5. Worksheet: Alpha Diversity

Thomas Zambiasi; Z620: Quantitative Biodiversity, Indiana University

25 January, 2023

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. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) to your name.
2. Complete as much of the worksheet as possible during 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. Answer questions in the worksheet. 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 ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For the assignment portion of the worksheet, 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, submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file `AlphaDiversity_Worskheet.Rmd` and the PDF output of `Knitr` (`AlphaDiversity_Worskheet.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 `5.AlphaDiversity` folder, and 4) Load the `vegan` R package (be sure to install first if you haven't already).

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

```
## [1] "C:/Users/tmzam/GitHub/QB2023_Zambiasi/2.Worksheets/5.AlphaDiversity"
```

```
setwd("C:/Users/tmzam/GitHub/QB2023_Zambiasi/2.Worksheets/5.AlphaDiversity")
library("vegan")
```

```
## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.6-2
```

2) LOADING DATA

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

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

```
## 'data.frame':    50 obs. of  225 variables:
##  - attr(*, "original.names")= chr [1:225] "Abarema.macradenium" "Acacia.melanoceras" "Acalypha.divers
```

3) SPECIES RICHNESS

Species richness (S) refers to 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` of the BCI data set, 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
```

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 four sites (i.e., rows) of the BCI matrix?

> ***Answer 1***: Both the specnumber() and S.obs() functions return the same values for observed richness at site 1 (which would be 93 species). The species richness at the first four sites would be 93 spp. at Site 1, 84 spp. at Site 2, 90 spp. at Site 3, and 94 spp. at Site 4.

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.

```
goods.c<-function(x=""){
  1-(rowSums(x==1)/rowSums(x))
}
goods.c(BCI)
```

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

Answer 2a: Good's Coverage can generate values between 0 and 1.

Answer 2b: If the number of singleton species equalled the total number of sampled species, the value of C in Good's Coverage would be 0 and we could conclude that no sampled species would have been seen more than once.

Answer 2c: Site 1 has a C value of 0.9308, so roughly 7% of the taxa were singletons at this site.

Answer 2d: Based on Good's Coverage, I would say that the BCI plots mostly contain species that occur more than once. All calculated C values fall between 0.87 and 0.95, indicating that singleton species make up only ~5-10% of the samples. Since larger sample sizes would tend to include more species (with rare spp. maybe seen only once), the sampling coverage at these plots does not appear to be very thorough.

Estimated richness

In the R code chunk below, do the following:

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

```
soilbac<-read.table("data/soilbac.txt",sep="\t",header=TRUE,row.names=1)
soilbac.t<-as.data.frame(t(soilbac))
soilbac1<-soilbac.t[10,]
S.obs(soilbac1)
```

```
## CF_2
## 1122
```

```
goods.c(soilbac1)
```

```
##      CF_2
## 0.6621102
```

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 sample soilbac1 (CF_2) had 1,122 recovered sequences.

Answer 3b: The observed richness of soilbac1 (CF_2) is 1,122.

Answer 3c: The coverage in the KBS sample was more extensive than that of the BCI sample. The Good's Coverage value at KBS was 0.6621, indicating that roughly 34% of the taxa observed were singletons. At BCI Site 1, the Good's Coverage value of 0.9308 indicates that only ~7% of the taxa were singletons.

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 `site1` and `soilbac1`.

```r
#chao1 function
chao1<-function(x=""){
  S.obs(x)+((sum(x==1)^2)/(2*sum(x==2)))
}
#chao2 function
chao2<-function(site="",SbyS=""){
  SbyS=as.data.frame(SbyS)
  x=SbyS[site,]
  SbyS.pa<-(SbyS>0)*1 #converts SbyS to p/a
  Q1=sum(colSums(SbyS.pa)==1) #spp. observed once
  Q2=sum(colSums(SbyS.pa)==2) #spp. observed twice
  chao2=S.obs(x)+((Q1^2)/(2*Q2))
  return(chao2)
}
#ace function
ace<-function(x="",thresh=10){
  x<-x[x>0] #excludes zero-abundance taxa
  S.abund<-length(which(x>thresh)) #richness of abundant taxa
  S.rare<-length(which(x<=thresh)) #richness of rare taxa
  singlt<-length(which(x==1)) #number of singleton taxa
  N.rare<-sum(x[which(x<=thresh)]) #abundance of rare indv.
  C.ace<-1-(singlt/N.rare) #coverage (prop. non-singleton rare indv.)
  i<-c(1:thresh)
  count<-function(i,y){
    length(y[y==i])
  }
  a.1<-sapply(i,count,x) #number indv. in richness i richness classes
  f.1<-(i*(i-1))*a.1 #k(k-1)kf, per Gotelli
  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)
}

#finding richness at BCI site 1
#first creating specific variable for Site 1
site1<-BCI[1,]
#w/ chao1
chao1(site1)
```

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

```r
#w/ chao2
chao2("1",BCI)
```

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

```r
#w/ ACE
ace(site1)
```

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

```
#finding richness in soilbac1 (CF_2)
#w/ chao1
chao1(soilbac1)
```

```
##     CF_2
## 2984.146
```

```
#w/ chao2
chao2("CF_2",soilbac.t)
```

```
##     CF_2
## 21103.39
```

```
#w/ ACE
ace(soilbac1)
```

```
## [1] 4908.339
```

Question 4: What is the difference between ACE and the Chao estimators? Do the estimators give consistent results? Which one would you choose to use and why?

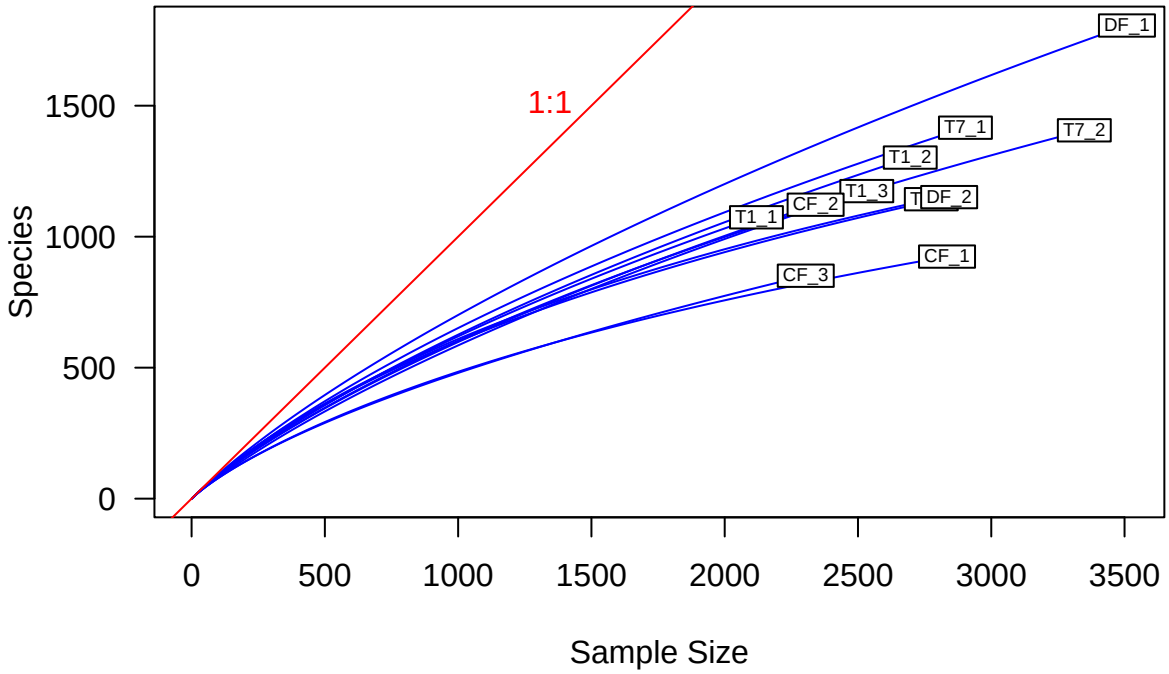
> ***Answer 4***: The ACE and Chao estimators both attempt to give a metric of species richness by combining observed richness with added estimates of rare species that may have been missed due to sampling constraints. The Chao estimators do this by just factoring in singleton and doubleton species (within a site for Chao1, across a siteXspp. matrix for Chao2). The ACE estimator adds to the observed richness based on its threshold of rare species of those with 10 or fewer individuals. These estimators all give different results; Chao1 and Chao2 give values that are relatively close but are measuring richness at different scales (single site v. multi-site). ACE gives a much larger value for its richness estimator than either of the Chao metrics. The choice of richness estimator probably depends on the context it's being used in, so in this case I would choose the Chao1 metric since both BCI Site 1 and the CF_2 plot in the soilbac data are single sites and I'm just trying to estimate spp. richness in one place.

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.

```r
#finding observed richness for all soilbac samples
soilbac.all<-S.obs(soilbac.t)
#getting size of smallest sample
small.soilbac<-min(rowSums(soilbac.t))
#rarefying every sample in soilbac to this level
sb.rarefy<-rarefy(soilbac.t,small.soilbac,se=TRUE)
#plotting rarefaction results with 1:1 line/label
rarecurve(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")
```



4) SPECIES EVNENNESS

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

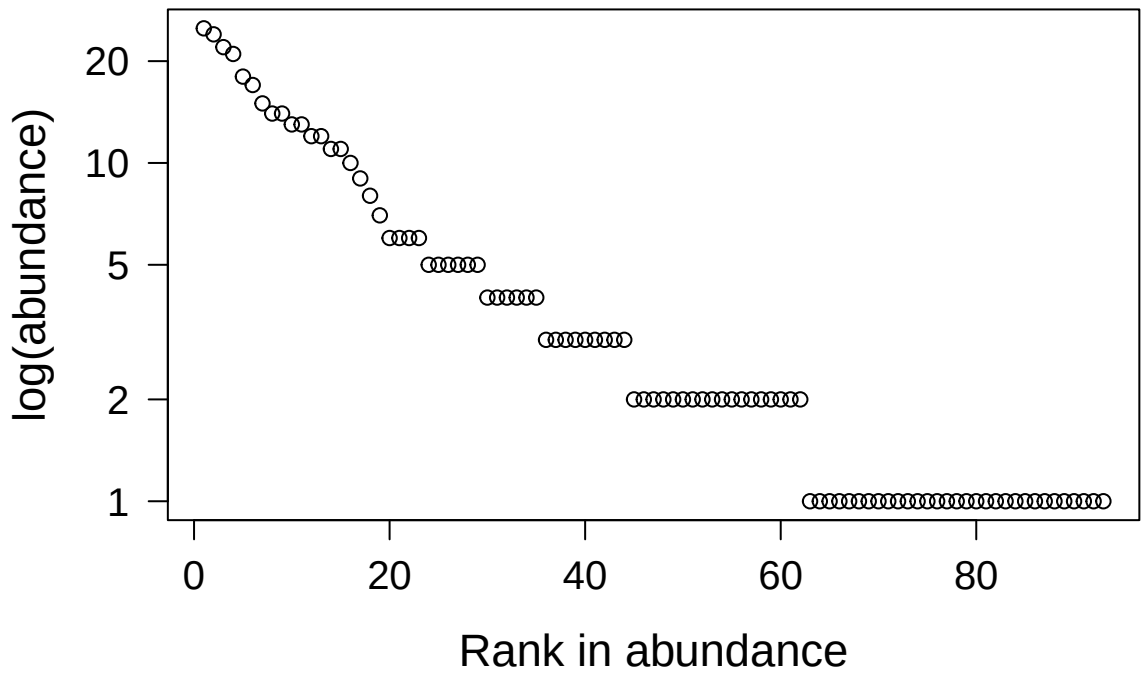
```
#function to create RAC
RAC<-function(x=""){
  x.ab=x[x>0] #this line makes sure only spp. with abundance > 0 are included
  x.ab.ranked=x.ab[order(x.ab,decreasing=TRUE)] #this line ranks spp. most to least abundant
  as.data.frame(lapply(x.ab.ranked,unlist))
  return(x.ab.ranked) #this line makes sure the output is the ranked vector
}
```

Now, let us 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)"

```
#creating ranks for BCI site1
s1.rac<-RAC(site1)
#creating RAC plot
s1.ranks<-as.vector(seq(1,length(s1.rac)))
opar<-par(no.readonly=TRUE) #saves default plot parameters
par(mar=c(5.1,5.1,4.1,2.1)) #creates new settings for par
plot(s1.ranks,log(s1.rac),type='p',axes=F, xlab="Rank in abundance",ylab="log(abundance)",las=1,cex.lab=
box() #adds a border to plot
axis(side=1,labels=T,cex.axis=1.25) #adds x-axis
axis(side=2,las=1,cex.axis=1.25,labels=c(1,2,5,10,20),at=log(c(1,2,5,10,20))) #adds log-scale y-axis
```




```
par<-opar #resets plot parameters
```

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***: Looking at species abundance data with a log-scaled axis helps to make some of the nuances in species evenness more clear. Many communities follow a similar phenomenon to the one seen in the graph, where many species are relatively rare and a smaller proportion have greater abundances. Without log-scaling the abundance axis, RAC plots would generally show a dramatic spike of the most abundant species, followed by a long (and low) plateau of rare species. The log-scaling compresses the abundance axis so that we can assess the evenness of the most abundant species in the community more clearly.

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`.

```
#writing function for Simpson's evenness
simpE<-function(x=""){
  S<-S.obs(x)
  x=as.data.frame(x)
  D<-diversity(x,"inv")
  E<-D/S
  return(E)
}

#calculating Simpson's evenness for BCI Site 1
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}.

```
#making function for S/W evenness
evar<-function(x){
  x<-as.vector(x[x>0])
  1-(2/pi)*atan(var(log(x)))
}

#calculating S/W evenness for BCI Site 1
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 Simpson's Evenness index gives a slightly lower evenness value (0.4238) than the value given by the Smith/Wilson Evenness index (0.5067). These values do not agree with one another; this is likely because the Simpson's index can reflect more information from notably abundant species in a community and make it seem more or less even than it really is. Because the Smith/Wilson index avoids this problem and has a larger value for evenness (closer to true evenness), I would conclude that the most abundant species at BCI Site 1 are not very even with one another or are much more abundant than the least abundant species. This skews the Simpson's index to make it appear that the community is less even than it is.

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".

```
#creating function for shannon index
shan.H<-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)
}

#using shannon index function to calculate diversity index for BCI Site 1
shan.H(site1)
```

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

```
#comparing to 'vegan' diversity function
diversity(site1,index="shannon")
```

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

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".

```
#creating function for simpson's diversity
simpD<-function(x=""){
  D=0
  N=sum(x)
  for(n_i in x){
    D=D+(n_i^2)/(N^2)
  }
  return(D)
}

#calculating simpson's diversity for BCI Site 1
simpD(site1)
```

```
## [1] 0.0253707
```

```
#inverse of simpson's diversity
Dinv<-1/(simpD(site1))
print(Dinv)
```

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

```
#1 - simpson's diversity
Dsub<-1-(simpD(site1))
print(Dsub)
```

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

```
#calculating simpson's diversity with 'vegan' diversity function to compare
diversity(site1,index="simp")
```

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

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.

```
#first need to create new vector of site 1 spp. with abundances greater than 0
rac2<-as.vector(site1[site1>0])

#first calculating inverse Simpson's D for site 1 (for comparison)
rac.invD<-diversity(rac2,"inv")
rac.invD
```

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

```
#now finding Fisher's alpha for BCI site 1
s1Fisher<-fisher.alpha(rac2)
s1Fisher
```

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

Question 7: 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 7***: Fisher's alpha is different from other diversity and evenness metrics like the Shannon index (H') or the Smith/Wilson index (Evar) because it consists of only one parameter (alpha) which comes from the data, while these other indices are calculated from more consistent formulas. Fisher's alpha accounts for the fact that it's unlikely that 100% of a community will be sampled, so it is estimating diversity of the whole community if it were to be fully sampled. These other metrics are calculating an index to represent diversity instead of estimating it.

6) HILL NUMBERS

Remember that we have learned about the advantages of Hill Numbers to measure and compare diversity among samples. We also learned to explore the effects of rare species in a community by examining diversity for a series of exponents q.

Question 8: Using `site1` of BCI and `vegan` package, a) calculate Hill numbers for q exponent 0, 1 and 2 (richness, exponential Shannon's entropy, and inverse Simpson's diversity). b) Interpret the effect of rare species in your community based on the response of diversity to increasing exponent q.

```
#code for calculating Hill numbers for BCI Site 1

#first is q=0, in which diversity is equal to sp. richness (calculate w/ specnumber)
site1q0<-specnumber(site1)
site1q0
```

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

```
#next is q=1, where diversity is equal to the exponential of shannon diversity index
site1q1<-exp(diversity(site1,index="shannon"))
site1q1
```

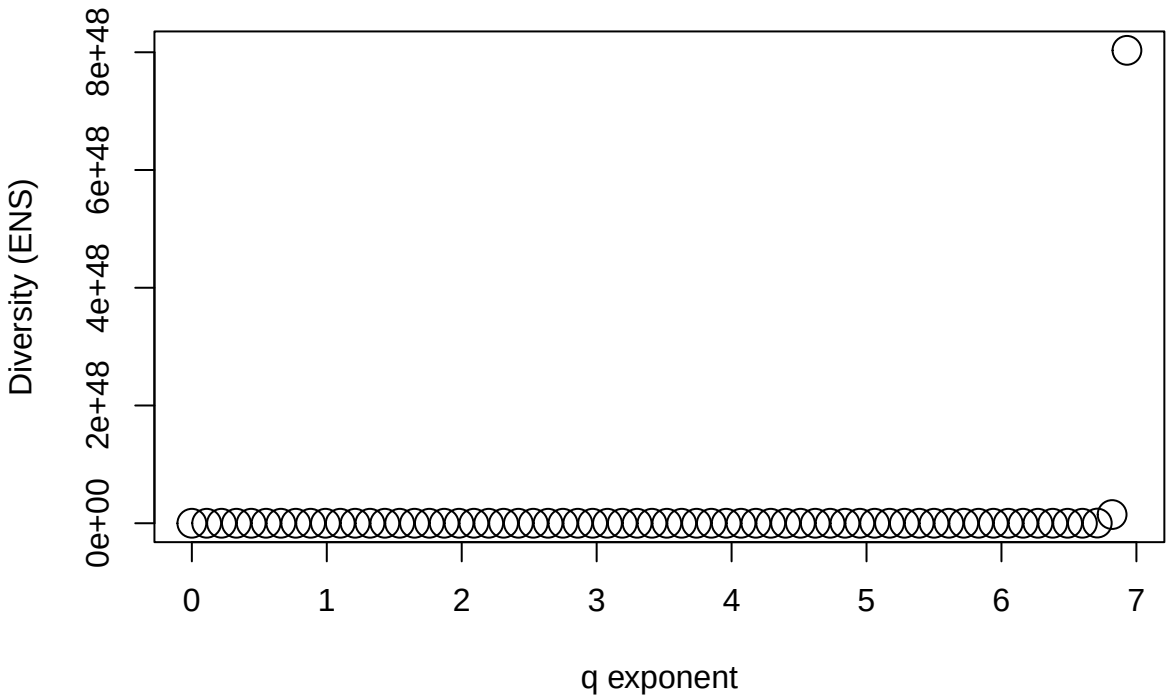
```
## [1] 55.6127
```

```
#last is q=2, where diversity is equal to reciprocal of simpson diversity
site1q2<-diversity(site1,index="invsimpson")
site1q2
```

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

```
#profile function for finding diversity from Hill numbers equation across range of q values
profile<-function(C){
  cbind(seq(0,7,by=0.11),
        unlist(lapply(seq(0,7,by=0.11),function(q)sum(apply(C,1,function(x)
          (x/sum(x))^q))^(1/1-q))))
}

s1Hill<-profile(site1)
plot(s1Hill[,1],s1Hill[,2],cex=2,xlab="q exponent",ylab="Diversity (ENS)")
```



Answer 8a: The Hill numbers for BCI Site 1 are 93 at q=0, 55.6127 at q=1, and 39.4156 at q=2. ***Answer 8b***: The diversity values drop fairly drastically with increases in the q exponent. Higher exponent values correspond to Shannon and Simpson diversity indices, both of which incorporate species evenness. These indices on their own are more heavily influenced by the most abundant species, potentially making a community seem less diverse than it really is. Overall, higher values of the exponent q give lower values for the effective number of species. *Note: I don't believe I completed this section correctly; I got a little confused and just calculated the Hill numbers for what the handout said each value of the q exponent was. I wanted to include the profile function because it seemed like that was the way to use the general Hill equation, but I wasn't quite sure how to interpret the output. I'd really appreciate any feedback on this part!*

##7) 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.

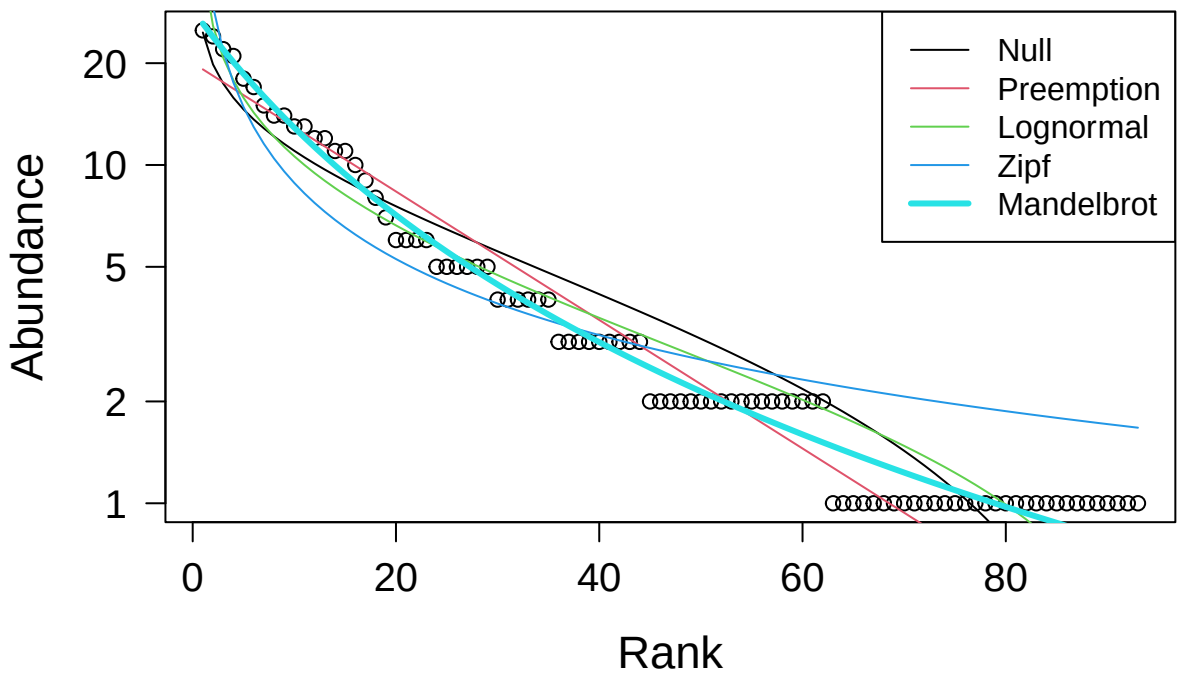
In the R code chunk below, please do the following:

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.

```
#using radfit function for BCI Site 1
s1rad<-radfit(site1)
s1rad
```

```
##
## 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
```

```
#plotting radfit results
plot.new()
plot(s1rad,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***: The best model for the BCI Site 1 rank abundance curve is the Mandelbrot model. This model is shown in the plot to follow the trend of the RAC most closely, and the radfit() output indicates that this model had the lowest values for Deviance (4.2271), AIC (286.1372), and BIC (293.7350). The model in the output with the lowest values for these metrics is the most suitable representation of the data. ***Answer 9b***: I would not be confident in making any inferences about what mechanisms might be determining community structure at BCI Site 1 from the outcomes of this test without knowing more context about the environment at this site (things such as nutrient availability, precipitation, temperature, herbivory, etc.). Having more of this contextual data and records of how the community may be changing over time (composition, changes in the RAC) would help with creating a much better hypothesis of what factors may affect the community.

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 niche preemption model assumes that the total abundance relates directly to the total amount of resources. Each additional species in a community takes up the same proportion (given as alpha) of resources; this continues until the resources are depleted. ***Answer 10b***: The preemption model is a straight line in the RAD plot because of the assumption that total abundance is in a linear relationship with the amount of available resources. If each additional species takes up the same amount of resources, then theoretically the differences in abundances between consecutive species in a RAC would be constant. This would produce the linear relationship we see in the RAD plot.

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***: It is important to look at the number of parameters in a model when assessing its fit for the data because we generally want to select the model that fits the data best with the fewest number of parameters. Including too many parameters may "over-fit" the model; in these instances, the model would perfectly describe the data it was created from but would not be a very good descriptor if applied to similar kinds of data that varied much from that original set.

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.

```
#creating function for Simpson's diversity with a  finite community
simpFin<-function(x=""){
  D=0
  N=sum(x)
  for(n_i in x){
```

```
    D=D+(n_i*(n_i-1))/(N*(N-1))
  }
  return(D)
}

#calculating Simpson's D (finite) for BCI Site 1
simpFin(site1)
```

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

```
#calculating 1-D for BCI Site 1
1-simpFin(site1)
```

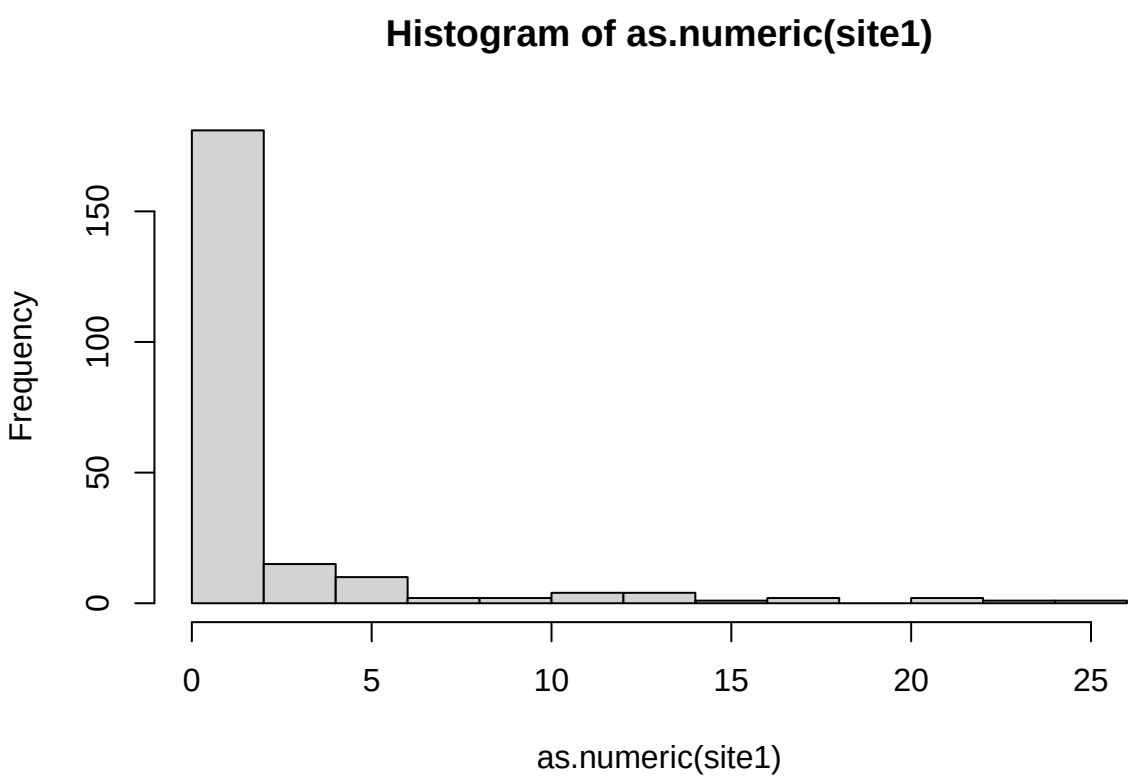
```
## [1] 0.9768097
```

```
#calculating inverse Simpson (finite) for BCI Site 1
1/simpFin(site1)
```

```
## [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.

```
#creating histogram for freq. distribution at BCI Site 1
hist(as.numeric(site1))
```



>The pattern in this histogram is very similar to what I might expect from a rank abundance curve, but with the axes flipped so that spp. abundance is on the x-axis. A vast majority of the species listed have abundances of zero, while fewer have abundances higher than 5 individuals.

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? >*Note: I loaded the dataset through the File menu because I could not get the read.csv() function to load it as anything but a single column, I think it's because of the extra header material*

   This dataset contains 8,044 species, although some do not have a formal name and are just listed as things such as "Surface Litter" or "UnSorted". *Another note: the file would not knit with the dataset being loaded in manually, so I've included a copy of it in my pull request. I found my number for 8,044 after loading the data, keeping "Yes" marked for headers in the wizard, setting the delimiter to comma, and using the function length(KBSmicroplotdata$species) to find the number of species in the dataset. Apologies again for the inconvenience with this answer.* Because of the current state of the dataset (I might need to rearrange some of this into a new data table to make an appropriate siteXspecies matrix) I couldn't create any code to find the number or sites, but the description of the data (https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-kbs.55.49) indicates that there are four different treatment levels that may correspond to sites (these being fertilized, unfertilized, disturbed, and undisturbed). *Apologies for the abbreviated answers to this question, I wanted to make sure I turned this in on time for this week and am planning on working through this dataset more thoroughly while developing my project plans in my group!*

SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed 5.AlphaDiversity_Worksheet.Rmd document, push it to GitHub, and create a pull request. Please make sure your updated repo include both the pdf and RMarkdown files.

Unless otherwise noted, this assignment is due on **Wednesday, January 25th, 2023 at 12:00 PM (noon)**.