7. Worksheet: Diversity Synthesis

Jonathan Enriquez Madrid; Z620: Quantitative Biodiversity, Indiana University

15 February, 2023

OVERVIEW

In this worksheet, you will conduct exercises that reinforce fundamental concepts of biodiversity. Specifically, you will construct a a site-by-species matrix by sampling confectionery taxa. With this primary data structure, you will then answer questions and generate figures using tools from previous weeks, along with wrangling techniques that we learned about in class.

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) to your name.
2. Complete as much of the worksheet as possible during class.
3. Refer to previous handouts to help with developing of questions and writing of code.
4. Answer questions in the worksheet. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For the assignment portion of the worksheet, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file `7.DiversitySynthesis_Worskheet.Rmd` and the PDF output of `Knitr` (`DiversitySynthesis_Worskheet.pdf`).

CONFECTIONARY EXERCISE GOALS

We will construct a site-by-species matrix using confectionery taxa (i.e, gummies). The instructors have created distinct **sources communities** that vary in the composition of gummy taxa with even and uneven communities. It might be fun to consider them as distinct geographical regions experiencing different environmental regimes, or different experimental units under different treatments. Each student will sample a source community and then use a taxonomic key to identify gummies and their abundances.

In the end, students will use the site-by-species matrix to:

1) explore their sampling efforts and their effects on species richness using **coverage** and **rarefaction** concept,

2) measure **alpha diversity** for each sub-sample collated from data with their peers from the same source community,

3) examine **beta diversity** between each source community using the data generated across each source community, and

4) use **data wrangling** tools they have learned during the class to accomplish the above goals.

SAMPLING PROTOCOL TO CONSTRUCT A SITE-BY-SPECIES MATRIX

1. Instructors will assign you to sample confectionery taxa from one of the two designated source community bucket (A and B).
2. After randomly sampling one unit (imagine as an equal biomass) from the source community, each student will count the total number of individuals (N), identify the taxa using the species key and quantify the abundance of each taxon.
3. Work with other students in your group to assemble data into a site-by-species matrix on the white board. One person needs to create a .csv or .txt file and share your group's site-by-species matrix with the class using GitHub. Make sure that you include a sample identifier (student name) and what community you sampled from.

GROUP BRAINSTORM

In smaller groups, take 15 minutes to brainstorm questions, code, statistical tests, and "fantasy figures" using the site-by-species matrix the class generated.

1. Using this data, explore how well your sampling effort was. You can use rarefaction and coverage tools you have learned earlier.
2. Investigate alpha diversity based on the methods you have learned in the rest of the handout and accompanying worksheet. For example, you can measure richness, Shannon diversity and Simpson index. You can also convert them to effective number of species using the Hill numbers concept.
3. Measure beta diversity using ordination and multivariate statistical methods. For example, you can create a PCoA plot, based on Bray-Curtis dissimilarity, of sites and communities using different shape and color codes. Use Permanova to test if there are differences between communities.

DATA ANALYSIS

1) Sampling coverage and rarefaction curves

Question 1: Using this data, explore how well your sampling effort was. Compare your sampling efforts with other groups. Do you think that your samples cover the actual diversity found in each source community? You can use rarefaction and coverage tools you have learned earlier.

Answer 1: Use the space below to generate a rarefaction curve/sample coverage based on the data we collected in class for each community. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

Answer 1: Based on the rarefaction curve, community A and community B seem to differ in how many number of species each holds. The curves for community A (lines 1-4) are higher on the graph than the curves for community B (lines 5-8) when looking at certain sample sizes, meaning that community A seems to hold more species than community B. Going deeper, and looking at the sites within the communities, we see that the sites also differ in the number of species they hold. Within community A, site 1 has the most number of species, followed by site 2, followed by site 3, and finally site 4 having the lowest number of species. Within community B, site 8 has the greatest number of species, followed by both site 6 and 7

(which seem to have around the same number of species), and finally site 5. Again, this is dependent on the sample size that one is looking at on the x-axis. As for the number of species in sites across communities, the site with the highest number of species is site 1, followed by site 2, followed by site 3 , followed by site 8, followed by sites 6 and 7, followed by site 4, and then finally site 5. This is a general observation from the rarefaction curves, without knowing if any of the differences are significant. As for coverage, the sites in each community seem to be sampled well. The lowest coverage in community A was site 2 at 0.72, while the highest coverage was at site 4 at 0.95. As for community B, the lowest coverage was for site 6 at 0.85, and the highest coverage was at site 5 at 0.92.

Clear environment, set working directory, & load packages

```
rm(list = ls())
getwd()
```

```
## [1] "C:/Users/jonat/GitHub/QB2023_Enriquez_Madrid/2.Worksheets/7.DiversitySynthesis"
```

```
setwd("C:/Users/jonat/GitHub/QB2023_Enriquez_Madrid/2.Worksheets/7.DiversitySynthesis")
package.list <- c('vegan','ggplot2','dplyr')
for (package in package.list) {
  if (!require(package, character.only = TRUE, quietly = TRUE)) {
    install.packages(package)
  }
  library(c(package), character.only = TRUE)
}
```

```
## This is vegan 2.6-4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Import data, subset data, & change to numeric

```
dat <- read.csv(file = "C:\\Users\\jonat\\OneDrive\\Documents\\Quantitative Biodiversity\\QB Data Wrang

dat1 <- dat[1:8,]#Get rid of columns with N/A
str(dat1)#Look at structure of data
```

```
## 'data.frame':    8 obs. of  32 variables:
##  $ Community: chr  "A" "A" "A" "A" ...
##  $ Name     : chr  "Erica" "Lauren" "Atalanta" "Anna" ...
##  $ SP1      : int  1 0 2 0 11 9 7 11
##  $ SP2      : int  2 4 0 4 3 10 16 8
##  $ SP3      : int  1 3 1 0 2 5 2 6
```

```
##  $ SP4      : int  2 0 0 4 1 10 1 3
##  $ SP5      : int  1 3 0 5 4 3 4 4
##  $ SP6      : int  2 2 6 1 2 1 2 1
##  $ SP7      : int  2 1 0 4 5 0 0 2
##  $ SP8      : int  2 2 0 0 0 1 5 4
##  $ SP9      : int  1 1 2 1 4 3 4 4
##  $ SP10     : int  1 1 1 0 5 0 2 5
##  $ SP11     : int  1 5 1 3 4 1 1 1
##  $ SP12     : int  1 4 2 3 4 1 4 3
##  $ SP13     : int  2 1 1 4 4 0 1 3
##  $ SP14     : int  5 3 3 2 4 4 1 3
##  $ SP15     : int  1 1 1 0 2 4 0 2
##  $ SP16     : int  5 1 1 4 2 1 2 3
##  $ SP17     : int  1 1 2 0 4 4 3 1
##  $ SP18     : int  0 0 0 2 1 4 3 2
##  $ SP19     : int  2 1 1 2 0 0 0 2
##  $ SP20     : int  2 0 2 2 1 1 6 0
##  $ SP21     : int  1 3 1 1 0 3 1 3
##  $ SP22     : int  1 1 2 0 0 2 1 2
##  $ SP23     : int  4 1 0 2 0 1 0 1
##  $ SP24     : int  2 1 2 4 1 0 1 1
##  $ SP25     : int  0 1 2 3 0 1 1 0
##  $ SP26     : int  3 2 5 2 0 1 3 0
##  $ SP27     : int  2 2 3 2 0 0 2 0
##  $ SP28     : int  4 2 2 2 0 1 0 1
##  $ SP29     : int  5 0 1 0 0 1 0 1
##  $ SP30     : int  2 1 2 4 1 2 3 0
```

```
dat1_species_only <- dat1[,3:32]#get rid of character columns "Community" and "Name"

dat1_species_only [1:30] = lapply(dat1_species_only[1:30], FUN = function(y){as.numeric(y)})#Convert sp

str(dat1_species_only)
```

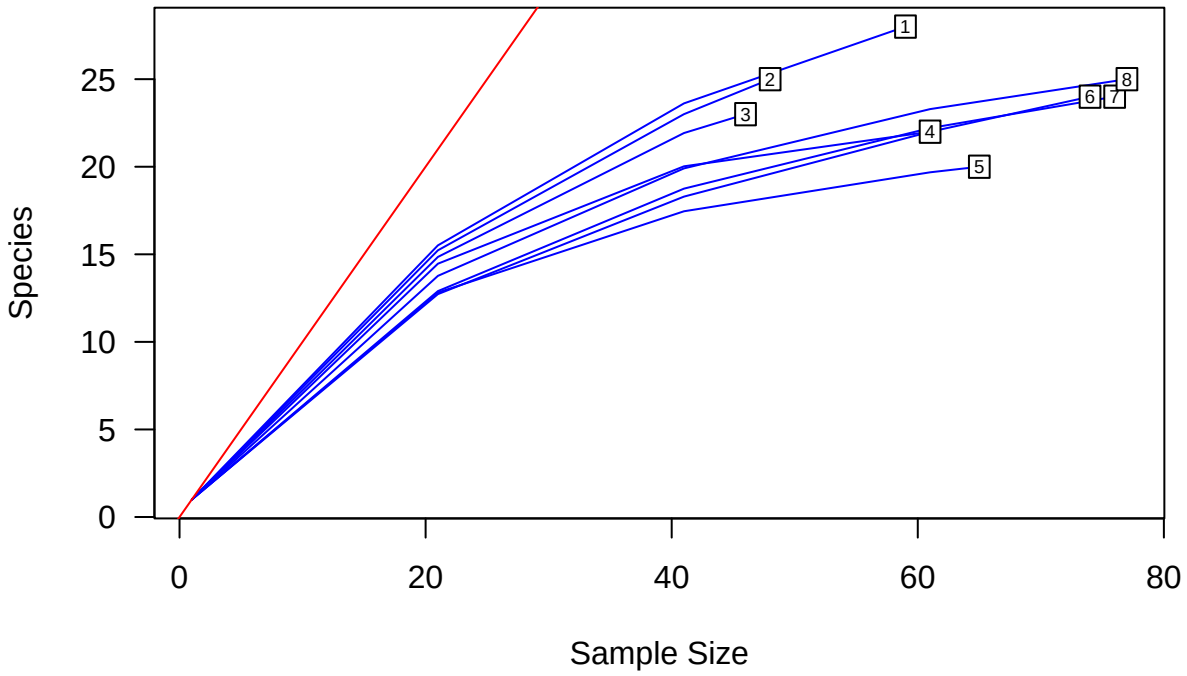
```
## 'data.frame':    8 obs. of  30 variables:
##  $ SP1 : num  1 0 2 0 11 9 7 11
##  $ SP2 : num  2 4 0 4 3 10 16 8
##  $ SP3 : num  1 3 1 0 2 5 2 6
##  $ SP4 : num  2 0 0 4 1 10 1 3
##  $ SP5 : num  1 3 0 5 4 3 4 4
##  $ SP6 : num  2 2 6 1 2 1 2 1
##  $ SP7 : num  2 1 0 4 5 0 0 2
##  $ SP8 : num  2 2 0 0 0 1 5 4
##  $ SP9 : num  1 1 2 1 4 3 4 4
##  $ SP10: num  1 1 1 0 5 0 2 5
##  $ SP11: num  1 5 1 3 4 1 1 1
##  $ SP12: num  1 4 2 3 4 1 4 3
##  $ SP13: num  2 1 1 4 4 0 1 3
##  $ SP14: num  5 3 3 2 4 4 1 3
##  $ SP15: num  1 1 1 0 2 4 0 2
##  $ SP16: num  5 1 1 4 2 1 2 3
##  $ SP17: num  1 1 2 0 4 4 3 1
##  $ SP18: num  0 0 0 2 1 4 3 2
##  $ SP19: num  2 1 1 2 0 0 0 2
```

```
##  $ SP20: num  2 0 2 2 1 1 6 0
##  $ SP21: num  1 3 1 1 0 3 1 3
##  $ SP22: num  1 1 2 0 0 2 1 2
##  $ SP23: num  4 1 0 2 0 1 0 1
##  $ SP24: num  2 1 2 4 1 0 1 1
##  $ SP25: num  0 1 2 3 0 1 1 0
##  $ SP26: num  3 2 5 2 0 1 3 0
##  $ SP27: num  2 2 3 2 0 0 2 0
##  $ SP28: num  4 2 2 2 0 1 0 1
##  $ SP29: num  5 0 1 0 0 1 0 1
##  $ SP30: num  2 1 2 4 1 2 3 0
```

```
dat_A <- dat1_species_only[1:4,]#Subset of data, community A
dat_B <- dat1_species_only[5:8,]#Subset of data, community B
```

Rarefaction curve

```
min.N <- min(rowSums(dat1_species_only))

dat.Rarefaction <- rarefy(x = dat1_species_only, sample = min.N, se = TRUE)#Rarefying species numbers

rarecurve(x = dat1_species_only, step = 20, col = "blue", cex = 0.6, las = 1)
abline(0, 1, col = 'red')
text(1500, 1500, "1:1", pos = 2, col = 'red')#Rarefaction curve/plot
```



Checking sample covergae

```r
Coverage <- function(x = ""){
  1 - (rowSums(x == 1) / rowSums(x))
}#Code for checking sample coverage

Coverage(dat_A)#Coverage of community A
```

```
##         1         2         3         4
## 0.8135593 0.7291667 0.8043478 0.9508197
```

```r
Coverage(dat_B)#Coverage of community B
```

```
##         5         6         7         8
## 0.9230769 0.8513514 0.8947368 0.9090909
```

2) Alpha diversity

Question 2: Compare alpha diversity measures within sites and among communities. You can calculate and plot richness, Shannon diversity, and Simpson index. You can also convert these indices to effective number of species using the Hill numbers concept by generating a diversity profile, which will make comparisons easier across sites.

What is the variation among the samples in your group and between the communities of other groups for the alpha diversity indices? Generate a hypothesis around the diversity metrics you chose and test your hypothesis. Interpret your findings.

Answer 2a - Analysis: Use the space below for code that is being used to analyze your data and test your hypotheses on your chosen alpha diversity tool. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

Answer 2a: The alpha diversity measure that I used was Shannon's diversity. I am testing the null hypothesis in where there should be no difference between sites in terms of their species richness and evenness. I checked Shannon's diversity for both communities A and B, and for each site within each community. A higher Shannon's diversity index indicates that there is a higher chance of finding a new species, and thus having greater species richness. Richness across sites, across communities, and within communities differ. In community A, site 1 has the largest Shannon's diversity index, followed by site 2, then site 4, and then finally site 3. For community B, site 8 has the largest Shannon's diversity index, followed by site 7, then site 6, and finally site 5. This shows us that the different sites have differences in their alpha diversity (species richness) and rejects the null hypothesis of there being no difference in richness between sites. However, these findings are just descriptive and we don't know if the differences are significant.

Alpha diversity analysis

```r
diversity(dat_A, index = "shannon")#Shannon diversity for community A
```

```
##        1        2        3        4
## 3.166063 3.053443 2.975329 2.998721
```

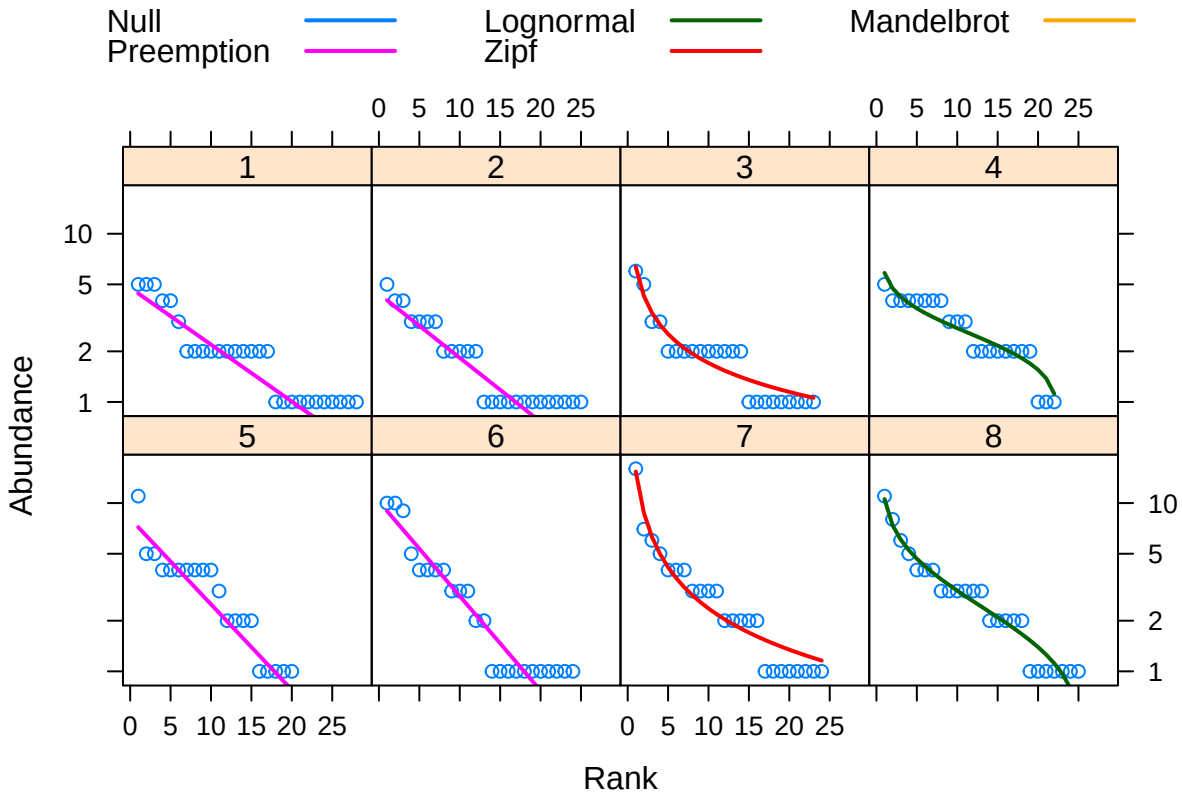
```r
diversity(dat_B, index = "shannon")#Shannon diversity for community B
```

```
##        5        6        7        8
## 2.787795 2.834947 2.836906 2.978193
```

Answer 2b - Plot: With your analysis, create one (and only one, although it can have multiple panels) *publication-quality* figure.

```
SEM <- function(x) {
  return(sd(x)/sqrt(length(x)))
}#Code for finding the SEM

com_div_all <- t(cbind.data.frame(dat_A, dat_B))#Combining community A and community B data

View(com_div_all)

com_div_sum_apply <- t(apply(com_div_all, 1, FUN = function(x)
{ c(mean = mean(x), SEM = sd(x)/sqrt(length(x)))}))#Finding mean and SEM of both community A & communit

RACresults <- radfit(dat1_species_only)#Running Rank abundance curve (RAC). RAC looks at
plot.new()
```

```
plot(RACresults, las = 1, cex.lab = 1.4, cex.axis = 1.25)#Plotting RAC
```



Answer 2c - Interpret results: Write an informative yet succinct (~5 sentences) caption that creates a "stand-alone" figure. Take a peek at figures and figure captions in a paper published in your favorite journal for inspiration.

Answer 2c: The figure presented shows the rank abundance curves (RAC) of each site within each community. Each RAC plot in the figure coincides with a site. Plot 1 is for site 1, plot 2 is for site 2, plot 3 is for site 3 and so on. The RAC plots depict species richness and evenness at each site. Richness is depicted by the y-axis, while evenness is depicted by the curve, where a steeper curve indicates less evenness. By looking across sites, we see that the RAC plots differ between sites, meaning that sites differ in their evenness and richness. Sites 1, 2, 5, and 6 are depicted by a preemtion model, site 3 and 7 depicted by a Zipf model, and sites 4 and 8 depicted by a lognormal model. Sites 1, 2, 5, and 6 are similar to each other, sites 3 and 7 are similar to each other, and sites 4 and 8 are similar to each other.

3) Beta diversity

Question 3: Measure beta diversity using ordination and multivariate statistics methods. You can create a PCoA plot, based on Bray-Curtis dissimilarity, of sites and communities using different shape and color codes. Then, you can use a Permanova to test if there are differences between communities. Generate a hypothesis around your chosen analysis and test your hypothesis. Interpret your findings.

Can you detect compositional differences between each source community sampled?

Answer 3a - Analysis: Use the space below for code that is being used to analyze your data and test your hypotheses on your chosen beta diversity tool. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

Answer 3a: I will test the null hypothesis that there is no difference in composition between each site and each community. I will use the Bray-Curtis dissimilarity metric to test if there are differences in composition

between sites and communities. The Bray-Curtis dissimilarity analysis shows that the sites do differ in their composition. The heatmap also shows that sites differ in their composition to one another. The PCoA plot shows that sites differ in their composition based on the community they are from as sites 1, 2, 3, and 4 (which make up Community A) are clustered together, and sites 5, 6, 7, and 8 (which make up community B) are clustered together. In addition, a Permanova shows that community has a significant affect on site, showing support for sites differing in their composition based on the community they are from.

Beta diversity analysis

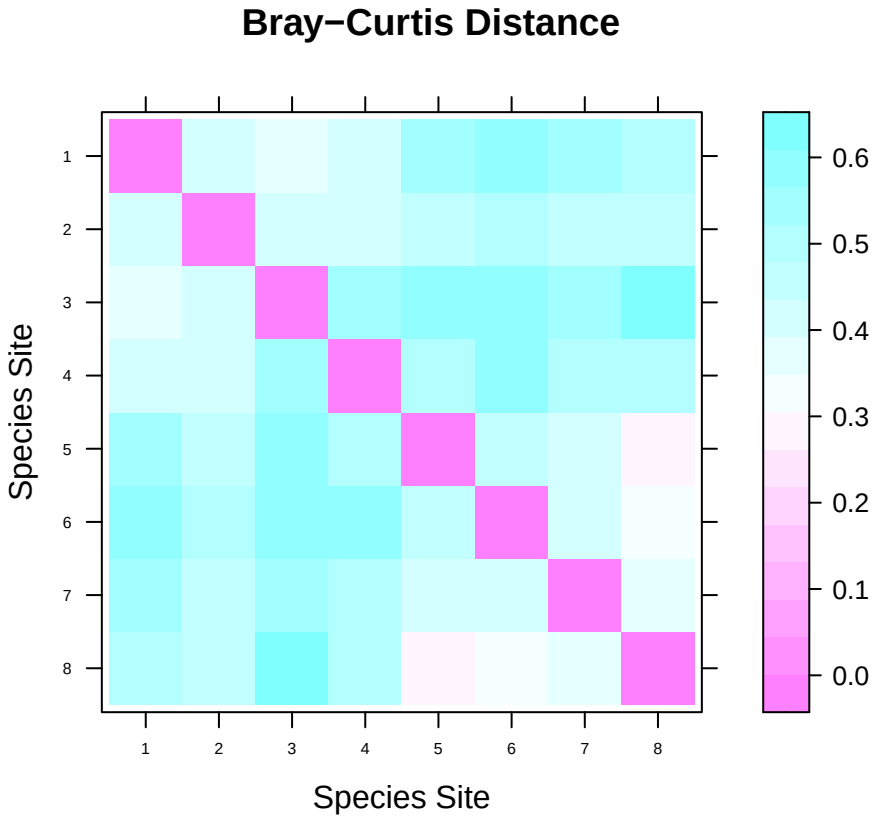
```
dat1_species_only.b <- vegdist(dat1_species_only, method = "bray")#Runs Bray-Curtis Dissimilarity (abun

library(vegan)
community <- c(rep("A", 4), rep("B", 4))#Creates 'factors' vector
adonis2(dat1_species_only ~ community, method = "bray", permutations = 999)#Runs Permanova on species b
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = dat1_species_only ~ community, permutations = 999, method = "bray")
##           Df SumOfSqs      R2      F Pr(>F)
## community  1  0.31895 0.39151 3.8605  0.022 *
## Residual   6  0.49571 0.60849
## Total      7  0.81466 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer 3b - Plot: With your analysis, create one (and only one, although it can have multiple panels) *publication-quality* figure.

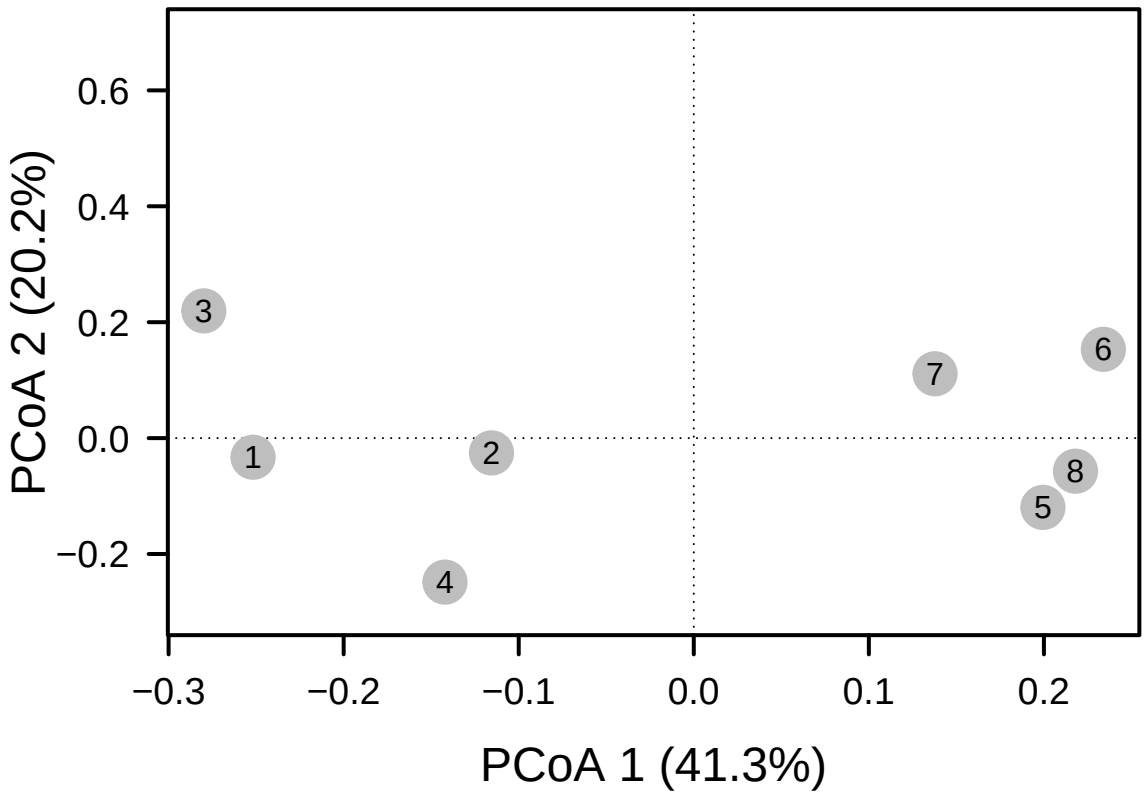
Heat map

```
library(lattice)
order <- rev(attr(dat1_species_only.b, "Labels"))#defines order of sites
levelplot(as.matrix(dat1_species_only.b) [, order], aspect = "iso", col.regions = ,
          xlab = "Species Site", ylab = "Species Site", scales = list(cex = 0.5),
          main = "Bray-Curtis Distance")#Code for making heat map
```



PCoA plot

```
dat1_species_only.pcoa <- cmdscale(dat1_species_only.b, eig = TRUE, k = 3)#Conducts PCoA

explainvar1 <- round(dat1_species_only.pcoa$eig[1] / sum(dat1_species_only.pcoa$eig), 3) * 100
explainvar2 <- round(dat1_species_only.pcoa$eig[2] / sum(dat1_species_only.pcoa$eig), 3) * 100
explainvar3 <- round(dat1_species_only.pcoa$eig[3] / sum(dat1_species_only.pcoa$eig), 3) *100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)#Quantifies the percent variation in the data set

par(mar = c(5, 5, 1, 2) + 0.1)#Defines plot parameters

plot(dat1_species_only.pcoa$points[ ,1], dat1_species_only.pcoa$points[ ,2], ylim = c(-0.3, 0.7),
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
cex.axis = 1.2, axes = FALSE)#Initiates plot

axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)#Adds axes

points(dat1_species_only.pcoa$points[ ,1], dat1_species_only.pcoa$points[ ,2], pch = 19, cex = 3, bg =
text(dat1_species_only.pcoa$points[ ,1], dat1_species_only.pcoa$points[ ,2], labels = row.names(dat1_sp
```



Answer 3c - Interpret results: Write an informative yet succinct (~5 sentences) caption that creates a "stand-alone" figure. Take a peek at figures and figure captions in a paper published in your favorite journal for inspiration.

Answer 3c: I'll be explaining the PCoA plot. The PCoA plot shows the different sites cluster together along the x-axis which explains 41.3% of the variation. Being that all the sites that are clustered to the left are all sites in community A, and all the sites that are clustered to the right are in community B, I would assume that the sites are clustering based on the community they are found, and that community is what is being portrayed on the x-axis. In addition, the Permanova shows that community has a significant affect on site, which is in accordance with what the PCoA plot shows.
SUBMITTING YOUR ASSIGNMENT Use Knitr to create a PDF of your completed 7.DiversitySynthesis_Worksheet.Rmd document, push it to GitHub, and create a pull request. Please make sure your updated repo includes both the pdf and RMarkdown files.

Unless otherwise noted, this assignment is due on **Wednesday, February 15th, 2023 at 12:00 PM (noon)**.