5. Worksheet: Alpha Diversity

Joshua Jones; Z620: Quantitative Biodiversity, Indiana University

03 April, 2021

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/joshu/quantbio/QB2021_Jones/2.Worksheets/5.AlphaDiversity"
```

```
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 4.0.4
```

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

```
## Warning: package 'permute' was built under R version 4.0.4
```

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

```
## This is vegan 2.5-7
```

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   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
```

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, they return the same values Site 1 = 93 taxa Site 2 = 84 taxa Site 3 = 90 taxa Site 4 = 94 taxa

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***:
> From 0 to 1 ***Answer 2b***:
> Coverage is 0 because all observed taxa are singletons ***Answer 2c***:
> .07 ***Answer 2d***:
> Each plot appears to have pretty good coverage, with the lowest estimated coverage only being ~88.9%

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[1,]

S.obs(soilbac1)
```

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

```
C(soilbac1)
```

```
##      T1_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:
1074 ***Answer 3b***: 1074 ***Answer 3c***:
The KBS sample appears to be much lower than the BCI sample

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)
}

S.ace <- function(x = "", thresh = 10){
  x <- x[x>0]
  S.abund <- length(which(x > thresh))
  S.rare <- length(which(x <= thresh))
  singlet <- length(which(x == 1))
  N.rare <- sum(x[which(x <= thresh)])
  C.ace <- 1 - (singlet / 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) + (singlet/C.ace) * max(G.ace,0)
  return(S.ace)
}


site1 <- BCI[1,]

print("Site 1 results")
```

```
## [1] "Site 1 results"
```

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

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

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

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

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

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

```
print("SoilBac1 Results")
```

```
## [1] "SoilBac1 Results"
```

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

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

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

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

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

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

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?

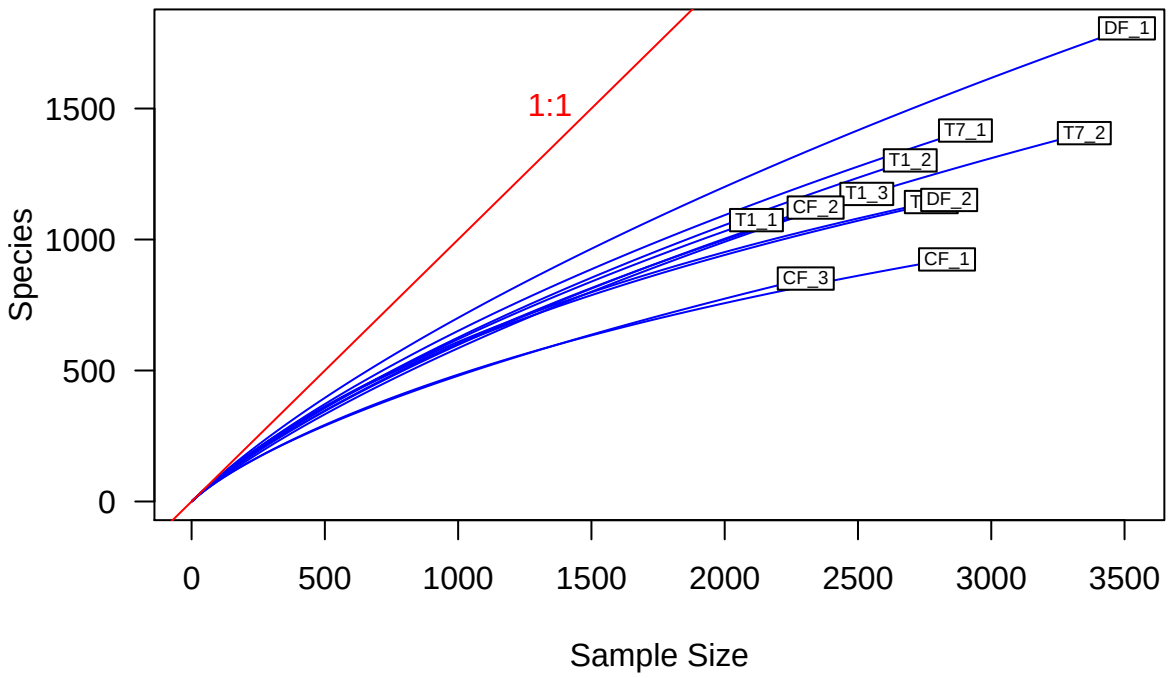
> ***Answer 4***: When estimating richness ACE factors in rare taxa at a higher cutoff. No, the estimates are variable across the different metrics. Which metric would be used would be dependent on the data, if my data had a lot of species at low abundance then I would use ACE, but otherwise I would use the Chao2 estimator.

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 = .6, las = 1)
abline(0, 1, col = 'red')
text(1500, 1500, "1:1", pos = 2, col = 'red')
```



##4) SPECIES EVENNESS Here, we consider how abundance varies among species, that is, **species evenness**.

Visualizing evenness: the rank abundance curve (RAC)

One of the most common ways to visualize evenness is in a **rank-abundance curve** (sometime referred to as a rank-abundance distribution or Whittaker plot). An RAC can be constructed by ranking species from the most abundant to the least abundant without respect to species labels (and hence no worries about 'ties' in abundance).

In the R code chunk below, do the following:

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

```
RAC <- function(x = ""){
  x = as.vector(x)
  x.ab = x[x >= 1]
  x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
  return(x.ab.ranked)
}
```

Now, let's examine the RAC for `site1` of the BCI data set.

In the R code chunk below, do the following:

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

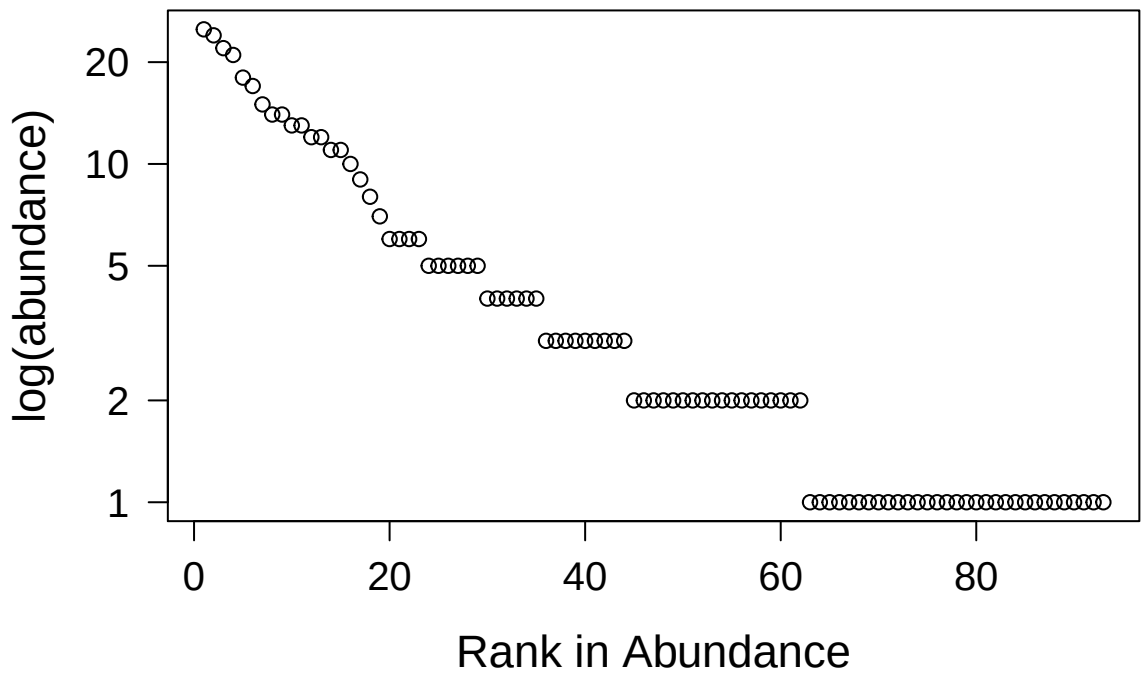
```
plot.new()
site1 <- BCI[1,]

rac <- RAC(x = site1)
ranks <- as.vector(seq(1, length(rac)))
opar <- par(no.readonly = TRUE)
par(mar = c(5.1,5.1,4.1,2.1))
plot(ranks, log(rac), type ='p', axes = F,
     xlab = "Rank in Abundance", ylab = "log(abundance)",
     las = 1, cex.lab = 1.4, cex.axis = 1.25)

box()
axis(side = 1, lavels = T, cex.axis = 1.25)
```

```
## Warning in axis(side = 1, lavels = T, cex.axis = 1.25): "lavels" is not a
## graphical parameter
```

```
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***: It allows us to see what proportion of the commuity are only present at relatively low or high abundances and therefore gives us more perspective on how even the community is as a whole.

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

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

In the R code chunk below, do the following:

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

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

SimpE(site1)
```

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

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

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

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

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

Evar(site1)
```

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

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

Answer 6: They are similar but not in complete agreeance, which makes sense as they are created using different equations, particularly that E(var) corrects as to reduce the effect of high abundance taxa while E(1/D) doesnt. Both of them show that the data is roughly half even.

##5) INTEGRATING RICHNESS AND EVENNESS: DIVERSITY METRICS

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

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

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

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

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


ShanH(site1)
```

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

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

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

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)

print(D.inv)
```

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

```
print(D.sub)
```

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

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

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

```
diversity(site1, "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.

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

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

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 $\boldsymbol{\alpha}$ actually estimates the diversity of the sites instead of just measuring the diversity revealed in the survey.

##6) MOVING BEYOND UNIVARIATE METRICS OF α DIVERSITY

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

Species abundance models

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

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

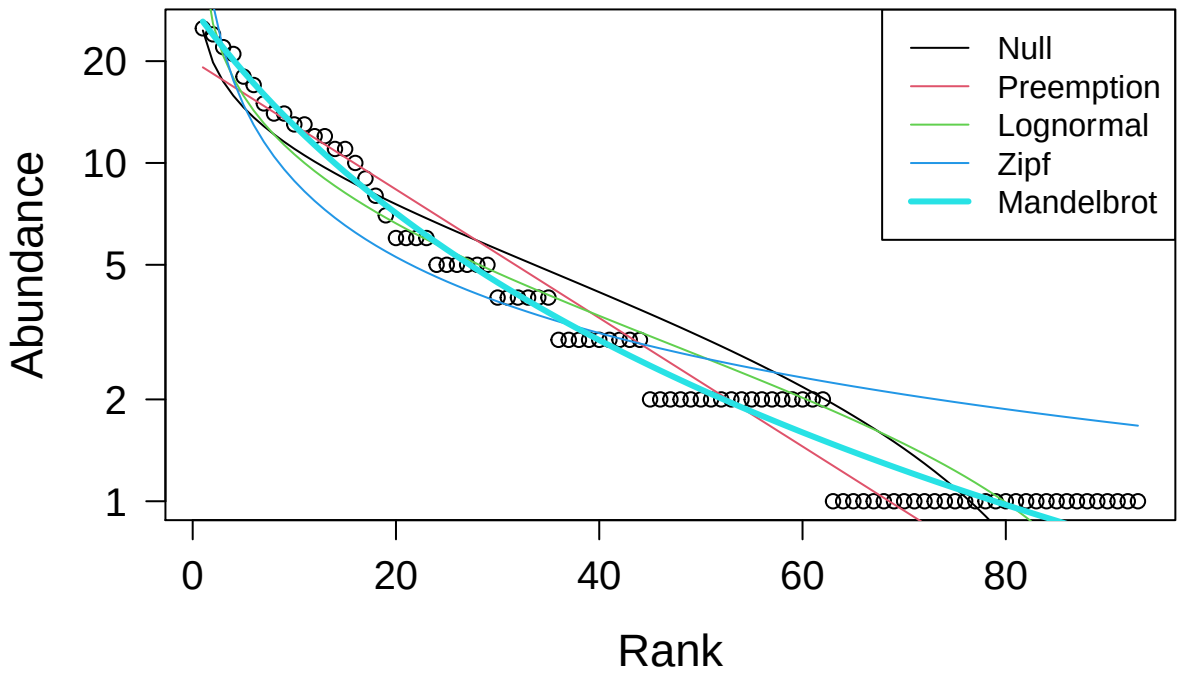
In the R code chunk below, please do the following:

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

```
RACresults <- radfit(site1)
RACresults
```

```
##
## RAD models, family poisson
## No. of species 93, total abundance 448
##
##            par1      par2     par3    Deviance AIC      BIC
## Null                                   39.5261 315.4362 315.4362
## Preemption  0.042797                    21.8939 299.8041 302.3367
## Lognormal   1.0687    1.0186           25.1528 305.0629 310.1281
## Zipf        0.11033  -0.74705          61.0465 340.9567 346.0219
## Mandelbrot  100.52   -2.312    24.084   4.2271 286.1372 293.7350
```

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



Question 8: 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 8a***:
> The Mandelbrot model seems to best fit the rank-abundance curve using both the AIC and the BIC metrics ***Answer 8b***: While I can't think of any direct inferences that can be made from this model fitting the data best, I can infer that whatever influences are determining the ecological community has resulted in a community similar to the preemption niche community (each next abundant taxa having a progressively lower abundance) but with a longer low abundance tail that would be expected from that sort of a community. Suggesting perhapse a unequal distribution of resources based on taxa abundance where the more abundant taxa don't consume their a proportional percent of the resources, allowing for rarer species to share and exist of the remainder of the resources.

Question 9: 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 organism with the highest abundance is also the organism that consumes a greater amount of the shared resouces proportionate to their comparative abundance. ***Answer 10b***: Because from its assumptions it projects that abundance across taxa will be a steady decline until it reaches the final taxa in the sample with an N ~ 1.

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***: This is important because over-parameterization of the model realtive to the amount of data in the sample may result in a model with a high fit but may also result in loss of reletavization of the model in explaining any further data or the system as a whole.

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.

```
site1simpD <- SimpD(site1)
print(site1simpD)
```

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

```
site1simpDminus <- 1 - SimpD(site1)
print(site1simpDminus)
```

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

```
site1invsimpD <- 1/SimpD(site1)
print(site1invsimpD)
```

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

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.

```
# Transposing site 1 to long form
site1.t <- t(site1)

# Filtering the data to remove all 0 abundances
site1.tz <- site1.t[apply(site1.t, 1, function(row) all(row != 0)),]

# Forming a histogram with remaining values
h <- hist(site1.tz, breaks = 25)
text(h$mids, h$counts, labels = h$counts, adj = c(.5, -.5))
```

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?

```
#Importing data
alpine <- read.csv("data/genus_table.csv")

#Quantifying sites
length(alpine)
```

```
## [1] 57
```

```
#Counting unique genera
length(unique(alpine$Genus))
```

```
## [1] 113
```

"length" returns 57, meaning that there are 57 total rows, but since there is a row for Genus names and another for a control sample, there are a total of 55 sites (counting sites with .2 after them but I don't yet know what these markers mean). "length" combined with "unique" lets me know that there are a total of 113 unique genera identified in the data (genera is the highest sensitivity reported) and by looking at the data we can see that are many "unknown" classified genera. Interesting, by building a rarefaction curve of my samples (not shown here) I can already see that there are 4 or 5 samples with an exteremely low number of total reads, too low to be rarefied to so they will most likely be discluded from the deeper analysis.

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, April 7th, 2021 at 12:00 PM (noon)**.

```
sessionInfo()
```

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] vegan_2.5-7     lattice_0.20-41 permute_0.9-5
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.27     MASS_7.3-53       grid_4.0.3        nlme_3.1-149
##  [5] magrittr_2.0.1    evaluate_0.14     highr_0.8         rlang_0.4.10
##  [9] stringi_1.5.3     Matrix_1.2-18     rmarkdown_2.6     splines_4.0.3
## [13] tools_4.0.3       stringr_1.4.0     xfun_0.20         yaml_2.2.1
## [17] parallel_4.0.3    compiler_4.0.3    cluster_2.1.0     mgcv_1.8-33
## [21] htmltools_0.5.1.1 knitr_1.31
```