**Working with site x species data in R**

**Introduction:**

Before beginning detailed analyses of a community data set, it's important to thoroughly familiarize yourself with the characteristics of the data. Simple graphical means are often the best way to begin.

**Reading in data:**

As in the last lesson, today we’ll once again be working with the bryceveg.R data matrix that contains vascular plant species abundance data (excluding trees) from 160 sampling locations in Bryce Canyon National Park, Utah. Each sampling location was a 375 m2 circular plot. A proxy of species abundance was used: instead of counting numbers of individuals, abundance was assigned as one of 8 mutually exclusive cover categories designating what % of the ground within a sampling plot was covered by that species. Those categories were:

Category name % cover Interpretation

Present 0% present but only in minute amounts

Trace 0-1% very rare

Class 1 1-5% relatively rare

Class 2 5-25% uncommon

Class 3 25-50% common

Class 4 50-75% very common

Class 5 75-95% extremely abundant

Class 6 95-100% dominant

No plots were covered 100% by a single species.

The data in bryceveg.R represent number codes that were given for each of these classes.

Species names in the dataset are abbreviated as six-letter abbreviations of the scientific name (the first three letters of the genus name and the first three letters of the specific epithet). Unfortunately, there is no convenient list of what species those abbreviations are referring to; you need to go to <https://www.nps.gov/im/ncpn/species-lists.htm>, select Bryce Canyon and vascular plants to get a list. So we’ll stick to the abbreviations.

Open RStudio and set your working directory:

setwd("C://path/to/your/class/folder/where/bryceveg.R/is”)

Remember that the slashes should go forward rather than backward.

We will be using the following packages. Note that each package must have been installed first (which was from an earlier lesson). Apply the library() function to each of these packages:

*labdsv*

*MASS*

*MVA*

*optpart*

*stats*

*vegan*

(You should use the library() function for every new R session.)

In the previous lesson, you read in bryceveg.R to create the data matrix veg. We will be working with those data again today. You could open up your working directory from last time’s session to access veg, but you’d have everything else along with it. So it’s cleaner and just as easy to just build veg again:

veg <- read.table("bryceveg.R",header=TRUE)

bryceveg.R was a text file that could just as easily have had a .txt file extension. You’d use the same read.table() syntax for either .R or .txt files.

To read in a .csv file (comma-delimited file, a very common database format), you’d use read.csv() instead.

To read in a .dbf (DBase) file, use read.dbf().

It is possible, although more difficult, to read Excel .xls or .xlsx files. Excel files can have problems with data formatting and storage, and the best advice is simply not to attempt to read .xls or .xlsx files into R. Instead, I recommend that you use Excel to export the spreadsheet to a .csv file, and then use read.csv().

**Viewing the data:**

Now that we've got the data into R, you can see that it is a site x species matrix. (You can click on veg in the Global Environments window to open up the file and view it easily.) With these data we can answer the following questions (which I am going to color-code to make it easier for you to navigate the subsequent pages):

* **In how many plots does each species occur?**
* **What is the mean cover of each species when it occurs (not including zeros for plots where it is absent)?**
* **Is the mean abundance of species correlated with the number of plots in which it occurs?**
* **How many species occur in each plot?**
* **Is the total abundance of vegetation correlated with the number of species in a plot?**

To answer many of these questions, we will first need to transform the data to make them scaled similarly to each other.

**Transformation of data:**

If the original data are strictly numeric and the transformation is a simple mathematical function, then we can simply operate on the whole matrix at once. For example, to square the values in veg (which would place more emphasis on dominant species) we could simply do:

vegsq <- veg^2

More commonly, however, we want to de-emphasize dominant species, which takes a square root transformation:

vegsqrt <- sqrt(veg)

In our case, however, our data are cover classes that represent intervals. The standard way to deal with interval data is to use a midpoint transformation, but there is no mathematical function for that. Instead, we can create a new data frame, called cover, and assign new values with the assignment operator. This preserves the original veg file.

cover <- veg

#to create a copy of veg

cover[veg==1] <- 3.0

#to convert class 1 to midpoint of 3.0 percent

cover[veg==2] <- 15.0

#to convert class 2 to midpoint of 15.0 percent

cover[veg==3] <- 37.5

#to convert class 3 to midpoint of 37.5 percent

cover[veg==4] <- 62.5

#to convert class 4 to midpoint of 62.5 percent

cover[veg==5] <- 85.0

#to convert class 5 to midpoint of 85.0 percent

cover[veg==6] <- 97.5

#to convert class 6 to midpoint of 97.5 percent

You can confirm the transformation worked by looking at a bit of the data (in this case, the first 15 rows and 10 columns):

cover[1:15,1:10]

junost ameuta arcpat arttri atrcan berfre ceamar cerled cermon chrdep

1 0 0.0 3.0 0.0 0 0 0.5 0.0 0 0

2 0 0.5 0.5 0.0 0 0 0.0 0.0 0 0

3 0 0.0 3.0 0.0 0 0 0.5 0.0 0 0

4 0 0.5 3.0 0.0 0 0 0.5 0.0 0 0

5 0 0.0 62.5 0.0 0 0 0.5 0.0 0 0

6 0 0.5 3.0 0.0 0 0 3.0 0.0 0 0

7 0 0.0 62.5 0.0 0 0 3.0 0.0 0 0

8 0 0.0 15.0 0.0 0 0 0.0 0.0 0 0

9 0 0.0 0.0 0.0 0 0 0.0 0.0 0 0

10 0 0.0 85.0 0.0 0 0 0.5 0.0 0 0

11 0 0.0 15.0 0.0 0 0 0.5 0.5 0 0

12 0 0.0 15.0 0.5 0 0 0.5 0.0 0 0

13 0 0.0 0.5 0.0 0 0 0.5 0.0 0 0

14 0 0.0 37.5 0.0 0 0 0.5 0.0 0 0

15 0 0.5 62.5 0.0 0 0 0.5 0.0 0 0

Now we can address the first question: **In how many plots does each species occur?**

spc.pres <- apply(veg>0,2,sum)

#to get number of presences for each species

#The first part of the function

#(veg>0) evaluates to TRUE/FALSE or 1/0),

#and it is the sum of ones and zeros that

#gets calculated.

plot(sort(spc.pres))

#to see a plot of the cumulative empirical density

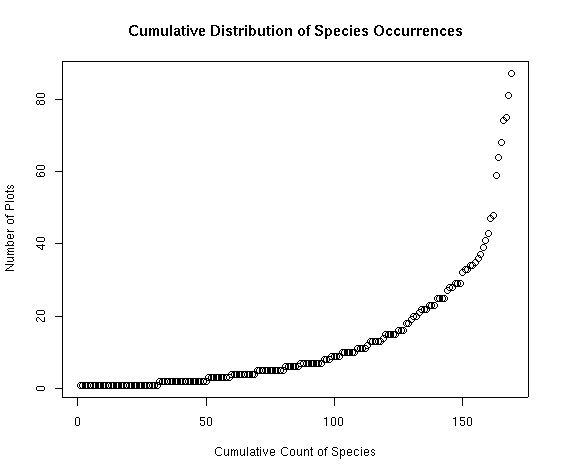
#function for species presences

We can label the plot with a title and more informative labels using the main=, xlab=, and ylab= arguments:

plot(sort(spc.pres),

main="Cumulative Distribution of Species Occurrences",

xlab='Cumulative Count of Species',ylab='Number of Plots')

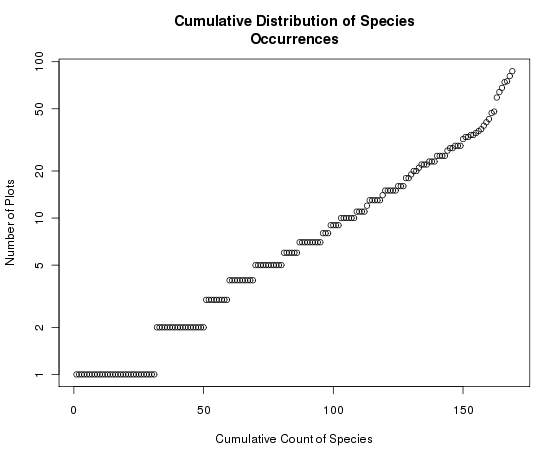


Recall from examining veg that there were 169 species overall and 160 sampling plots. So this figure shows that it took sampling about 90 plots before all 169 species were encountered. This is a fairly typical pattern of the distribution of species occurrences: it is highly skewed in a curvilinear fashion. We can correct for that by log-transforming the Y axis to achieve a semi-log plot:

plot(sort(spc.pres),log='y',

main="Cumulative Distribution of Species Occurrences",

xlab='Cumulative Count of Species',ylab='Number of Plots')

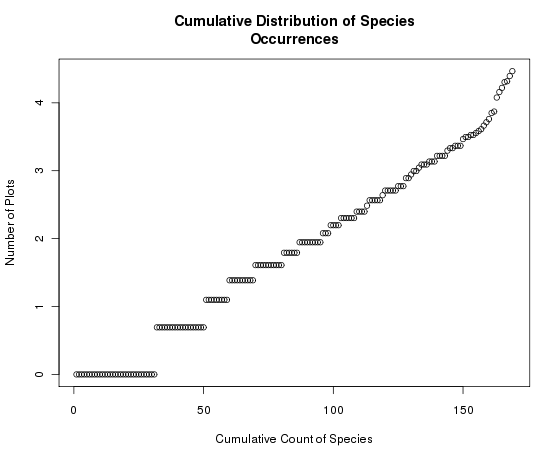


Notice how the Y axis is now log-scaled but retains the data in the original scale. If we had plotted the log species abundances we would get

plot(sort(log(spc.pres)),

main="Cumulative Distribution of Species Occurrences",

xlab='Cumulative Count of Species',ylab='Number of Plots')



Which is the same as the previous graph but with a different Y axis. Notice here that the graph itself has not changed, but the Y axis is now scaled 0-4 (log of spc.pres) instead of 1-100. Generally, we want to transform the axis and keep the units in the original scale for ease of interpretation, so the previous graph was preferable. Examine that graph.

Along the X axis is the number of species at or below a specific Y value. For example, approximately 100 out of 169 species were found in 10 or fewer plots; 130 of the species were found in 20 or fewer plots. (Thus, most species were relatively rare. This is a very common species abundance pattern.)

To see the exact distributions of how many species were found in 10 or fewer plots, use the seq() function as follows:

seq(1,ncol(veg))[sort(spc.pres)==10]

[1] 103 104 105 106 107 108

This is how you interpret the resulting vector: the first 102 species occur fewer than 10 times, species 103-108 occur exactly 10 times, and species 109-169 occur more often than 10 times.

Let's look at this expression a little closer. The second half (the logical subscript, i.e., [sort(spc.pres)==10])means "those species that occur 10 times". (That’s what the double equals sign signifies.) To see which species those are:

spc.pres[spc.pres==10]

agrdas asthum astmeg erifla eriumb tradub

10 10 10 10 10 10

Generally speaking, most species are rare; few species are common. To see this well-known pattern in this dataset, enter

sort(spc.pres)

chrdep shearg arcuva atrcon agrscr phlpra anemul artcar astchi astten dessop

1 1 1 1 1 1 1 1 1 1 1

echtri erieat erisub genaff gerric heddru ivesab leueri ligpor litinc lygspi

1 1 1 1 1 1 1 1 1 1 1

. . .

. . .

. . .

tetcan erirac artarb juncom chrvis oryhym sithys pacmyr ceamar purtri berrep

35 36 37 39 41 43 47 48 59 64 68

arcpat senmul symore carrss

74 75 81 87

where ellipses (…) represent omitted entries for the purpose of illustration here.

OK, so now we know that we have a lot of species that do not occur at most sites, and a few species that do. **What is the mean cover of each species when it occurs (not including zeros for plots where it is absent)?**

Here we use the apply() function to calculate the sums of the columns of the matrix veg, and then divide by the number of plots in which that species occurs.

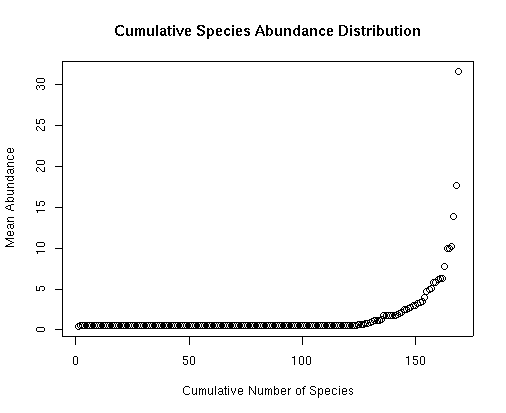
tmp <- apply(cover,2,sum)

spc.mean <- tmp/spc.pres   
#to get the average cover for each species

This code divided a vector of numbers (tmp) by another vector of numbers (spc.pres) to get yet a third vector (spc.mean). Now examine it:

plot(sort(spc.mean),main="Cumulative Species Abundance Distribution",

xlab="Cumulative Number of Species",ylab="Mean Abundance")



Notice that this plot of abundance is even more extreme than the previous plot of occurrence. Most species only occur at very low abundance (i.e., cover), whereas relatively few are abundant (only four species have an average cover >10%).

Instead of dot plots like above, we can also plot the data in histogram form:

hist(spc.pres)



Species-abundance distributions frequently follow a log-normal distribution. We can easily convert the plot to log distributions by:

hist(log(spc.pres))



Clearly, even after log transformation, the number of species with very few occurrences is higher than we would expect for any sort of normal distribution. (Students who took my Community Ecology class may remember the “veil line” effect of sample size; more samples may “fix” that large left-hand bar.)

In addition to the universal pattern that most species are rare/few are common, another truism about species-abundance relationships is that widespread species tend to be more abundant than species that are found in fewer locations. We can examine this in our data by answering the following question: **Is the mean abundance of species correlated with the number of plots in which it occurs?**

To answer this question we can plot the species mean abundance (spc.mean) as a function of the species occurrences, i.e., presences (spc.pres), as follows.

plot(spc.pres,spc.mean)

to examine the relationship between species distribution and abundance. To see which dot is which species we can use the identify() function:

identify(spc.pres,spc.mean,names(veg))

Click on some dots and then hit Esc; the species abbreviation will appear next to the dots you clicked on. This is what it looks like for quite a few of the dots selected:



The distribution shown is interesting. The widespread species are on the right side of the X axis, and most (e.g. carrss, symor, senmul, berrep) have low mean abundances. This is in contrast to what is more commonly seen, that widespread species are abundant (and, conversely, species that are restricted in distribution are also few in number). Only arcpat is both widespread and abundant. Similarly, some other species (e.g. arttri, artarb) are much less widespread but abundant when present.

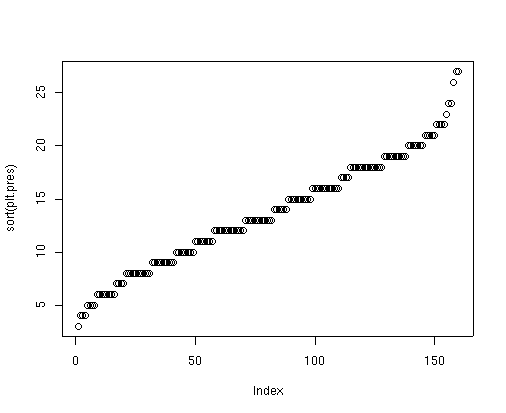
**How many species occur in each plot?** And, relatedly, **is the total abundance of vegetation correlated with the number of species in a plot?**

To calculate this, we will again use the apply function, this time applied to the rows rather than columns of veg.

plt.pres<-apply(cover>0,1,sum)

#to calculate the number of species in each plot

plot(sort(plt.pres))



So this graph tells us that the number of species per plot ranged from 3 to 27.

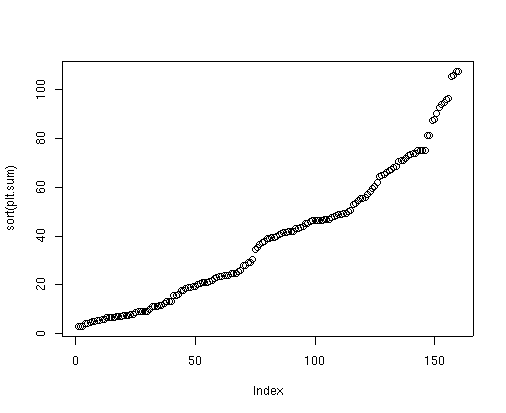
To graph the total cover on each plot, first calculate the vector of plot cover as follows:

plt.sum <- apply(cover,1,sum)

#to calculate the total cover on each plot

and then plot as:

plot(sort(plt.sum))



This plot tells us that total cover per plot ranged from 3% to just greater than 100% (107.5). (Because the species are estimated individually and then summed, it is possible to get more than 100% cover in a single plot.)

Finally, to answer our question about the relationship between total cover and number of species/plot:

plot(plt.pres,plt.sum)



Given the lack of a pattern in this graph, there is not much of a relationship between the number of species per plot and the total cover. Quite often, species-rich areas also have lots of individuals; however, our we measured abundance as cover rather than numbers of individuals, so perhaps that accounts for our atypical result.

**So to summarize patterns in the Bryce Canyon site x species dataset:**

The data appear to follow the well-known pattern that most species are rare, few are common (because most species occurred in few plots). Relatedly, the average “abundance” (mean cover) of most species was fairly low; only four species had an average cover >10%. Surprisingly, though, the mean abundance of species was generally not associated with how widespread they were. There were 3-27 species (out of 169) per plot. This species richness was not associated with abundance, meaning that more diverse plots didn’t also have greater abundance (but perhaps our data weren’t appropriate to examine this since they were cover data rather than counts of individuals).

**Assignment:** due 0800 Monday, 22 February

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

Use the head() function to see what the 6 column names are (or just look at the file by clicking on it in the Global Environments window). Recalling info from a previous lesson, you should see that gb is a data frame because it includes both numerical and categorical data. Here is what each of the column titles means:

Species = scientific name

Quantity = number of individuals (i.e., abundance)

Sample = alphanumeric code for each of the 18 sampling locations (i.e., plots)

Abbr = abbreviation of species’ scientific name

Max.Ht = maximum height of the vegetation in the sampling location (plot)

Habitat = three habitat types (edge [ecotone], grass [grassland], wood [woodland]) were sampled, with 6 plots per habitat type

Today we will focus on the species data; next time we’ll do analyses involving habitat.

Notice that these data are not in the “rows = samples, columns = species and environmental data” format that we have been using. We can still work with it, though!

Notice that the data are true abundance (number of individuals) data; let’s say that we want a site x species matrix (where the Sample column includes the site codes), but where the data are species presence/abundance rather than abundance. You can use the table() function to create a presence-absence matrix, creating a new R object that is a contingency table:

gb.pa = table(\_\_\_\_\_, \_\_\_\_\_)

**Q1. If I want the species to be columns and sites to be rows, what two column names would I enter in what order the blanks above?** (Recall that you also have to reference the original data frame name. And since the species names are long, use their abbreviations instead.)

Use the head() function to make sure the matrix looks correct.

Notice how the data are now all 1 (presence) or 0 (absence). The sum of each row thus represents the species richness for each sample. The sum of each column is the frequency of the species across all the samples. You can find each of these via the following:

rowSums(gb.pa)

colSums(gb.pa)

**Q2. Which sample has the most species (and how many species)?**

What if you were interested in determining which species was the most widespread? Write code and answer this question: **Q3. Which species is the most widespread (i.e., found in the most samples)?**

(For fun, do some of the other analyses we did today on the ground beetle data.)

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 site x species