Week2_Supplementary

Jay T. Lennon

January 17, 2015

1) SETUP

Retrieve and Set Your Working Directory

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

```
#install.packages("vegan")
#instead, use try("vegan")
require("vegan")
```

```
## Warning: package 'vegan' was built under R version 3.2.5
```

```
## Warning: package 'permute' was built under R version 3.2.5
```

```
## Warning: package 'lattice' was built under R version 3.2.5
```

Another commonly used estimator is **ACE**, which stands for abundance-based coverage estimator. While Chao1 makes inferences based on the number of singletons and doubletons, ACE implements a threshold to look at the abundance of other *rare* species. By convention, ACE defines rare species as taxa that have 10 or fewer individuals. Consequently, whether one uses the ACE estimator depends on whether one's samples tend to have many species of few individuals. If so, the ACE estimator may ignore the majority of sampled species. Now we will write a function for the ACE estimator:

```
S.ace <- function(x = "", thresh = 10){
  x <- x[x>0]                           # excludes zero-abundance taxa
  S.abund <- length(which(x > thresh))  # richness of abundant taxa
  S.rare  <- length(which(x <= thresh)) # richness of rare taxa
  singlt  <- length(which(x == 1))      # number of singleton taxa
  N.rare  <- sum(x[which(x <= thresh)]) # abundance of rare individuals
  C.ace   <- 1 - (singlt / N.rare)      # coverage (prop non-singlt rare inds)
  i       <- c(1:thresh)                # threshold abundance range
  count   <- function(i, y){            # counter to go through i range
    length(y[y == i])
  }
  a.1     <- sapply(i, count, x)        # number of individuals in richness i richness classes
  f.1     <- (i * (i - 1)) * a.1        # k(k-1)kf sensu Gotelli
  G.ace   <- (S.rare/C.ace)*(sum(f.1)/(N.rare*(N.rare-1)))
  S.ace   <- S.abund + (S.rare/C.ace) + (singlt/C.ace) * max(G.ace,0)
  return(S.ace)
}
```

Notes + `estimateR` is a function in the `vegan` pakcage + It will spit out observed richness, along with Chao1, ACE, and their associated confidence intervals. + You can look more into the code in R packages using commands like this: vegan:::estimateR.default + Try it out!

More on EVAR

E_{var} uses the arctangent, which varies between -$\pi/2$ and $\pi/2$ and without being periodic like waves of the sine and cosine functions. Multiplying the arctangent by 2/pi forces the result to take values between 0 and 1. Finally, subtracting this from one allows low evenness to be associated with values near 0 and high evenness to be associated with values near 1. We can confirm this with a more explicit R chunk:

```r
data(BCI)
site1 <- BCI[1, ]
site1 <- site1[site1 !=0]
P <- log(site1) # log-tranform the abundances of the RAC and assign them to a vector P
AvgAb <- mean(P) # find the average of the log abundances
X <- 0 # assign zero to variable X
Evar <- 0 # declare a scalar varible Evar

for (x in P) { # making use of a 'for' loop. for loops are an elementary control structure in all progr
  X = X + (x - AvgAb)^2 / (length(P) - 1)
}

Evar = 1 - (2/pi)*atan(X) # these operations make the value of Evar range between 0 and 1
Evar # print Evar for Site1
```

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

Fisher's α

R.A. Fisher (1943) derived one of the first and most successful models for how abundance varies among species, i.e., the log-series distribution. This model has only a single fitted parameter, i.e., α, Because α is a fitted parameter, it is less straightforward to estimate and we will not attempt to code a function for it, here. Fisher's α has often been used as a diversity metric and is the root of α-diversity and, according to the authors of the vegan package, it is asymptotically similar to inverse Simpson's. Let's do this comparison using the RAC from site 1 of the BCI site-by-species matrix.

```r
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)
  }

rac <- RAC(x = site1)
invD <- diversity(rac, "inv")
invD
```

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

```r
Fisher <- fisher.alpha(rac)
Fisher
```

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

As we can see, the two measurements are somewhat similar. They would converge if our community was much greater in total abundance and richness. However, discussion of Fisher's α introduces a new concept, that is, of estimating diversity instead of just calculating a diversity metric. The difference being that an estimate of diversity implicitly or explicitly accounts for samplign error, that is, the fact that when samplign most ecological communities that we are not observing every single individual.

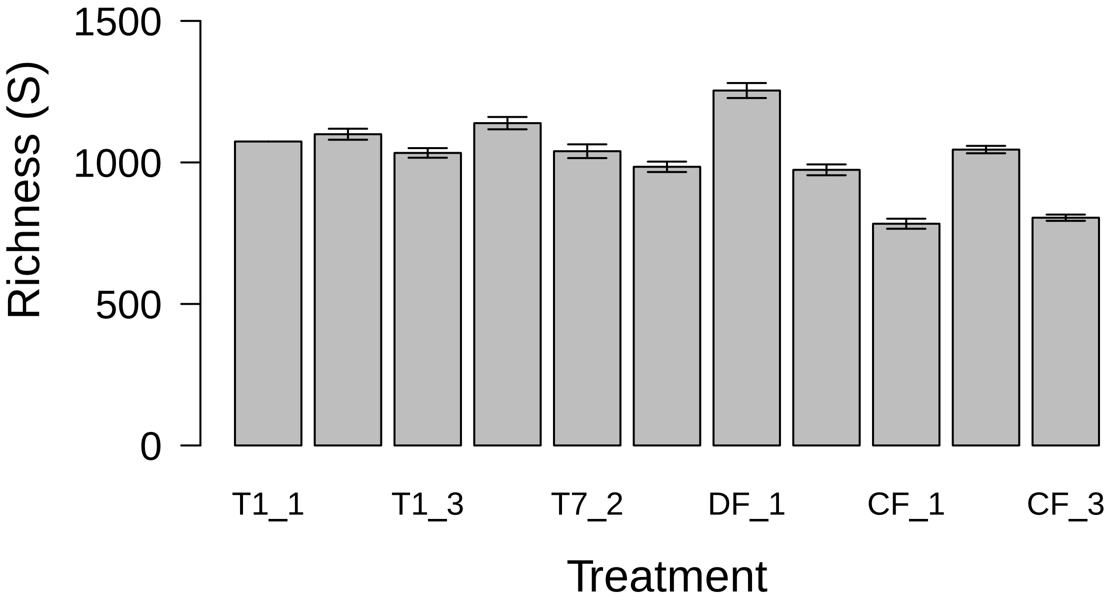
We can use the information from the `rarefy` function to create a barplot that compares each site. Remember, we can calculate 95% confidence intervals using $95\%CI = \bar{x} \pm SEM \times 1.96$.

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

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

soilbac.S <- S.obs(soilbac)
min.N <- min(rowSums(soilbac))
S.rarefy <- rarefy(x = soilbac, sample = min.N, se = TRUE)


opar <- par(no.readonly = TRUE)
par(mar=c(5.1, 6.1, 4.1, 2.1))
S.plot <- barplot(S.rarefy[1, ], xlab = "Treatment", ylab = NULL,
                  ylim =c(0, round(max(soilbac.S), digits = 0)),
                  pch = 15, las = 1, cex = 1, cex.lab = 1.4, cex.axis = 1.25)
arrows(x0 = S.plot, y0 = S.rarefy[1, ], y1 = S.rarefy[1, ] - (S.rarefy[2, ] * 1.96),
       angle = 90, length=0.1, lwd = 1)
arrows(x0 = S.plot, y0 = S.rarefy[1, ], y1 = S.rarefy[1, ] + (S.rarefy[2, ] * 1.96),
       angle = 90, length=0.1, lwd = 1)
title(ylab = "Richness (S)", line = 4, cex.lab = 1.4)
```



```
par(opar)
```

Notice that we did a few things differently here. Why did we have to plot the y-axis label manually? What did the `par(mar=c())` function do?

But How Well Did You Sample Your Site?

Accurate estimates of richness are influenced by sampling effort and biases. 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 and estimate Good's Coverage for `site1` of BCI:

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

Question 4: 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_1 equaled N?
c. What portion of taxa in `site1` were represented as singletons?
d. Have the researchers at BCI done a good job of sampling `site1`?

> ***Answer 4a***:
> ***Answer 4b***:
> ***Answer 4c***:
> ***Answer 4d***:

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 desireable measures of evennness, 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 calcualte, 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 $\pi/2$ without being periodic like sine waves. Multiplying the arctangent by $2/\pi$ forces the result 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. In the end, an E_{var} function 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
```

Question 9: Compare estimates of evenness for `site1` of BCI using $E_{1/D}$ and E_{var}. Do they agree? What can you infer from the results?

> ***Answer 9***:

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 $H' = -\sum p_i ln(p_i)$. 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:

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

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

```
rac <- RAC(x = site1)
diversity(rac, index = "shannon")
```

7) 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 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(rac)
```

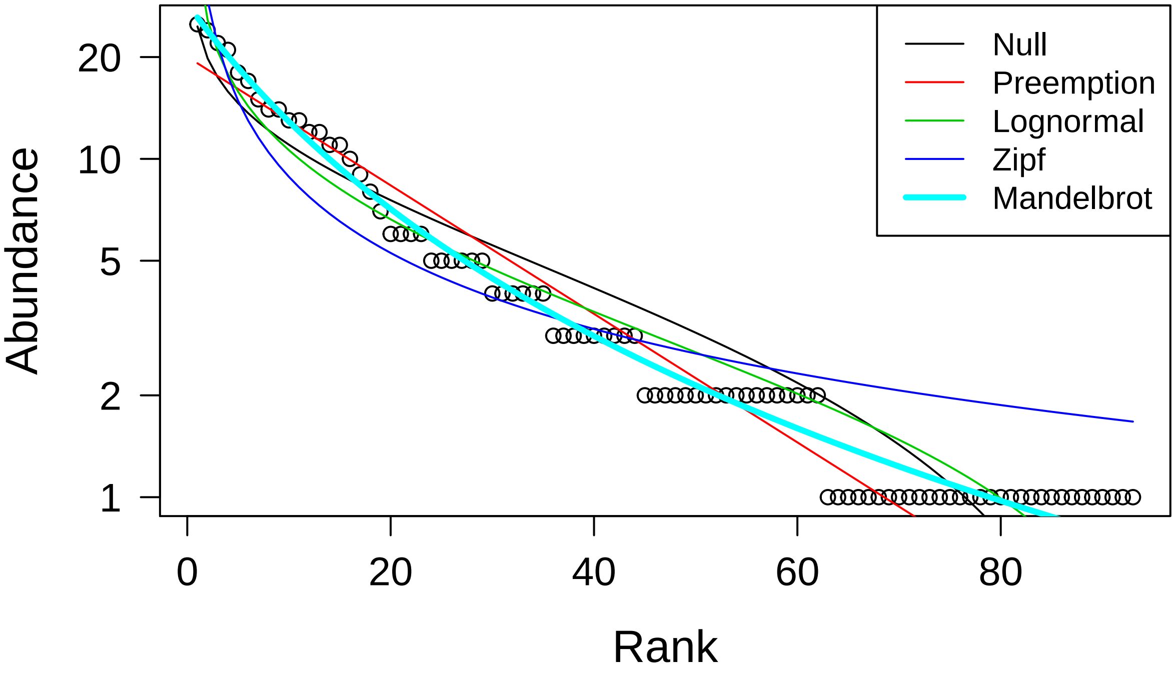
From the output, you can see that `vegan` fits five models to our rank-abundance curve: *Null*, *Preemption*, *Lognormal*, *Zipf*, and *Mandelbrot*. Before explaining what these models represent, let's run through the `vegan` output:

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 idea of residual sums of squares.
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 asssign 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:

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



Question 11: Based on the output of `radfit()` and plotting above, discuss which model best fits our rank-abundance curve for `site1`?

> ***Answer 11***:

Interpreting the RAC models in `vegan`:

Null:

A **broken stick model** (Pielou 1975) where the expected abundance of a species at rank r is $a_r = \frac{N}{S} \cdot \sum\limits_{x=r}^{S} \frac{1}{x}$. 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 estimated 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 $a_r = N \cdot \alpha \cdot (1 - \alpha)^{(r-1)}$.

Question 12: 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 12a:
Answer 12b:

Lognormal:

Many statistical models assume that the distribution of values are normally distributed. When applied to species abundances, this means that they conform to the shape of a symmetrical bell curve, more precisely known as the Gaussian distribution. In contrast, the log-Normal model assumes that the logarithmic abundances are normally distributed. The expected abundance of a species at rank r is then: $a_r = e^{log(\mu)+log(\sigma)\cdot Z}$, where Z is the Normal deviate. A Normal deviate is simply the number of standard deviations a score is from the mean of its population. The log-normal model was introduced into ecology by Frank Preston in 1948 and is one of the most widely successful species abundance models.

Zipf:

The Zipf model is based on Zipf's Law, an well-known observation that many types of ranked data are fit by a simple scaling law (a.k.a., power law). In short, the abundance of a species in the RAC is inversely proportional to its rank. 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, a generalization of the Zipf model made by the mathematician and father of fractal geometry, Benoit Mandelbrot. This model adds one parameter (β) to the Zipf model. The expected abundance of a species (a) at rank r is $a_r = N \cdot c \cdot (r + \beta)^{\gamma}$. Here, the p_1 parameter of the Zipf model changes into a meaningless scaling constant c.