## Loading required package: vegan

## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.4-2
```

2) LOADING DATA

In the R code chunk below, do the following: 1) Load your dataset, and 2) Display the structure of the dataset (if the structure is long, use `max.level=0` to show just basic information).

```
data("BCI")
str(BCI, max.level = 0)
```

```
## 'data.frame':    50 obs. of  225 variables:
##   [list output truncated]
##  - attr(*, "original.names")= chr  "Abarema.macradenium" "Acacia.melanoceras" "Acalypha.diversifolia
```

3) SPECIES RICHNESS

Species richness (S) is simply the number of species in a system or the number of species observed in a sample.

Observed Richness

In the R code chunk below, do the following:

1. Write a function called `S.obs` to calculate observed richness
2. Use your function to determine the number of species in `site1`, and
3. Compare the output of your function to the output of the `specnumber()` function in vegan.

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

S.obs(BCI[1,])
```

```
##  1
## 93
```

```
specnumber(BCI[1,])
```

```
##  1
## 93
```

```
S.obs(BCI[1:4,])
```

```
##  1  2  3  4
## 93 84 90 94
```

Question 1: Does `specnumber()` from `vegan` return the same value for observed richness in `site1` as our function `S.obs`? What is the species richness of the first 4 sites (i.e., rows) of the BCI matrix?

> ***Answer 1***: Yes! My function returns the same value for obsreved richness as the specnumber() function from vegan, which is 93. The first 4 sites have a richness of 93, 84, 90, and 94 in order, or 361 in total when added together (though we probably don't really care about that).

Coverage. How Well Did You Sample Your Site?

In the R code chunk below, do the following:

1. Write a function to calculate Good's Coverage, and

2. Use that function to calculate coverage for all sites in the BCI matrix.

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

coverage <- C(BCI)
coverage
```

```
##         1         2         3         4         5         6         7
## 0.9308036 0.9287356 0.9200864 0.9468504 0.9287129 0.9174757 0.9326923
##         8         9        10        11        12        13        14
## 0.9443155 0.9095355 0.9275362 0.9152120 0.9071038 0.9242054 0.9132420
##        15        16        17        18        19        20        21
## 0.9350649 0.9267735 0.8950131 0.9193084 0.8891455 0.9114219 0.8946078
##        22        23        24        25        26        27        28
## 0.9066986 0.8705882 0.9030612 0.9095023 0.9115479 0.9088729 0.9198966
##        29        30        31        32        33        34        35
## 0.8983516 0.9221053 0.9382423 0.9411765 0.9220183 0.9239374 0.9267887
##        36        37        38        39        40        41        42
## 0.9186047 0.9379310 0.9306488 0.9268868 0.9386503 0.8880597 0.9299517
##        43        44        45        46        47        48        49
## 0.9140049 0.9168704 0.9234234 0.9348837 0.8847059 0.9228916 0.9086651
##        50
## 0.9143519
```

```
# Portion of taxa in site 1 represented by singetons

Singletons <- function(x=""){ (rowSums(x==1) / rowSums(x))}
Singletons(BCI[1,])*100
```

```
##        1
## 6.919643
```

Question 2: Answer the following questions about coverage:

a. What is the range of values that can be generated by Good's Coverage?
b. What would we conclude from Good's Coverage if n_i equaled N?
c. What portion of taxa in `site1` were represented by singletons?
d. Make some observations about coverage at the BCI plots.

Answer 2a:
The range of values that can be generated by Good's Coverage is 0 to 1. If all species were seen as singletons, that would lead to 1 - N/N, so 1-1. If no species seen were singletons, that would be 1 - 0/N, or 1-0.

Answer 2b:
We would conclude that the site was not sampled very well at all. Seeing every species only once might actually mean that there really is only one member of each species present in the sampling area, but more likely it means that we missed a lot of individuals while sampling.

Answer 2c:
From above, the portion of taxa in site1 represented by singletons is 6.9%, so about 7% of taxa found were found as singletons.

Answer 2d:
Overall, the coverage across all of the plots is quite good in the BCI dataset. All of the sites have values above 0.85 for Good's coverage, meaning that no more than 15% of the taxa discovered at any site were discovered as singletons. To me this means that the researchers who collected this species incidence data did a good job making sure they sampled effectively and limited the

number of species seen that were only seen once.

Estimated Richness

In the R code chunk below, do the following:

1. Load the microbial dataset (located in the `/Week2-Alpha/data` folder),
2. Transform and transpose the data as needed (see handout),
3. Create a vector (`soilbac1`) with the bacterial OTU abundances at any site in the dataset,
4. Calculate the observed richness at that particular site, and
5. Calculate the coverage at that particular site

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

soil.rich.site1 <- S.obs(soilbac1)
soil.rich.site1
```

```
## T1_1
## 1074
```

```
soil.coverage.site1 <- C(soilbac1)
soil.coverage.site1
```

```
##      T1_1
## 0.6479471
```

```
# N of site 1?

sum(soilbac1)
```

```
## [1] 2119
```

Question 3: 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 3a:
The total number of sequences recovered from the sample 'soilbac1' is 2119.

Answer 3b: The observed richness of 'soilbac1' is 1074.

Answer 3c:
The coverage of soilbac1 is much lower than that of site1 from BCI. We found a coverage of 0.65 for the soil sample at site 1, and a coverage of 0.93 at site 1 of the BCI sample. This makes some sense to me, as I might naively expect to find more rare soil microbial species in a sample then I would expect to find rare tree species.

Richness Estimators

In the R code chunk below, do the following:

1. Write a function to calculate **Chao1**,

2. Write a function to calculate **Chao2**,
3. Write a function to calculate **ACE**, and
4. Use these functions to estimate richness at both `site1` and `soilbac1`.

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

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)
  }

S.ace <- function(x = "", thresh = 10){
  x <- x[x>0]
  S.abund <- length(which(x > thresh))
  S.rare <- length(which(x <= thresh))
  singlt <- length(which(x == 1))
  N.rare <- sum(x[which(x <= thresh)])
  C.ace <- 1 - (singlt / N.rare)
  i <- c(1:thresh)
  count <- function(i, y) {
  length(y[y == i]) }
  a.1 <- sapply (i, count, x)
  f.1 <- (i * (i - 1)) * a.1
  G.ace <- (S.rare/C.ace)*(sum(f.1)/(N.rare*(N.rare-1)))
  S.ace <- S.abund + (S.rare/C.ace) + (singlt/C.ace) * max(G.ace,0)
  return(S.ace)
}

BCI1 <- BCI[1,]

S.chao1(BCI1)
```

```
##        1
## 119.6944
```

```
S.chao1(soilbac1)
```

```
##     T1_1
## 2628.514
```

```
S.chao2(1, BCI)
```

```
##        1
## 104.6053
```

```
S.chao2(1, soilbac.t)
```

```
##     T1_1
## 21055.39
```

```
S.ace(BCI1)
```

```
## [1] 159.3404
```

```
S.ace(soilbac1)
```

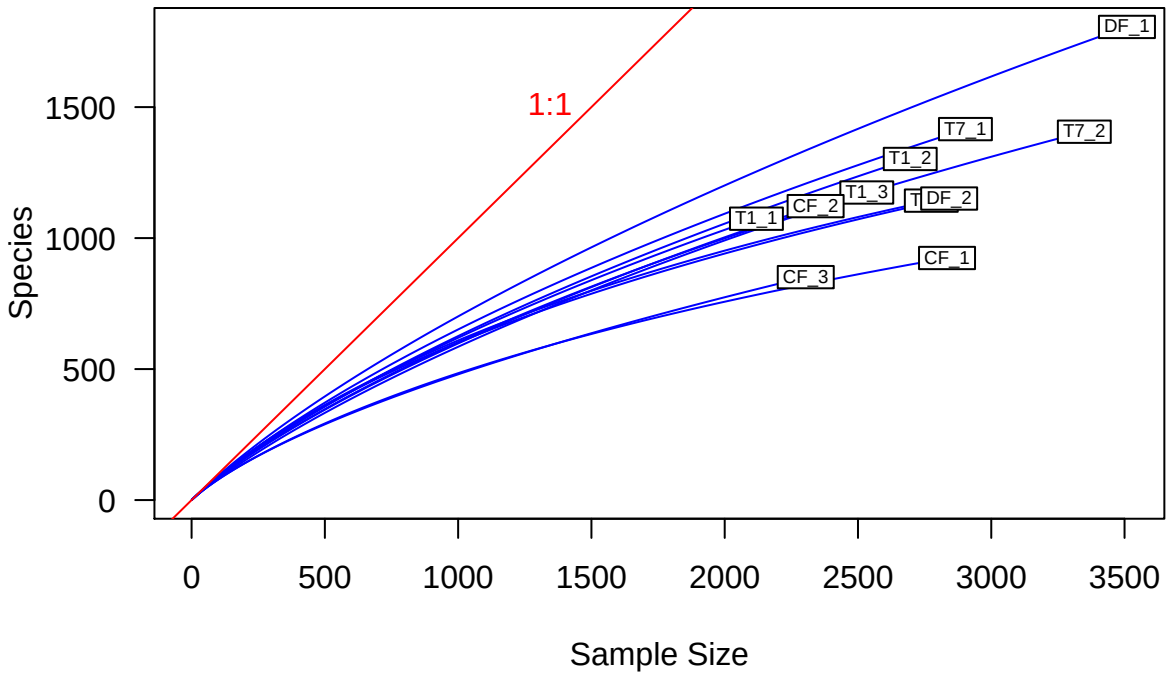
```
## [1] 4465.983
```

Rarefaction

In the R code chunk below, please do the following:

1. Calculate observed richness for all samples in `soilbac`,
2. Determine the size of the smallest sample,
3. Use the `rarefy()` function to rarefy each sample to this level,
4. Plot the rarefaction results, and
5. Add the 1:1 line and label. 00034762140003476214

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



Question 4: What is the difference between ACE and the Chao estimators?

> ***Answer 4***: The Chao estimators both rely the presence species singletons and doubletons to make their estimates about species richness at sites being examined. ACE, on the other hand, uses a definable threshold number of individuals of a species to make estimates based on the abundance of taxa that are still rare, but have more than one or two individuals present and also those that are common (above the defined threshold.)

4) SPECIES EVENNESS

Here, we consider how abundance varies among species, that is, **species evenness**.

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). An 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).

In the R code chunk below, do the following:

1. Write a function to construct a RAC,
2. Be sure your function removes species that have zero abundances,
3. Order the vector (RAC) from greatest (most abundant) to least (least abundant), and
4. Return the ranked vector

```
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.

In the R code chunk below, do the following:

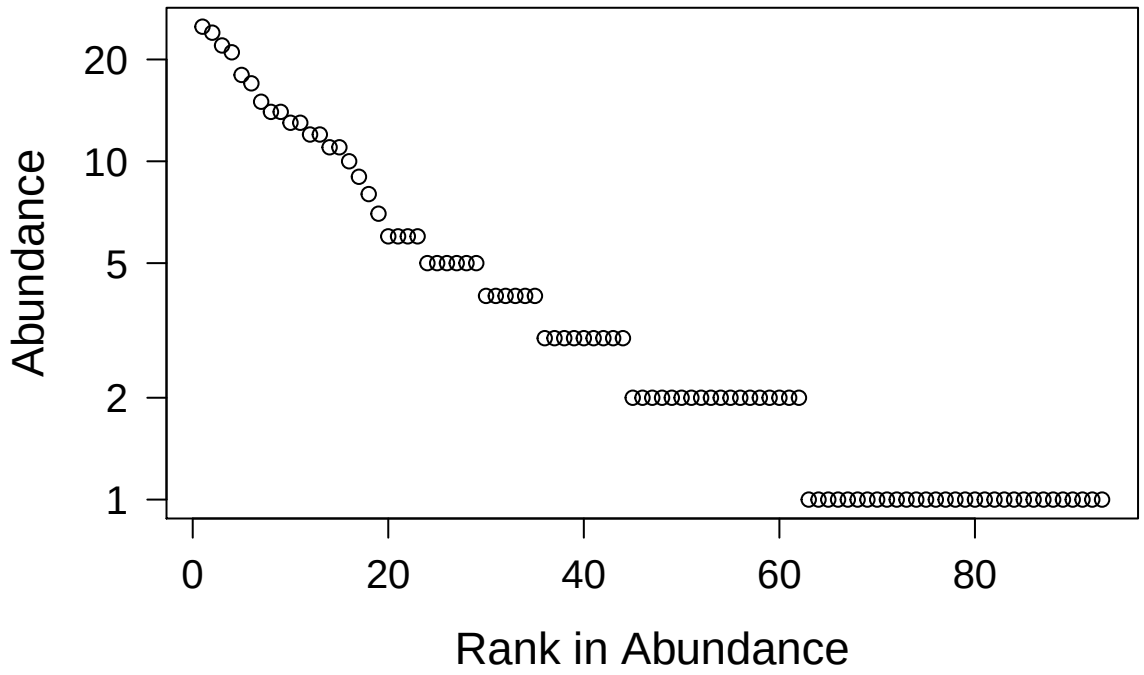
1. Create a sequence of ranks and plot the RAC with natural-log-transformed abundances,
2. Label the x-axis "Rank in abundance" and the y-axis "log(abundance)"

```
plot.new()
site1 <- BCI1

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 5: What effect does visualizing species abundance data on a log-scaled axis have on how we interpret evenness in the RAC?

> ***Answer 5***: In my opinion, using a log scale makes it much easier to visualize large differences (so long as you remember that it is actually on a log scale!), and this is still true when the data we are dealing with is species abundances. THe pattern of many rarer species, and few very common species in an environment makes it difficult to accruately represent abundance data if we are using non-transformed axes, either the rare species would appear to be at 0 abundance, to keep the graph compact, or the entire graph would need to be extrC <- function(x="”){ 1 - (rowSums(x==1) / rowSums(x))}emely tall to fit everyone together! This log scaled axis we used makes it much easier to compare the variation in abundance between species.

Now that we have visualized unevennes, it is time to quantify it using Simpson's evenness ($E_{1/D}$) and Smith and Wilson's evenness index (E_{var}).

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

In the R code chunk below, do the following:

1. Write the function to calculate $E_{1/D}$, and
2. Calculate $E_{1/D}$ for `site1`.

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

SimpE(site1)
```

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

Smith and Wilson's evenness index (E_{var})

In the R code chunk below, please do the following:

1. Write the function to calculate E_{var},
2. Calculate E_{var} for `site1`, and
3. Compare $E_{1/D}$ and E_{var}.

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

Evar(site1)
```

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

Question 6: Compare estimates of evenness for `site1` of BCI using $E_{1/D}$ and E_{var}. Do they agree? If so, why? If not, why? What can you infer from the results.

> ***Answer 6***: The measures of Simpson's evenness and Smith and Wilson's evenness index are pretty close to agreement, off by about 0.08, but there is still some disagreememnt there. From the handout, it seems that Simpson's is criticized for being biased by the most abundant species in a population so it may be that Smith and Wilson's is the more accurate measure of evenness when there are a number of very abundant species present, which seems to be the case for the BCI data if we go off of the rank abundance curve we generated above.
> Overall, it seems like the BCI data is relatively even and that Simpson's may not be the best estimator of evenness due to the presence of a number of very abundant species.

5) INTEGRATING RICHNESS AND EVENNESS: DIVERSITY METRICS

So far, we have introduced two primary aspects of diversity, i.e., richness and evenness. Here, we will use popular indices to estimate diversity, which explicitly incorporate richness and evenness We will write our own diversity functions and compare them against the functions in `vegan`.

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

In the R code chunk below, please do the following:

1. Provide the code for calculating H' (Shannon's diversity),
2. Compare this estimate with the output of `vegan`'s diversity function using method = "shannon".

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

```
    return(H)
}

diversity(site1, index = "shannon")
```

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

```
ShanH(site1)
```

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

```
# The same!
```

Simpson's diversity (or dominance)

In the R code chunk below, please do the following:

1. Provide the code for calculating D (Simpson's diversity),
2. Calculate both the inverse (1/D) and 1 - D,
3. Compare this estimate with the output of `vegan's` diversity function using method = "simp".

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

D.inv <- 1/SimpD(site1)
D.sub <- 1-SimpD(site1)
D.inv
```

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

```
D.sub
```

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

```
diversity(site1, "inv")
```

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

```
diversity(site1, "simp")
```

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

Question 7: Compare estimates of evenness for `site1` of BCI using $E_{H'}$ and E_{var}. Do they agree? If so, why? If not, why? What can you infer from the results.

> ***Answer 7***: The estimates of evenness for site1 of BCI found using Simpson and Shannon's methods are not that close value wise, but it seems like they are both pointing to the same conclusion: that site1 is quite diverse. Simpson's, by itself, is the probability of two samples from the data being the same species when taken at random. The inverse Simpson's we calculated was quite high, meaning this probability was low (also supported by the fact that 1 - this probability is 0.975). A similar conclusion was also reached through the use of Shannon's which found that,

roughly speaking, the probability of two subsequent samples from this dataset being different species was very high (implying high diversity and evenness among the sample).

Fisher's $\boldsymbol{\alpha}$

In the R code chunk below, please do the following:

1. Provide the code for calculating Fisher's $\boldsymbol{\alpha}$,
2. Calculate Fisher's $\boldsymbol{\alpha}$ for `site1` of BCI.

```
Fisher <- function(x = ""){
  fisher.alpha(x)}

Fisher(site1)
```

```
##        1
## 35.67297
```

Question 8: How is Fisher's $\boldsymbol{\alpha}$ different from $E_{H'}$ and E_{var}? What does Fisher's $\boldsymbol{\alpha}$ take into account that $E_{H'}$ and E_{var} do not?

> ***Answer 8***: From what I can find, Fisher's seems to be different from the other two diversity metrics we used in that it is only based on two constants, and the number of species seen with a certain abundance (n). Simpson's and Shannon's on the other hand, are based on proportions of individuals found belonging to certain species, not actual counts. What this seems to amount to is that the output of Fisher's is an actual predicted value of the number of species at different levels of abundance, rather than an abstract value representing general diversity, as given by the other two metrics calculated (I think!).

6) MOVING BEYOND UNIVARIATE METRICS OF α DIVERSITY

The diversity metrics that we just learned about attempt to integrate richness and evenness into a single, univariate metric. Although useful, information is invariably lost in this process. If we go back to the rank-abundance curve, we can retrieve additional information – and in some cases – make inferences about the processes influencing the structure of an ecological system.

Species abundance models

The RAC is a simple data structure that is both a vector of abundances. It is also a row in the site-by-species matrix (minus the zeros, i.e., absences).

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.

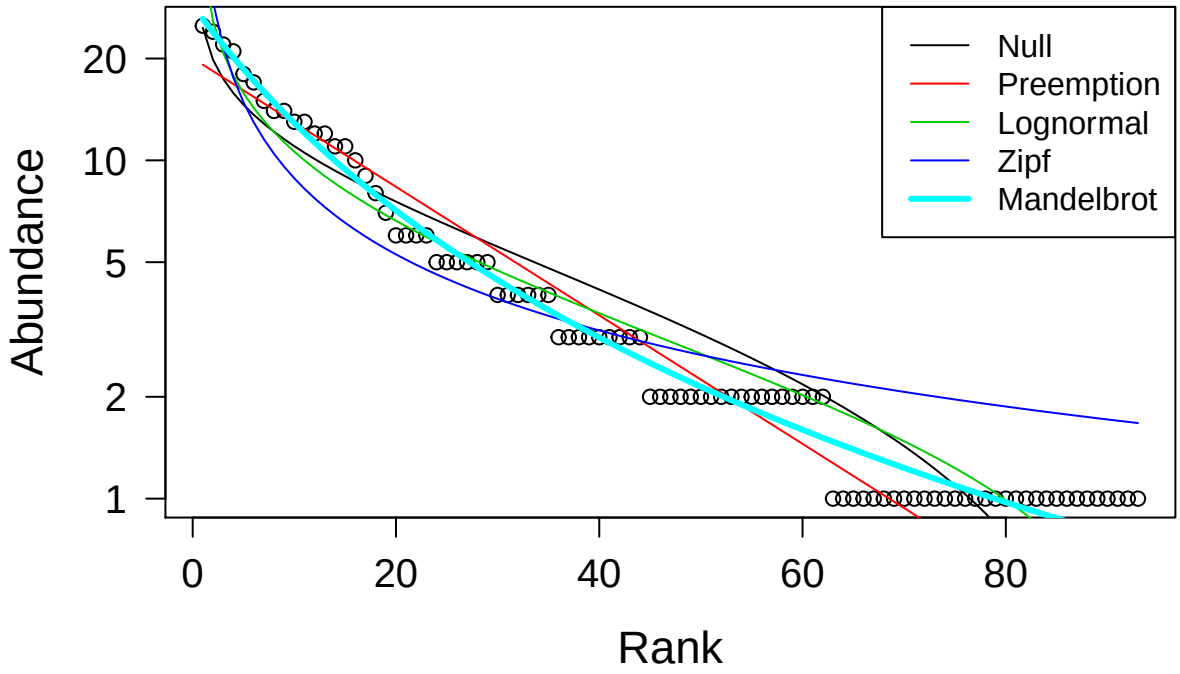
In the R code chunk below, please do the following:

1. Use the `radfit()` function in the `vegan` package to fit the predictions of various species abundance models to the RAC of `site1` in BCI,
2. Display the results of the `radfit()` function, and
3. Plot the results of the `radfit()` function using the code provided in the handout.

```
RACresults <- radfit(site1)
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.new()
plot(RACresults, las = 1, cex.lab = 1.4, cex.axis = 1.25)
```



Question 9: Answer the following questions about the rank abundance curves: a) Based on the output of `radfit()` and plotting above, discuss which model best fits our rank-abundance curve for `site1`? b) Can we make any inferences about the forces, processes, and/or mechanisms influencing the structure of our system, e.g., an ecological community?

Answer 9a:
Based on the plot above, in conjunction with the numerical output from radfit(), I believe it is pretty clear that the Zipf-Mandelbrot model best fits our rank-abundance curve generated for site1. The Mandelbrot line seems to most consistently pass through the points of our RAC, and this model also has the lowest calculated deviance (least amount of variance unexplained by the model), and also the lowest AIC and BIC values (which according to the handout is a good thing! :))

Answer 9b: Based on the RAC generated, it seems safe to infer that there is a fair amount of competition occuring in the system being observed. Competition for resources between species would go a long way to explaining the presence of a few, very abundant species, and many more rare species. The species which are the best competitors, are outcompeting their competition in this system, gaining the lion's share of resources, and becoming common. I feel like competiton in this particular case of site1 makes the most sense, as the data is from the presence/absence of various tree species in a tropical forest. It is easy to imagine that the competition for resources

such as light (race to the top of the canopy/be forced into using the few unexploited light spots on the forest floor), water, and nutrients must be extremely intense. Thus any species with even a slight competitive advantage could become far more successful when compared to less fit species.

Question 10: Answer the following questions about the preemption model: 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 li000347621400034762140003476214ne in the RAD plot?

Answer 10a:
The preemption model seems to assume that the relationship between total abundance, and total resources that can be preempted is inversely proportional. As the total abundance in a community increases, the amount of resources left to be preempted decreases linearly. When there are no species present, all resources can still be preempted, when the first species arrives, a large number of resources are still up for grabs, but when the community is full, there are no more resources left for subsequent species to steal away.

Answer 10b: The niche preemption model looks like a straight line in the RAD plot because according to the model each subsequent species in rank order has exactly half the abundance as the one before. Each species' abundance within the community is defined as proportional to the amount of resources they are able to obtain, which is also directly proportional to their rank order, or the order in which they "arrived" within the community. So, for an example: if the first species to arrive in a region takes .5 of the available resources for themself, the next species to arrive would take .25 of the leftovers, the next species would take .125, the next .075, and so on. Because of this, the relationship between abundance and rank will be a line of constantly negative slope.

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: I would assume that it is important to account for the number of parameters a given model uses when judging how well it explains some data because it seems like it would be too easy to jam a bunch of very specific parameters into a model to explain one dataset really well, but then have limited or no broad generalizability to other datasets. That is, given a particular set of data, some very specific parameters could be devised that would "explain" it or other datasets just like it perfectly well, but they would be too specific and thus useless in other situations that they were not tailor-made to worki in.

SYNTHESIS

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 $D = \sum \frac{n_i(n_i-1)}{N(N-1)}$. Assuming a finite community, calculate Simpson's D, 1 - D, and Simpson's inverse (i.e. 1/D) for `site 1` of the BCI site-by-species matrix.

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

SimpD.finite(site1)
```

```
## [1] 0.02319032
```

```
D.inv.finite <- 1/SimpD.finite(site1)
D.sub.finite <- 1-SimpD.finite(site1)
D.inv.finite
```

```
## [1] 43.12145
```

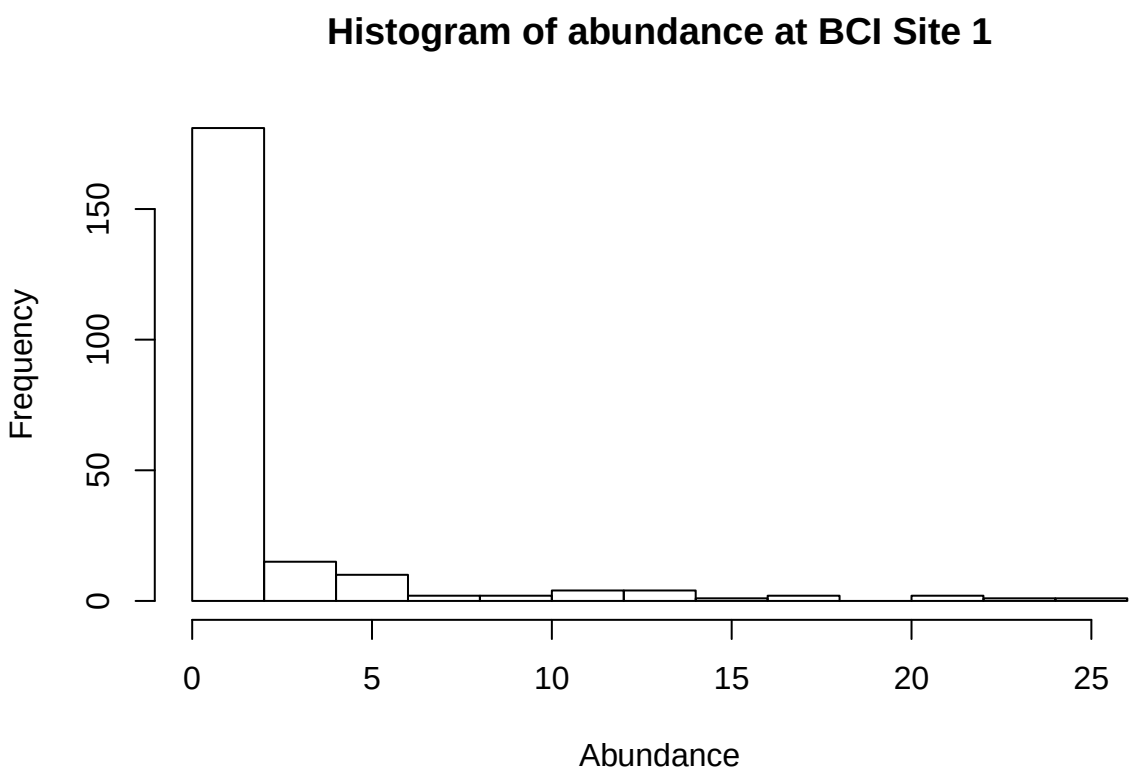
```
D.sub.finite
```

```
## [1] 0.9768097
```

2. Along with the rank-abundance curve (RAC), another way to visualize the distribution of abundance among species is with a histogram (a.k.a., frequency distribution) that shows the frequency of different abundance classes. For example, in a given sample, there may be 10 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.

```
site1.hist <- as.numeric(site1)

hist(site1.hist, plot = TRUE, xlab = "Abundance", main = "Histogram of abundance at BCI Site 1")
```



The general pattern shown in the histogram plotted above is very similar to that seen in the earlier RAC we plotted, just with the axes flipped. Here the y-axis shows the number of species seen with a certain number of individuals, and the x-axis shows what number of individuals that actually is. Said another way, in the site1 dataset there were a whole bunch of species that weren't seen, so the histogram shows a large peak on the left side of the graph, at 0. On the other hand, there were very few species who were really abundant, so the right side of the graph is quite low.

3. We asked you to find a biodiversity dataset with your partner. This data could be one of your own

or it could be something that you obtained from the literature. Load that dataset. How many sites are there? How many species are there in the entire site-by-species matrix? Any other interesting observations based on what you learned this week?

```
my.data <- read.table("/media/removable/USB Drive/Z620.jay.non.GitHub/Endophyte communities and trees f

dim(my.data)
```

```
## [1]  97 847
```

```
# 97 sites (rows) and 847 species (columns)

my.data.otu.only <- my.data[,2:847]

S.obs(my.data.otu.only)
```

```
##  [1] 63 55 52 66 59 44 59 56 56 46 49 53 71 53 55 68 57 65 40 38 46 61 45
## [24] 38 48 54 49 32 57 51 62 67 53 61 47 60 63 43 50 54 55 30 49 52 39 59
## [47] 31 44 34 40 43 34 20 38 42 73 73 54 60 47 58 26 53 44 56 44 23 41 33
## [70] 51 33 57 45 49 47 38 40 37 26 36 35 39 57 66 69 51 54 34 37 47 47 22
## [93] 48 55 52 49 48
```

```
specnumber(my.data.otu.only)
```

```
##  [1] 63 55 52 66 59 44 59 56 56 46 49 53 71 53 55 68 57 65 40 38 46 61 45
## [24] 38 48 54 49 32 57 51 62 67 53 61 47 60 63 43 50 54 55 30 49 52 39 59
## [47] 31 44 34 40 43 34 20 38 42 73 73 54 60 47 58 26 53 44 56 44 23 41 33
## [70] 51 33 57 45 49 47 38 40 37 26 36 35 39 57 66 69 51 54 34 37 47 47 22
## [93] 48 55 52 49 48
```

Despite there being 847 distinct fungal species present in my site-by-species matrix, no one site seems to have over 90 (maybe even 80?) species present. So richness is pretty low at every site, at least in terms of the vast number of possible species tested for (maybe they wouldn't be expected to be at every site anyways). Because the data are given as OTUs and not actual counts, I'm not sure that I can do any estimates of coverage or richness with the data in its current form (or at least I can't think of how to right now).

SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed alpha_assignment.Rmd document, push it to GitHub, and create a pull request. Please make sure your updated repo include both the HTML and RMarkdown files.

Unless otherwise noted, this assignment is due on **Wednesday, January 25th, 2015 at 12:00 PM (noon)**.