Assignment: Local (alpha) Diversity

Michelle Benavidez; Z620: Quantitative Biodiversity, Indiana University

25 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] "C:/Users/Michelle/GitHub/QB2017_Benavidez/Week2-Alpha"
```

```
setwd("C:/Users/Michelle/GitHub/QB2017_Benavidez/Week2-Alpha")
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}
site1=BCI[1,]
S.obs(site1)
```

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

```
specnumber(site1)
```

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

```
sites=BCI[1:4,]
specnumber(sites)
```

```
##  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***: specnumber and the funtion S.obs do return the same number for observes richness. Richness for the first four sites totals 361.

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 - (sum(x == 1) / rowSums(x)) }
C.all=function(x = ""){1 - (rowSums(x == 1) / rowSums(x)) }
C.all(BCI)
```

```
##         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
```

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***: Good's Coverage ranges from 0 - 1.

Answer 2b***: If the number of singletons equaled the number of individuals in the sample, then Good's Coverage would equal 0. This is likely an indication that a substantial number of OTUs were not discovered in the sampling area. (i.e. The site was not sampled well.)

Answer 2c***: Approximtely 7% of the sampled individuals are represented by singletones.

Answer 2d***: Coverage ranges from 0.87 to 0.95 which is relativly good. It seems as if all site were sampled relatively well.

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=T,row.names=1)
soilbac.t=as.data.frame(t(soilbac))
soilbac1=soilbac.t[1,]
specnumber(soilbac1)
```

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

```
C(soilbac1)
```

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

```
rowSums(soilbac1)
```

```
## T1_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***: 2119 observations

Answer 3b***: 1074 unique sequences

Answer 3c***: Coverage at BCI (~0.93) reflects better sampleing coverage than that for KBS (~0.65)

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.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)
}

S.chao1(site1)
```

```
##        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(site1)
```

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

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

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

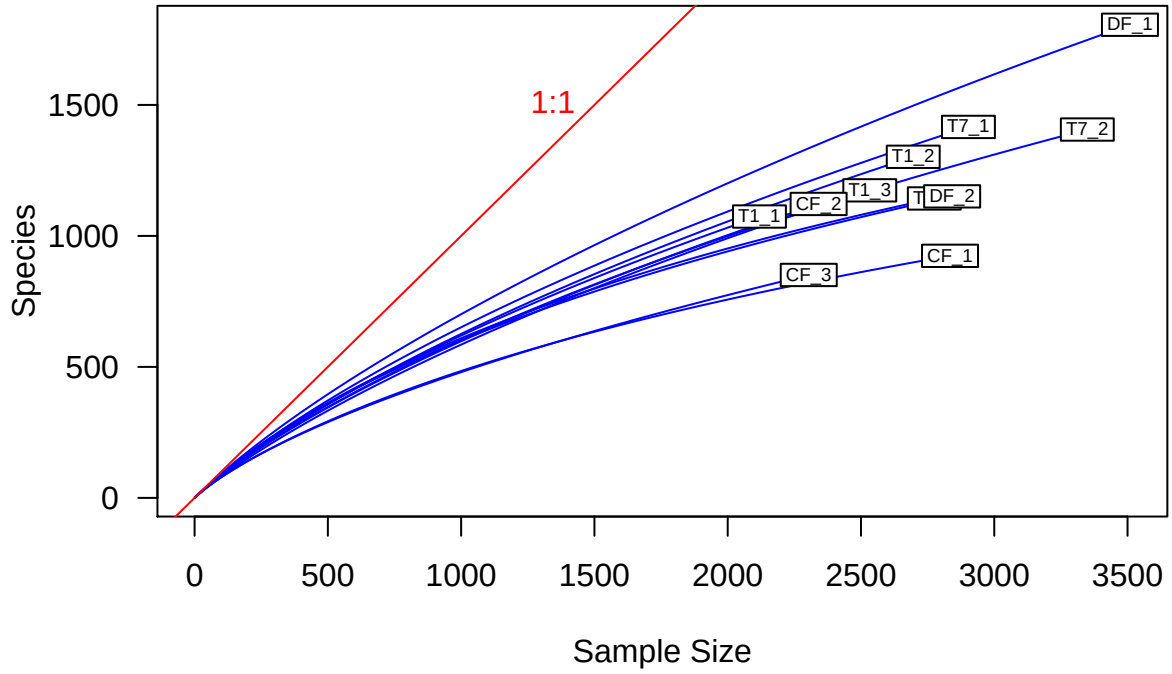
Rarefaction

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.

```
soilbac.S=S.obs(soilbac.t)
min.N=min(rowSums(soilbac.t))
S.rarefy=rarefy(x=soilbac.t,sample=min.N,se=T)
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***: Unlike Chao estimators which only estimate singletons and doublbletons, ACE accounts for rare species (i.e. species numbering less than or equal to 10); therefore, one should use ACE when there is a variety of taxa with only a few individuals.

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

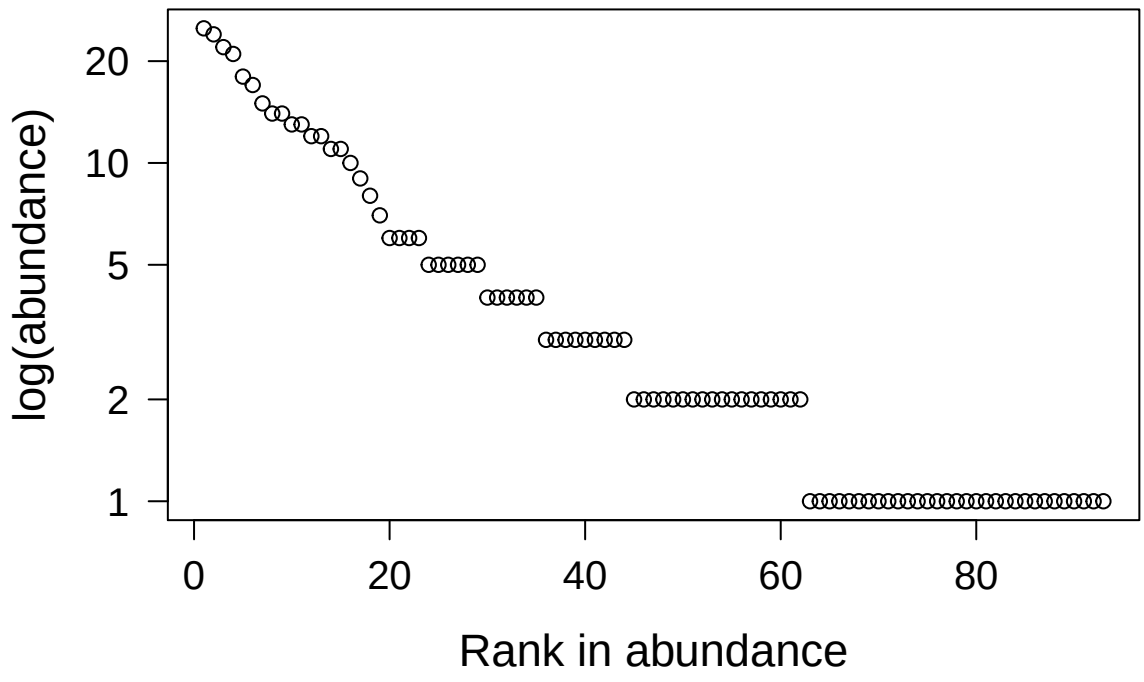
```
RAC=function(x = ""){
x = as.vector(x)
x.ab = x[x > 0]
x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
return(x.ab.ranked)
}
```

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=BCI[1, ]
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 = "log(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***: The log-scale visiualization gives as idea of how evenly abundance is distributed among observations. Further it gives us insight on what ranking of species are most responsible for the uneveness.

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`.

```
par <- opar
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***: Reponses from Simpson (~0.42) are not equal to Smith and Wilson (~0.51); therefore they two tests do not necessarily agree. This disagreement is likely due to Simpson being biased towards differences in the most abundant species. Smith and Wilson provides a more robust meaure of evenness than Simpson. Due to the unevenness in dataset Smith and Wilson provide a better measure. Based on results of Smith and Wilson, eveness is moderatly even.

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)
}
```

```
ShanH(site1)
```

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

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

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

```
EH=4.018412/log(225)
EH
```

```
## [1] 0.7419382
```

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***: E_{var} (~0.51) does not agree with E_H (~0.74). This is because the former only calculates evenness while the latter incorporates evenness and richness together thereby providing a measure of uncertainity.

Fisher's α

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.

```
rac=as.vector(site1[site1>0])
invD=diversity(rac,"inv")
invD
```

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

```
Fisher=fisher.alpha(rac)
Fisher
```

```
## [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***: Fisher's alpha is different from Shannon's and Simth & Wilson's because it uses the log seeries distribution. In addition, Fisher's is also different from Smith & Wilson's and Shannon's in that Fisher's takes into account sampling error.

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

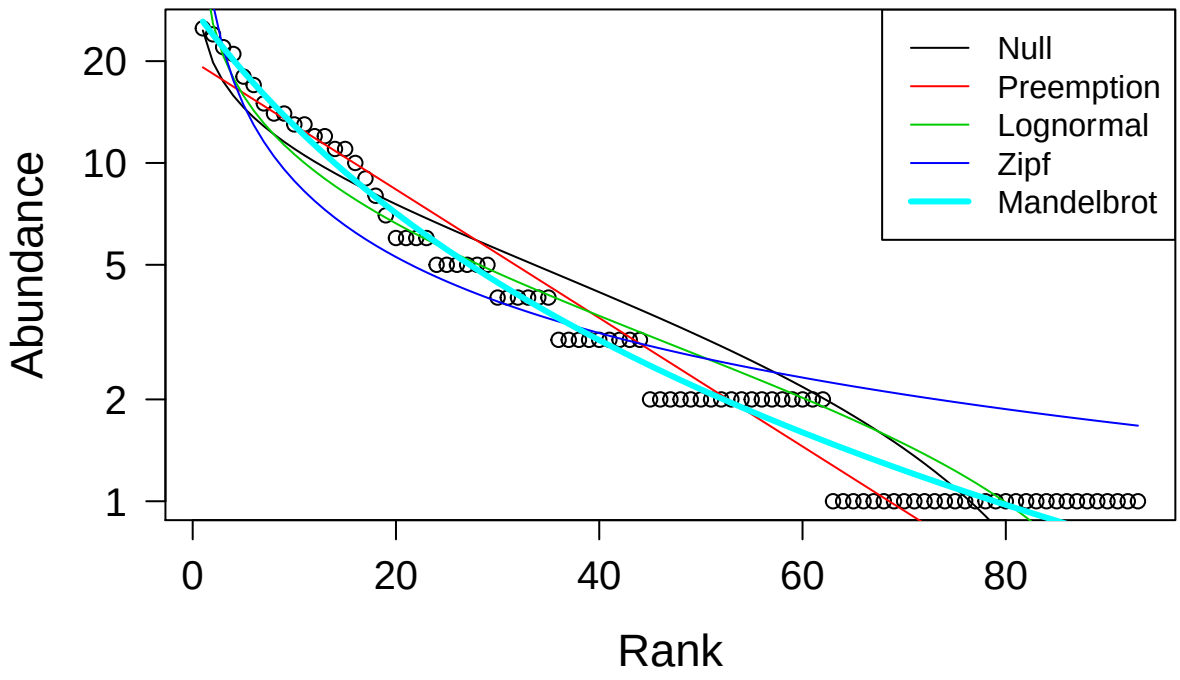
The RAC is a simple data structure that is both a vector of abundances. It is also a row in the site-by-species matrix (minus the zeros, i.e., absences).

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)
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 visualization of the results, it appears as if the Mandelbrot is the best fit for our site1 RAC. This model falls closely to the middle of most observations in the data set and does not deviate significantly from the distribution. Answer 9b***: The community seems to have low evenness. The steep slope at the beginning suggests that more abundant species (low rank) are at much higher abundances than lower ranked species. Looking at individual species and understanding life history of those species may give us a better understanding on what type of mechanisms may be at play.

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 line in the RAD plot?

> Answer 10a***: The preemption model assumes a geometric relationship between rank and abundance; therefore, the total resources that can be preempted for each new species is directy proportional to the previously colonized species. Answer 10b***: The preemption model looks like a straight line compared to the other models on the RAC because of the geometric relationship it assumes. For example, if the first colonizing species represents 1, then the second would represent 0.5, the third 0.25, etc... so it looks like a linear relationship when graphed.

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***: You should have a number of parameters that is appropriate for your data so you

can maximize adequate fit of the model. Too many parameters for your data and you may overfit your model. Too few parameters and your estimator may not be flexiable enough to identify trends in the data.

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. > Answer: Simpson's D: [1] 0.9768097; Simpson's D Inverse: [1] 43.12145
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. > Answer: Data appears to be right-skewed with zero being the most common number This is followed by a sharp decline in frequency which then preceeds to steadily decline.
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. > We have data for seven years - for these questions, I am entering data for the earilest of those years (1982).

How many sites are there? > Answer: 1250

How many species are there in the entire site-by-species matrix? > Answer: 306

Any other interesting observations based on what you learned this week? > Answer: Coverage seems to be above 0.75 for all sites.

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)**.