6. Worksheet: Diversity Sampling

Mark Hibbins; Z620: Quantitative Biodiversity, Indiana University

29 January, 2019

OVERVIEW

In this worksheet, you will use the jelly bean site-by-species matrix generated from **6. Diversity Sampling**. Along with tools outlined in the **5. Local (alpha) Diversity** and **7. Control Structures** handouts, you will develop questions, visualize data, and test hypotheses related to sampling effects and its effect on estimates of within-sample biodiversity.

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 handout 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 `6.DiversitySampling_Worskheet.Rmd` and the PDF output of `Knitr` (`DiversitySampling_Worskheet.pdf`).

1) Group brainstorming

With your team partner and perhaps other students in the class, spend 15 minutes or so brainstorming questions, code, "fantasy figures", and statistical tests that could be used to test questions with the class's data represented in the site-by-species matrix that you have generated.

2) Code

Use the space below for code that is being used to analyze your data and test your hypotheses. Create one (and only one, although it can have multiple panels) *publication quality* figure. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

Workspace setup

```
#Workspace setup

setwd("/Users/mark/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/6.DiversitySa
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 3.4.4
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.5-3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
##
## 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
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.4.4
```

```
JellyBeans <- read_delim("~/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/6.Di
    "\t", escape_double = FALSE, trim_ws = TRUE) #load class collected data
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Group = col_character(),
##   Site = col_character()
## )
## See spec(...) for full column specifications.
```

```
JellyBeans_Source <- read_delim("~/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Workshee
    "\t", escape_double = FALSE, trim_ws = TRUE) #load source data
```

```
## Parsed with column specification:
## cols(
##   Flavor = col_character(),
##   Mix = col_character(),
##   Count = col_double(),
##   Description = col_character()
## )
```

Question 1: How well did our classification and sampling conventions cover the jellybean community?

```
good_C = function(x) { #function for Good's coverage
  1 - (sum(x == 1) / sum(x))
}

jellybean_sbys_full <- JellyBeans[,3:30] #subset matrix that just has the species data
```

```r
coverage_all <- apply(jellybean_sbys_full, 1, good_C) #estimate coverage for all sites
coverage_all
```

```
## [1] 0.8250000 0.8533333 0.8933333 0.8875000 0.9125000 0.8902439 0.9104478
## [8] 0.9027778 0.8813559
```

```r
mean(coverage_all) #mean coverage across all sites
```

```
## [1] 0.8840547
```

Answer: Overall coverage appears to be hovering around 88%, which is pretty good.

Question 2: Does coverage differ between groups A, B, and the full data?

```r
jellybean_sbys_A <- subset(JellyBeans, Group == 'A', select=c(3:30)) #split dataframes into A and B
jellybean_sbys_B <- subset(JellyBeans, Group == 'B', select=c(3:30))

coverage_A <- apply(jellybean_sbys_A, 1, good_C) #Coverage across sites for group A
mean(coverage_A) #mean coverage for group A
```

```
## [1] 0.8729377
```

```r
coverage_B <- apply(jellybean_sbys_B, 1, good_C) #Coverage across sites for group B
mean(coverage_B) #mean coverage for group B
```

```
## [1] 0.8979509
```

```r
t.test(coverage_A, coverage_B) #difference between A and B
```

```
##
##  Welch Two Sample t-test
##
## data:  coverage_A and coverage_B
## t = -1.5026, df = 5.9977, p-value = 0.1837
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06575042  0.01572391
## sample estimates:
## mean of x mean of y
## 0.8729377 0.8979509
```

```r
t.test(coverage_A, coverage_all) #difference between A and full
```

```
##
##  Welch Two Sample t-test
##
## data:  coverage_A and coverage_all
## t = -0.63788, df = 7.3859, p-value = 0.5428
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05189508  0.02966108
## sample estimates:
## mean of x mean of y
## 0.8729377 0.8840547
```

```r
t.test(coverage_B, coverage_all) #difference between B and full
```

```
##
##  Welch Two Sample t-test
##
## data:  coverage_B and coverage_all
## t = 1.127, df = 10.014, p-value = 0.286
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01357117  0.04136367
## sample estimates:
## mean of x mean of y
## 0.8979509 0.8840547
```

> ***Answer***: While the estimates of coverage are slightly different, the differences are not statistically significant. Therefore the effects of sampling do not appear strong enough to change our conclusions about coverage.

Question 3: Does diversity (Shannon's diversity) differ between groups A, B, and the full data?

```
ShanH <- function(x){ #Function for Shannon's diversity index
  H = 0
  for(n_i in x){
    if(n_i > 0){
      p = n_i / sum(x)
      H = H - p*log(p)
    }
  }
  return(H)
}

diversity_A <- apply(jellybean_sbys_A, 1, ShanH) #Shannon's diversity across sites for group A
mean(diversity_A) #mean Shannon's diversity for group A
```

```
## [1] 2.619364
```

```
diversity_B <- apply(jellybean_sbys_B, 1, ShanH) #Shannon's diversity across sites for group B
mean(diversity_B) #Shannon's diversity for group B
```

```
## [1] 2.755517
```

```
diversity_all <- apply(jellybean_sbys_full, 1, ShanH) #Shannon's diversity across all sites
mean(diversity_all) #Overall Shannon's diversity
```

```
## [1] 2.679876
```

```
t.test(diversity_A, diversity_B) #Difference between A and B
```

```
##
##  Welch Two Sample t-test
##
## data:  diversity_A and diversity_B
## t = -2.1839, df = 6.8868, p-value = 0.0659
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.28406709  0.01176121
## sample estimates:
```

```
## mean of x mean of y
##  2.619364  2.755517
```

```
t.test(diversity_A, diversity_all) #Difference between A and total
```

```
##
##  Welch Two Sample t-test
##
## data:  diversity_A and diversity_all
## t = -0.96139, df = 8.7132, p-value = 0.3623
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.20361675  0.08259191
## sample estimates:
## mean of x mean of y
##  2.619364  2.679876
```

```
t.test(diversity_B, diversity_all) #Difference between B and total
```

```
##
##  Welch Two Sample t-test
##
## data:  diversity_B and diversity_all
## t = 1.4014, df = 8.9302, p-value = 0.1949
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04660548  0.19788653
## sample estimates:
## mean of x mean of y
##  2.755517  2.679876
```

> ***Answer***: Neither group partition is different from the full dataset, but there actually is a marginally significant difference ($p = 0.0659$) between groups A and B. This effectively highlights how biases introduced by sampling can affect our inferences of species diversity.

Figure: Overlapping species abundance distributions for groups A and B

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.4.3
```

```
library(ggplot2)
```

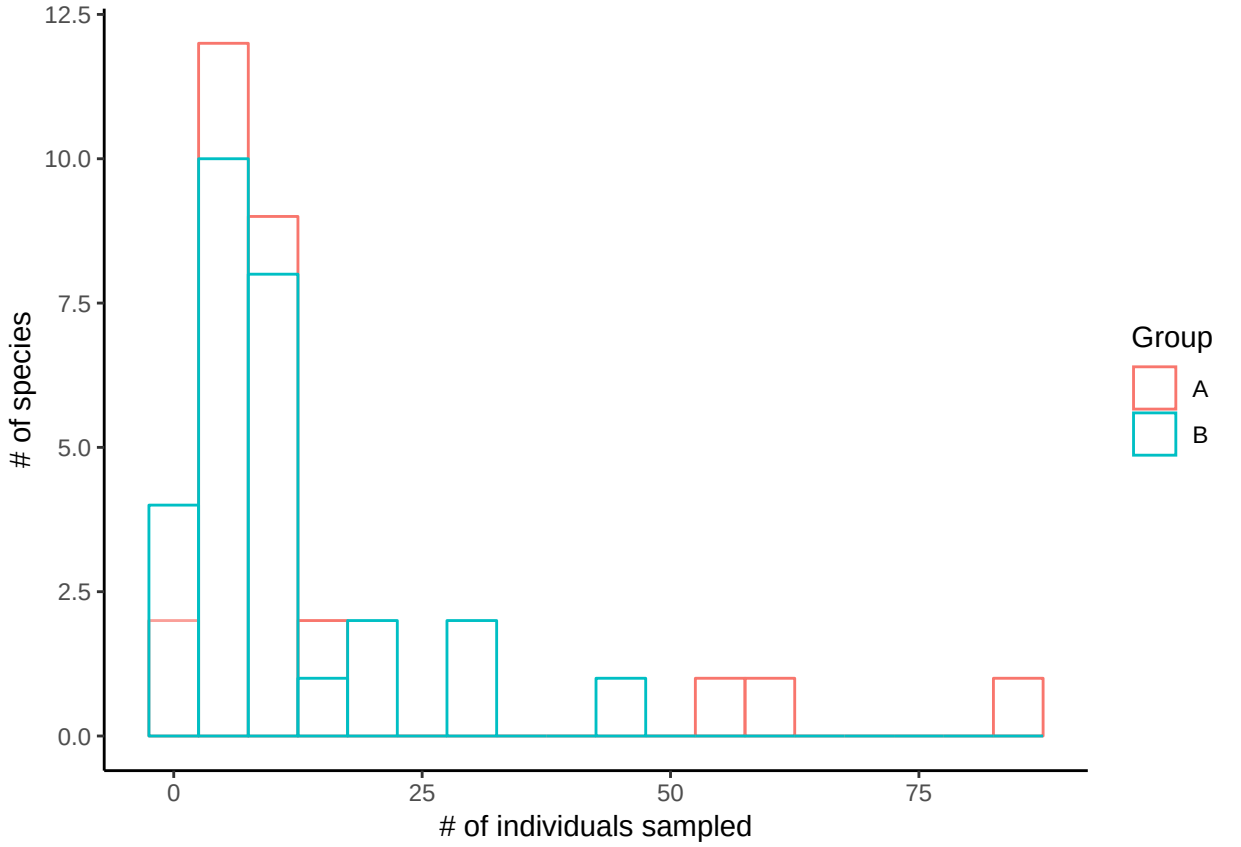
```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
jellybeans_long <- melt(JellyBeans) #convert categorical variables into factors
```

```
## Using Group, Site as id variables
```

```
jellybeans_coll <- aggregate(value ~ Group + variable,
                             jellybeans_long, sum) #collapse dataframe into sums by factor

SAD_AB <- ggplot(jellybeans_coll, aes(x=value)) #sets up plot object

SAD_AB <- SAD_AB + geom_histogram(aes(color=Group), #adds the histogram
                                  fill = 'white',
                                  alpha = 0.3,
```

```
                                  position = 'identity',
                                  binwidth = 5)

SAD_AB <- SAD_AB + theme_bw() + theme(panel.border = element_blank(), #cleans up the grid space
                                      panel.grid.major = element_blank(),
                                      panel.grid.minor = element_blank(),
                                      axis.line = element_line(colour = "black"))

SAD_AB <- SAD_AB + labs(x = '# of individuals sampled', #updates the axis labels
                        y = '# of species')

SAD_AB
```



3) Figure caption

Write an informative yet succicnt (~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.

> Species abundance distribution across all sites, subsetted into groups A (red) and B (blue). Groups A and B are overlaid onto the same bin where the distribution is overlapping. For both groups, the majority of species are sampled modestly, with counts between 5 and 35. Group B has more species sampled very rarely (0 to 5 times), and that are sampled at moderate frequency (20 to 30 times). All the instances with the highest sampling (50 to 90 individuals) belong to group A.

SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed 6.DiversitySampling_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, January 30th, 2017 at 12:00 PM (noon)**.