# 7. Worksheet: Diversity Synthesis

Joy O'Brien; 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: Observed richness among communities ranges from 20-28 (meaning that is the number of species detected within the sample). My observed richness was the lowest (20 species). In terms of Good's coverage, this value ranged from 0.73 to 0.95 indicating that there was variation in sampling efforts among students. The rare curve shown below indicates that as sample size increased, the amount of different species detected started to become more stationary between 20-25 species. The differences that we see here between sites is largely due to differences sample size which dimishes the chance of detecting all species within the community.

**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.

```
# Set up
rm(list = ls())
getwd()
```

```
## [1] "/Users/joyobrien/GitHub/QB2023_OBrien/2.Worksheets/7.DiversitySynthesis"
```

```
setwd("~/GitHub/QB2023_OBrien/2.Worksheets/7.DiversitySynthesis")
require("vegan")
```

```
## Loading required package: vegan
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

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

```
# Load packages
package.list <- c('vegan', 'tidyverse', 'ggplot2', 'dplyr', 'broom')
for (package in package.list) {
  if(!require(package, character.only = TRUE, quietly = TRUE)) {
    install.packages(package)
  }
  library(c(package), character.only = TRUE)
}
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

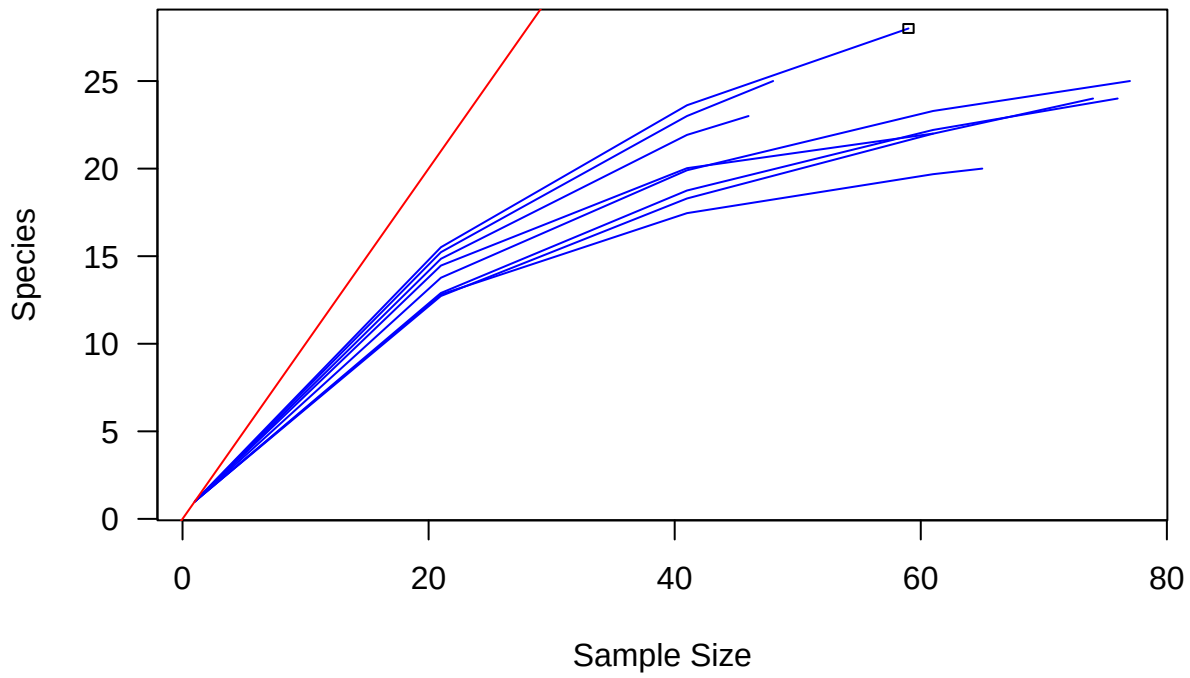
```
# Sampling coverage and rarefaction curves

# Load necessary packages
library(vegan)

# Load data
gummy <- read.csv("/Users/joyobrien/GitHub/QB2023_OBrien/2.Worksheets/7.DiversitySynthesis/class_data.cs

# Removing the first two column from the dataset so we can work with site by species matrix
new.gummy <- gummy[ ,3:32]

# Calculating observed richness
specnumber(new.gummy)
```

```
## [1] 28 25 23 22 20 24 24 25
```

```r
# Function for Good's coverage
C <- function(x = ""){
  1 - rowSums(x == 1) / rowSums(x)
}
# Calculating Good's Coverage
C(new.gummy)
```

```
## [1] 0.8135593 0.7291667 0.8043478 0.9508197 0.9230769 0.8513514 0.8947368
## [8] 0.9090909
```

```r
# Idenfify sample with fewest occupancies
min.N <- min(rowSums(new.gummy))

# Rarifying to rarefy each sample to this level
S.rarefy <- rarefy(x = new.gummy, sample = min.N, se = TRUE)

# Plotting rarefaction results
rarecurve(x = new.gummy, step = 20, col = "blue", cex = 0.6, las = 1)
abline(0, 1, col = 'red')
text(1500, 1500, "1:1", pos = 2, col = 'red')
```




**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.

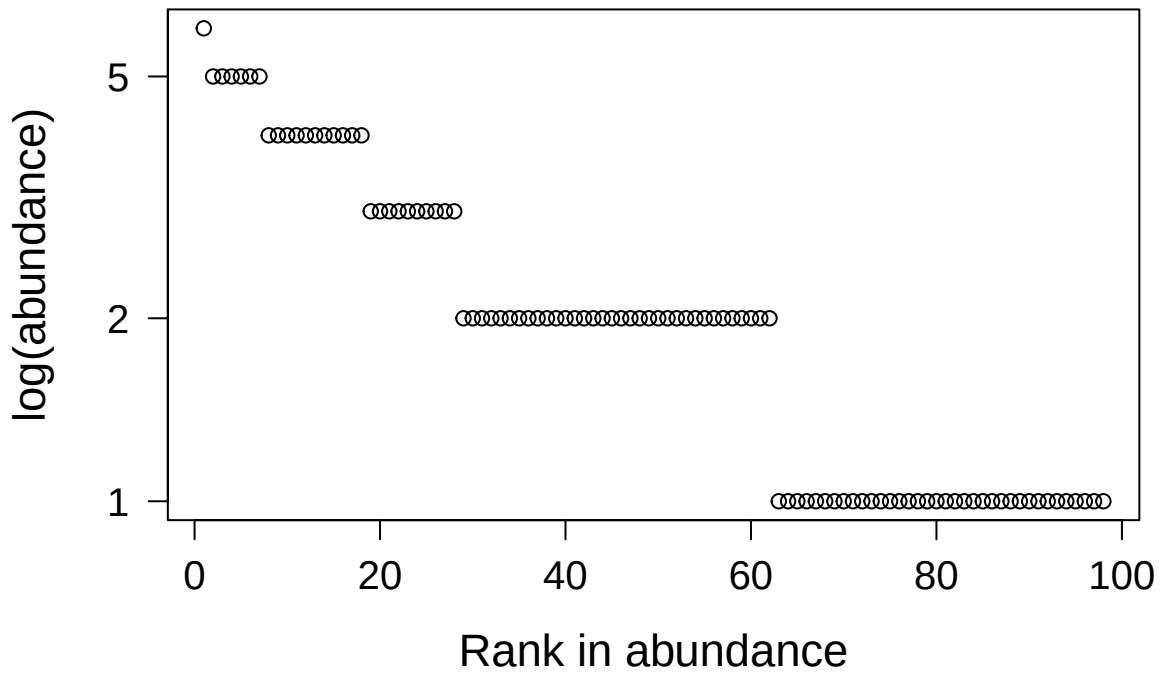
> Answer: My site has a shannon diversity metric of 2.78 which is the lowest shannon diversity value from community B (ranged from 2.78 to 2.97). Since my site is located within community B, I hypothesize that Community A is more diverse than Community B. Shannon diversity indices for community A range between 2.9 and 3.1 which is larger than those for community B. Additionally, the rank abundance curves for community A and community B both indicate that the abundance among species is unequally distributed which explains the variation in Shannon diversity values.

**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.

```
# Wrangling the data
str(gummy)
```

```
## '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
```

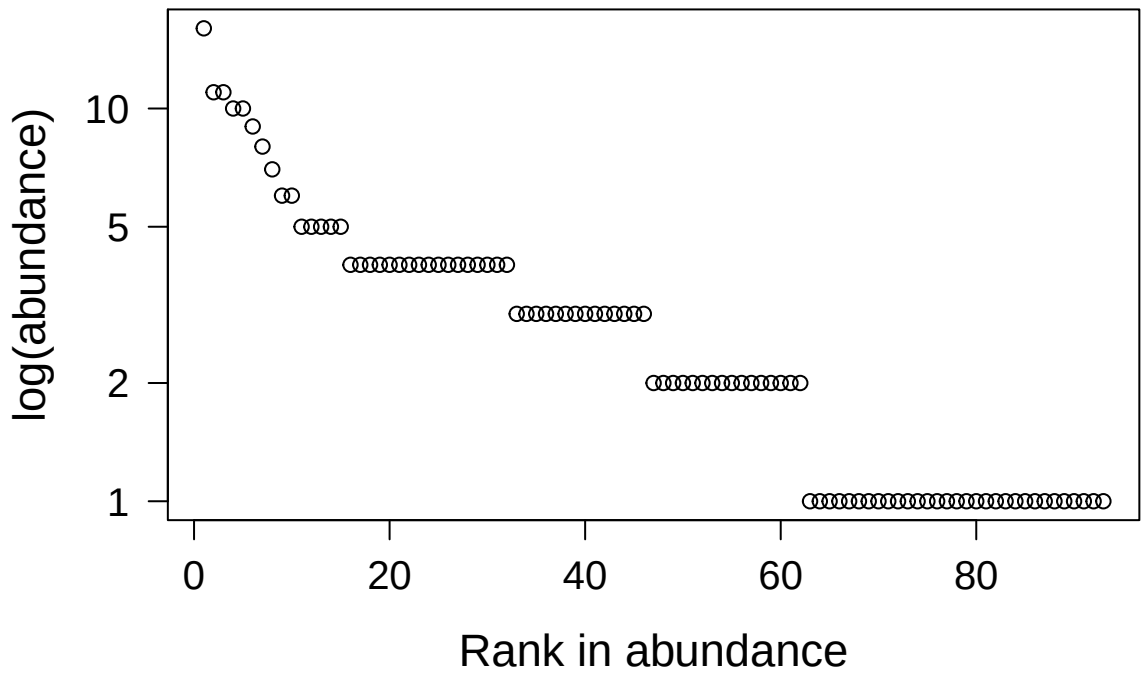
```r
# Convert species numbers to numeric
new.gummy [1:30] = lapply(new.gummy[1:30], FUN =
                          function(y){as.numeric(y)})
# Sub-setting community data
A <- new.gummy[1:4, ]
B <- new.gummy[5:8, ]

# Calculating Shannon diversity
# Community A
A_div <- diversity(A, index = "shannon")
A_divmat <- as.matrix(A_div)
AH <- mean(A_divmat)
# Community B
B_div <- diversity(B, index = "shannon")
B_divmat <- as.matrix(B_div)
BH <- mean(B_divmat)

# Combine shannon estimates for communities A and B into one data frame
diversity <- merge(AH, BH)
# RAC

RAC <- function(x = ""){
  x.ab = x[x > 0]
  x.ab.ranked = x.ab[order(x.ab, decreasing = TRUE)]
  as.data.frame(lapply(x.ab.ranked, unlist))
  return(x.ab.ranked)
}
# RAC for community A
rac <- RAC(x = A)
ranks <- as.vector(seq(1, length(rac)))
opar <- par(no.readonly = TRUE)
par(mar = c(5.1, 5.1, 4.1, 2.1))
plot(ranks, log(rac), type = 'p', axes = F,
     xlab = "Rank in abundance", ylab = "log(abundance)",
     las = 1, cex.lab = 1.4, cex.axis = 1.25)

box()
axis(side = 1, labels = T, cex.axis = 1.25)
axis(side = 2, las = 1, cex.axis = 1.25,
     labels = c(1, 2, 5, 10, 20), at = log(c(1, 2,5, 10, 20)))
```




```
par <- opar
```

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

```
# RAC for community B
rac <- RAC(x = B)
ranks <- as.vector(seq(1, length(rac)))
opar <- par(no.readonly = TRUE)
par(mar = c(5.1, 5.1, 4.1, 2.1))
plot(ranks, log(rac), type = 'p', axes = F,
     xlab = "Rank in abundance", ylab = "log(abundance)",
     las = 1, cex.lab = 1.4, cex.axis = 1.25)

box()
axis(side = 1, labels = T, cex.axis = 1.25)
axis(side = 2, las = 1, cex.axis = 1.25,
     labels = c(1, 2, 5, 10, 20), at = log(c(1, 2,5, 10, 20)))
```




```
par <- opar
```

**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: Figure 1. Rank abundance curve for Community B detailing species richness and evenness. Community B contains 30 different species and 4 sites that were randomly sampled (sample size varied as did species composition). The x-axis represents the species based on abundance with the most abundant species given a rank of 1. The y-axis represents relative abundance of species. The slope we see here indicates that there is low evenness in the community due to high-ranking species having the highest abundance when compared to lower ranking species. This is a common trend in community ecology.

### 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: Yes. I hypothesized for the beta diversity analysis is that there will be compositional differences between sites within gummy communities (given that we know there are diversity differences between communities). Based on analysis, there are significant compositional differences between communities A and B as indicated via a Permanova test (p-value = 0.029). We can also see the differences in community composition with the PCoA plot shown below. PCoA axis 1 explains 41.3 % of the variation and PCoA axis 2 explains 20.2% of the variation. Based on the permanova results and the way that the sites are clustering in the ordination,

community type may explain the compositional differences within the sites. I also noticed that sites 1, 2, 3, and 4 shown in the PCoA are more loosely clustered than the sites from community B, this may support my hypothesis from the previous section that community A is more diverse than community B.

**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.

```
# Obtaining species only
species <- gummy[ ,3:32]
str(species)
```

```
## 'data.frame':    8 obs. of  30 variables:
##  $ 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
```

```
# Convert species numbers to numeric
species [1:30] = lapply(species[1:30], FUN =
                          function(y){as.numeric(y)})
str(species)
```

```
## '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
```

```
# Calculate distance using Bray-Curtis
species.d <- vegdist(species, method = "bray")

# Name communities via a factors vector
community <- c(rep("A", 4), rep("B", 4))
# Permanova
adonis2(species ~ community, method = "bray", permutations = 999)
```

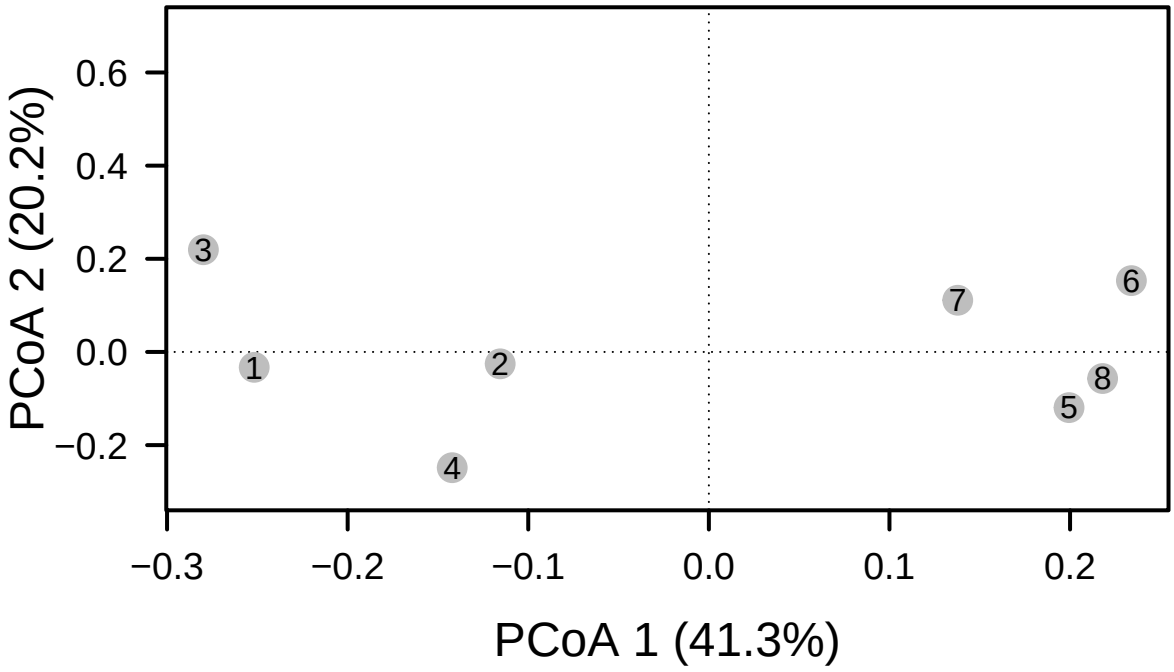
```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = species ~ community, permutations = 999, method = "bray")
##           Df SumOfSqs      R2      F Pr(>F)
## community  1  0.31895 0.39151 3.8605  0.028 *
## 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.

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(lattice)

order <- rev(attr(species, "Labels"))

species.d <- as.matrix(species.d)

gummy.pcoa <- cmdscale(species.d, eig = TRUE, k = 3)
# Quantify percent variation explained by the first three axes
explainvar1 <- round(gummy.pcoa$eig[1] / sum(gummy.pcoa$eig), 3) * 100
explainvar2<- round(gummy.pcoa$eig[2] / sum(gummy.pcoa$eig), 3) * 100
explainvar3 <- round(gummy.pcoa$eig[3] / sum(gummy.pcoa$eig), 3) * 100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)

# Define plot parameters
par(mar =  c(5.1, 4.1, 4.1, 2.1) + 0.1)

# Initiate plot
plot(gummy.pcoa$points[ ,1], gummy.pcoa$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)

# Add axes
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)

# Add points and labels
points(gummy.pcoa$points[ ,1], gummy.pcoa$points[ ,2],
       pch = 19, cex = 2, bg = "gray", col = "gray")
text(gummy.pcoa$points[ ,1], gummy.pcoa$points[ ,2],
     labels = row.names(gummy.pcoa$points))
```




**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. > Figure 2. PCoA plot of gummy species community composition within sites. Sites within Community A are labeled 1-4 and sites within Community B are labeled 5-8. The PCoA axes represent the percent variation explained. We speculate that PCoA axis 1 explains the site variation in communities A and B since they form distinct clusters based on sites. Community A clusters on the left while community B clusters on the right.

## 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 15$^{th}$, 2023 at 12:00 PM (noon)**.