Handout: Local (alpha) Diversity

Z620: Quantitative Biodiversity, Indiana University

January 20, 2017

OVERVIEW

In this exercise, we will explore aspects of local or site-specific diversity, also known as alpha (α) diversity. We will quantify the two primary components of α-diversity (**richness** and **evenness**) using the site-by-species matrix. From there, we will discuss ways to integrate richness and evenness, which will include univariate metrics of diversity and an investigation of the **species-abundance distribution (SAD)**.

1) SETUP

Retrieve and Set Your Working Directory

```
rm(list=ls())
getwd()
setwd("~/GitHub/QuantitativeBiodiversity/QB-2017/Week2-Alpha")
```

Install Packages

In this exercise, we will use a popular R package called `vegan`, which contains tools commonly used in ecological research, including analyses of α-diversity. Jari Oksanen has created an excellent tutorial that provides an overview of the `vegan` package: http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

Let's install the `vegan` package and its dependencies. If you are running `vegan` for the first time, you will need to install it with the `install.packages` function. Otherwise, we recommend you just use the `require` function to load `vegan` and its dependencies.

```
#install.packages("vegan")
require("vegan")
```

2) LOADING DATA

We will start by using the tropical forest site-by-species matrix from Barro-Colorado Island (BCI). BCI is a 1,560-hectare island in the center of the Panama Canal that is administered by the Smithsonian Tropical Research Institution (http://www.stri.si.edu/english/research/facilities/terrestrial/barro_colorado/). Conveniently, the `vegan` package contains a version of the BCI dataset. The BCI dataset is a census of all trees of at least 10 cm in diameter at breast height (DBH) that occur in 50 one-hectare plots. More information on the BCI dataset can be found elsewhere (http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/). You can also learn more about the BCI dataset associated with the `vegan` package by typing `help(BCI)` in the command line. Let's load the BCI data using the `data` function:

```
data(BCI)
```

3) SPECIES RICHNESS

Species richness (S) is the number of species in a system or the number of species observed in a sample. Species richness is also the foremost aspect of diversity. In fact, it is usually what most people are referring to when they talk about α-diversity. Calculating species richness for a sample is often straightforward, i.e., count the number of species present in a sample. However, estimating species richness for a community from an incomplete sample requires assumptions about the nature of the sampling effort, which could include biases or issues related to coverage. In this part of the handout, you will see that there are several ways to estimate richness, which attempt to account for the number of species that were not detected.

Observed Richness

The simplest way to calculate species richness is to just add up the number of species that were detected in a sample. It is important to note that, in most site-by-species matrices, not every species will be present at each site. Consequently, we cannot simply count the columns of the site-by-species matrix to obtain the value of observed richness.

Let's write a function that calculates observed species richness of a site. Functions are central to the use of most any computing software and, in short, are simply pieces of code that operate on one or more variables or objects (e.g., scalars, lists, matrices, etc.). In fact, you've already used several native R functions, e.g., dim(BCI).

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

We can then call the function by typing `S.obs()` and placing the name of our vector in the parentheses. There is also a function in the `vegan` package called `specnumber()` that calculates observed richness.

Coverage: How Well Did You Sample Your Site?

Accurate estimates of richness are influenced by sampling effort and sampling bias. Even when the sampling effort is unbiased, the more individuals that are censused, the more likely you are to encounter new species. One index that provides an estimate of how well a site is sampled is **Good's Coverage (C)**, which is defined as $C = 1 - \frac{n_1}{N}$, where n_1 is the number of *singleton species* (species only detected once), and N is the total number of individuals in the sample. Examining the equation for Good's Coverage reveals that the fraction is simply the portion of N represented by singleton species. Subtracting this from 1 give the portion of N belonging to species sampled more than once.

Let's write a function for Good's Coverage:

```
C <- function(x = ""){
  1 - (sum(x == 1) / rowSums(x))
  }
```

Estimated Richness

There are few systems on Earth that have been better surveyed than the trees at BCI. For most ecological systems, sample size is much smaller than N and many taxa can easily go undetected. Consequently, while observed richness can be easily calculated, true richness must be estimated. To address the case of estimating richness, we are going to introduce a dataset derived from 16S rRNA gene sequences for bacteria, which were collected from the Kellogg Biological Station Long-Term Ecological Site (http://lter.kbs.msu.edu/).

Because it is sometimes hard to apply the species concept to bacteria, sequences are often binned into **operational taxonomic units (OTU)** based on the percent of their gene sequence similarity. Therefore, we will technically be dealing with a site-by-OTU matrix for some of the following exercises.

In the following R chunk, we load the bacterial dataset, transpose it, and identify a sample that happens to be the first column of data ("TL_1") in the site-by-species matrix, which we will use for subsequent analysis.

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

There are number of statistical techniques developed to estimate richness based on the coverage of species in a sample (e.g., Hughes et al. 2001 – http://aem.asm.org/content/67/10/4399). We can think of these estimators as ways to **extrapolate** richness based on information contained in our sample. Here, we will highlight two commonly used richness estimators developed by Anne Chao. Both estimators fall into the category of being non-parametric, which means that they are not based on any assumptions that the data come from a particular underlying distribution (e.g. Normal distribution).

Chao1 is an ***abundance-based estimator*** and is useful for examining richness of a single site. Chao1 is calculated using observed richness (`S.obs`), the observed number of **singletons** (species with an observed abundance of 1), and the observed number of **doubletons** (species with observed abundance of 2). Because it requires singleton and doubleton data to calculate, Chao1 cannot be used on site-by-species matrix where abundances have been relativized, i.e., where the abundance of each species is divided by the row sum.

Let's write a function for Chao1:

```
S.chao1 <- function(x = ""){
  S.obs(x) + (sum(x == 1)^2) / (2 * sum(x == 2))
  }
```

It's important to note that it's not always necessary to write new functions from scratch. Often a diversity metric builds off of a simpler diversity metric, allowing for us to save some time by re-using our existing code. We can see that in the S.chao1 function, where we call on the S.obs() function to estimate the number of species in the site. From there we write some new code by using sum(x == y) to calculate the number of species in site x with y abundance, raising the number of singletons to the second power, and dividing the squared singleton count by twice the number of doubletons in the sample.

Chao2 is an ***incidence-based estimator*** that uses presence-absence data for examining richness across multiple sites. Chao2 is calculated using observed richness (`S.obs`). However, in Chao2, **singletons** and **doubletons** refer to species observed once and twice, respectively, *among sites or samples*.

Let's write a function for Chao2. In this function, the first argument is site, supplied either as row number or row title (e.g., 1 or "T1_1") and the second argument is the name of the site-by-species matrix (e.g., soilbac):

```
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)
  }
```

Notice that this function is a bit more complicated than previous functions we have written, so we will take time to deconstruct it. We can see that the function S.chao2 takes two arguments, i.e. the site of interest and

the site-by-species matrix. The function first converts the site-by-species matrix into a data frame. From there we select the site of interest and count the number of singletons and doubletons among all sites. The code in the function following the line defining doubletons should look very familiar. In fact, it is nearly identical to the S.chao1 operator, the only difference being that we are dealing with singletons and doubletons across all sites rather than within a single site.

In Chapter 5, pages 56 and 57, of "Biological Diversity: frontier in measurement and assessment" you will see that Chao1 and Chao2 are mathematically very similar, and mainly differ in whether we are considering the abundance of species at one site (Chao1) or the presence of species across multiple sites (Chao2).

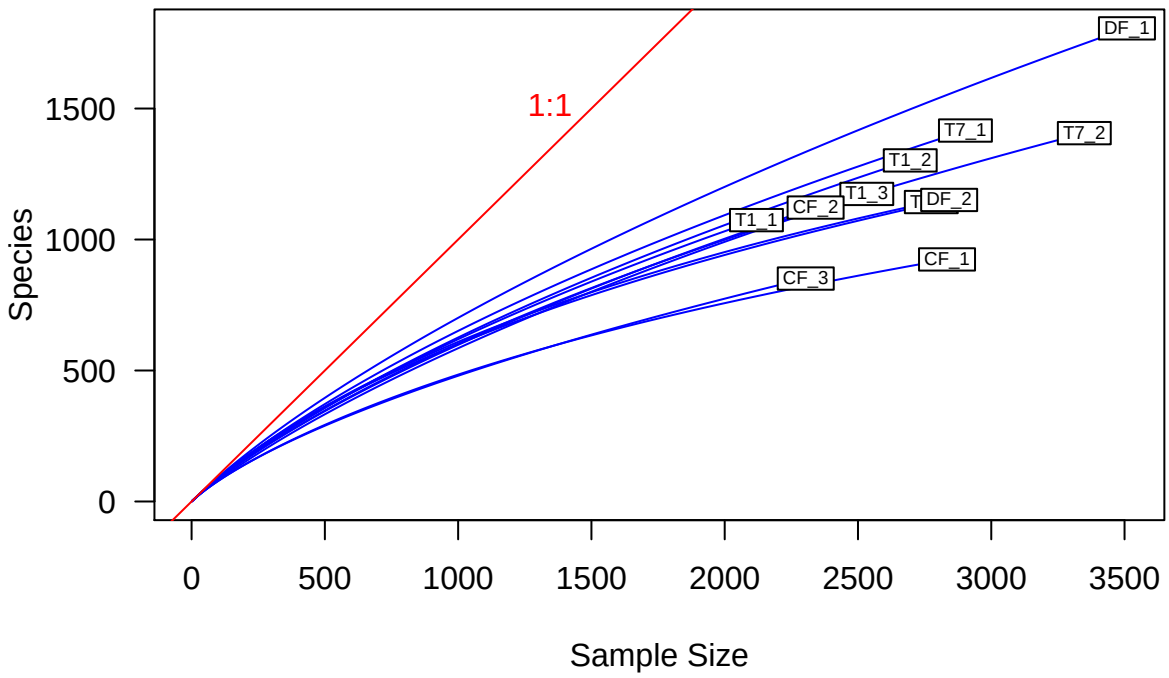
Rarefaction

Often researchers want to compare the richness of two or more ecological communities. Ideally, we would sample our sites with equal effort, all of our sites would have equal densities of individuals, and all individuals would have equal probabilities of being detected. However, these conditions are rarely met in ecological studies. Because sites with greater abundance generally have greater species richness, differences in richness may largely be due to differences in sample size, i.e., N.

A common way to reduce the bias associated with different N is to **rarify** samples down to a "lowest common denominator", an approach known as rarefaction. For example, if there are 2 sites, one with $N = 100$ and one with $N = 50$, we could randomly sample (without replacement) 50 individuals from the site with greater N. Generating many random samples (each of 50 individuals and without replacement) from the larger site will allow us to calculate the mean and standard error of expected S. From here we will be able to tell whether the S of the smaller sample falls within the confidence-intervals for expected S of the larger sample and hence, whether the difference in S between the two sites is simply due to a difference in N. While this demonstrates how rarefaction works in concept, the expected richness in a sample rarefied from N to n individuals is solved mathematically (see Hurlbert, 1971; Vegan Tutorial).

In the following section, we will calculate observed richness for soil bacteria collected from 11 of the KBS sites, where T1 = agriculture, T7 = grassland, DF = deciduous forest, and CF = coniferous forest. Then, we will identify the sample with the fewest sequences (i.e., N) and use `vegan` to rarefy to that sampling effort. Last, we will use commands from `vegan` to construct rarefaction curves for the different samples.

```
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 EVENNESS

There is more to α-diversity than just the number of species in a sample. Specifically, it is important to consider how abundance varies among species (or other taxonomic groups), that is, **evenness**. Many important biodiversity issues such as species coexistence, community stability, the detection of rare taxa, and biological invasions relate to evenness.

Visualizing Evenness: The Rank-Abundance Curve (RAC)

One of the most common ways to visualize evenness is with a **rank-abundance curve**, which is sometimes referred to as a rank-abundance distribution (RAD). A RAC can be constructed by ranking species from the most-to-least abundant.

Let's write a function to construct an RAC. First, we will remove species that have zero abundances. Then, we will order the vector from greatest (most abundant) to least (least abundant).

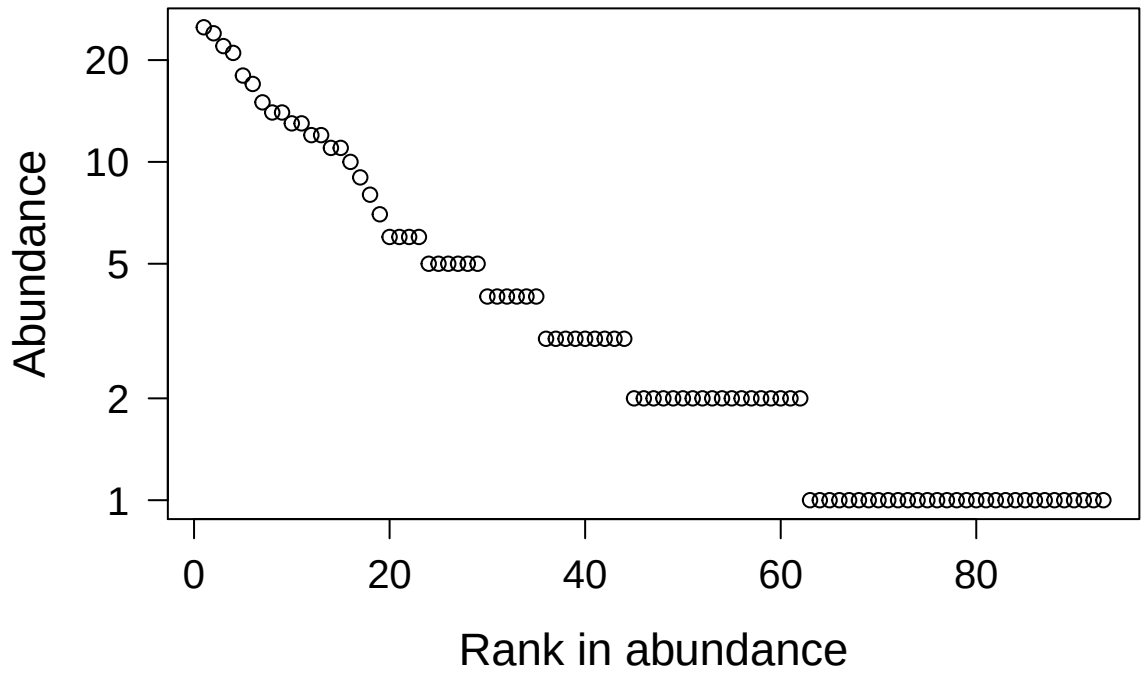
```
RAC <- function(x = ""){
  x = as.vector(x)
  x.ab = x[x > 0]
  x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
  return(x.ab.ranked)
  }
```

Now, let's examine the RAC for `site1` of the BCI data set. We will do this by creating a sequence of ranks and plotting the RAC with natural-log-transformed abundances.

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

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

box()                                                   # Manually adds border
axis(side = 1, labels = T, cex.axis = 1.25)             # Manually adds X-axis
axis(side = 2, las = 1, cex.axis = 1.25,                # Manually adds Log-Scaled Y-axis
     labels = c(1, 2, 5, 10, 20), at = log(c(1, 2, 5, 10, 20)))
```




```
par <- opar                                             # Resets plotting parameters
```

It is clear from looking at the RAC for `site1` that the abundance among species is unequally distributed. This sort of uneven distribution of abundance is one of the most ubiquitous patterns in ecology and has provoked a long history of study, theories, and explanations (see McGill et al. 2007) (http://www.ncbi.nlm.nih.gov/pubmed/17845298).

Now that we have visualized species evenness, it is time to quantify it. Based on decades of work, researchers have identified desirable features of an evenness metric. One of these features is that the values generated

by the metric should be relatively easy to intuit. For this reason, useful metrics are often bound between a minimum evenness of 0 and a maximum evenness of 1. Another important feature is that evenness values should be independent of richness S. That is, we don't want evenness to simply be a reflection of S. Here, we will introduce two metrics of evenness that meet the above criteria: Simpson's evenness ($E_{1/D}$) and Smith and Wilson's evenness index (E_{var}).

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

Simpson's evenness metric essentially reflects the sample variance of the rank-abundance curve, and is calculated as $E_{1/D} = \frac{1}{S} \sum \frac{N(N-1)}{n_i(n_i-1)}$, where S is species richness, N is total abundance, n_i is the ith species.

The `vegan` package does not have a function for Simpson's evenness but we can derive it using **Simpson's diversity (D)**, which estimates the probability that the next sampled individual belongs to a different species. We will speak more on Simpson's diversity in the next section. In the following R chunk we will write a function that estimates Simpson's evenness using the `diversity` function in `vegan` and observed richness.

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

Now, let's calculate Simpson's evenness for `site 1` in the BCI site-by-species matrix.

```
site1 <- BCI[1, ]
SimpE(site1)
```

We can see that Simpson's evenness for `site1` is moderately even (~0.42). However, Simpson's evenness has been criticized for being biased by the most abundant species. That is, the value of the metric is reportedly sensitive to differences in the few most abundant species. Let's examine the value of evenness for `site1` using a slightly more involved metric that is less biased by abundant species.

Smith and Wilson's Evenness Index (E_{var})

After reviewing existing metrics, Smith and Wilson (1996) derived a more robust measure of evenness, which they called E_{var}. This metric is standardized to take values between 0 (no evenness) and 1 (perfect evenness). Abundances are transformed to their natural logarithms to decrease bias towards the most abundant species, that is, the potential for a metric's value to be influenced more by large numbers than small ones. E_{var}, like all desirable measures of evenness, is independent of richness (S). The metric is calculated as: $E_{var} = 1 - 2/\pi \cdot \arctan(\sigma^2)$, where σ^2 is the sample variance (`var(log(x))`).

While seemingly more involved to calculate, E_{var} simply reduces to finding the sample variance of the log-transformed abundances and then standardizing it to take values between 0 and 1 using elementary trigonometry. Specifically, E_{var} uses the arctangent, which varies between -$\pi/2$ and $\pi/2$ and causes the metric to take values between 0 and 1. Subtracting this from one allows low evenness to be associated with values near 0 and high evenness to be associated with values near 1. A function for E_{var} can be written as follows:

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

Now let's use the E_{var} function to estimate evenness for `site1` of the BCI site-by-species matrix.

```
Evar(site1)
```

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

5) INTEGRATING RICHNESS AND EVENNESS: DIVERSITY METRICS

So far, we have introduced two primary aspects of diversity, i.e., richness and evenness. While we often examine each of these independently, the interaction between richness and evenness is important. Here, we will estimate popular indices of 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)

Shannon's diversity metric is derived from Shannon's information entropy, and is essentially a measure of uncertainty. This metric is used across the natural sciences and is calculated as $H' = -\sum p_i ln(p_i)$ where p_i is the proportion of individuals found in the ith species. A greater value of Shannon's diversity implies a greater chance that the next sampled individual will belong to a different species. Let's calculate Shannon's diversity for the RAC of `site1` in the BCI site-by-species matrix and then compare it to the `vegan` estimate:

```
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)
}
```

The basic structure of `ShanH` should look familiar, but there are some new features in this function. Namely, it includes a **for loop** and an **if statement**. The for loop iterates through each item (i.e., n_i) in the argument that is passed to the function. Since this function, like the Chao functions above, takes a single site (i.e., x) as an argument, the for loop iterates through the abundance of each species (i.e., n_i) in the site. In the next line, the if statement asks whether or not the abundance of a species meets a given criterion. Here, we are asking whether the abundance of a species is greater than zero. This is important because Shannon's H involves calculating the natural logarithm of the abundance, which is undefined for the value zero. If the abundance is greater than zero the function calculates the relative frequency (i.e., p) of the species in the site, takes the natural logarithm of the frequency, multiplies the logarithm by the frequency, and subtracts it from H Each species contributes to H, which starts off at zero (i.e., H = 0), based on its proportion in the sample. Therefore, H grows each time it loops through a species. Once this operation has been performed for all species in the site the value H is returned to the user.

Now we will use `vegan` to estimate Shannon's index:

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

Simpson's Diversity (or Dominance)

Simpsons diversity is a straightforward metric and is calculated as $D = \sum p_i^2$ where again, p_i is the proportion of individuals found in the ith species. Simpson's index is commonly expressed as $1/D$ so that index values naturally increase with diversity. Let's calculate Simpson's diversity for `site1` and then compare it to the `vegan` estimate:

```
SimpD <- function(x = ""){
  D = 0
  N = sum(x)
  for (n_i in x){
    D = D + (n_i^2)/(N^2)
  }
  return(D)
}
```

And now let's express Simpsons's diversity as $1/D$ (invD) and 1-D:

```
D.inv <- 1/SimpD(site1)
```

Now that we have written functions to understand how some of these metrics are quantified, we can also use `vegan` to estimate Simpsons's index:

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

6) MOVING BEYOND UNIVARIATE METRICS OF α-DIVERSIY

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 data reducing 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.

The uneven shape of the RAC is one of the most intensively studied patterns in ecology, and underpins all or most ecological theories of biodiversity. Predicting the form of the RAC is the first test that any biodiversity theory must pass and there are there are dozens of models that have attempted to explain the uneven form of the RAC across ecological systems. These models attempt to predict the form of the RAC according to mechanisms and processes that are believed to be important to the assembly and structure of ecological systems.

Again, we are going to make use of `vegan`. Specifically, we are going to use the `radfit()` function to fit the predictions of various species abundance models to the RAC of `site1` in BCI

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

From the output, you can see that `vegan` fits five models to our rank-abundance curve: *Null*, *Preemption*, *Lognormal*, *Zipf*, and *Mandelbrot*. Before explaining what these models represent, let's run through the `vegan` output:

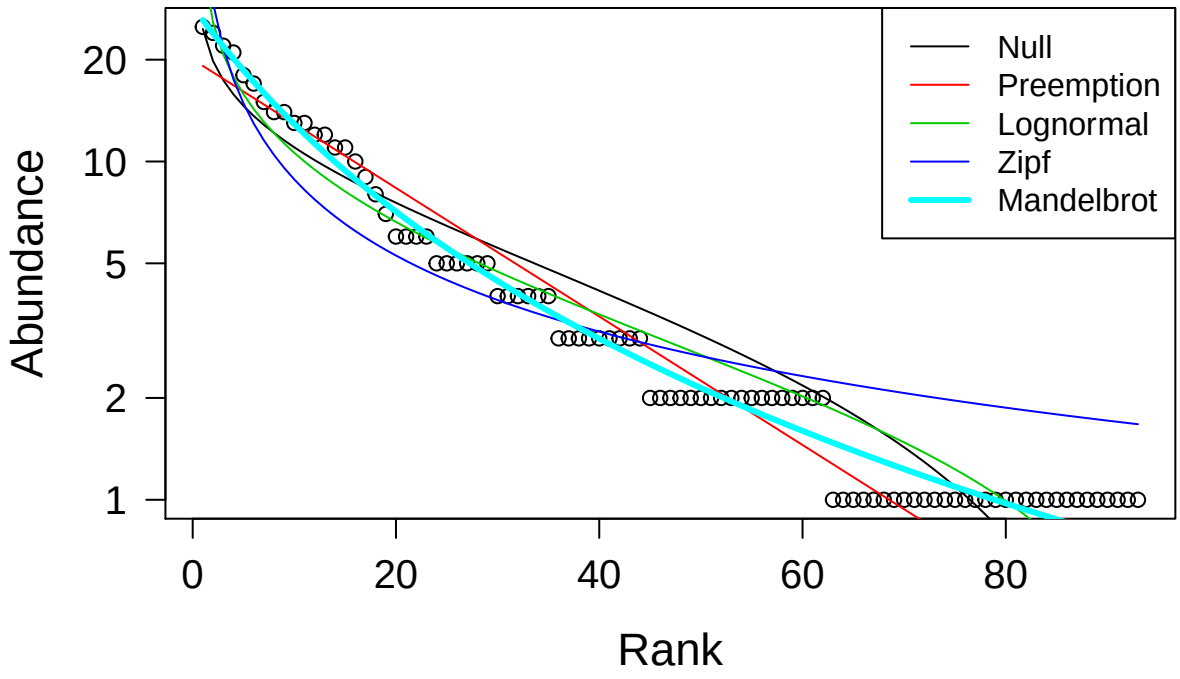
1. Next to "RAD models", we see "family poisson", which tells us that by default, `vegan` assumes Poisson distributed error terms.
2. Below this, we see that `vegan` returns the number of species (S) and the number of individuals (N) for the empirical RAC.

3. Next, we see a table of information, the first columns of which are par1, par2, and par3. These columns pertain to model parameters and reveal that the different models use different numbers of parameters; the null model uses none.
4. Next, we see a column for Deviance, which is a quality of fit statistic based on the residual sums of squares, which is a measure of the amount of variance that is not explained by a given model.
5. After Deviance, we see columns for AIC and BIC, which are the estimated **Akaike Information Criterion** and the **Bayesian Information Criterion**, respectively.

Notes on AIC and BIC: AIC and BIC are commonly used for model selection. In other words, they help us identify a model that is best supported by our data. Obviously, the more parameters a model has, the better it will fit a data set. However, it's not necessarily desirable to have an over-parameterized model. So, AIC and BIC assign penalties that correspond with the number of parameters that a model uses. In the end, the "best" model has the lowest AIC or BIC value.

Now, let's visualize our results by plotting the empirical RAC and the predicted RAC for each model:

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



A quick interpretation of the RAC models in vegan:

Null: A **broken stick model** (Pielou 1975) where the expected abundance of a species at rank r is $a_r = \frac{N}{S} \cdot \sum_{x=r}^{S} \frac{1}{x}$. N is the total number of individuals and S is the total number of species. This gives a constraint-based null model where the N individuals are randomly distributed among S species, and there are

no fitted parameters. Null models often reveal that realistic patterns can be expected from random sampling, and have been extremely useful in ecology and evolution.

Preemption: The **niche preemption model** (a.k.a., geometric series or Motomura model): Envision an environment occupied by a single species. Now, envision that a second species colonizes the environment and takes some portion of resources equal to α. Then, envision that a third species colonizes the environment and takes a portion of resources equal to α away from the second species. Imagine this process continuing until N is zero. The only fitted parameter is the preemption coefficient α, which gives the decay rate of abundance per rank. The expected abundance (a) of species at rank r is $a_r = N \cdot \alpha \cdot (1-\alpha)^{(r-1)}$.

Lognormal: Many statistical tests and models assume that data are normally distributed. However, species abundances are rarely ever normally distributed and there is almost always an excess of rare species. The lognormal model presumes that species abundances become normally distributed under *logarithmic transformation*. Log-transforming species abundances decreases the disparity between rare and dominant species, e.g., $0 = \log(1)$ and $3 = \log(1000)$. The lognormal model was introduced into ecology by Frank Preston in 1948 and is one of the most widely successful species abundance models. The expected abundance (a) of species at rank r is $\exp[\log(\mu) + \log(\sigma)\Phi]$, where Φ is a standard normal function, μ is the fitted mean abundance, and σ is the fitted standard deviation of abundances.

Zipf: The Zipf model is based on Zipf's Law (see: https://en.wikipedia.org/wiki/Zipf's_law). In short, the abundance of a species in the RAC is inversely proportional to its rank in abundance. The expected abundance (a) of species at rank r is: $a_r = N \cdot p_1 \cdot r^{\gamma}$, where p_1 is the fitted proportion of the most abundant species, and γ is a decay coefficient.

Mandelbrot: Shortened name for the Zipf–Mandelbrot model. This model adds one parameter (β) to the Zipf model. Ecologically, β can be thought of as a parameter that controls the evenness of the predicted RAC. For example, a positive β increases the evenness among highly abundant species while a negative value decreases evenness. The expected abundance of a species (a) at rank r is $a_r = N \cdot c \cdot (r+\beta)^{\gamma}$.