The most basic attributes of a community are **diversity = richness + evenness**

**richness** (S) = number of species

**evenness** = relative abundance of each species (i.e., individuals/species)

We will go through all three of these components (diversity, richness, evenness) today.

Example of why using a diversity index might be useful:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Species 1** | **Species 2** | **Species 3** | **Species 4** | **Species 5** | **Row sum** |
| **Site 1** | 20 | 20 | 20 | 20 | 20 | 100 |
| **Site 2** | 96 | 1 | 1 | 1 | 1 | 100 |

Species richness (S) and total number of individuals (N) are identical for both sites:

S = 5

N = 100 individuals

But site 1 is much more even and therefore more diverse, relatively speaking. So we need something beyond S or N to denote this difference: we need a diversity index. (FYI, a diversity metric is the same as a diversity index.) I’m including the math for most of the ones I discuss below so that you can see what goes into them, but you’ll never need to do raw calculations because R will do all the calculations.

**Richness indices** – seldom do we truly know the value of S; more often, we must estimate the number of species present in a community. We can do so by either using counts of species in a single site (abundance-based methods) or using the frequency of species across many sites (incidence-based methods). Incidence-based methods are particularly useful if you have only presence/absence data or otherwise cannot make accurate counts of numbers of individuals (as might be the case for clonal species, or for species where it is difficult to keep track of numbers of individuals in massive swarms).

All of the methods below result in estimates of richness that are *usually* higher than the number of species present in your dataset: they are estimating what the “true” richness is in the community based on the assumption that your data are a representative sample. Occasionally, however, the estimate may be lower than the number you actually recorded (this is especially problematic for the Jackknife metrics discussed below).

**- Chao1 (abundance-based)** and **Chao2 (incidence-based):**

where Sobs is the observed number of species, a1 and a2 are the numbers of singletons and doubletons respectively (i.e., species represented by only 1 or 2 individuals respectively), and Q1 and Q2 are species that occur at only 1 or 2 sites respectively.

Developed by Anne Chao, these nonparametric metrics are especially useful for datasets with many species that are rare (i.e., have low abundances or occurrences). The idea behind them is that if a community is being sampled and rare species (singletons) are still being discovered, then there are likely still more rare species yet to be found; as soon as all species have been recovered at least twice (doubletons), there are likely no more species to be found and so your estimate of diversity is complete. They assume the data follow a Poisson distribution, and their estimates of diversity are very conservative (are minimum estimates of diversity). However, these metrics have been shown to be quite imprecise (Brose and Martinez 2004).

Chao and colleagues also developed another abundance-based diversity metric called **ACE (Abundance Coverage Estimator)** and another incidence-based one called **ICE (Incidence Coverage Estimator)**. I won’t go into the equations for them, as they are rather complex, but they are based on data beyond singletons and doubletons and are less subject to the problems that Chao1 and Chao2 have with consistently underestimating species richness, especially at low sample sizes (Magurran 2004).

**- Jackknife (1st order and 2nd order) (both are incidence-based):**

where Sobs is the observed number of species, N is the number of sites, and Q1 and Q2 are species that occur at only 1 or 2 sites respectively. These nonparametric metrics can both overestimate or underestimate the true diversity of a community. They perform poorly when there are lots of 0s in your data, typically resulting in a value far higher than the true richness present. For very small sample sizes, they may yield a number that is lower than the S that you actually observed in your data!

**- Bootstrap (incidence-based):**

where p is the frequency of a species (i.e., its incidence: the proportion of sites at which it occurs) and Sobs and N are as defined above. It is most useful when the number of sites sampled is large (>100). Like the Jackknife indices, it performs poorly when there are lots of 0s in your data but is less prone to overestimate richness.

Magurran (2004) summarizes the performance of these (and other) metrics, reviewing several studies that compared various metrics but finding no consensus as to which was “best.” Therefore, you should consider your sample size, the structure of your data (do you have many 0s?), and whether you have abundance or incidence data and then pick metrics accordingly. **It would be prudent to use several metrics and then use the consensus outcome.**

**Diversity** **indices** - standard benchmarks for evaluating management efficacy, disturbance, and other factors. There are *many* kinds, with different ones emphasizing different aspects (e.g. rarity over commonness). Here are some commonly used ones.

**- Shannon-Wiener Index, H’:**



where *pi* = proportion of total individuals that belong to the ith of S species

This is one of the most commonly used indices, yet it is fraught with problems! (It’s highly sensitive to number of individuals sampled, highly influenced by rarity, and can be difficult to interpret.)

- **Simpson’s Index, l**:



Ranges from 0 (high diversity) to 1 (low), representing the probability that two individuals drawn at random belong to the same species. This 0-1 range is counterintuitive (lower number for higher diversity), so the inverse is often used (inverse Simpson’s Index): D = 1/l. ALTERNATIVELY, sometimes you’ll see D as 1 – l. This latter form is not called the inverse Simpson’s Index, of course, but just like the inverse form it creates a more intuitive index where larger values indicate higher diversity; this form is what the packages *BiodiversityR* and *vegan* use.

**- Fisher’s Alpha, a:**

This index is often used because it allows you to compare communities that differ in number of individuals (i.e., it is theoretically independent of sample size, although this holds in practice only if the number of individuals sampled exceeds 1000). However, it also assumes that species’ abundances follow a log-series distribution, which is often not the case. Therefore, I am not going to cover it further.

And there are MANY others.

**Evenness indices -**

**- Pielou Evenness, J:**

J = H′/H’max = H′/ln(S)

J compares an actual computed diversity index value to its maximum possible value, H’max (which occurs when all species are equally common), which is equal to ln(S). It often uses the Shannon-Wiener Index as the diversity index value (as in the equation above) and so is sometimes called the Shannon Evenness Index. However, you can also use the Simpson’s Index (just plug in l for H’ above).

This index ranges from 0 (not even) to 1 (completely even). Smaller values indicate that there is a great deal of variation in abundances between different taxa within the community. This index is highly sensitive to sample size and is also highly sensitive to rare taxa.

And many others.

There are MANY diversity and evenness indices (see Magurran 1988, 2004—these books are great overviews on this subject). So, which should you use? I find the documentation for *vegan*’s diversity() on this to be informative and amusing. In general, I recommend using multiple indices to see whether there are overall trends that emerge. Because different diversity indices weight species’ abundances differently (i.e., treat rare species differently) and have different data assumptions, the **conclusion of whether one community is more diverse than another can depend on the diversity measure used. Thus, use of multiple indices is warranted to determine what trends exist.** Most R packages will give you output from multiple diversity indices simultaneously, making it easy for you to see whether there are consistent trends. If you get output from a particular metric that seems contrary to the other metrics, then you should investigate that more: it is possible that your data violated assumptions for that metric. If the output does not exhibit consistent trends, then that too needs further investigation.

**Scales of diversity:**

There are different scales of diversity, ranging from local to biogeographic, that are due to different processes (local diversity - result of organism responses to local conditions; regional/biogeographic diversity - result of speciation):

**alpha () diversity** = number of species within a site or habitat

-quantified by the indices discussed above

**beta () diversity** = turnover of diversity between habitats, representing the extent to which species composition varies by habitat

-obviously influenced by scale ( decreases as area sampled increases [in space or time])

where is average alpha diversity; relatively higher values indicate the fewer shared species there are across habitats in the landscape (i.e., low community similarity) whereas a value of 0 would indicate total overlap across habitats (no turnover of species)

-can also be quantified by similarity indices

**gamma (g) diversity** = the number of species found within a region (sometimes called “landscape diversity”)

**epsilon (e) diversity** = diversity of a biogeographic province or biome

**Exercises:**

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

*BiodiversityR*

*labdsv*

*lattice* – this is a package we haven’t used before, so install it first

*MASS*

*MVA*

*optpart*

*picante*

*plyr* – this is a package we haven’t used before, so install it first

*stats*

*vegan*

From the SAEC\_data folder 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.)

As you can see, it’s a generic site x species matrix, with only the first column (the sites) being named and grouped as a group (“Sites”). This isn’t exactly ideal but is quite common. (Compare this dataset with the tidier Ground\_beetles\_habitat.csv, for example.)

*Species richness:*

To calculate species richness for each site, we can use the apply() function, as we did in the site x species lesson:

divS <- apply(div[,-1]>0,1,sum)

In this line, the [,-1] tells R that the first column should be ignored and not treated as data (since it’s a list of site names). The rest of the argument tells R that, for each row, it will count the number of cells that have non-zero values (i.e., species presence). The result is a vector of species richness at each site (named divS):

[1] 3 4 4 2 4 3 3 3 3 2

To pair this list with each site’s name:

div$Sites <- as.factor(div$Sites)

levels(div$Sites)

#this yields a list of the site names:

#[1] "Site1" "Site10" "Site2" "Site3" "Site4" "Site5" "Site6" "Site7" "Site8" "Site9"

The as.factor() argument is needed to convert numeric entries to categorical ones. If you didn’t do this, the levels() line would yield a result of NULL. HOWEVER, variables treated as factors in R are arranged alphabetically, meaning that Site 10 would come after Site 1 because that is in alphabetical order. R doesn’t know how to put an alphanumeric code (letters and numbers) into numerical order if the letters come first, as in this case. So that means that it has the species diversity values in order by rows but applies the tally in alphabetical order, meaning that there is a frameshift error. The levels (Site1 to Site 10) assigned as factors are being arranged incorrectly when trying to extract the levels. To correct this:

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

Site\_names

levels(Site\_names)

names(divS) <- levels(Site\_names)

divS

#yields a list of species richness by site in correct order

**Which site has the highest richness?**

*Richness indices – incidence-based:*

Read in GBbiol.csv as an object named gb.biol, with row.names = 1 and GBsite.csv as gb.site with row.names = 1. Recall from a previous lesson that these two datasets are a site x species matrix and a site x environment data frame, respectively. The GBbiol.csv is a presence-absence matrix, so we can only use incidence-based indices on it (namely, Chao2, Jackknife1, Jackknife2, and Bootstrap).

*vegan*'s specpool() gives various estimates of species richness for the entire pool of sites (i.e., all sites combined; recall that these data were collected at 18 sites belonging to 3 habitat types):

specpool(gb.biol)

Species chao chao.se jack1 jack1.se jack2 boot boot.se n

All 48 53.03704 4.685246 55.55556 3.832931 57.64706 51.81422 3.021546 18

The entries with .se indicate the standard error for that metric. Recall that since GBbiol.csv was presence/absence data, the Chao metric is the Chao2.

You can see that even though the data had 48 species present, each of the four indices indicated that there were more species (from 51.8 to 57.6) likely present that were not sampled.

We can also examine richness across the three habitat types by using both GBbiol.csv and GBsite.csv and pooling by habitat type:

gb.sp <- specpool(gb.biol, pool = gb.site$Habitat)

**For which habitat type was richness most underestimated?**

If you want to examine richness in a particular habitat type (e.g. grass):

gb.sp["Grass",]

If you want to obtain a particular metric for all habitat types:

gb.sp[,c("chao", "chao.se")]

To create and plot confidence intervals of the indices, we can make use of something similar to creation of the species accumulation curves we did last week. The poolaccum() function creates species accumulation curves for estimates of total species richness based on iterative sampling:

(gb.sp = poolaccum(gb.biol))

plot(gb.sp)

The resulting plots are for the five columns of output of the specpool() function—your recorded S, Chao, Jackknife1, Jackknife2, and Bootstrap—with 95% confidence intervals based on sample size (i.e., number of sites sampled). From these plots, you can see which indices behave weirdly (looking at you, Chao2…) or consistently overestimate or underestimate.

*Richness indices – abundance-based:*

Read in Ground\_beetles\_abundance.csv as an R object named GBA, with row.names = 1.

*vegan*'s estimateR() function is used for abundance data:

GB.sp = estimateR(GBA)

#To view the results:

GB.sp

(Because everything in R is case-sensitive, GB.sp is a different object from gb.sp that we had earlier today!)

The result is a matrix that has the Chao1 and ACE indices for each of the 18 sites.

To create and plot confidence intervals of abundance-based indices, we use estaccumR():

GB.sp2 = estaccumR(GBA)

GB.sp2

plot(GB.sp2, col="black",

strip=function(..., bg) strip.default(..., bg="gray90"), display=c("chao", "ace"))

Notice how both Chao1 and ACE have wider confidence intervals at lower sample sizes (as might be expected), meaning that estimates of richness from low numbers of samples may be inaccurate.

*Diversity indices:*

Many R packages can be used. For example, here’s how to calculate the Shannon-Wiener Index in *vegan* for the div dataset:

diversity(div[-1], MARGIN = 2, index="shannon")

The MARGIN argument within diversity() tells R to use frequencies of species. If you already have a presence/absence matrix, you don’t need this.

The [-1] tells R to ignore the first column, which is categorical (site names). If you don’t have a column like that, you can ignore that as well and just say:

diversity(*yourRobject’sname*, index=”shannon")

But because of the structure of div\_data.csv, we have to give explicit instructions to R.

Alternatively, you could also use:

diversity(div[2:11], MARGIN = 2, index=”shannon")

which tells R to only use columns 2-11 in the calculation. The result is the same as before.

And here’s how to do it in R using *plyr* to manipulate the data frame:

ddply(div,~Sites,function(x) + data.frame(SHANNON=diversity(x[-1], index="shannon")))

They yield the same answers. (Notice how the sites are listed: R doesn’t know 10 comes after 9, it thinks 10 is the next item after 1.)

**Which site is the most diverse?**

*vegan* has options for “shannon", “simpson”, “invsimpson”, and “fisher.alpha”.

R has options for “shannon", “brillouin” (not discussed, as it’s rare; it is similar to the Shannon-Wiener Index and is used when samples were not collected randomly), and “simpson”. It is easiest to calculate them when you arrange the data frame with *plyr*.

There are other packages that calculate these and other diversity indices as well (*hillR*, *diverse*, etc.). *BiodiversityR* has some nice options. To explore, we’ll use the GBA object you created a few pages ago and its paired gb.site object to compare diversity across habitat types, using *BiodiversityR*’s diversitycomp() function:

diversitycomp(GBA, y = gb.site, factor1 = "Habitat", index = "Shannon")

But you’ll get an error message!

Error in diversitycomp(GBA, y = gb.site, factor1 = "Habitat", index = "Shannon") :

specified factor1 'Habitat' is not a factor

In addition: Warning message:

In if ((method %in% METHOD) == F) { :

the condition has length > 1 and only the first element will be used

Remember from the site x environment lesson that you have to tell R when variables exist as strings so that it can treat them as variables and quantify them and not just treat them as words. So re-read in GBsite.csv and re-do:

gb.site2 <- read.csv("GBsite.csv", row.names = 1, stringsAsFactors = TRUE)

diversitycomp(GBA, y = gb.site2, factor1 = "Habitat", index = "Shannon")

**Which site is most diverse?**

If you want to examine each site’s alpha diversity, *BiodiversityR*’s diversityresult() command is used:

diversityresult(GBA, y=gb.site2, factor="Habitat",

level="Edge", index="Shannon", method=c("each site"))

Note: if you want Simpson’s Index, *BiodiversityR* and *vegan* already convert it to 1 – l.

Since we had 6 replicate sites in each of the three habitats, we can create a boxplot showing average diversity:

GBA\_Shannon <- diversityresult(GBA, index="Shannon", method=c("each site"))

GBA\_Shannon$Habitat = gb.site2$Habitat

str(GBA\_Shannon)

#This line tells you what words to use in the code below and in #what order

boxplot(Shannon~Habitat, data=GBA\_Shannon, ylab="Shannon’s Diversity Index", xlab="Habitat (6 sites each)")

(There are also other plotting options, of course, but a lot of them require you to calculate each component, such as error bars or 95% CI, manually. We will not do so here.)

(If you have replicates, you can use standard statistics to compare average diversity [or richness, or evenness]. But even if you do not, you can still compare diversity between two sites using two approaches:

1. You can use a modified t-test…but here’s the bad news: you have to program each mathematical step [formulae to calculate the t value, sample variances, and degrees of freedom] in R because there is no package that I know of that does so for you. Furthermore, each index uses different formulae for each of these. And then you have to compare the calculated values to a table of critical t values.
2. Because t-tests are subject to the usual parametric assumptions of normality and homogeneity of variances, a **bootstrapping** approach is generally preferred. This is a process that creates a distribution from your data by iteratively sampling with replacement. What that means is that you determine how many individuals are in your two samples, randomly select that number of individuals [with replacement each time you draw a new individua]) from your original data so that you have a new sample with the same number of individuals, do that many times [1000 is standard], calculate the diversity index of interest from the new samples, and then assess the mean and variance of the new [bootstrapped] samples. R can do this in a package called *boot* with the command boot().

All of this is to say that these steps are tedious, so we won’t cover them further here. For more information, I recommend Gardener 2014.)

*Evenness indices:*

Let’s once again start with div.

As above, we can use *vegan* or R with *plyr*:

S <- apply(div[,-1]>0,1,sum)

J <- diversity(div[-1], index="shannon")/log(S)

OR

ddply(div,~Sites,function(x) + data.frame(SHANNON=diversity(x[-1], index="shannon")/log(sum(x[-1]>0))))

**Important note! In R, log() is used for natural logarithms**, which we normally denote as ln(). To calculate a base-10 logarithm in R, use log10().

**Which site has the greatest evenness?**

Now let’s switch to GBA and work with *BiodiversityR*, which calls Pielou’s J “Jevenness”:

diversitycomp(GBA, y=gb.site2, factor1="Habitat", index="Jevenness")

diversityresult(GBA, y=gb.site2, factor="Habitat", level="Edge",

index="Jevenness",method="each site")

*Scales of diversity:*

To examine beta diversity as the ratio of regional diversity to average within-habitat diversity, we’ll use the gb.biol and gb.site objects you created earlier today.

First, calculate gamma diversity as the total number of species across the three habitat types. That uses *vegan*’s specpool() function:

gamma <- specpool(gb.biol)

#To just extract the number of species and not the indices calculated by specpool():

gamma$Species

To calculate the mean alpha diversity, you can use the tapply() function:

alpha <- tapply(specnumber(gb.biol), gb.site$Habitat, FUN = mean)

alpha

We already knew that Grass had the most species (and, indeed, that this number was likely underestimated) and was most diverse. Now you can calculate beta diversity:

beta <- (gamma$Species/alpha) - 1

beta

The result is three numbers, one for each habitat type. These numbers indicate that the Wood habitat has the fewest unique species (highest beta diversity indicates lowest similarity in species composition); indeed, if you examine the raw data, you’ll see that every species in Wood was found in at least one of the other two habitat types.

*vegan* also has 24 other indices of beta diversity with its betadiver() function; each is known by an abbreviation. To find out the formulas for each, use betadiver(help="TRUE"):

1 "w" = (b+c)/(2\*a+b+c)

2 "-1" = (b+c)/(2\*a+b+c)

3 "c" = (b+c)/2

4 "wb" = b+c

5 "r" = 2\*b\*c/((a+b+c)^2-2\*b\*c)

6 "I" = log(2\*a+b+c) - 2\*a\*log(2)/(2\*a+b+c) - ((a+b)\*log(a+b) +

(a+c)\*log(a+c)) / (2\*a+b+c)

7 "e" = exp(log(2\*a+b+c) - 2\*a\*log(2)/(2\*a+b+c) - ((a+b)\*log(a+b) +

(a+c)\*log(a+c)) / (2\*a+b+c))-1

8 "t" = (b+c)/(2\*a+b+c)

9 "me" = (b+c)/(2\*a+b+c)

10 "j" = a/(a+b+c)

11 "sor" = 2\*a/(2\*a+b+c)

12 "m" = (2\*a+b+c)\*(b+c)/(a+b+c)

13 "-2" = pmin(b,c)/(pmax(b,c)+a)

14 "co" = (a\*c+a\*b+2\*b\*c)/(2\*(a+b)\*(a+c))

15 "cc" = (b+c)/(a+b+c)

16 "g" = (b+c)/(a+b+c)

17 "-3" = pmin(b,c)/(a+b+c)

18 "l" = (b+c)/2

19 "19" = 2\*(b\*c+1)/(a+b+c)/(a+b+c-1)

20 "hk" = (b+c)/(2\*a+b+c)

21 "rlb" = a/(a+c)

22 "sim" = pmin(b,c)/(pmin(b,c)+a)

23 "gl" = 2\*abs(b-c)/(2\*a+b+c)

24 "z" = (log(2)-log(2\*a+b+c)+log(a+b+c))/log(2)

These are kinds of **similarity indices (which is a bit of a misnomer, since most actually quantify dissimilarity)**, and all of the ones above are used for presence/absence data. They come from Table 1 in Koleff et al. (2003). In the formulae above, a stands for the number of species shared between two communities being compared, b = the number of species unique to community 1, and c = the number of species unique to community 2.

Whittaker’s “w” is perhaps the most widely used index of beta diversity, followed closely by Jaccard’s “j” and Sørensen’s “sor”. Koleff et al. (2003), however, recommend “sim” over those other indices.

Each index does pairwise comparisons. If we wanted to examine similarity across our three habitat types, that’s three comparisons that have to be made (Edge vs. Wood, Edge vs. Grass, and Wood vs. Grass). The result of betadiver() is a matrix with these three comparisons, with the output expressed as dissimilarity:

gbt <- rowsum(gb.biol, group = gb.site$Habitat)

betadiver(gbt, method="w")

Edge Grass

Grass 0.46875

Wood 0.45000 0.64000

Based on Whittaker’s w, Grass is more dissimilar to Wood than it is to Edge (or expressed a different way, Grass is more similar to Edge than it is to Wood).

But are those differences statistically significant? Because the output of betadiver() is a matrix of dissimilarity values, you can do a kind of nonparametric ANOVA on it called an ANOSIM (analysis of similarity). You start by making an overall dissimilarity matrix for all 18 sites and then do the ANOSIM on that. (Briefly, the steps in the analysis are: 1. You have your matrix of dissimilarity scores for every pair of sites. 2. ANOSIM converts the dissimilarities to ranks. 3. It calculates the R statistic as the ratio between dissimilarities between sites within a group and the dissimilarities between sites that are in different groups. The closer this value is to 1, the more that the sites within a group are similar to each other and dissimilar to sites in other groups. 4. The significance of the R-statistic is determined by permuting the membership of sites in groups.

gb.b <- betadiver(gb.biol, method = "w")

gb.ano <- anosim(gb.b, gb.site$Habitat)

gb.ano

Call:

anosim(x = gb.b, grouping = gb.site$Habitat)

Dissimilarity: beta.w

ANOSIM statistic R: 0.99877

Significance: 0.001

Permutation: free

Number of permutations: 999

In our example, the R statistic is very high, and the p-value is significant. So there are significant differences in beta diversity by habitat type.

(There are also similarity indices for use with abundance data, such as Bray-Curtis or Morisita’s indices. And there are MANY other ways of examining beta diversity; Gardener 2014 does a pretty thorough job of covering many of them. Indeed, most of his book is spent covering various aspects of beta diversity, including many types of indices and graphing options.)

Use Zhou et al. 2013 as real-world application of Shannon index

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>.

**References:**

Brose, U., and N.D. Martinez. 2004. Estimating the richness of species with variable mobility. Oikos 105:292-300.

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

Koleff, P., K.J. Gaston, and J.J. Lennon. 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 72:367-382.

Magurran, A.E. 1988. *Ecological Diversity and Its Measurement*. Princeton University Press, Princeton, NJ.

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

**Assignment:** due 0800 Monday, 8 March

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.

For your assignment, you will be using two new datasets, butterfly\_pres\_abs.csv and butterfly\_sites.csv. These are sister datasets (one is site x species, the other is site x environment, both for the same sites) on adult butterfly occurrence at 15 sites in Ohio in three habitat types: fens, bogs, and meadows. (Fens and bogs are types of wetlands; fens are fed by groundwater whereas bogs are fed by rainwater. Bogs are usually more acidic and thus have more acid-tolerant species present. Meadows are non-wetland sites.) There were five replicate sites surveyed in each of the three habitat types. Adults were surveyed over the course of a year at these sites using a standard survey method called a Pollard walk (a sort of three-dimensional transect that is traditional in butterfly surveys). In butterfly\_pres\_abs.csv, the species are listed in alphabetical order; 1 = presence, 0 = absence.

Using the skills you learned today, provide a quantitative assessment of butterfly diversity patterns by habitat type. Specifically, answer the following questions:

**Q1. How many species of butterflies were found?**

**Q2. How many species were found in each habitat type? Which habitat type had the most species?**

**Q3. Which richness indicator gives the “best” approximation of the observed overall species richness?**

**Q4. Which habitat type was the most diverse? Which was the least diverse? *Be thorough.* Provide graphical support for your answers.**

**Q5. Which habitat type had individuals that were the most evenly distributed across species?**

**Q6. What was the landscape-scaled species richness of butterflies?**

**Q7. Which habitat type had highest alpha diversity?**

**Q8. Which sites were most similar in species composition? *Be thorough*. Provide statistical support for your answers.**

**BONUS:** **Were any sites outliers with respect to butterfly richness?**

Make an RMarkdown Word file of your work and turn that in. Be sure to include your answers to the questions asked!