Rarefaction Example

- Consider this dataset:

<table>
<thead>
<tr>
<th></th>
<th>sp1</th>
<th>sp2</th>
<th>sp3</th>
<th>sp4</th>
<th>sp5</th>
<th>sp6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sam2</td>
<td>138</td>
<td>5</td>
<td>2</td>
<td>260</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>sam3</td>
<td>115</td>
<td>7</td>
<td>12</td>
<td>325</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>sam4</td>
<td>125</td>
<td>3</td>
<td>4</td>
<td>190</td>
<td>4</td>
<td>27</td>
</tr>
</tbody>
</table>

- Where is diversity highest?
  - Shannon
    - sam1 sam2 sam3 sam4
      - 4 6 6 6
    - 1.0375911 0.9176461 0.9908044 1.0397044

- What about rarefied diversity?
  - rarefy(community, sample=10)

<table>
<thead>
<tr>
<th></th>
<th>sp1</th>
<th>sp2</th>
<th>sp3</th>
<th>sp4</th>
<th>sp5</th>
<th>sp6</th>
<th>∑  S</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>sam2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>sam3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>sam4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

- Original matrix:

<table>
<thead>
<tr>
<th></th>
<th>sp1</th>
<th>sp2</th>
<th>sp3</th>
<th>sp4</th>
<th>sp5</th>
<th>sp6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sam2</td>
<td>138</td>
<td>5</td>
<td>2</td>
<td>260</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>sam3</td>
<td>115</td>
<td>7</td>
<td>12</td>
<td>325</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>sam4</td>
<td>125</td>
<td>3</td>
<td>4</td>
<td>190</td>
<td>4</td>
<td>27</td>
</tr>
</tbody>
</table>

- Rarefied matrix

<table>
<thead>
<tr>
<th></th>
<th>sp1</th>
<th>sp2</th>
<th>sp3</th>
<th>sp4</th>
<th>sp5</th>
<th>sp6</th>
<th>∑  S</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>sam2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>sam3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>sam4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

- Repeated 1,000 times, the average S

<table>
<thead>
<tr>
<th></th>
<th>sp1</th>
<th>sp2</th>
<th>sp3</th>
<th>sp4</th>
<th>sp5</th>
<th>sp6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sam2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sam3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>sam4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Species Abundance Curves

- Plot of rank abundance (x-axis) vs abundance or P_i (y-axis).
- More diverse communities lack numerically dominant species, flatter line.

Functions:
- rankabundance, rankabuncomp in the BiodiversityR package
- rankabundance(community)
  - Can use either raw abundance or proportion data
- rankabuncomp allows for comparison among factors
- rankabunplot will plot the results

```
rank_data <- rankabundance(Brier_Ck)
rankabunplot(rank_data)
```
Community Similarity or Dissimilarity

- Community similarity indices quantify similarity among two samples.
- For a full community matrix do all possible pairwise comparisons among communities.

- Symmetrical vs. asymmetrical metrics
  - Are shared zeros indicators of actual similarity?
  - Community data typically uses asymmetric metric

- Qualitative vs. quantitative
  - Ordinal vs categorical, bionmical etc.

- Q mode vs. R mode
  - Q: Are actual objects being compared (question is how similar are A and B)
    - E.g.: how similar is community A and B
  - R: Are relationships or dependence of measures of interest (question is if A is correlated with B)
    - E.g. Is species X abundance correlated with temperature

- Functions:
  - `vegdist` (vegan package), `dsvdis` (labdsv package), `daisy` (cluster package), `designdist` (vegan)

- Very often raw abundance data can be used
  - Variable depending on properties of metric, look at how each is calculated

- Most are bound (0-1 range)
  - Conversion to similarity or dissimilarity
    - `sim_matrix<-vegdist(Brier_Ck)`
    - `dsim_matrix<-1-sim_matrix`
Quantitative Indices of Similarity (0-1.0)

\[ Ruzicka(PSI) = \sum_{i=1}^{s} \min P_i \]
Where \( P_i \) is the proportion of the community composed of species i.

\[ Bray - Curtis = \sum_{i=1}^{s} \frac{(x_{ij} - x_{ik})}{x_{ij} + x_{ik}} \]
Where \( x_i \) is the abundance of species i in community j.

Both are bound
Typically log transform data

Quantitative Indices of Similarity (unbound)

\[ Euclidian = \sqrt{\sum_{i=1}^{s} (x_{ij} - x_{ik})^2} \]

\[ Manhattan = \sum_{i=1}^{s} |x_{ij} - x_{ik}| \]

Typically done without transformation. Some use presence/absence matrix with these metrics. No upper limit.
Qualitative Indices of Similarity (0-1.0)

Steinhaus = 1 - \( \frac{a}{a + b + c} \)

Sorensen = 1 - \( \frac{2a}{(2a + b + c)} \)

Where:
- \( a \) = number of species in both communities,
- \( b \) = number of species unique to community 1,
- \( c \) = number of species unique to community 2.

Both convert data to presence/absence.

Can be converted to dissimilarity by 1-Steinhaus or Sorensen.

Defining your own function

- Function `designdist` in the vegan package allows you to define any similarity index.

...review of 24 measures of beta diversity
**Assignment**

- We will be using the bee gut microbiome data from the paper read for class today. The data is available in datadryad:
  - [https://doi.org/10.5061/dryad.r02r1](https://doi.org/10.5061/dryad.r02r1)
- The data is presented in three files:
  - Frequency of OTUs ("species") by library("samples")
  - Taxonomic information for each OTU (species)
  - Sample information (which treatment, etc.)
- I did some filtering and processing to match what was written in the paper (eliminated mitochondria and chloroplasts, combined the "Naturals" groups etc.)
- The processed data are given to you as:
  - `otu_table.csv` – samples in rows by OTU in columns
  - `sample_meta.csv` – Factors for colony ID, Population, and Treatment

**Assignment**

- 1. Eliminate species (OTUs, the columns) with zero occurrences. What is the new total OTUs? Does that match what is in the paper?
- 2. Calculate the total OTUs per sample. What is the minimum of this among all the samples? Does this match what is in the paper?
- 3. Use the `diversitycomp` function to calculate species richness, Shannon diversity, and evenness for samples pooled (method="pooled") among treatments (factor1="Treatment").
- 4. Plot a species accumulation curve (random method) for the whole dataset. Use `specaccum` function.
- 5. Calculate a Bray-Curtis similarity matrix for the whole dataset. What is the mean Bray-Curtis dissimilarity?
- 6. Create a rarefied matrix using `rarefy`, with the sample size set to the minimum total number of OTUs in a sample (item #2 above).
- 7. Repeat 3, 4, and 5 above with the rarefied matrix. Discuss differences in your synthesis.