ENTMLGY 6707 Entomological Techniques and Data Analysis

Measures of species diversity

Characterizing insect communities

For a long time humans have been interested in documenting and describing the differences in abundance and diversity of species. Insects are no exception. Insects are among the most abundant and diverse organisms on Earth. Of course, this is one reason why we study them!

Community ecologists seek to understand how multiple species interact with each other and their environment in a specified time and place. Based on the study questions, this typically involves a focal habitat or set of species. Understanding patterns of species diversity along environmental gradients or across spatial and/or temporal scales continues to be a major focus of community ecology research. Because of this continued interest, a plethora of tools have been developed to characterize communities based on dimensions of diversity.

Measures of taxonomic diversity

Traditional approaches to characterize communities have focused on taxonomic diversity. This involves measuring unique types (e.g., species, genera, families, orders) and comparing among samples, sites, etc. Although there are many different ways to characterize diversity, assessing patterns of abundance, richness, diversity, and evenness are some of the most common approaches. Importantly, community ecologists generally use a combination of metrics to assess patterns of diversity.

We will use an open source data set to demonstrate how to calculate multiple diversity metrics. These data are ant species collected in the Hemlock Removal Experiment at the NSF Harvard Forest Long-term Ecological Research (LTER) site (See more here: https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=HF118). This experiment includes four treatments (Hemlock girdled, Hemlock logged, Hemlock control, and Hardwood control) each replicated across two (n = 2) 90 x 90 m plots. The goal of this experiment was to investigate the effects of hemlock mortality from the invasive hemlock woolly adelgid and preemptive hemlock removal via logging. The specific study that we are using focuses on ant community responses to these environmental changes, but there have been many studies conducted within this experiment.

Because the data are openly available on the NSF LTER website, we can directly load them into R.

```
ants <- read.csv(file="https://pasta.lternet.edu/package/data/eml/knb-lter-hfr/118/34/90ca76
```

Notice that the ant data are in `long format`. Each row contains the count of one species at a specific site and date. Ant species are combined into one column called 'code'. In order to calculate the diversity metrics, we want the data represented in `wide format`. In this case, each species and their associated counts will be represented in a separate column. Each row will represent the counts of all ant species at a specific site and date. We can reshape the data set using the `reshape2` package.

```
ant.matrix <- dcast(ants, year + block + plot + treatment + trap.type ~ code,
                    sum, value.var = "abundance", na.rm = TRUE)

# change variables to factors
ant.matrix$year <- as.factor(ant.matrix$year)
ant.matrix$block <- as.factor(ant.matrix$block)
ant.matrix$plot <- as.factor(ant.matrix$plot)
ant.matrix$treatment <- as.factor(ant.matrix$treatment)
ant.matrix$trap.type <- as.factor(ant.matrix$trap.type)

# shorten treatment names
levels(ant.matrix$treatment)
```

```
[1] "Girdled"         "HardwoodControl" "HemlockControl"  "Logged"
```

```
levels(ant.matrix$treatment)[levels(ant.matrix$treatment)=="HardwoodControl"] <- "Hardwood"
levels(ant.matrix$treatment)[levels(ant.matrix$treatment)=="HemlockControl"] <- "Hemlock"

levels(ant.matrix$year)
```

```
 [1] "2003" "2004" "2005" "2006" "2007" "2008" "2009" "2010" "2011" "2012"
[11] "2013" "2014" "2015" "2018"
```

The data set includes ant species collected over multiple years. Let's subset the data to focus on ants collected via pitfall traps in 2015 and 2018. Since it is likely that some of the ant species represented in the full data set (2003-2018) were not collected in these years (especially rarer species), let's also remove the columns of species that were not collected.

```
yr1 <- ant.matrix %>% filter(year == "2015") %>% droplevels()
yr2 <- ant.matrix %>% filter(year == "2018") %>% droplevels()

ants2 <- rbind(yr1, yr2)

ants3 <- ants2[, colSums(ants2 !=0) > 0]
```

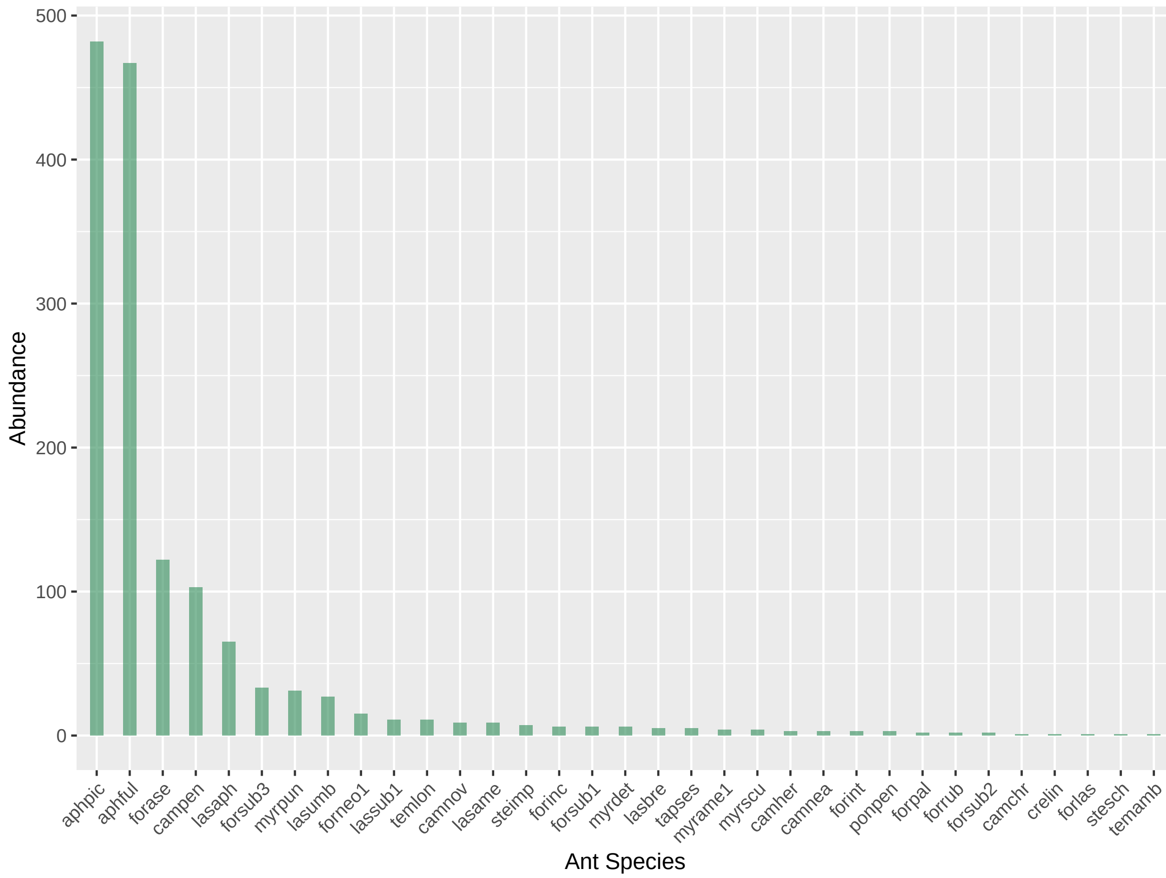
Now we are left with a data set in `wide format` that has 33 ant species collected via pitfall traps in the four treatments. Let's characterize the community using diversity metrics!

Abundance

Abundance measures the number of individuals of each distinct type. Descriptive information about differences in species abundances provides valuable information to characterize the community. Typically, a community has a few abundant species and many rare species. Therefore, understanding the distribution of abundance among species in the community provides information about species dominance.

```
ant.dom <- colSums(ants3[,6:38]) # sum abundances across sites
ant.dom <- as.data.frame(ant.dom) # change to data frame
names(ant.dom)[1] <- "count" # rename the abundance column
ant.dom$species <- rownames(ant.dom) # make a column of species
ant.dom$species <- as.factor(ant.dom$species) # change species to a factor
```

```
ant.dom %>% mutate(species = fct_reorder(species, desc(count)))  %>%
  ggplot( aes(x = species, y = count)) +
  geom_bar(stat="identity", fill="seagreen", alpha=.6, width=.4) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) +
  xlab("Ant Species") +
  ylab("Abundance")
```



In the figure above, we can see that there are two dominant species within the data set, *Aphaenogaster picea* and *Aphaenogaster fulva*. The remaining 31 ant species were collected in much lower abundance. If I were discussing these data in a presentation or in a manuscript, my first results slide or paragraph would highlight descriptive information about the community, including the total number of individuals collected, the number of species collected (and sometimes the number of different genera), and the most abundant species collected. If there were nonnative species collected, I would highlight that information as well.

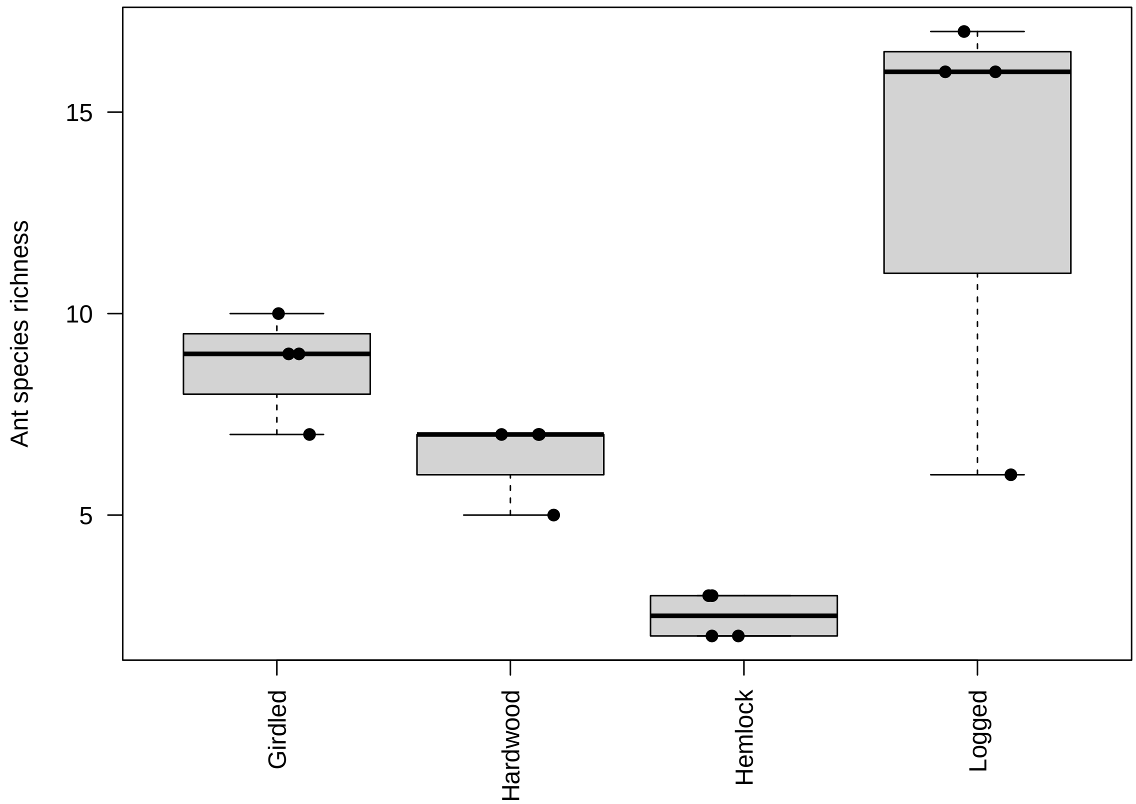
Richness

Richness measures the number of distinct types in a community and perhaps is the most basic measure of diversity. Even with the most intensive and comprehensive sampling designs, it is very unlikely that we can count all insect species within a given habitat or site (perhaps this is more achievable for large taxa like mammals). Therefore, we have to estimate species richness by collecting samples. The accuracy of these estimates are dependent on sampling type and effort, which are tied to the experimental design and sampling methodology. We will highlight two common ways to estimate species richness.

Count the number of species

The most basic method is to simply count the number of species collected per trap, site, etc. Although this metric is sensitive to differences in sampling effort among sites (i.e., increased sampling effort often results in more species observed), it weights all species equally, independent from their relative abundance. In other words, common and rare species are weighted equally. This basic metric can be easily calculated using the `vegan` package. Once you calculate richness and add it as a variable into the data set, species richness can be used as a response variable in a model.

```
ants3$sp.rich <- specnumber(ants3[,6:38])
par(mar=c(6,4,2,2))
boxplot(sp.rich ~ treatment, data = ants3,
        xlab = "", ylab = "Ant species richness",
        cex.axis = 1, las = 2)
stripchart(sp.rich ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```



Species accumulation curves

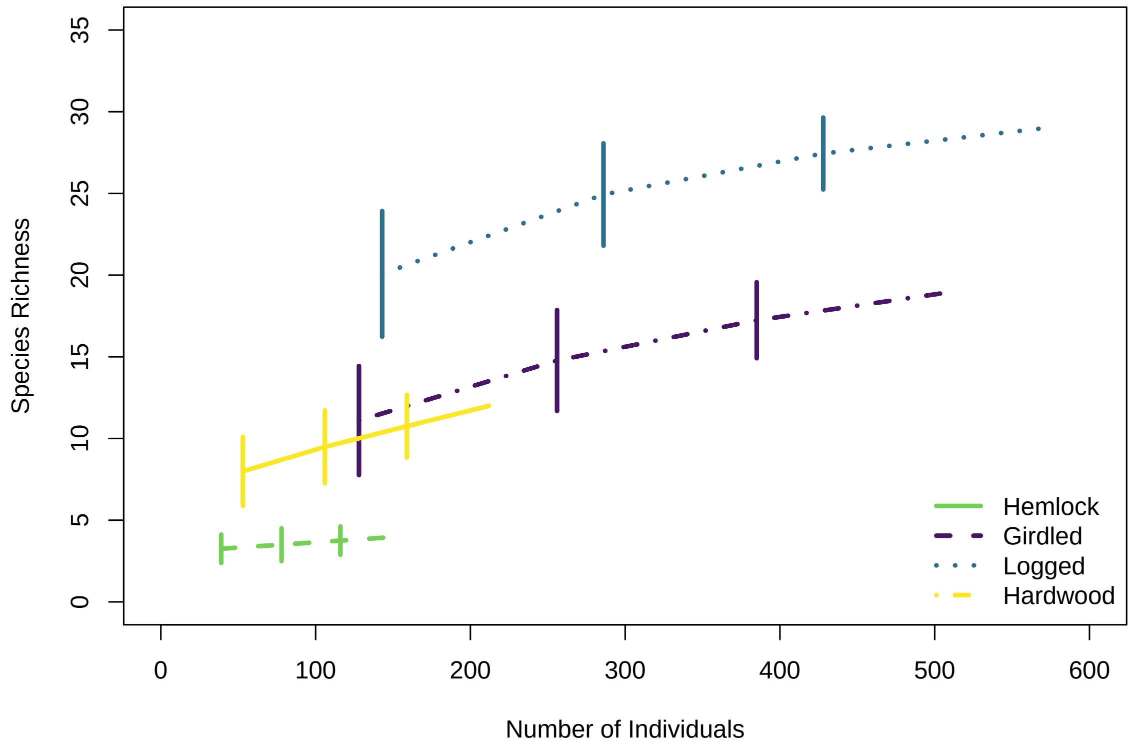
A second method is to estimate species richness with accumulation curves. Since it is unlikely that we will detect all species present at a site, this method estimates species richness using information from the data set about new species found in subsequent samples or with the collection of subsequent individuals. Therefore, accumulation curves can be used to estimate the number of species by standardizing among sites or by the number of individuals. Standardization by sites is helpful when you have an unbalanced experimental design, while standardizing by individuals is helpful when samples have high variability in the number of individuals collected. The estimation occurs via a permutation procedure that repeatedly subsamples the data, and calculates a mean and standard deviation. There are many different accumulator functions that can be used. If you plan to use this method in your research, I would strongly suggest learning about these different functions and selecting the one that best fits your data. I typically use the rarefaction function. Rarefaction is useful when samples have large differences in abundance. This method standardizes the sample sizes during the permutation procedure so that species richness among sites or treatments is compared at equivalent abundances. We can use the `vegan` package to estimate ant species richness among the four hemlock treatments.

```
girdled <- ants3[which(ants3$treatment == "Girdled"),]
hardwood <- ants3[which(ants3$treatment == "Hardwood"),]
hemlock <- ants3[which(ants3$treatment == "Hemlock"),]
logged <- ants3[which(ants3$treatment == "Logged"),]
```

```
sp.girdled <- specaccum(girdled[6:38], method = "rarefaction",
                        permutations = 100, gamma = "jack2")
sp.hardwood <- specaccum(hardwood[6:38], method = "rarefaction",
                         permutations = 100, gamma = "jack2")
sp.hemlock <- specaccum(hemlock[6:38], method = "rarefaction",
                        permutations = 100, gamma = "jack2")
sp.logged <- specaccum(logged[6:38], method = "rarefaction",
                       permutations = 100, gamma = "jack2")

plot(sp.girdled, pch = 19, col = "#481567FF", xvar = c("individuals"),
     lty = 4, lwd = 3,
     ylab = "Species Richness", xlab = "Number of Individuals",
     xlim = c(0, 600), ylim = c(0, 35))
plot(sp.hardwood, add = TRUE, pch = 15, xvar = c("individuals"),
     lty = 1, lwd = 3, col = "#FDE725FF")
plot(sp.hemlock, add = TRUE, pch = 4, xvar = c("individuals"),
     lty = 2, lwd = 3, col = "#73D055FF")
plot(sp.logged, add = TRUE, pch = 9, xvar = c("individuals"),
     lty = 3, lwd = 3, col = "#2D708EFF")

legend("bottomright", legend = c("Hemlock", "Girdled", "Logged", "Hardwood"),
       lty = c(1,2,3,4), cex = 1, bty = "n", lwd = 3,
       col = c("#73D055FF", "#481567FF", "#2D708EFF", "#FDE725FF"))
```



The figure above tells us a few things. First, ants were more abundant and rich in the hemlock logged and girdled treatments compared to the hardwood and hemlock controls. Second, we see that the accumulation curve for hemlock control is rather flat, which suggests that the sampling effort was able to capture the majority of species estimated to be in that treatment. The accumulation curves for the other three treatments appear to still be increasing (i.e., do not appear to have reached an asymptote), especially the hardwood treatment. Therefore, it is likely additional sampling will yield observations of new species not currently represented in the data set. Here, the vertical lines represent the standard deviation.

Species accumulation curves can be paired with jackknife estimates that calculate the extrapolated species richness using first-order or second-order estimator functions. Jackknife estimates use the observed number of species collected at a site as well as the observed numbers of singleton species (species with only one observation) or doubleton species (species with two observations). First-order estimates only take into account the singletons, while second-order estimates take into account both singletons and doubletons. We can calculate first- and second-order jackknife estimates with the `fossil` package. As an example, let's calculate the first-order and second-order jackknife estimates for the hemlock logged treatment and compare those values to the observed species richness.

```r
# hardwood
jack1(hardwood[6:38], taxa.row = FALSE, abund = TRUE)
```

```
[1] 16.97642
```

```r
jack2(hardwood[6:38], taxa.row = FALSE, abund = TRUE)
```

```
[1] 21.85849
```

```r
specnumber(ants3[,6:38], groups = ants3$treatment)
```

```
 Girdled Hardwood  Hemlock   Logged 
      19       12        4       29 
```

The observed richness for the hardwood treatment is 12 ant species. Estimated species richness based on first-order and second-order jackknife estimates are 16.9 and 21.8 species, respectively. This aligns with our conclusion from the species accumulation curve above. Sampling effort during the years 2015 and 2018 was able to capture 55-71% of the estimated number of ant species present in the community.

Diversity

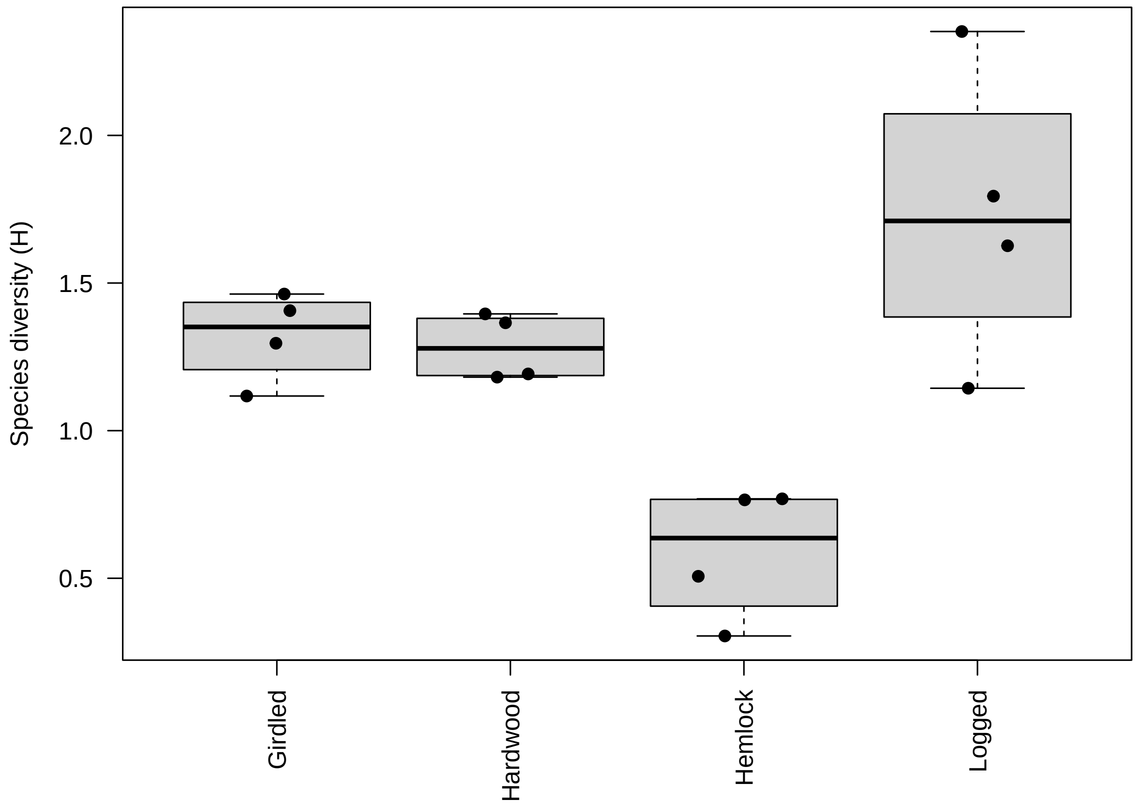
Diversity measures the number of types and their evenness (i.e., relative abundance; more on evenness below). For example, two sites may have the same number of species (i.e., the same species richness), but there may be large differences in the relative abundances of those species. Therefore, communities that have higher species richness and higher species evenness (i.e., more equal relative abundances among species) are generally considered more diverse.

There are many, many metrics that have been developed to estimate species diversity. These metrics tend to differ based on the weighting of species richness versus species evenness. We can use the `vegan` and `hillR` packages to calculate several of the common diversity indices.

Shannon diversity index

The Shannon diversity index (H), also known as Shannon-Wiener, Shannon-Weaver, or Shannon Entropy, uses the abundance of each species in a sample to determine the proportion that each species contributes to the total. The proportion of each species is calculated, those proportions are multiplied by the natural log of the proportion, and the absolute value of the total is the index value.

```
ants3$sh.div <- diversity(ants3[,6:38], index = "shannon")
par(mar=c(6,4,2,2))
boxplot(sh.div ~ treatment, data = ants3,
        xlab = "", ylab = "Species diversity (H)",
        cex.axis = 1, las = 2)
stripchart(sh.div ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```

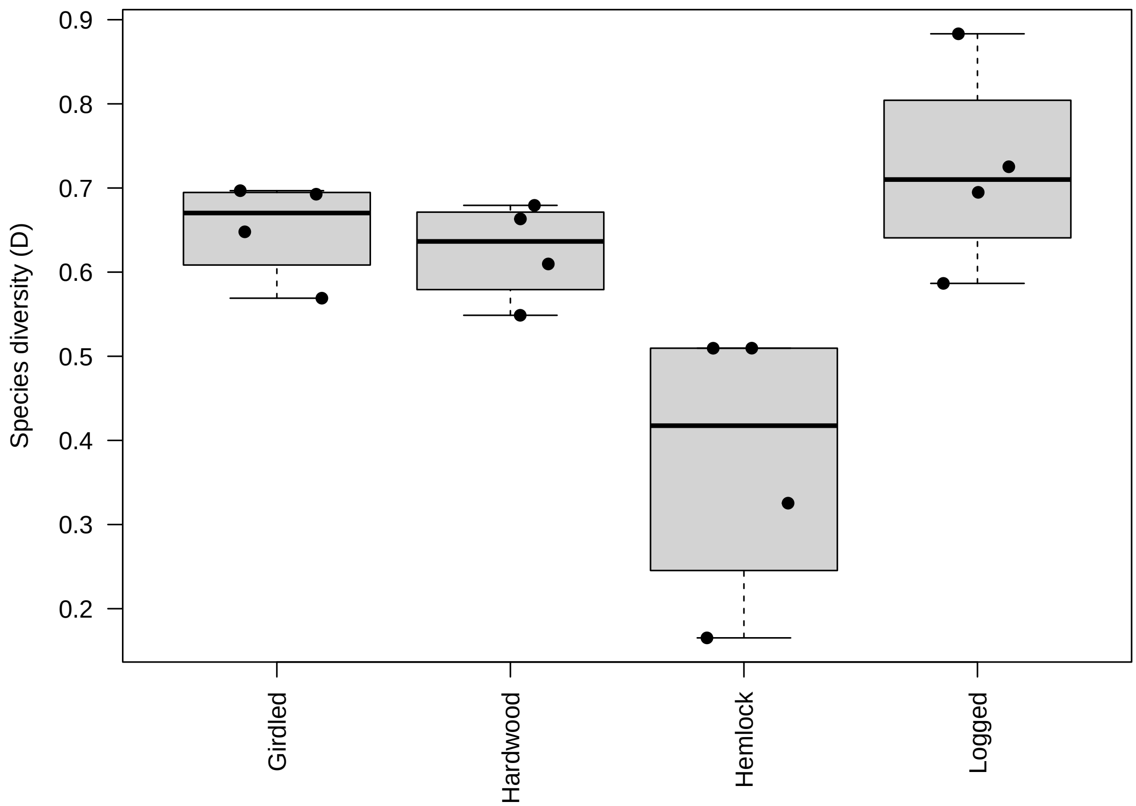


Simpson diversity index

The Simpson diversity index (D) uses the abundance of each species in a sample to determine the proportion that each species contributes to the total. For this metric, the proportion of each species is calculated, those proportions are squared, and then totaled to get the index value. It represents the probability that two randomly selected individuals will be of the same species. Based on this calculation, as the number of species in the community increases, the probability of selecting the same species will decrease. This means that diversity decreases as D increases, which is not intuitive. To correct for this, the Simpson index is often revised by taking 1 - D, the Gini-Simpson index. Values of D range between 0 and 1.

```
ants3$sp.div <- diversity(ants3[,6:38], index = "simpson")
par(mar=c(6,4,2,2))
boxplot(sp.div ~ treatment, data = ants3,
        xlab = "", ylab = "Species diversity (D)",
        cex.axis = 1, las = 2)
```

```r
stripchart(sp.div ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```



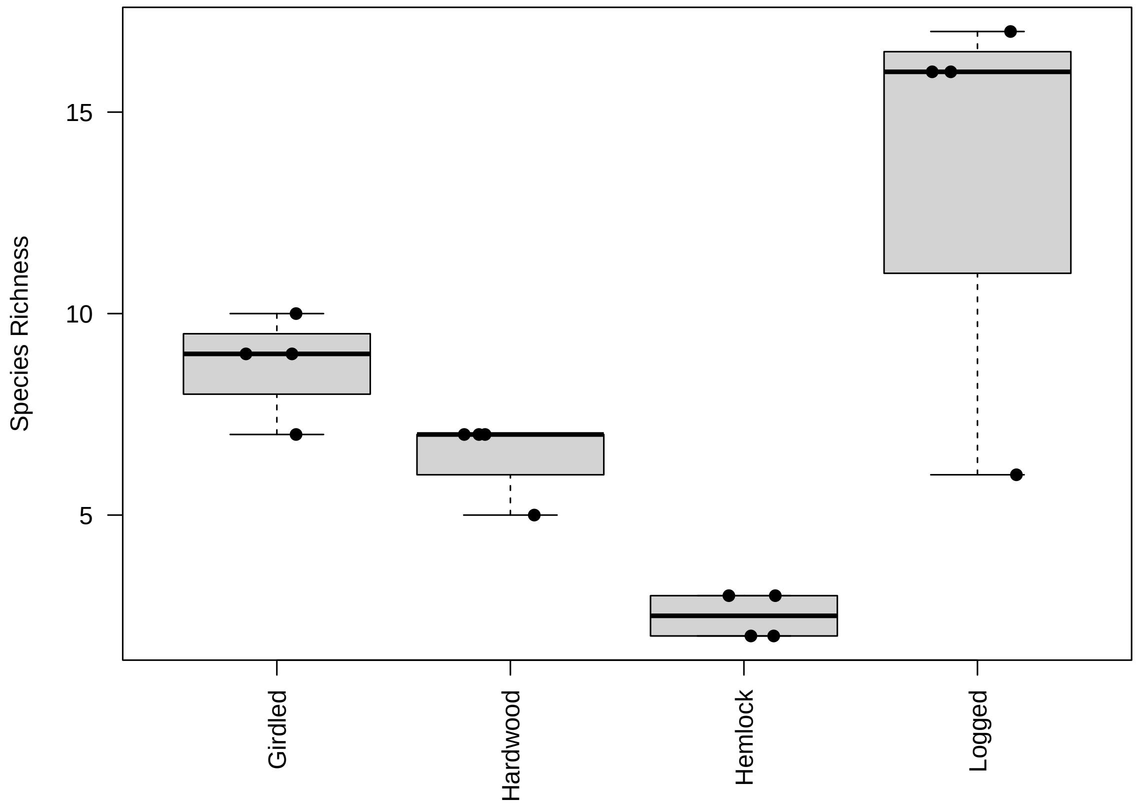
Effective Number of Species aka Hill numbers

As I have mentioned, there are many indices that have been developed to estimate species diversity of communities. Folks continue to develop new indices or revise old indices, and a complete review of all indices is outside of the scope of this course. However, the Shannon and Simpson diversity indices have recently been revised, and these revised metrics have become common in the literature.

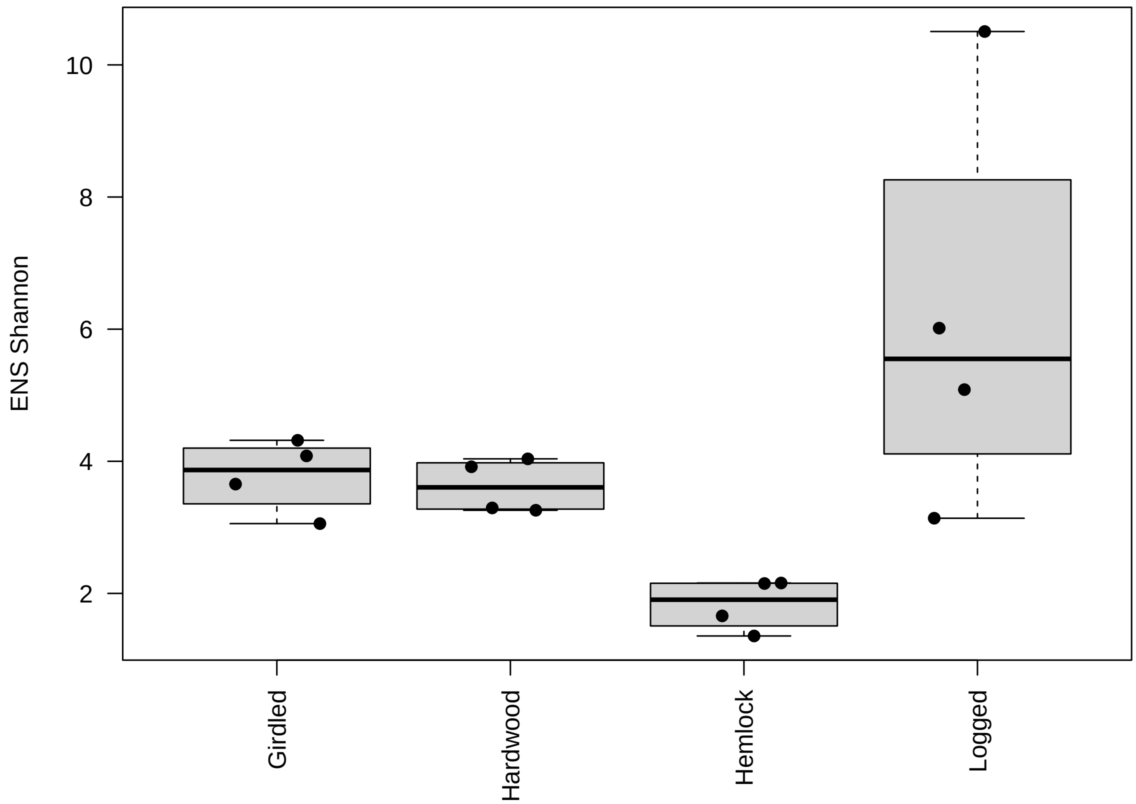
Lou Jost, among others, have argued that the Shannon and Simpson diversity indices are misleading because they are not measured in units of species. The Shannon index is derived from information theory, and the units are *bits*. The Simpson diversity index is a probability. Therefore, Jost and others suggest each index should be converted into the `effective number of species`, and provide simple formulas to do so. Effective number of species (ENS) indices are thought to more intuitively reflect the species richness and diversity of communities. In a perfectly even community, ENS is equal to the species richness. In uneven communities, ENS is always smaller than the species richness. To calculate ENS, the exponential of Shannon diversity can be calculated or the reciprocal of Simpson diversity.

Mark Hill argued that these ENS diversity metrics can be united into a single family of generalized entropy metrics that are parameterized by the variable q, and this concept was recently revived by Jost and others. These metrics are referred to as `Hill numbers`. q essentially controls the sensitivity of each metric to species relative abundances, which quantifies how much the measures discount rare species when calculating diversity. This family of metrics includes species richness ($q = 0$), Shannon ENS ($q = 1$), and Simpson ENS ($q = 2$). All three metrics of Hill numbers can be calculated using the package `hillR`.

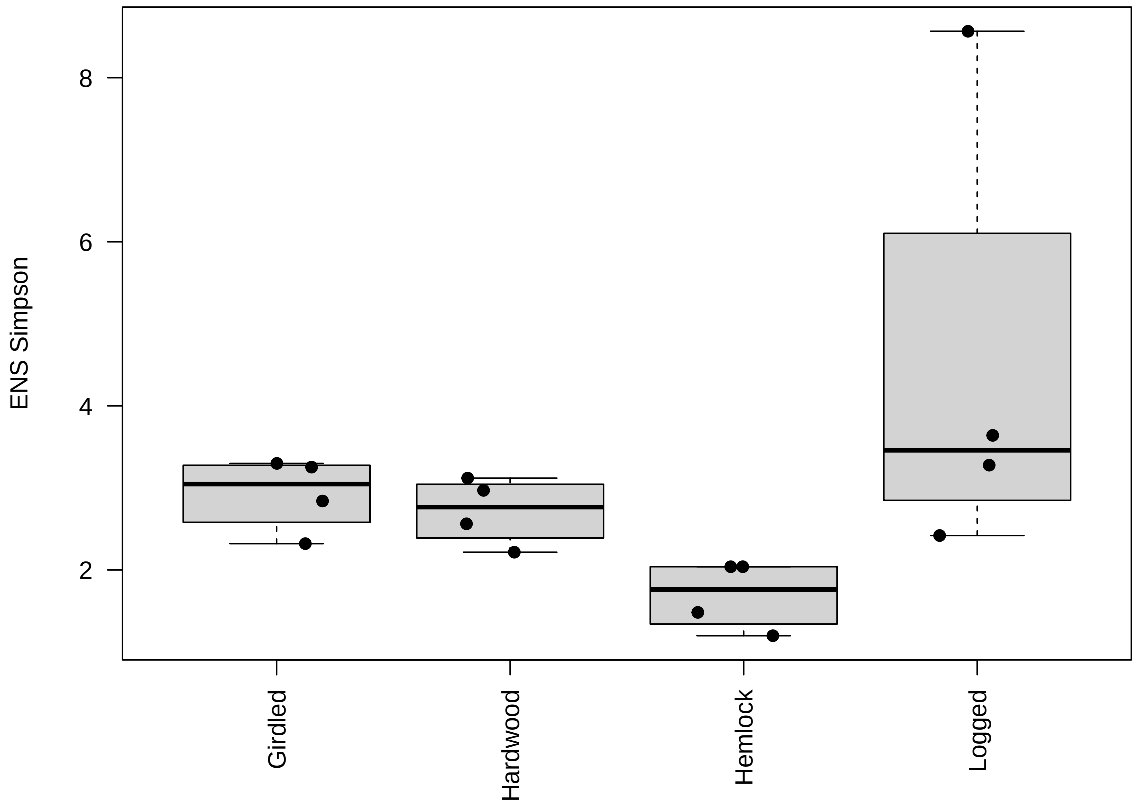
```
ants3$hill.rich <- hill_taxa(ants3[,6:38], q = 0, MARGIN = 1)
par(mar=c(6,4,2,2))
boxplot(hill.rich ~ treatment, data = ants3,
        xlab = "", ylab = "Species Richness",
        cex.axis = 1, las = 2)
stripchart(hill.rich ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```




```r
ants3$hill.shan <- hill_taxa(ants3[,6:38], q = 1, MARGIN = 1)
par(mar=c(6,4,2,2))
boxplot(hill.shan ~ treatment, data = ants3,
        xlab = "", ylab = "ENS Shannon",
        cex.axis = 1, las = 2)
stripchart(hill.shan ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```




```
ants3$hill.simp <- hill_taxa(ants3[,6:38], q = 2, MARGIN = 1)
par(mar=c(6,4,2,2))
boxplot(hill.simp ~ treatment, data = ants3,
        xlab = "", ylab = "ENS Simpson",
        cex.axis = 1, las = 2)
stripchart(hill.simp ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```



Hill, M.O. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology, 54: 427-432.

Jost, L. 2006. Entropy and diversity. Oikos, 113: 363-375.

Chao, A., Chiu, C.H. and Jost, L., 2014. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. Annual Review of Ecology, Evolution, and Systematics, 45: 297-324.

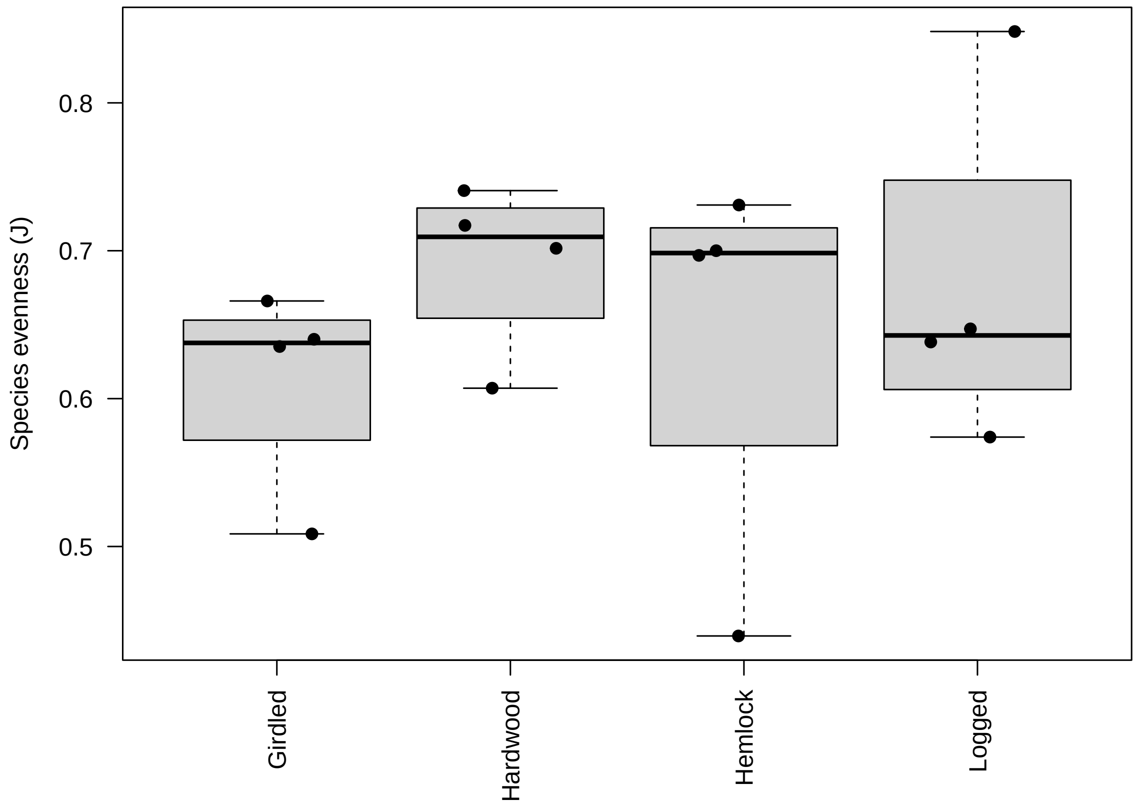
Evenness

Evenness measures the relative abundance of each type. This metric reflects how even a community is based on the species abundance distribution. Communities with high evenness tend to have the majority of species represented at fairly equal abundances. Communities with low evenness have one or two very abundant species, while the remaining species are present at low abundances. Therefore, this metric provides some information about species dominance in communities. There are three commonly used species evenness metrics, and each are derived from a species diversity index. All three metrics can be calculated using the `chemodiv` package.

Pielou's *J*

Pielou's J is derived from the Shannon diversity index. It is calculated as a ratio of the Shannon index to the maximum Shannon index (i.e., where all species have the same relative abundance). This metric ranges from 0-1, where 1 means that all species in the community have the same relative abundances.

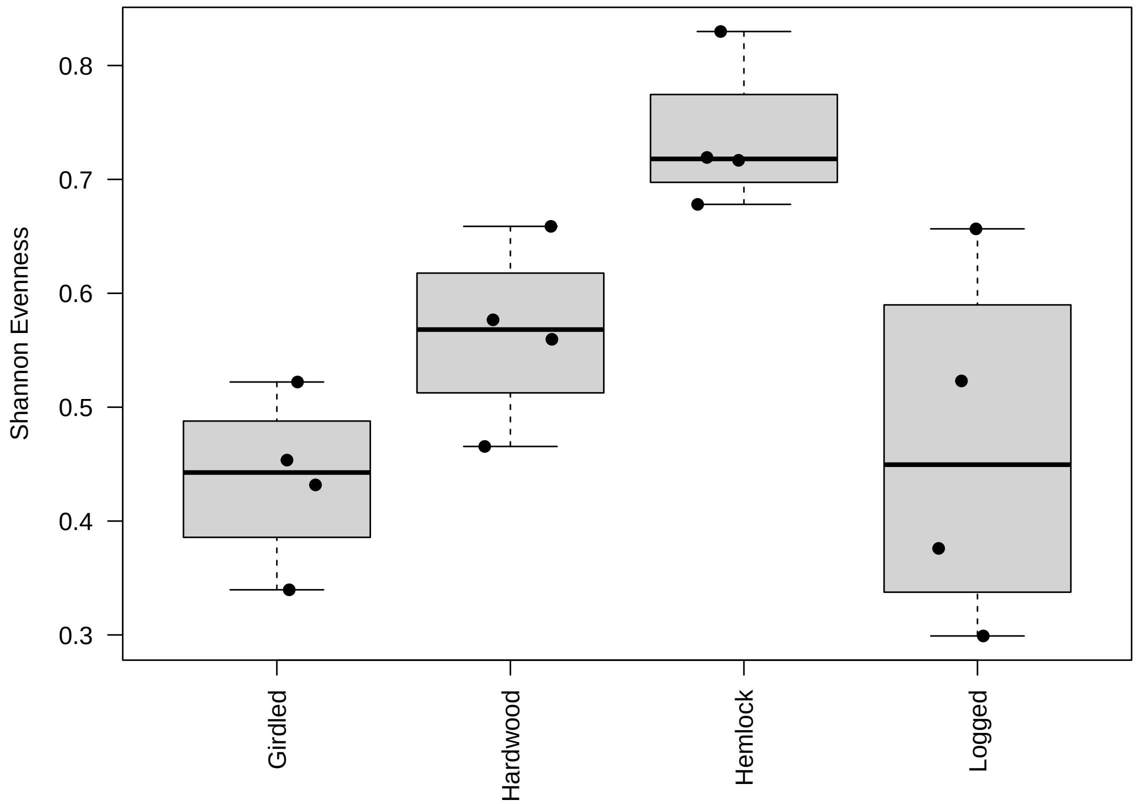
```
ants3$sp.evenJ <- calcDiv(ants3[,6:38], type = "PielouEven")
par(mar=c(6,4,2,2))
boxplot(sp.evenJ$PielouEven ~ treatment, data = ants3,
        xlab = "", ylab = "Species evenness (J)",
        cex.axis = 1, las = 2)
stripchart(sp.evenJ$PielouEven ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```



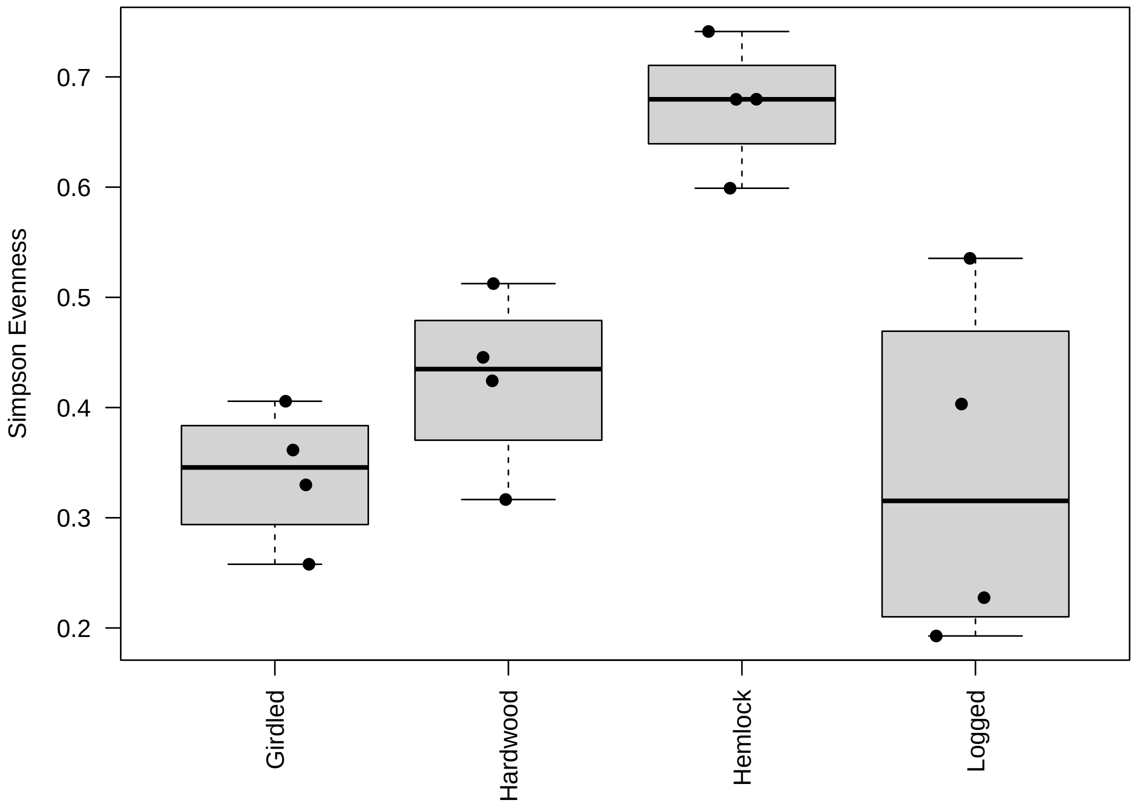
Species evenness derived from Hill numbers

Species evenness can be derived from ENS Shannon and ENS Simpson indices. These metrics are calculated by dividing the ENS indices by the observed number of species.

```
ants3$sp.evenSh <- calcDiv(ants3[,6:38], type = "HillEven", q = 1)
par(mar=c(6,4,2,2))
boxplot(sp.evenSh$HillEven ~ treatment, data = ants3,
        xlab = "", ylab = "Shannon Evenness",
        cex.axis = 1, las = 2)
stripchart(sp.evenSh$HillEven ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```




```
ants3$sp.evenSp <- calcDiv(ants3[,6:38], type = "HillEven", q = 2)
par(mar=c(6,4,2,2))
boxplot(sp.evenSp$HillEven ~ treatment, data = ants3,
        xlab = "", ylab = "Simpson Evenness",
        cex.axis = 1, las = 2)
stripchart(sp.evenSp$HillEven ~ treatment, data = ants3, pch = 19, add = TRUE,
           vertical = TRUE, method = "jitter", jitter = 0.2)
```


Simpson Evenness
0.7
0.6
0.5
0.4
0.3
0.2
Girdled
Hardwood
Hemlock
Logged

R Activity

1. We will use another open source data set from the NSF Harvard Forest Long-term Ecological Research (LTER) site. These data are spiders collected in the Hemlock Removal Experiment. Remember, this experiment includes four treatments (Hemlock girdled, Hemlock logged, Hemlock control, and Hardwood control) each replicated across two ($n = 2$) 90 x 90 m plots. Load the data into R. We will characterize spider communities among these four treatments and between sampling methods.

2. Before calculating the diversity metrics, you will have to do some data wrangling. First, create a new variable `abundance` by summing the counts of adult male and female spiders. Next, change the data set from `long` format to `wide` format using spider genus as the taxonomic resolution (i.e., each column should be a spider genus). Be sure that the new data frame includes the appropriate predictor and nuisance variables to assess differences among treatments and sampling methods. Then, remove any columns that do not have count data (some genera are indicated as immatures with imm), as well as three columns with unidentified spiders (LinytoID, LinyToID, and unk_toID). Lastly, change the variables `block`, `plot`, `treatment`, and `sampling method` to factors. Provide a summary of the data set. How many spider genera were collected?

3. Provide a short description of the data set. What are the response variables? What are the predictor variables? Nuisance variables? How many observations are there in the data set? Any missing data?

4. What are the three most abundant spider genera?

5. Calculate spider `genera richness` for each sample, add this new variable to the data set, and create a boxplot that shows spider `genera richness` as a function of `treatment`.

6. Calculate another diversity metric of your choice and create a boxplot that shows spider `genera diversity` as a function of `treatment`.

7. Create a boxplot that shows spider `genera diversity` as a function of `sampling method`.

8. Fit a regression modeling spider `genera diversity` as a function of `treatment` and `sampling method`, include the interaction. You will need to use materials from previous lectures and activities to determine the appropriate structure of the model. Provide the appropriate outputs to evaluate the model.

9. Write 1-2 sentences interpreting the results.