5. Worksheet: Alpha Diversity_Bryan Guevara

Student Name; Z620: Quantitative Biodiversity, Indiana University

29 January, 2025

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 `Week-2/` folder folder, and 4) Load the `vegan` R package (be sure to install first if you have not already).

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

```
## [1] "/cloud/project/QB2025_Guevara/Week2-Alpha"
```

```
setwd("/cloud/project/QB2025_Guevara/Week2-Alpha")
install.packages("vegan")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.6-8
```

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
```

```
site1 <- BCI[1, ]
```

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   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##  93  84  90  94 101  85  82  88  90  94  87  84  93  98  93  93  93  89 109 100
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##  99  91  99  95 105  91  99  85  86  97  77  88  86  92  83  92  88  82  84  80
##  41  42  43  44  45  46  47  48  49  50
## 102  87  86  81  81  86 102  91  91  93
```

```
specnumber(BCI)
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##  93  84  90  94 101  85  82  88  90  94  87  84  93  98  93  93  93  89 109 100
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##  99  91  99  95 105  91  99  85  86  97  77  88  86  92  83  92  88  82  84  80
##  41  42  43  44  45  46  47  48  49  50
## 102  87  86  81  81  86 102  91  91  93
```

```
help(specnumber)
```

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***: Yes, it does appear that observed richness in site1 from specnumber matches the observed richness we acquire from our function 'S.obs'. The species richness from the first four sites are 93, 84, 90, and 94.

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))}
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: The range of values that can be generated is any number between 0 and 1 as Good's Coverage is a proportion as we take the ratio of singleton species to total number of individuals in our sample and subtract this value from 1.

Answer 2b: If ni = N then that would result in a 1:1 ratio and we have a Good's Coverage value of 0. This would indicate basically a complete lack of coverage.

Answer 2c: In site 1, approximately 6.92% of the taxa in site1 are represented by singletons as the propotion of represented by singletons would be = 1 - C or 1 - 0.9308.

Answer 2d: Site 23 has the greatest portion of taxa represented by singletons across all the observed sites while site 4 seems to have the lowest. Most sites have ~10% of their taxa represented by singletons.

Estimated richness

In the R code chunk below, do the following:

1. Load the microbial dataset (located in the `Week-2/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

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

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

```r
#observed richness of T1_1
S.obs(soilbac1)
```

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

```r
#Good's Coverage of site T1_1
C(soilbac1)
```

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

```r
dim(soilbac1)
```

```
## [1]     1 13310
```

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: There were a total of 2119 sequences recovered from our soilbac1 sample.

Answer 3b: The observed richness of soilbac1 seems to be 1074

Answer 3c: The coverage betwen the BCI sample and the KBS sample are very different with the KBS sample having a much smaller overall coverage

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
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)
}

#Estimated richness of site1 and soilbac1
S.chao1(soilbac1)
```

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

```
S.chao1(site1)
```

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

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

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

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

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

```
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)
}

#Estimated richness using ACE
S.ace(soilbac1)
```

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

```
S.ace(site1)
```

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

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 difference between ACE and Chao estimators is that ACE uses a threshold to looik at abundance of rare species and defines rare species as taxa that have 10 or fewer individuals. The estimators do not give the most consisten results as they each use different parameters in order to estimate richness (Chao1 makes inferences based on number of singletons and doubletons of species at one site ; Chao2 uses the presence/absence of species across multipel sites; ACE uses a threshold to identifyu abundance of rare species). For our soilbac1 dataset, I would probably avoid ACE because it might underrestimate richness because it assumes a sufficient enough sampling coverage and soilbac1 seems to have rather poor coverage as indicated by the Good's Coverage value meaning that there are a lot rare species. Thus, for soilbac1, I would probably use Chao2 as it considers the presence of species across multiple sites rather than focusing on a singular site as that would be more representative of how rich a species might be in a given ecosystem. I would prefer a holistic POV rather than a the richness of a single site within an ecosystem. For site1 from site1, the coverage is a lot better so I could consider using ACE in

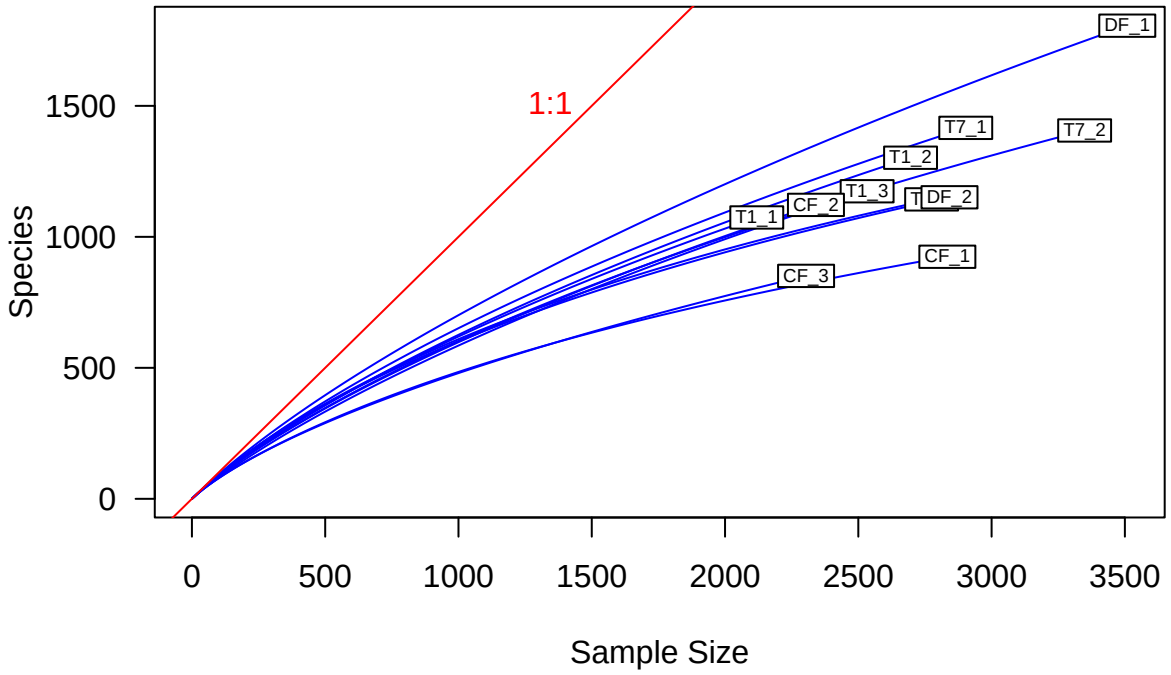
this case as there are relatively fewer rarer species in this site compared to soilbac1.

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 = 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')
```



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

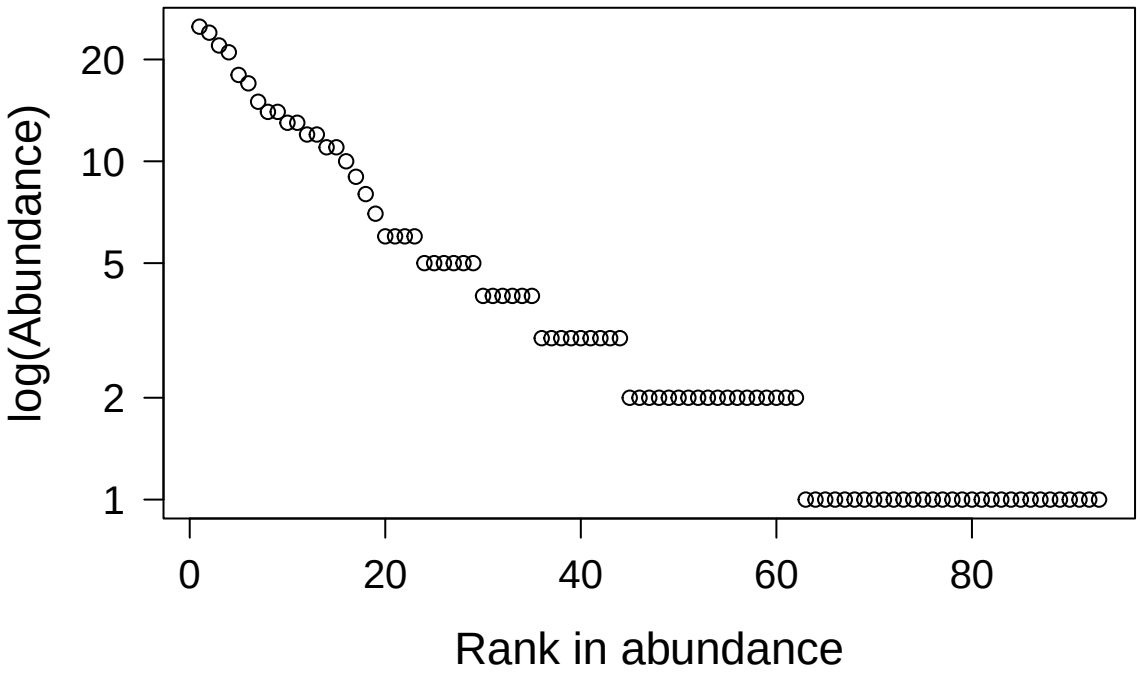
```
RAC <- function(x = ""){
  x.ab = x[x > 0]
  x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
  as.data.frame(lapply(x.ab.ranked, unlist))
  return(x.ab.ranked)
}

#x.ab = x[x > 0] removes species that have an abundance of zero
#x.ab.ranked orders the vectors from greatest (most abundant) to least (least abundant)
```

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)"

```
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***: From what we can see on the graph, when we log scale things, a community that might have high dominance of a few species might look less steep (like around rank 20) which can give the impression of greater evenness when there is in fact less evenness because there are some rather abundant species even though they might not be as abundant as the species ranking from 1-10. In a community with relatively strong evenness, log scaling could hide some minor variations in actual evenness, skewing how we might interpret evenness using our RAC.

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)
}

site1 <- BCI[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}.

```
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 estimates for evenness for site1 do not necessarily agree with Simpson's estimate of evenness is slightly smaller than Smith and Wilson's. They do not agree entirely because Simpson's is more sensitive to differences in the few most abundant species within the measured community. Smith and Wilson's uses the sample variance of the log-transformed abundances and then standardizes it resulting in a different value from Simpson's. From both of the results, I can infer that the evenness at site1 really isn't all that great as the estimates are ~0.5 or lower. I would categorize them as moderately even with 1 being the most even and values of 0 being associated with low evenness.

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
```

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)

diversity(site1, "inv")
```

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

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

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

Fisher's α

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

1. Provide the code for calculating Fisher's α,
2. Calculate Fisher's α 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 7: How is Fisher's α different from $E_{H'}$ and E_{var}? What does Fisher's α take into account that $E_{H'}$ and E_{var} do not?

> ***Answer 7***: Fisher's alpha estimates diversity rather just calculating a diversity metric directly acounting for sampling error. We are not observing every single individual as we would for Smith and Wilson's Evenness Index or for Shannon's diversity.

6) HILL NUMBERS

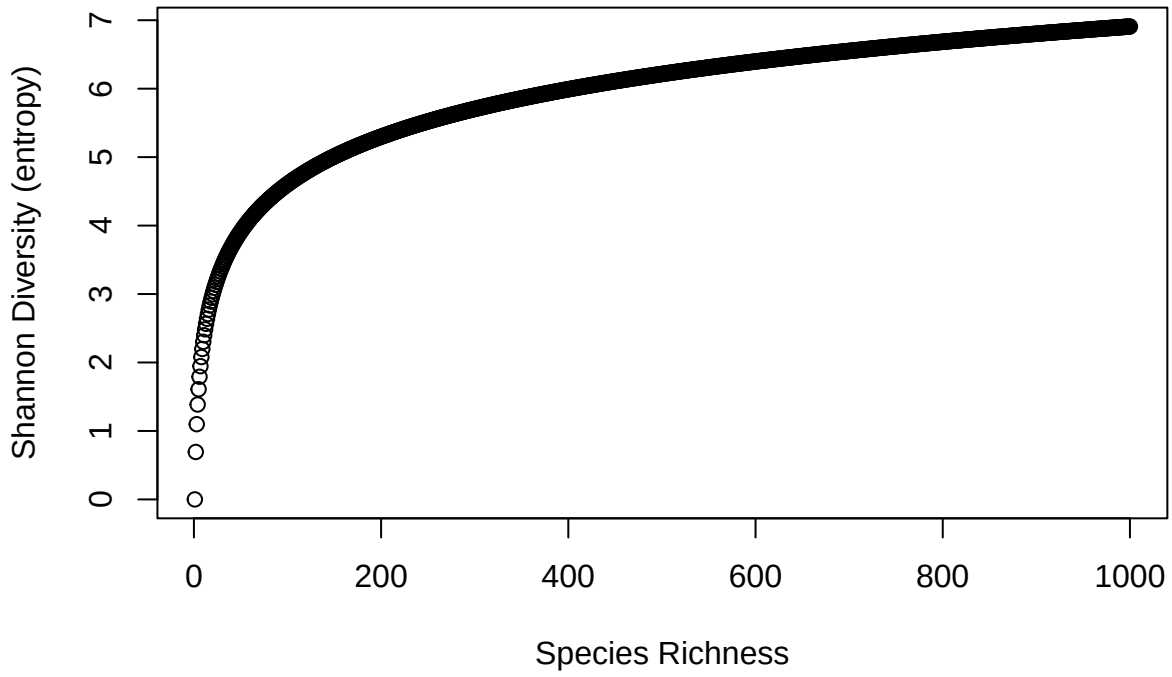
Remember that we have learned about the advantages of Hill Numbers to measure and compare diversity among samples. We also learned to explore the effects of rare species in a community by examining diversity for a series of exponents q.

```r
#Simulate communities
C1 <- data.frame(t(rep(1, 500))); colnames(C1) <- paste("sp", 1:500)
C2 <- data.frame(t(c(rep(1, 250)))); colnames(C2) <- paste("sp", 1:250)

#Calculate shannon diversity
H1 <- diversity(C1, index = "shannon")
H2 <- diversity(C2, index = "shannon")
H1;H2
```

```
## [1] 6.214608
```

```
## [1] 5.521461
```

```r
#Calculating Shannon's entropy for each richness level
H_all <- matrix(ncol = 2, nrow = 1000)
for(i in 1:1000){
  C <- data.frame(t(rep(1, i)))
  colnames(C) = paste("sp", 1:i)
H_all[i,1] <- i
H_all[i,2] <- diversity(C, index = "shannon") }

plot(H_all[,1], H_all[,2], xlab = "Species Richness", ylab = "Shannon Diversity (entropy)")
```



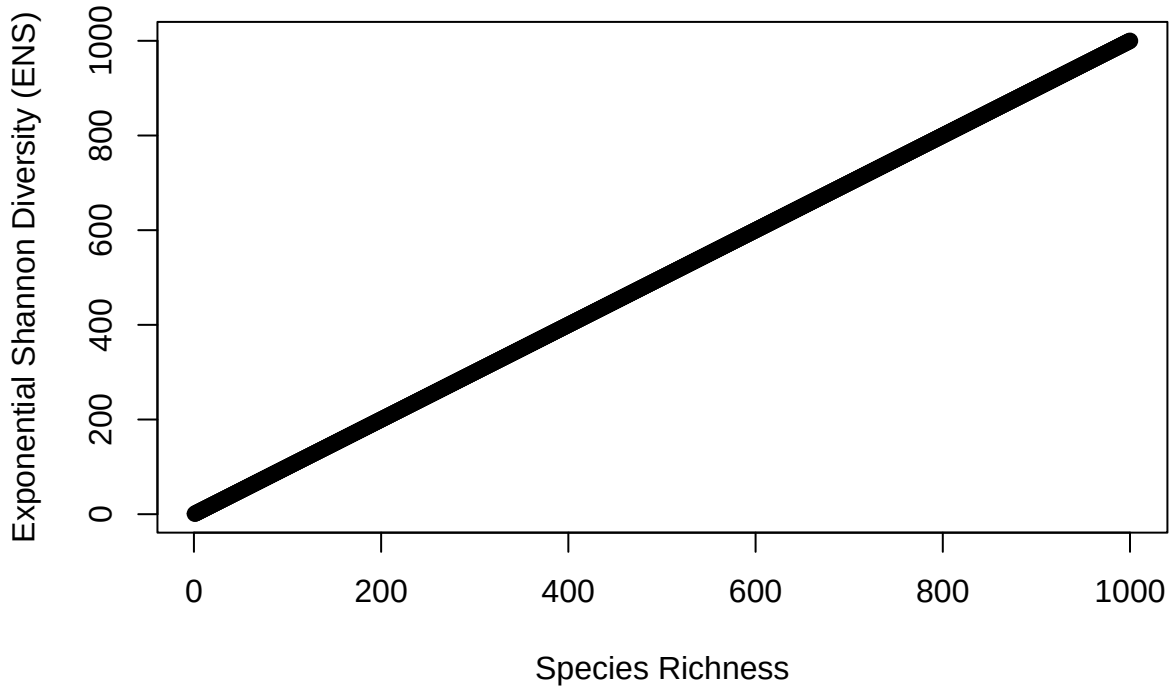

```r
#Calculate exponential Shannon's entropy (equal to the Hill number q = 1)
H1_Hill <- exp(diversity(C1, index = "shannon"))
H2_Hill <- exp(diversity(C2, index = "shannon"))
H1_Hill; H2_Hill
```

```
## [1] 500
```

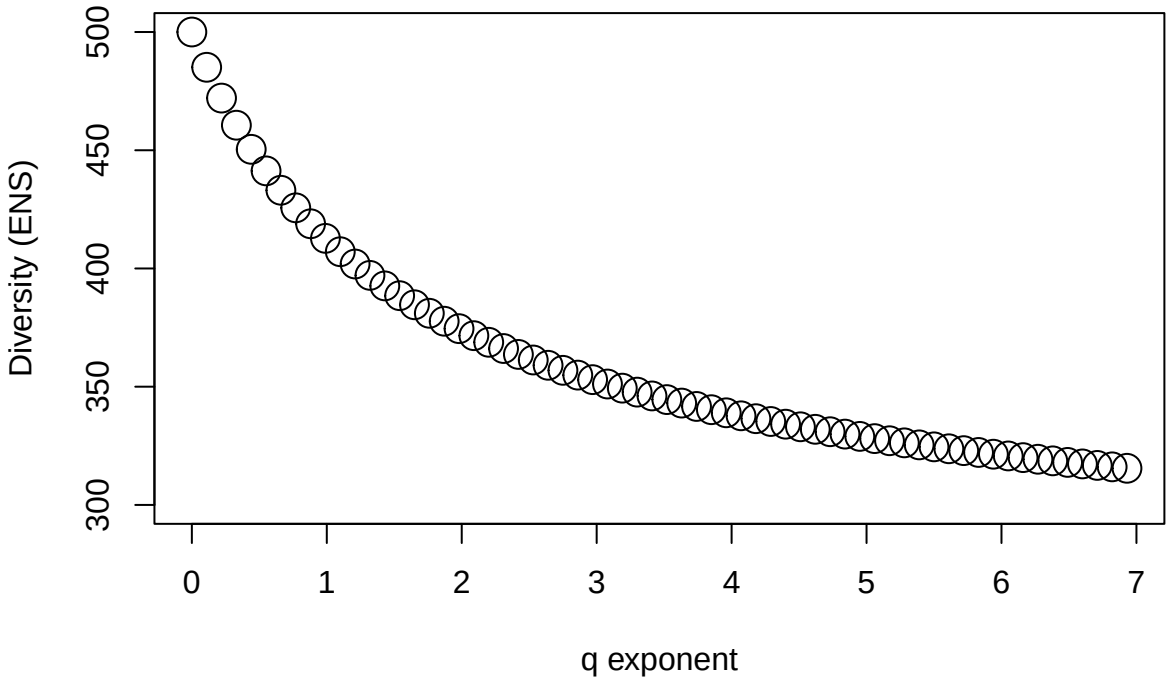
```
## [1] 250
```

```r
#Calculate for each richness level to compare Shannon entropy with Hill number 1 (exponential Shannon's
H_all_Hill <- matrix(ncol = 2, nrow = 1000)
for(i in 1:1000) {
```

```
  C = data.frame(t(rep(1,i)))
  colnames(C) = paste("sp", 1:i)
  H_all_Hill[i, 1] = i
  H_all_Hill[i, 2] = exp(diversity(C, index = "shannon"))}

plot(H_all_Hill[,1], H_all_Hill[,2], xlab = "Species Richness",
     ylab = "Exponential Shannon Diversity (ENS)")
```




```
#Function profile to calculate diversity from thje equation above along a continuum 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)))))}

set.seed(42)
C3 <- data.frame(t(sample(1:1000, 500))); colnames(C3) = paste("sp", 1:500)
C3_profile <- profile(C3)
plot(C3_profile[,1], C3_profile[,2], ylim=c(300,500), cex = 2,
     xlab = "q exponent", ylab = "Diversity (ENS)")
```



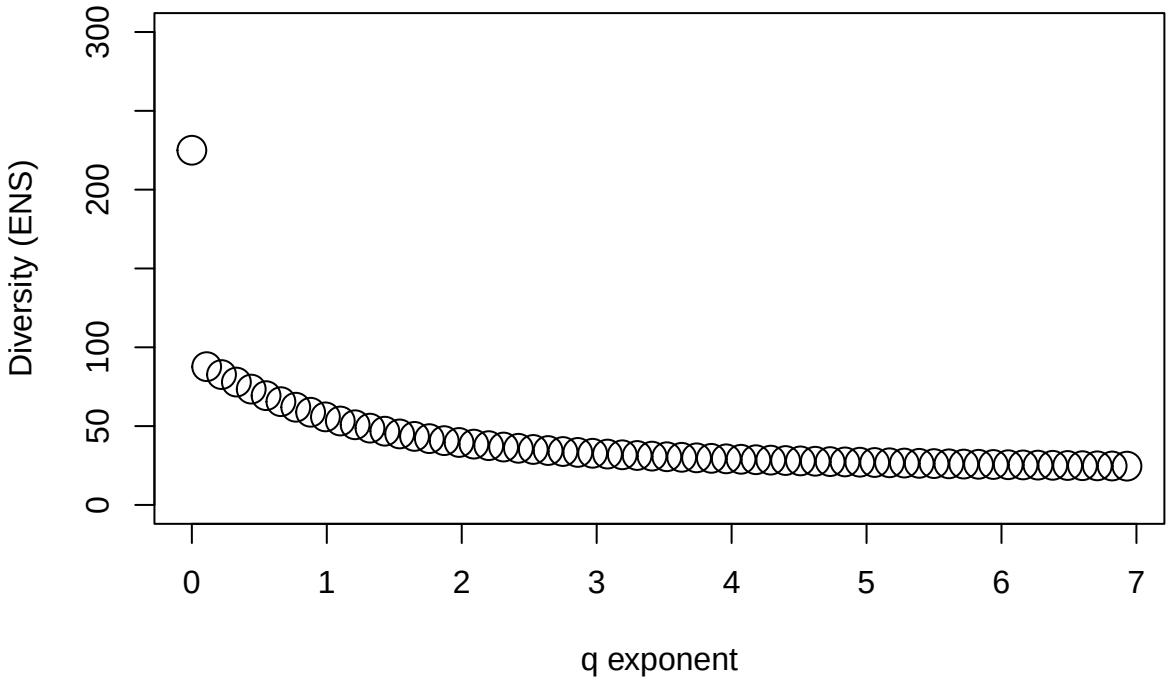

```
#For question 8 on site1

H3 <- diversity(site1, index = "shannon")
H3
```

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

```
H3_Hill <- exp(diversity(site1, index = "shannon"))
H3_Hill
```

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

```
profile_H3 <- function(site1) {
  cbind(seq(0,7, by = 0.11),
        unlist(lapply(seq(0,7, by = 0.11), function(q) sum(apply(site1, 1, function(x)
          (x/sum(x))^q))^(1/(1-q)))))}

site1_profile <- profile_H3(site1)
plot(site1_profile[,1], site1_profile[,2], ylim=c(0,300), cex = 2,
     xlab = "q exponent", ylab = "Diversity (ENS)")
```



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.

> ***Answer 8a***: It seems that when q = 0 (diversity is species richness), our value is 225. When q = 1 (diversity is exponenetial Shannon diversity), our value is roughly 55.86. When q = 2 (where diversity is reciprocal of Simpson diversity), our value is rougly 39.61 ***Answer 8b***: There are many rare species in our community based on our species richness (q = 0) but once we start considering abundance using our Shannon's and Simpson's diversity metrics, their impact on diversity becomes relatively small.

##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).

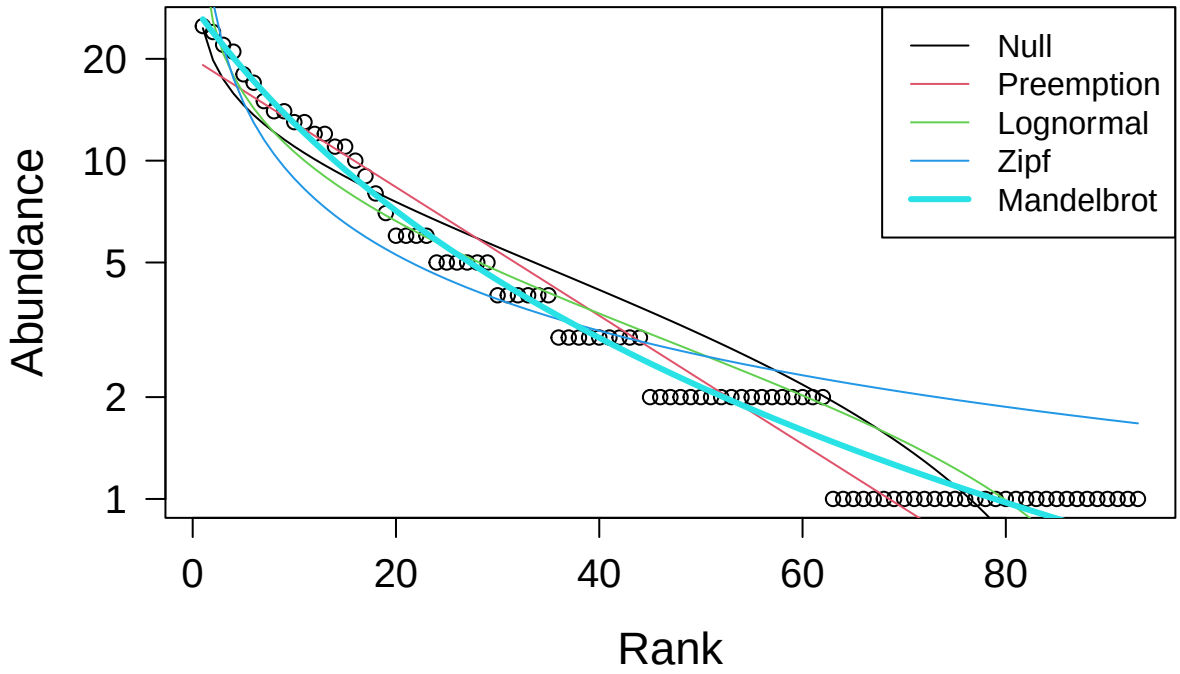
Predicting the form of the RAC is the first test that any biodiversity theory must pass and there are no less than 20 models that have attempted to explain the uneven form of the RAC across ecological systems.

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.

```
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***: It seems that the Mandelbrot model best fits our RAc ***Answer 9b***: Because the RAC best fits the Mandelbrot model, we can infer that there is some sort of niche partitioning or structuring going on as the Mandelbrot model suggests some sort of self-organizing structure amongst 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 line in the RAD plot?

> ***Answer 10a***: The preemption model from above assumes that total abundance is linked to the resources available within the ecosystem, with species acquiring resources in a hierarchical or sequential manner as they are the species that have arrived first thus "preempting" the resources as the rest of species in lowers ranks have less abundance because there are less resources due to these early species. ***Answer 10b***: It looks like a straight line because it assumes that each subsequential species acquires a fixed proportion of the remaining resources that the first arrivers did not acquire, leading to a linear decline in species abundances.

Question 10: 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***: The more parameters a model uses, the better it will fit a given data set so that we can better determine the best fitting model when we perform a radfit() like we did here.

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(site1)
```

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

```
#Simpson's inverse for site1
D.inv
```

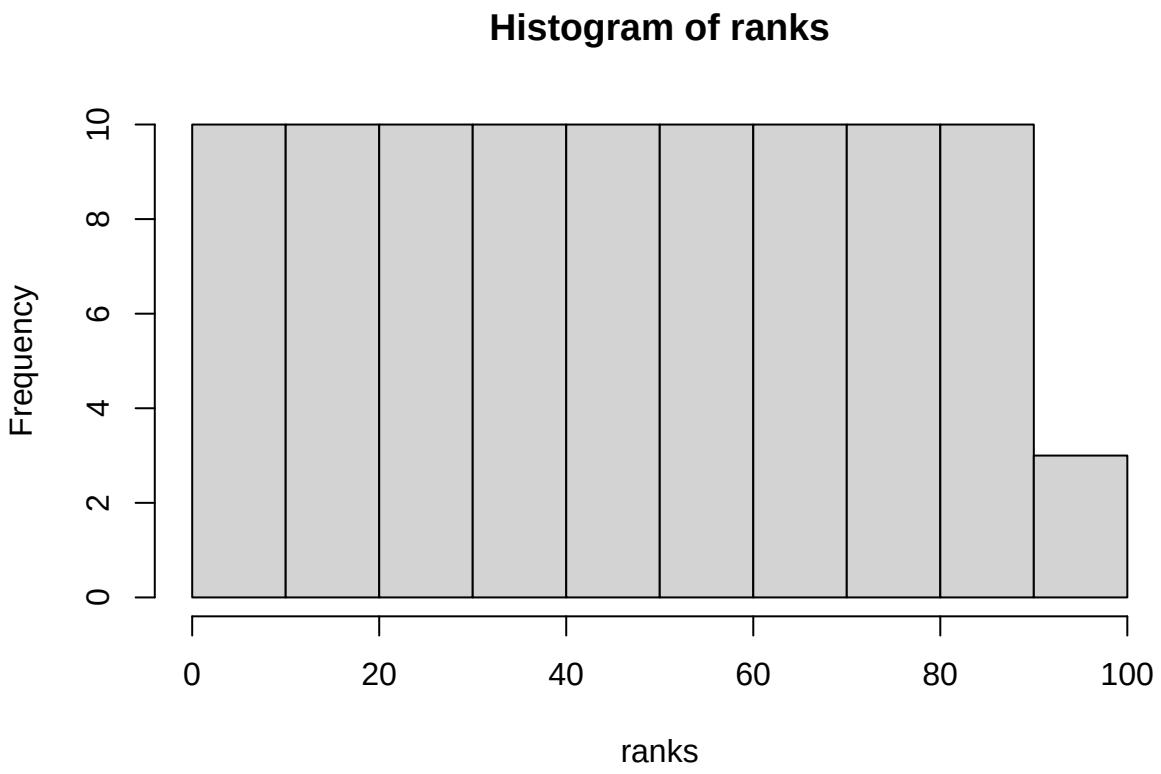
```
## [1] 39.41555
```

```
#Simpson's 1-D
D.sub
```

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

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.

```
plot.new()
hist(ranks)
```



>Answer to Synthesis #2 >When I perform the histogram, the general pattern that I am seeing is one in which the majority of the species have a relatively high and equal frequency of abundance at site 1. This would lead to faulty and misleading interpretation of the diversity and evenness of the species that we see within site 1.

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?

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 29th, 2025 at 12:00 PM (noon)**.