5. Worksheet: Alpha Diversity

Aishwarya Vaidya; Z620: Quantitative Biodiversity, Indiana University

06 February, 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).

```
getwd()
```

```
## [1] "/cloud/project/QB2025_Vaidya/Week2-Alpha"
```

```
setwd("/cloud/project/QB2025_Vaidya/Week2-Alpha")
require(vegan)
```

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

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`.

```
data(BCI)
S.obs <- function(site) {
  return(sum(site > 0))
}
 observed_richness <- S.obs(BCI[1, ])
vegan_richness <- specnumber(BCI[1, ])
vegan_richness
```

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

```
vegan_richness <- specnumber(BCI[1, ])
vegan_richness
```

```
##  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: Yes both first four sites present same values of 1 2 3 4 93 84 90 94

```
S.obs <- function(site) {return(sum(site > 0))}
 observed_richness <- S.obs(BCI[1, ])
 vegan_richness <- specnumber(BCI[1, ])
 observed_richness
```

```
## [1] 93
```

```
 richness_custom <- apply(BCI[1:4, ], 1, S.obs)
 richness_custom
```

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

```r
 richness_vegan <- specnumber(BCI[1:4, ])
 richness_vegan
```

```
##  1  2  3  4
## 93 84 90 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. >

```r
 Goods_Coverage <- function(site) { N <- sum(site)
F1 <- sum(site == 1)
if (N == 0) {
    return(NA) }
    else { return(1 - (F1 / N)) }}

coverage_values <- apply(BCI, 1, Goods_Coverage)
coverage_values
```

```
##         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 ranges between 0 and 1.

Answer 2b: Good's Coverage is defined as: C=1-(F1/N) Since F1 is the number of singletons i.e observed only once , while N is the total number of individuals observed, then F=0 means C=1, suggesting perhaps a complet dataset where none appear only once suggesting high completedness.

Answer 2c: Thus a lower fraction of o.33 observed for the code below suggests the specieswas observed several times.

```r
site1 <- BCI[1, ]
F1 <- sum(site1 == 1)
N <- sum(site1 > 0)
singleton_fraction <- F1 / N
singleton_fraction
```

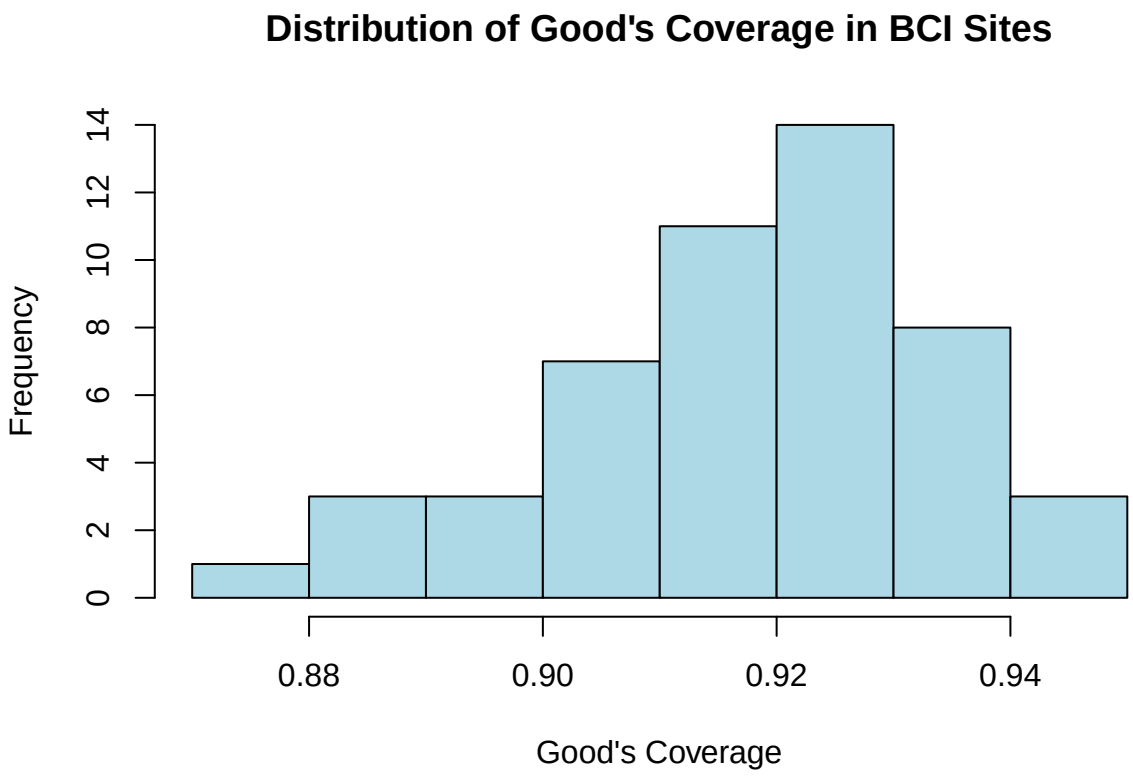
```
## [1] 0.3333333
```

Answer 2d: Since majority of them are more than 0.8 on the Good's coverage

```
Goods_Coverage <- function(site) {
N <- sum(site)
F1 <- sum(site == 1)
if (N == 0) {
    return(NA)
} else {
    return(1 - (F1 / N))
  }
}

coverage_values <- apply(BCI, 1, Goods_Coverage)
summary(coverage_values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8706  0.9095  0.9200  0.9182  0.9287  0.9469
```

```
hist(coverage_values, main = "Distribution of Good's Coverage in BCI Sites",
     xlab = "Good's Coverage", col = "lightblue", border = "black")
```



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

```
soilbac <- read.table("/cloud/project/QB2025_Vaidya/Week2-Alpha/data/soilbac.txt", sep = "\t", header =
soilbac.t <- as.data.frame(t(soilbac))
soilbac1 <- soilbac.t[1, ]
N_soilbac1 <- sum(soilbac.t$T1_1)
N_soilbac1
```

```
## [1] 0
```

```
observed_richness <- sum(soilbac1 > 0)
observed_richness
```

```
## [1] 1074
```

```
singleton_count <- sum(soilbac1 == 1)
total_reads <- sum(soilbac1)
coverage <- 1 - (singleton_count / total_reads)
coverage
```

```
## [1] 0.6479471
```

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: 0

```
soilbac.t <- as.data.frame(t(soilbac))
soilbac1 <- soilbac.t[1, ]
N_soilbac1 <- sum(soilbac.t$T1_1)
N_soilbac1
```

```
## [1] 0
```

Answer 3b: 1074

```
observed_richness <- sum(soilbac1 > 0)
observed_richness
```

```
## [1] 1074
```

Answer 3c: It is 0 for BCI sample (site1) while 0.648 for KBS sample soilbac1.

```
t.BCI <- t(BCI)
 site1 <- t.BCI[1, ]
 singleton_count_site1 <- sum(site1 == 1)
 total_reads_site1 <- sum(site1)
 coverage_site1 <- 1 - (singleton_count_site1 / total_reads_site1)
 coverage_site1
```

```
## [1] 0
```

```
 soilbac.t <- as.data.frame(t(soilbac))
  soilbac1 <- soilbac.t[1, ]
 singleton_count_soilbac1 <- sum(soilbac1 == 1)
```

```
total_reads_soilbac1 <- sum(soilbac1)
coverage_soilbac1 <- 1 - (singleton_count_soilbac1 / total_reads_soilbac1)
coverage_soilbac1
```

```
## [1] 0.6479471
```

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`.

```
data(BCI)
BCI.t <- as.data.frame(t(BCI))
 site1 <- t.BCI[1, ]

 S.chao1 <- function(x) {
  S.obs <- specnumber(x)
  Q1 <- sum(x == 1)
  Q2 <- sum(x == 2)
  if (Q2 == 0) {
    return(S.obs)
  } else {
    return(S.obs + (Q1^2) / (2 * Q2))
  }
}

S.chao2 <- function(site, SbyS) {
  SbyS <- as.data.frame(SbyS)  # Ensure SbyS is a data frame
  x <- SbyS[site, , drop = FALSE]  # Ensure x remains a data frame

  SbyS.pa <- (SbyS > 0) * 1  # Convert presence/absence to 1s and 0s
  Q1 <- sum(colSums(SbyS.pa) == 1, na.rm = TRUE)  # Singletons
  Q2 <- sum(colSums(SbyS.pa) == 2, na.rm = TRUE)  # Doubletons
  S.obs <- specnumber(x)  # Observed species richness

  # Handle cases where Q2 = 0 to prevent division by zero
  if (is.na(Q2) || Q2 == 0) {
    S.chao2 <- S.obs
  } else {
    S.chao2 <- S.obs + (Q1^2) / (2 * Q2)
  }

  return(S.chao2)
}
S.ace <- function(x, thresh = 10) {
  # Ensure x is numeric and positive
  x <- x[x > 0]

  # Calculate species abundance categories
```

```
  S.abund <- sum(x > thresh)    # Number of abundant species
  S.rare <- sum(x <= thresh)    # Number of rare species
  singlt <- sum(x == 1)         # Number of singletons
  N.rare <- sum(x[x <= thresh])  # Total abundance of rare species

  # Handle potential zero-division issues
  if (N.rare == 0) {
    return(S.abund)  # Return abundant species count if no rare species
  }

  C.ace <- 1 - (singlt / N.rare)
  if (C.ace <= 0) {
    return(S.abund + S.rare)  # Avoid division by zero
  }

  # Count function for rare species
  count <- function(i, y) sum(y == i)
  i <- 1:thresh
  a.1 <- sapply(i, count, x)

  f.1 <- sum(i * (i - 1) * a.1)  # Sum for G.ace calculation

  # Avoid division by zero in G.ace
  if (N.rare * (N.rare - 1) == 0) {
    G.ace <- 0
  } else {
    G.ace <- (S.rare / C.ace) + (singlt / C.ace) * max(f.1 / (N.rare * (N.rare - 1)), 0)
  }

  # Calculate final ACE richness estimate
  S.ace <- S.abund + G.ace
  return(S.ace)
}
```

```
 cat("Richness estimates for site1 (BCI):\n")
```

```
## Richness estimates for site1 (BCI):
```

```
 cat("Chao1:", S.chao1(site1), "\n")
```

```
## Chao1: 1
```

```
cat("Chao2:", S.chao2("site1", t.BCI), "\n")
```

```
## Chao2: NA
```

```
cat("ACE:", S.ace(site1), "\n")
```

```
## ACE: 1
```

```
cat("Richness estimates for soilbac1:\n")
```

```
## Richness estimates for soilbac1:
```

```
cat("Chao1:", S.chao1(soilbac1), "\n")
```

```
## Chao1: 2628.514
```

```
cat("Chao2:", S.chao2(1, soilbac.t), "\n")
```

```
## Chao2: 21055.39
```

```
cat("ACE:", S.ace(soilbac1), "\n")
```

```
## ACE: 1901.294
```

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***: Chao1 estimates species richness by counting rare species in abundance data, while Chao2 looks at presence-absence across samples. ACE, on the other hand, separates common and rare species using a set cutoff (usually 10) and considers how often species appear. These methods can give different results where Chao estimators are more cautious and work best when data is incomplete, while ACE is more reliable when species vary in how often they show up. I am guessing the right method might also depend on the type of data, Chao1 is better when many species are missing, while ACE works well when species are found at different frequencies.

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.

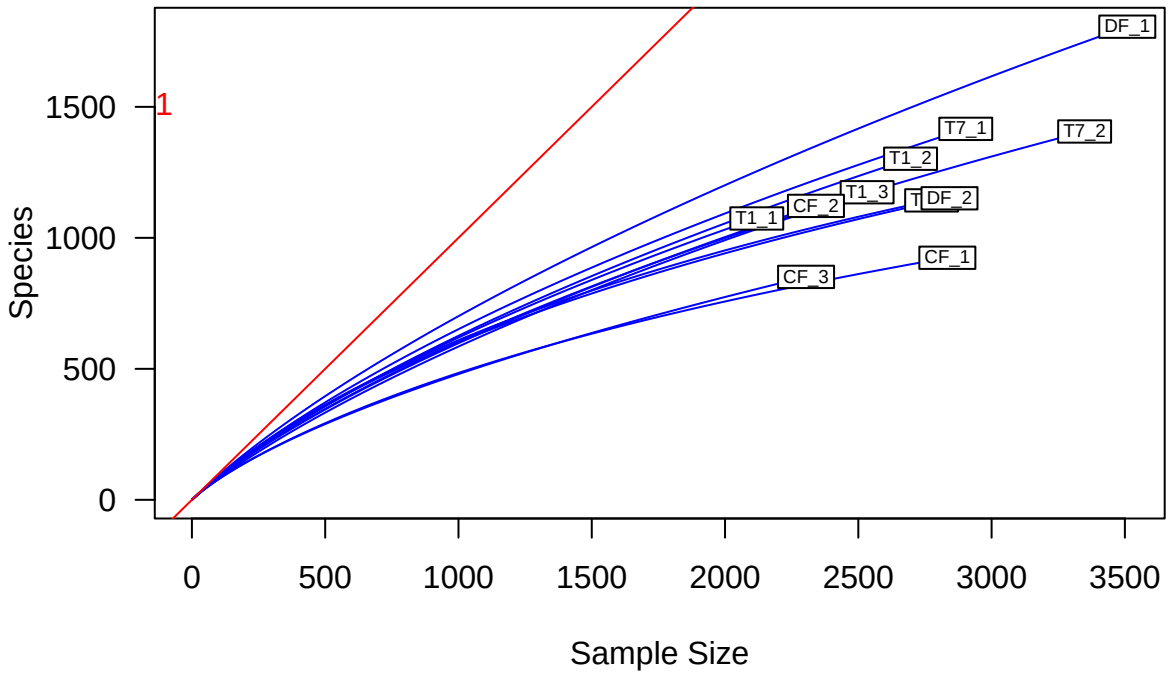
```
S.obs <- function(x = ""){
rowSums(x > 0) + 1
}
S.obs(soilbac.t)
```

```
## T1_1 T1_2 T1_3 T7_1 T7_2 T7_3 DF_1 DF_2 CF_1 CF_2 CF_3
## 1075 1303 1175 1417 1407 1144 1807 1152  925 1123  852
```

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

```
## [1] 0
```

```
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, "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

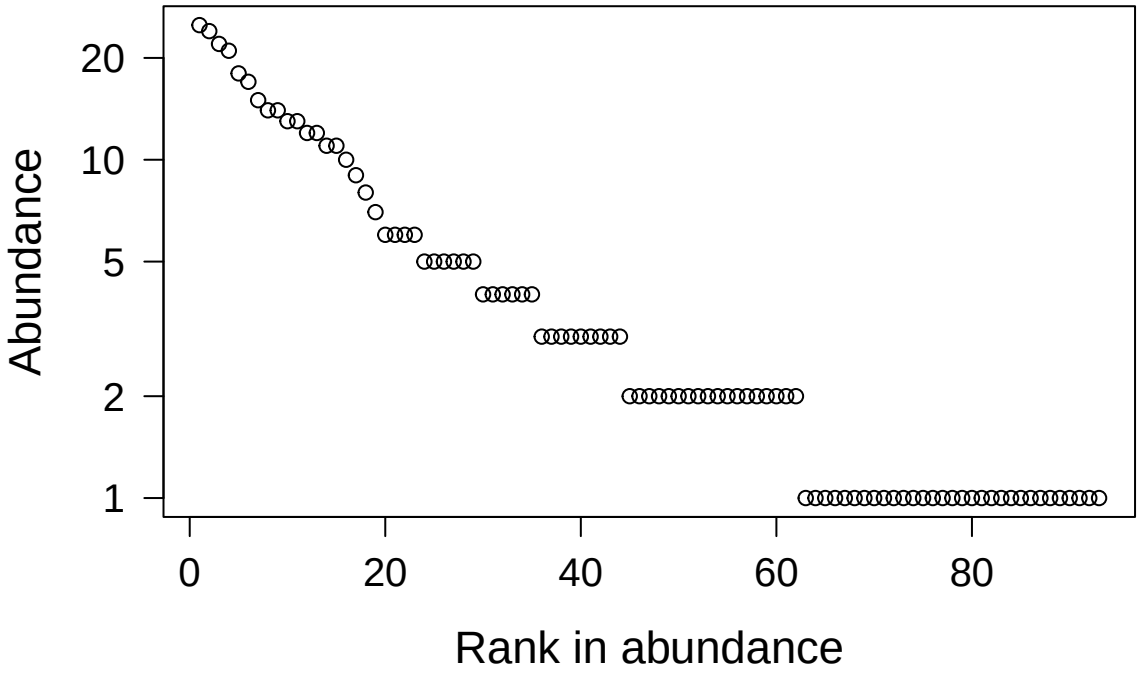
```
 RAC <- function(x = ""){
x.ab = x[x>0]
x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
as.data.frame(lapply(x.ab.ranked, unlist))
return(x.ab.ranked)
}

construct_RAC <- function(abundance_vector) {
+ abundance_vector <- abundance_vector[abundance_vector > 0]
+ ranked_vector <- sort(abundance_vector, decreasing = TRUE)
+ return(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)"

```
 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 = "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)))
```



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***: Using a log scale on a Rank Abundance Curve (RAC) makes it easier to see how evenly species are distributed in a community. Without a log scale, the most common species can dominate the graph, hiding the less common ones. With the log scale, the differences between rare species become clearer. A steep slope on a log scale means some species are way more abundant than others (low evenness), while a flatter slope means the species are more evenly distributed (high evenness). Basically, it helps us see the whole picture, not just the key players within.

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
## 0.4193144
```

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 values SimpE = 0.42 and Evar = 0.51 both measure how evenly species are spread out in the community. SimpE suggests that the species are somewhat uneven, with a few being more dominant. On the other hand, Evar is a bit higher, meaning the species are more evenly spread out when you consider both the number of species and their abundance. So, overall, Evar shows a slightly more balanced distribution of species in site1 compared to SimpE. Since $E_{1/D}$ is less than E_{var}, it might mean few species are taking over, even if there are lots of species overall.

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
```

```
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
```

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
```

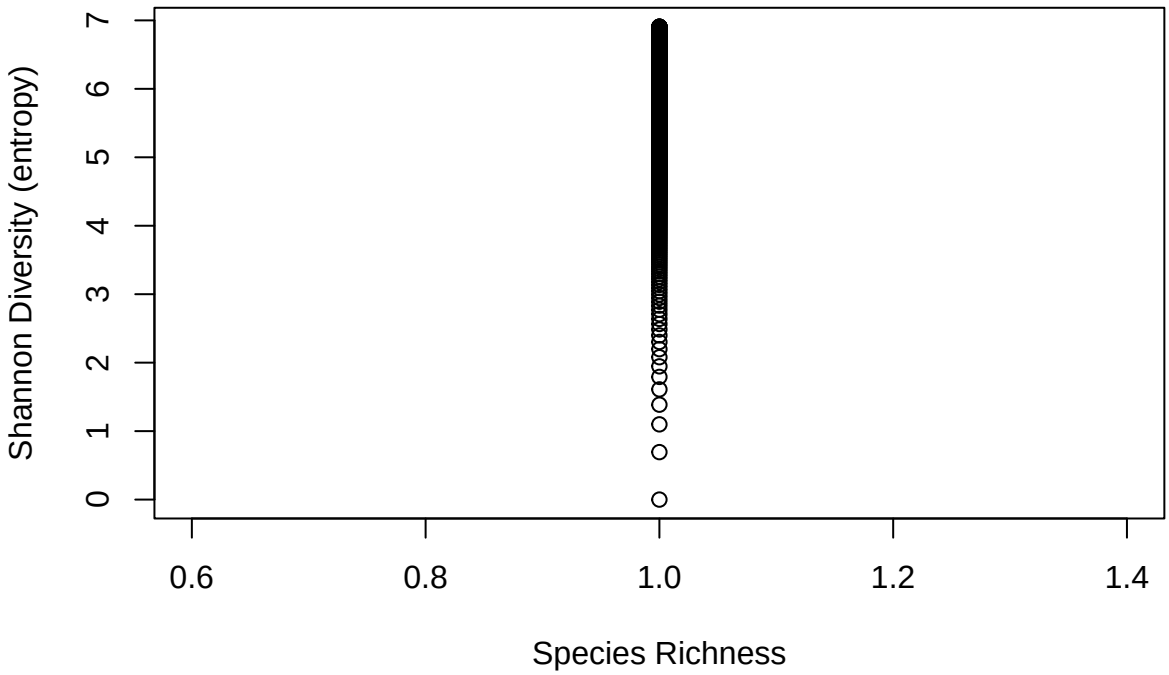
```
## [1] 35.67297
```

```
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)
H1 <- diversity (C1, index = 'shannon')
H2 <- diversity (C2, index ='shannon')
H1; H2
```

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

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

```
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] <- 1
H_all[i, 2] <- diversity(C, index = 'shannon')
}
plot(H_all[ ,1], H_all[ , 2],xlab = "Species Richness", ylab = "Shannon Diversity (entropy)")
```



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***:Fishers $\boldsymbol{\alpha}$ is lower than $E_{H'}$ and E_{var}, suggests there arent that many species to begin with, as Fisher's $\boldsymbol{\alpha}$ generally looks at the number of different species, while $E_{H'}$ is thrown off by the number of dominant species, while E_{var} denotes how evenly the data is distributed. Fisher's alpha talks about how many species are part of the community, thus being concerned about the richness of the data, while the other two are mainly concerned about the evenness or normality of the data, without being much concerned about the total number of species.

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. Fishers

> ***Answer 8a***: q=0(richness) is 93; q=1(exponential Shannon's entropy) is 55.6 while q=2 (inverse Simpson's diversity) is 0.025. Since q0 is more than q1, many rare species exist contributing to the species richness but not as much to the species evenness. Now, since q1 is also more than q2, even within the abundant species, there are far less that are significantly dominant.Since q1 and q2 are very far apart, they are not dominated by rare species.

> ***Answer 8b***: As the hill numbers reduce to increasing exponent q Fishers means rare species make up large portion of the community but they dont contribute the species abundance.

##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

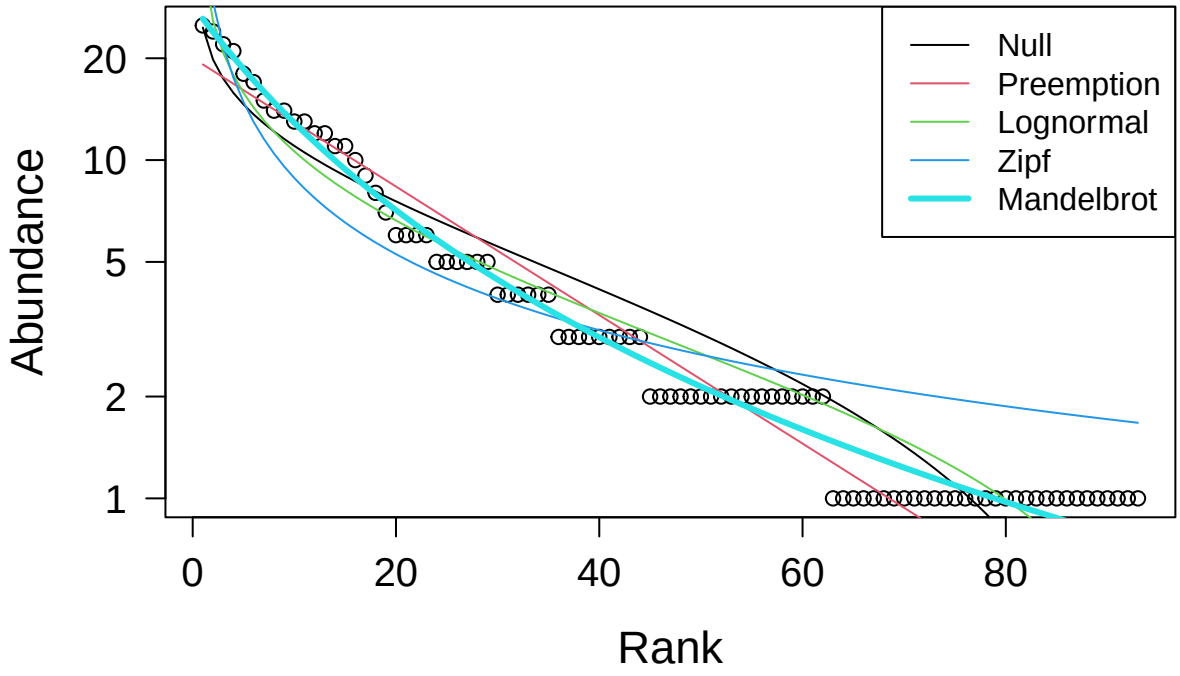
The RAC is a simple data structure that is both a vector of abundances. It is also a row in the site-by-species matrix (minus the zeros, i.e., absences).

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)
```




```
AIC(RACresults)
```

```
##       Null Preemption  Lognormal       Zipf Mandelbrot
##   315.4362   299.8041   305.0629   340.9567   286.1372
```

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 RAC results, it suggests which of the species given describe RAC abundance well. Since of these Mandelbrot has the lowest AIC values, it is the best model, suggesting a strong dominance effect where very few species are dominant while others are rare. Since MAndelbrot is the best fit model it suggests dominant species outcompete the weaker ones. ***Answer 9b***: The Zipf-Mandelbrot model helps explain why some species dominate certain ecosystems while others are rare, mainly due to dispersal limits, patchy habitats, and historical factors. Basically, not all species can spread everywhere i.e some get stuck in specific areas because

of barriers like mountains, rivers, or human activity. This leads to clusters of species thriving in some spots while unable to spread to many others. A pattern of this kind is common in places like rainforests, coral reefs, islands, and urban landscapes, where species are shaped by who got there first and how easily they can move around. Since this RAC best fits Mandelbrot model, it could mean that the species distribution is more about movement challenges and history than just competition alone.

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: It assumes that in nature, when a species move into an ecosystem, it occupies a set portion of the available resources, with each new species arriving getting a progressively smaller share. The total number of individuals (i.e N) in an ecosystem are thus basically limited by resources there are, to ultimately occupy. ***Answer 10b***: Because the way species take resources follows a predictable shrinking pattern, with each new species getting less than the one before it. This when graphed on a log scale, the exponential drop turns into a straight-line decline. The steeper the line, the more the first species occupied the resources to themselves.

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: It is important to account for the number of parameters in a model because a model with more parameters can always fit the data better, but that doesn't mean it's actually a better explanation of what's happening. A model should balance accuracy with simplicity and so we considered AIC to measure the best fit model that ensures we don't just pick the one that overfits the data but instead choose the one that actually captures the underlying pattern in a meaningful way.

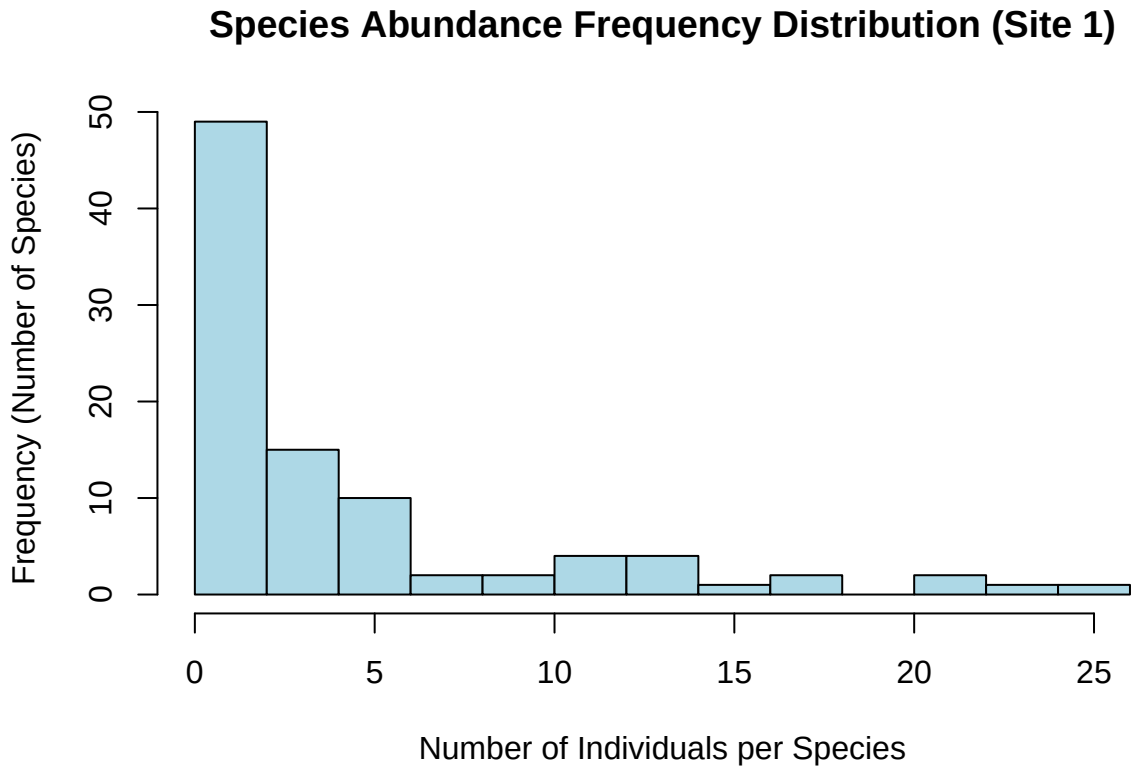
SYNTHESIS

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. >Simpson's D: 0.02319032;1 - D (Evenness): 0.9768097 ;Inverse Simpson's Index (1/D): 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. > Data points for site1 of BCI are right skewed with a higher left peak and a smaller rigth tail. Left peak show a lot of species with low abundance, mostly represented by just one or two individuals, and a few species with a much larger number of individuals within the long tail stretching to the right, where the few dominant species are. >This kind of pattern suggests that competition or environmental factors might allow a few species to dominate the ecosystem, while many others only exist in small numbers. Low abundance species (with high peak on left) are represented by only one or two individuals and consists of rare species, while high abundant species may have more individuals, creating a long right tail on the histogram. These species dominate the community. This suggests an ecosystem where competition might lead to dominance by a few species, or where environmental factors, such as habitat preference or resource availability, enable certain species to thrive while others struggle. This type of distribution could be seen in environments that are subject to intense competition, predation, or environmental

stress, where only a few species are able to establish large populations. Overall, this kind of histogram shows an ecosystem where certain species are scarce, and others are somewhat more plentiful.

```
site1 <- BCI[1, ]
site1_abundances <- site1[site1 > 0]


hist(site1_abundances,
     main = "Species Abundance Frequency Distribution (Site 1)",
     xlab = "Number of Individuals per Species",
     ylab = "Frequency (Number of Species)",
     col = "lightblue",
     border = "black",
     breaks = 10
)
```



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. >load("/cloud/project/QB2025_Vaidya/longdataBac_objects2_datadryad.rda") How many sites are there? >179 study sites How many species are there in the entire site-by-species matrix? >num_species <- specnumber(bac_by_site) Any other interesting observations based on what you learned this week? > It was interesting to see that ShanH could be used when dealing with quite diverse environments while SimpD to be used to give an estimate where there is dominance of few species.

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)**.