**Abundance patterns:**

Abundance data typically come from direct counts of individuals or assays of percent cover (e.g. for plants or sessile animals like barnacles). Less common are indirect counts (e.g. counting number of nests or fecal pellets rather than the animals themselves) or other proxies that assume that there is a 1:1 relationship between counts of the proxies and counts of the organisms themselves.

Studies that use abundance data **assume that all species are equally detectable**, but this is rarely true: Some species are large and obvious whereas others are small and cryptic. They also tacitly **assume that all species are collected using the same method** at the same temporal and spatial scale. (Using different methods can be an “apples vs. oranges” problem.) These two assumptions are interrelated: species that are small and/or secretive require more intense searching and therefore require a more fine-scaled sampling unit (such as a quadrat) than do large and obvious species that can be detected with coarser methods (such as aerial photography). Because it is not feasible to cover the same area with (for example) quadrats as with aerial photographs (or vice versa), in such cases it is necessary to standardize abundance counts for each species by expressing them per unit area (i.e., as density). You can also standardize species abundances by subtracting the mean and dividing by the standard deviation.

As was seen in the site x species lesson, species abundance data are usually skewed and therefore are often log-transformed. One commonly used version of this is log(*x*+*k*), where *x* is species abundance and *k* is a constant, usually 1. The constant is added to avoid calculation of the logarithm of zero (species absence from a site or sample), since log(0) is undefined. Be advised that even though the log transformation is commonly used, it will give different results if different units are used. For example, the log(*x*+*k*) transformation will give different answers if you record biomass in grams vs. in kilograms! Most other transformations do not suffer from this problem.

Because abundance is one component of diversity, it is important to recognize that if you are comparing diversity between two communities, that you standardize by abundance. This is done via a technique called **rarefaction**, whereby the community with the larger number of individuals in the sample is reduced (“rarefy” means “to thin”) to equal the number of individuals in the smaller community sample. Rarefaction creates a random subset of the larger community that is the same size as the smaller community and then uses this to estimate the expected number of species that you would have encountered in the larger community if you had only sampled that smaller number of individuals. Rarefaction thus generates equivalent abundances based on different sample sizes; rarefaction assumes that total abundance imbalances between communities are due to sampling differences and not due to differences in actual abundances (due to some aspect that makes one community not truly comparable to another, such as differences in disturbance, habitat structure, land-use history, or other features that create an apples vs. oranges situation).

Once you are satisfied that your abundance data are sound (transform if necessary, rarefy if necessary), then you can examine one of the most fundamental traits of communities: patterns of abundance across species. Recall two lessons ago (“Patterns in community data”) that there are relationships between species occurrence and abundance, but abundance data themselves also exhibit well-documented patterns, including the truism that “most species are rare, few are common.” Indeed, in the “Patterns in community data” lesson, I showed a plot of species on the X-axis ranked from those with the most individuals to those with the fewest (down to species represented by singletons) against the abundance of each species on the Y-axis, resulting in a **rank-abundance curve** or **dominance-diversity plot** (also known as a Whittaker plot, or a species-abundance distribution). Rank-abundance curves usually have negative slopes when plotted with these axes; if you arrange the species in decreasing order of abundance and the graph is flat, that indicates high levels of evenness. But as we explored in the “Patterns in community data” lesson, most communities exhibit skew in their species-abundance distributions.

Different communities can be compared by comparing the shape (slope) of their rank-abundance curves. There are over 2 dozen shapes that species-abundance distributions can take, each resulting from a different process that can be mathematically modelled. It can be difficult to distinguish among these models, and each has its limitations. Wilson (1991), Magurran (2004), and McGill (2011) have much more information about various shapes of rank-abundance curves, and I also recommend Krebs (1999) for some worked examples.

*vegan* has five models, each with characteristic rank-abundance curve shapes:

* Pre-emption (geometric series)
* Broken stick
* Log-normal
* Mandelbrot (also called Zipf-Mandelbrot)
* Zipf

(Gardener 2014 has the mathematical formulae for each if you’re interested in the hairy details.)

These and other rank-abundance models fall into two camps: one group is based on assumptions of niche partitioning and competition that limits species’ abundances, and the other is based on the statistical behavior of large samples.

Pre-emption:

The first-arriving species pre-empt resources, which leaves fewer resources for individuals of other species, thereby limiting later arrivers and overall diversity. The result is a curve with a constant ratio between successive units (e.g. 1st unit is 2x [or 3x or whatever] of the 2nd, which is itself 2x of the 3rd, etc.). (This kind of constant-ratio function is called a geometric series.) The result is a short, steep curve with no tail/skew. This pattern tends to hold for communities in settings where there is only one or a few dominating environmental factors controlling colonization and establishment (e.g. light limitation for understory plants, or soil nutrient limitation for desert microbes). It is often found in communities at early successional stages.

If your community has more singletons than any other abundance class, then this model assumes a more log-series shape (less steep). The pre-emption model is an ecological model that assumes that species arrival dictates which species will get a larger share of resources and thus which species will come to dominate the community; the log series is an ecological model that assumes that the most competitive species will get a larger share of resources regardless of when it arrives in the community.

Broken stick:

This model assumes that resources are partitioned in proportion to the abundance of species, with no severe dominance seen because the niche is being partitioned simultaneously; the result is a shallow, almost flat line of very even abundances. It is often used as a null model since we are examining departures from a perfectly even community. The broken-stick model is an ecological model, often used as a null model because it is the most parsimonious: it assumes that there are environmental resource gradients that species partition in a simple way.

Log-normal:

This model is thought to be the most common one in nature and is characterized by having a few species that are very common but with most species VERY rare (i.e., highly right-skewed), often stair-stepped, with singletons less numerous than species of intermediate abundance(there is a sizeable area just after singletons–this “middle class” creates a less steep slope than for geometric but more so than for broken stick). This shape can emerge with increased sampling effort (as more rare species are counted in your community). The log-normal distribution is usually considered the default for undisturbed, natural communities; disturbance drives the distribution to become more like the geometric series. The log-normal model is a statistical model because the logarithmic abundances of many species are distributed normally, requiring no ecological reason to do so.

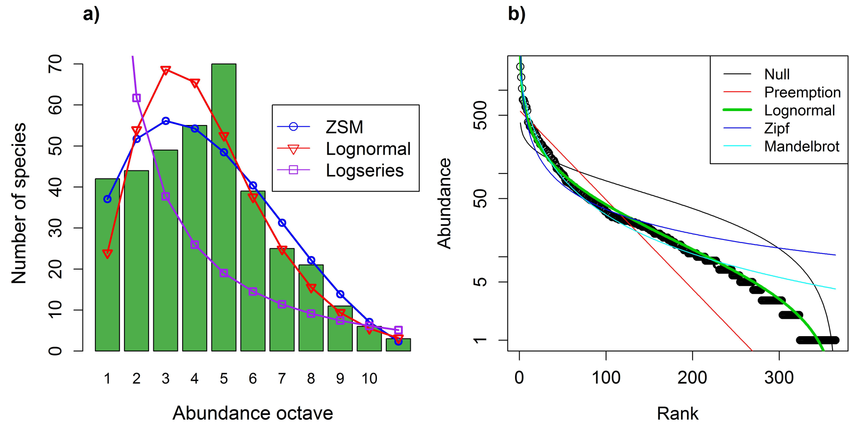
Mandelbrot:

This model is sometimes called the Zipf-Mandelbrot model because it is a modification of the Zipf model made by the father of fractal geometry, the mathematician Benoit Mandelbrot. In this model, abundance depends on presence of previous species as well as physical conditions; later colonizers of a community have more specific niche requirements and thus are rarer than pioneer/early successional species. The result is a community with a few very abundant species and many minor species of comparable abundance. This model assumes that the same species would be present at the same point in succession in similar habitat, which is ecologically unrealistic.

Zipf:

Forerunner of Mandelbrot but with fewer parameters (and with the same assumptions and artificiality as Mandelbrot), only relying on physical conditions and not the presence of pioneer species. This model is based on Zipf’s Law, an empirical law that points out that for many types of data, the rank-frequency distribution is an inverse relation fit by a simple scaling law (power law). In a Zipf model, the abundance of a species is inversely proportional to its rank.

Here are some examples of each, from Matthews and Whittaker (2014); the black dots are their actual datapoints:



Unless you have a very clear idea about what factors are affecting your community, the usual process is to general several models simultaneously and then compare them (visually as well as quantitatively via statistics). *vegan*'s radfit() function can be used to do these things IF your data are integers (i.e., count data). (If you have some different assay of abundance, such as percent cover, then you’ll have to modify the process somewhat; see Gardener 2014 for more info and instructions.)

**Exercises:**

Open a new RStudio session and set your working directory with the following libraries:

*BiodiversityR*

*labdsv*

*MASS*

*MVA*

*optpart*

*plyr*

*stats*

*vegan*

From the course website, download the dataset named div\_data.csv to your course folder. Read in that file into your R session and call it div. (Header = TRUE since the first row in the file is a list of column names.) Examine the data in that file.

First, determine richness at each site (using code you learned in the “Site x species” lesson):

sp.pres <- apply(div>0,2,sum)

sp.pres

Sites Sp1 Sp2 Sp3 Sp4 Sp5 Sp6 Sp7 Sp8 Sp9 Sp10

10 1 1 3 5 6 1 4 5 3 2

Sites 5, 4, and 8 have the highest richness values. Keep this in mind for later.

What about in terms of abundance? This is straightforward with ddply() from *plyr*:

div1 <- div

#This keeps the original R object intact. div1 is a copy that we can play with.

ddply(div,~Sites,function(x) + data.frame(ABUNDANCE=sum(x[-1])))

Sites ABUNDANCE

1 Site1 75

2 Site10 106

3 Site2 93

4 Site3 85

5 Site4 54

6 Site5 70

7 Site6 63

8 Site7 46

9 Site8 56

10 Site9 111

On the basis of abundance, Sites 7, 4, and 8 might be considered the poorest (support the fewest individuals).

*Rarefaction:*

If we want to compare our 10 sites equitably, we need to set them to equivalent sizes. Site 7 is the smallest, with 46 individuals. You can tell that either by looking at the data, or by

rowSums(div[2:11])

(The 2:11 indicates the rows with actual data; recall that column 1 was sites.)

With the rarefy() function we can set all sites to 46 individuals:

ddply(div,~Sites,function(x) + data.frame(RAREFY=rarefy(x[-1], sample=46, MARGIN=1)))

MARGIN = 1 indicates that the function proceeds by row (if = 2, it would go by column).

The result is a list of abundances by sites that allows for a more equivalent comparison. **How do these results compare with our non-rarefied analyses above?**

Here’s another way to do it:

rarefy(div[2:11])

The [2:11] indicates the columns where the numbers are (column 1 is categorical, consisting of site names).

But here’s what results:

Error in rarefy(div[2:11]) : the size of 'sample' must be given --

Hint: Smallest site maximum 46

So R indicates the value that you need to rarefy to, in this case 46:

R <- rarefy(div[2:11], sample = 46)

names(R) = levels(div$Sites)

R

Or for easier reading of the results:

R <- rarefy(div[2:11], sample = 46)

names(R) <- factor(div$Sites, levels = c("Site1","Site2","Site3","Site4","Site5","Site6","Site7","Site8","Site9", "Site10"))

R

The result is the rarefied species richness for each site if each site had only 46 individuals. You can see that sites 2, 3, and 5 now have the highest richness whereas before it was 5, 4, and 8.

*Rank-abundance curves:*

Read in the Ground\_beetles\_abundance.csv file as an object named GBA (with row.names = 1). Then apply the radfit() function:

gb.rad <- radfit(GBA)

gb.rad

The result is a table of values for each of the 18 sites from Ground\_beetles\_abundance.csv as follows (don’t worry about any warning messages):

Deviance for RAD models:

E1 E2 etc.

Null 828.4633 546.8046

Preemption 86.1171 74.6780

Lognormal 96.6051 144.1771

Zipf 105.5441 184.9127

Mandelbrot 39.9992 74.7680

The RAD stands for “rank abundance distribution.” “Null” is the broken stick, and the other four models are indicated by their usual names. “Deviance” is this function’s default way of assessing model fit and is based on residual sums of squares; from the above results, it looks like the Mandelbrot model has the best fit to the data from site E1 and the pre-emption model is best for E2 because they exhibit the lowest deviance values. However, deviance can be affected by the number of parameters it is examining (i.e., sample size: number of species and number of individuals). The more parameters a model has, the better it will fit a data set. However, it is possible to have an over-parameterized model (over-fit), and deviance is subject to that. Other goodness-of-fit statistics, such as AIC (Akaike’s Information Criterion) and BIC (Bayesian Information Criterion), assign penalties based on the number of parameters that a model uses. The best model has the lowest AIC or BIC value. AIC is very widely used in ecology and wildlife biology.

You can use the summary() command to give more details about each sample; the output will be long, since there were 18 sites:

summary(gb.rad)

Here is what the results look like for just the first site listed:

\*\*\* E1 \*\*\*

RAD models, family poisson

No. of species 17, total abundance 715

par1 par2 par3 Deviance AIC BIC

Null 828.463 888.755 888.755

Preemption 0.5215 86.117 148.409 149.242

Lognormal 1.5238 2.4142 96.605 160.897 162.563

Zipf 0.63709 -2.0258 105.544 169.836 171.502

Mandelbrot 3399.6 -5.3947 3.9929 39.999 106.291 108.791

Based on AIC, the Mandelbrot model is the best fit for site E1 (just as deviance indicated in an earlier paragraph). This means that this site’s community of ground beetles has a few very abundant species and then lots of species of comparable abundance.

If you want to use one goodness-of-fit statistic (for example, AIC) to examine all of the sites, then you can use the following:

sapply(gb.rad, function(x) unlist(lapply(x$models, AIC)))

E1 E2 E3 E4 E5 E6 Etc.

Null 888.7550 604.7068 589.1388 965.4858 973.9772 771.2113

Preemption 148.4088 134.5803 108.4166 248.3355 183.4645 146.9783

Lognormal 160.8968 206.0793 171.8131 233.6363 232.5519 210.6583

Zipf 169.8358 246.8149 206.6991 257.8194 281.9487 260.2791

Mandelbrot 106.2909 138.5802 112.4166 160.2413 144.2947 150.9779

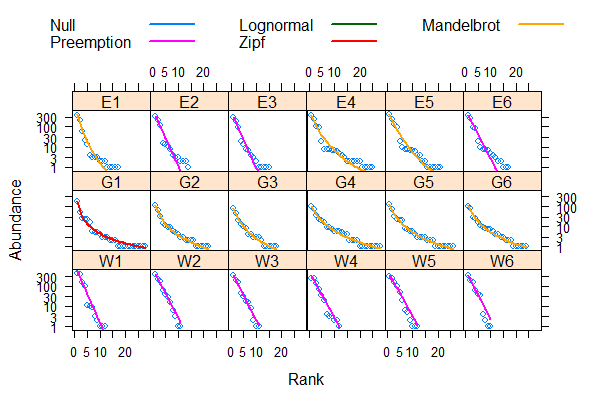
From this sample of just the six E (ecotone) sites, you can see that the pre-emption and Mandelbrot models provide the best fits (indicated in yellow above). Then once you identify the best-fit models, you should think about the interpretation of your findings. In this case, sites E2, E3, and E6 may be experiencing some kind of disturbance more than the other E sites; they are all at an early successional stage (that’s what an ecotone is), so it’s not surprising that the pre-emption model was the best fit for so many of them. The pre-emption model indicates that they are very uneven communities (i.e., dominated by one or a few species, which should then warrant more investigation—which species were they?). For the others with the Mandelbrot fit, their communities are a bit more even; do the dominant species at E2, E3, and E6 occur at the other sites? If so, what is limiting their abundance there? (So you can see how doing these kinds of analyses can give you insights into your data and can point you in the direction of deeper investigations.)

Sometimes, one or more rank-abundance models may fail, meaning that they cannot be made to fit the shape of the data. Zipf and Mandelbrot are especially prone to this. If the fitting fails, NA is returned.

Now let’s examine the results graphically:

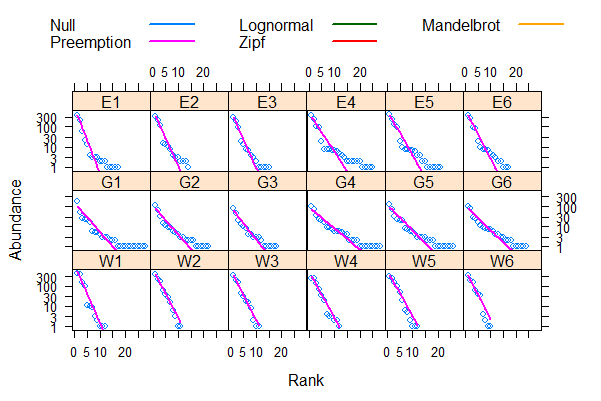
plot(gb.rad)

The result is a plot for each site, with the model with the lowest AIC fitted to the data:



If you want to examine only a particular type of model and fit that to each site (even if it didn’t have the lowest AIC), then you can do this (e.g. for the pre-emption model):

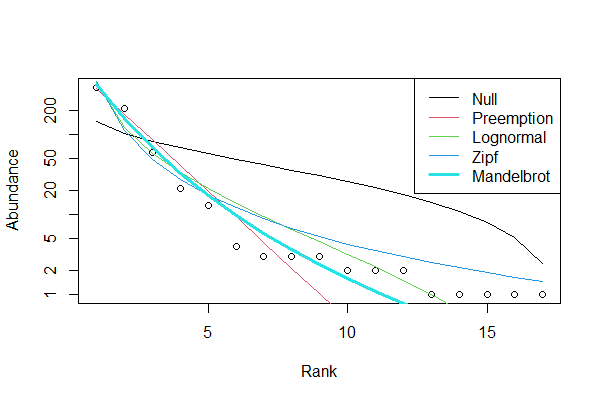
plot(gb.rad, model = "Preemption")



Take a look at site G1 in this graph and compare it to the previous one; you can see the better model fit (Zipf) in the first graph.

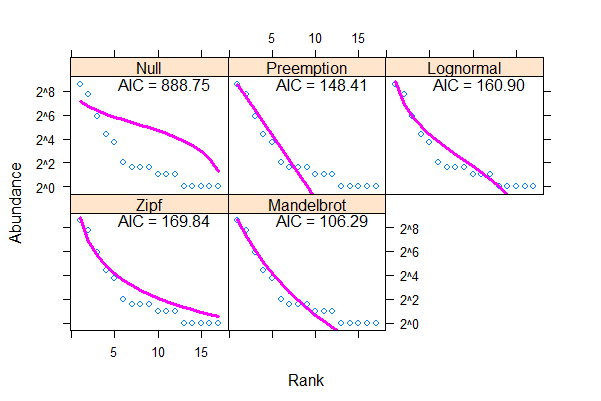
Now let’s look in more detail at a single site (e.g. E1). This code gives you all five rank-abundance models on the same graph:

plot(gb.rad$E1)



whereas this code gives you five panels, each with only one model:

radlattice(gb.rad$E1)

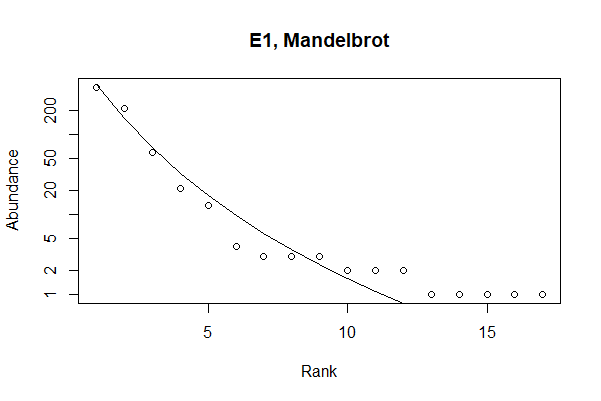


(Unfortunately, the colors are set by the plot() function internally.)

Now let’s identify the points on each plot to species, to visualize ones that are rare vs. common. This should be done on a per-site basis, so once again let’s focus on site E1. We discovered earlier that the Mandelbrot model was the best-fit one for E1’s data, so you first build only that model for E1.

gb.mand = rad.zipfbrot(GBA["E1",])

A = plot(gb.mand, main = "E1, Mandelbrot")



Each one of those dots is a species, so to determine which dot is which species, you can use the identify() function:

identify(A, cex=0.75)

You’ve used this function before in the site x species lesson. You just click on each dot and when done, hit the Escape (Esc) key. (If you do all 17 points, the function will end automatically after the last one.) In the code above, the cex=0.75 indicates the label size. You can play around with it, and you don’t have to click exactly on each dot: you can click near each dot and that is where its label will be placed.

*vegan* also has two functions that fit a lognormal model to your data (prestonfit() and prestondistr()). You can do ?prestonfit and ?prestondistr in R to find out more info.

In the video for today, I discuss a couple of papers from the primary scientific literature that use the topics covered today (rarefaction, dominance-diversity plots) (and one that uses topics from last week, diversity indices) to address their research topics. These papers (Avila-Cabadilla et al. 2009, Zhou et al. 2013) illustrate practical applications of these topics.

**References:**

Avila-Cabadilla, L.D., K.E. Stoner, M. Henry, and M. Yolotl Alvarez Añorve. 2009. Composition, structure and diversity of phyllostomid bat assemblages in different successional stages of a tropical dry forest. Forest Ecology and Management 258:986-996. <https://doi.org/10.1016/j.foreco.2008.12.011>.

Gardener, M. 2014. *Community Ecology: Analytical Methods using R and Excel*. Pelagic Publishing, Exeter, UK.

Krebs, C.J. 1999. *Ecological Methodology* (2nd ed.). Addison-Welsey Educational Publ., Inc., Menlo Park, CA.

Magurran, A.E. 2004. *Measuring Biological Diversity*. Blackwell Publishing, Malden, MA.

Matthews, T.J., and R.J. Whittaker. 2014. Fitting and comparing competing models of the species abundance distribution: assessment and prospect. Frontiers of Biogeography 6. <https://doi.org/10.21425/F5FBG20607>.

McGill, B.J. 2011. Species abundance distributions. Pp. 105-122 in: *Biological Diversity: Frontiers in Measurement and Assessment* (A.E. Magurran and B.J. McGill, eds.). Oxford University Press, New York, NY.

Wilson, J.B. 1991. Methods for fitting dominance/diversity curves. J. Veg. Sci. 2:35-46.

Zhou, Y., et al. 2013. Biogeography of the ecosystems of a healthy human body. Genome Biology 14:R1. <http://genomebiology.com/content/14/1/R1>.

**Assignment:** due 0800 Monday, March 15

Start a fresh RStudio session. Remember to set your working directory to your course folder and use the same package libraries as we used today.

Load Ground\_beetles\_abundance.csv as an object named GBA (indicate row.names = 1). This file contains abundances of beetles at each of 18 sites. Notice carefully how this data table is laid out: there is no column with site names (unlike div).

**Q1. What are the top three sites with the greatest richness?**

**Q2. What are the top three sites with the greatest abundance?**

Now rarefy to the smallest abundance. **Q3. After rarefaction, are the richness patterns consistent with your answer to Q1?**

Now take those top 3 most speciose sites (after rarefaction) and plot their rank-abundance curves. **Q4. What is the most appropriate statistical rank-abundance model for each of the three most speciose communities?** (I don’t want results from the other sites!)

**Provide only the best-fit plots for each of these three sites.**

**Q5. How would you interpret your findings? (Also think about this with respect to the habitat type(s) of the sites involved.)**

(Ignore all the warning messages!)

Make an RMarkdown Word file of your work and turn that in. Be sure to include your answers to the questions asked! Turn in your assignment as a Word document via email to [iroro.tanshi@ttu.edu](mailto:iroro.tanshi@ttu.edu) no later than 8:00 a.m. on Monday of next week. In your email, please include the following as the Subject line:

Assignment on abundance