Rarefying a dataset (most common use)
• Original matrix (community) with row sums

<table>
<thead>
<tr>
<th>Sample</th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
<th>Species 4</th>
<th>Species 5</th>
<th>Species 6</th>
<th>Species 7</th>
<th>Species 8</th>
<th>Species 9</th>
<th>Species 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>13</td>
<td>13</td>
<td>42</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>28</td>
<td>117</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 2</td>
<td>30</td>
<td>33</td>
<td>14</td>
<td>19</td>
<td>24</td>
<td>26</td>
<td>5</td>
<td>151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 3</td>
<td>2</td>
<td>21</td>
<td>22</td>
<td>11</td>
<td>13</td>
<td>3</td>
<td>13</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 4</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>17</td>
<td>23</td>
<td>14</td>
<td>1</td>
<td>79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 5</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>34</td>
<td>34</td>
<td>21</td>
<td>115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 6</td>
<td>30</td>
<td>10</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>18</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 7</td>
<td>32</td>
<td>4</td>
<td>11</td>
<td>33</td>
<td>26</td>
<td>29</td>
<td>11</td>
<td>146</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 8</td>
<td>16</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>20</td>
<td>11</td>
<td>77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 9</td>
<td>24</td>
<td>1</td>
<td>34</td>
<td>43</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td>133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 10</td>
<td>24</td>
<td>8</td>
<td>13</td>
<td>7</td>
<td>31</td>
<td>28</td>
<td>42</td>
<td>153</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Totals
<table>
<thead>
<tr>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
<th>Species 4</th>
<th>Species 5</th>
<th>Species 6</th>
<th>Species 7</th>
<th>Species 8</th>
<th>Species 9</th>
<th>Species 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>31</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>26</td>
<td>90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• raref(your_dataset, 90)
• Note that row sums are <=90

Rarefaction to estimate S for group of samples (specaccum)

<table>
<thead>
<tr>
<th>Site</th>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
<th>Species 4</th>
<th>Species 5</th>
<th>Species 6</th>
<th>Species 7</th>
<th>Species 8</th>
<th>Species 9</th>
<th>Species 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.65623</td>
<td>0.20289</td>
<td>0.05186</td>
<td>0.01221</td>
<td>0.00224</td>
<td>2.4e-04</td>
<td>0.00001</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

Values of diversity as groups of individuals (16 at a time, the row average) are selected.
Estimates diversity, does not produce a rarefied dataset.

Distance matrix analyses

What can you do with a large triangular similarity matrix?
Recall that each element in the triangular matrix is a similarity measure between two objects (samples in this case).
What if your hypothesis is that two groups of samples are different?

Earlier approaches
• Can be used as descriptive statistics with subjective levels of “similar” or “not similar”.
• Next week – cluster analyses and dendrograms. Visual but also largely descriptive.
• Future classes – ordinations (also descriptive!)
• Traditional statistical methods are not appropriate due to a lack of independence among samples.
Patterns of interest

Three techniques to test for these patterns in a similarity matrix.

- **ANOSIM** – Analysis of similarity
- **MRPP** – Multi response permutation procedure
- **Multivariate Analysis of Variance for Distance Matrices**
- **Analysis of Molecular Variation (AMOVA)**

- These all work for any application where you can calculate some distance among units and want to look for differences among groups of units.
  - Eg. morphological data, spatial layout data, behavioral data, molecular data

**ANOSIM**

- Dissimilarity matrix divided up by grouping variable(s)
- Between groups vs among comparisons of interest
  - Rank all similarities
  - If groups are more similar than random (null), then mean rank similarity within a group should be less than among groups.

\[
R = \frac{(B - W)}{N(N - 1)} \div 4
\]

- **B** – mean between group ranks
- **W** – mean within group ranks
- **N** – number of samples

**ANOSIM**

- Calculate observed R statistic
  - Ranges from -1 to 1
  - 0 = no pattern
  - 1 = all within group ranks less than among group ranks
  - -1 all within group ranks greater than among groups

Test of significance

- How do you test the significance of the R value?
- “What is the probability of obtaining an R value equal to or greater than the observed?”
- Permutation test:
  - Randomize the observed data, calculate R, repeat
  - Compare observed R to distribution of randomized R
- Use caution in how you set up your randomization!

Need variable (factor) grouping your data

- Load the data as you normally would.
- You will need to separate the community data from the variables describing your groups.
- This code yields four objects
  - Alldata: everything in the original data file
  - Year and creek: variables with year and creek factors
  - Community: community matrix

<table>
<thead>
<tr>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.6</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>6</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>7</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>8</td>
<td>0.7</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>24</td>
<td>41</td>
<td>32</td>
<td>24</td>
<td>7</td>
<td>14</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>41</td>
<td>32</td>
<td>24</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>32</td>
<td>32</td>
<td>24</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>45</td>
<td>41</td>
<td>32</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>24</td>
<td>24</td>
<td>32</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>7</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Similarity matrix, sites 1-5 and 6-10 represent groups.
Null: groups are more similar than at random.

Mean similarity within sites 1-5 = 0.8
Mean similarity within sites 6-10 = 0.7
Mean similarity between = 0.6

R = (29.6 - 15.3)/(10(9)/4)
R = 0.64

Permutation test:

- Randomize the observed data, calculate R, repeat
- Compare observed R to distribution of randomized R

Mean rank within sites 1-5 and sites 6-10 = 15.3
Mean similarity between = 29.6
ANOSIM (vegan package)

R Code:

```r
# dissimilarity (community proportion, method="bray")
dist = vegdist(community_proportion, method='bray')
# year_ano = anosim (dist, year, permutations=5000)
year_ano = anosim(dist,year, permutations=5000)
plot(year_ano)
plot a frequency distribution of the permuted R values
hist(year_ano$perm)
```

Options

- # permutations
- Strata (block groups for permutations)
- Requires dissimilarity matrix
- You can either supply a dissimilarity (triangular) matrix, or community matrix and specify what kind of dissimilarity to produce (method=)

Output

- Observed R, vector of permuted Rs, significance of observed R

ANOSIM output

Dissimilarity: bray
ANOSIM statistic R: 0.4576
Significance: 0.001
Based on 999 permutations

![ANOSIM output](image)

ANOSIM

- Can use any similarity measure on virtually any data
- Converts to rank similarity (information loss)
- Two way designs can be tested (see Primer ver. 6)
- Use of ranks means differences in the amount of variability (dispersion) not detected
- Pairwise tests often not built in (but see Primer)
- Often done in conjunction with Non Metric Multidimensional scaling (NMDS) ordination

ANOSIM

From the writer of the procedure in R: “I don’t quite trust this method. Somebody should study its performance carefully. The function returns a lot of information to ease further scrutiny.”

- Very popular technique. However, there are more powerful alternatives.
Multiple Response Permutation Procedure (MRPP)

- Conceptually similar to ANOSIM but does not use ranks
- Calculated statistic is $\Delta$ = weighted average within group similarity
- Permutations – randomize matrix, recalculate $\Delta$
- Significance of $\Delta$ assessed by distribution of permuted $\Delta$ scores.
- Code
  ```r
  dist = vegdist(community_proportion, method="bray")
  year_mrpp = mrpp(dist, year, permutations=5000)
  ```
- Options
  - As in ANOSIM - permutations, strata
  - Weights – three methods for weighting sample size for $\Delta$ calculation

MRPP

- Avoiding ranks means
  - MRPP can detect differences in mean as well as dispersion.
  - Biases in distance metric will be more pronounced than in ANOSIM
- Output
  - Observed $\Delta$
  - Vector of permuted $\Delta$ and expected $\Delta$
  - Significance of observed $\Delta$
  - $A$ - estimate of the proportion of distances explained by the factor


Multivariate Analysis of Variance Using Distance Matrices

- Recently described, generally superior to ANOSIM in all ways, MRPP in most ways
- Robust alternative to MANOVA, sometimes called permutational MANOVA
- Using this technique with one dependent variable and a Euclidean distance matrix should yield same results as a traditional ANOVA
- Also similar to AMOVA

SS_t = sum of squared differences between all observations and the overall centroid.

SS_w = sum of squared differences between group observations and group centroid.

SS_a = SS_t - SS_w

Pseudo-F = ratio of SS_a to SS_w
Multivariate Analysis of Variance Using Distance Matrices

- Permute distance matrix to generate distribution of Pseudo-F and probability of observed pseudo-F.
- Traditional ANOVA output (pseudo-F) and partitioning of variance
- Higher level ANOVA designs, including continuous variable factors and interactions
- Use of continuous variables
- Pairwise comparisons?

Options

- Permutations
- Distance metric (procedure works with raw community dataset not a triangular matrix)
- Strata
- Model, eg. community ~ site * time

A note about R models

- There is standard R language for model formulas
  - Assume the following variables
    - Y – dependent variable
    - X1, X2 – continuous independent variables
    - F1, F2 – discrete independent variables (factors)
      - `Y~X`
        - Linear regression
      - `Y~X+F`
        - Multiple regression with interaction
      - `Y~F1+F2`
        - Single factor ANOVA
      - `Y~F1:F2`
        - Two factor ANOVA without interaction
      - `Y~F1*F2`
        - Two factor ANOVA with interaction
- See help.start() section 11 – Statistical Models in R
AMOVA

- Very similar to analysis of variance using distance metric (in fact, using a Euclidean distance matrix should yield identical results).

- First developed for analyzing mtDNA haplotypes
  - Distance matrix was pairwise steps in a network
  - Grouping variable was population


AMOVA

- Two amova functions, both named “amova”
- Package ade4
  - Haplotypic approach – requires 1) matrix of haplotype frequency by population, 2) Euclidian distance among haplotypes in network, and 3) assignment of populations to groups
  - Separate function randtest tests significance through permutation.
- Package pegas
  - More general approach – requires genetic distance matrix and a factor.

Reading

- Sample script and dataset
- Two papers
- Text: Chapter 3, information on similarity measures
- For information on formulas in R, review chapter 11 - help.start()

Assignment

- New dataset (spaeth.csv)
- Rarefy dataset to 50 individuals per sample (rrarefy)
- Perform ANOSIM and MRP to test for community differences
  - Test for year and creek differences on proportional data after rarefaction

- Multivariate Analysis of Variance for Distance Matrix
  - Use proportional after rarefaction
  - Test for year and creek differences as above but in one analysis and include an interaction term
  - Test for year and season differences while only doing randomizations within creek