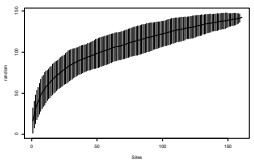
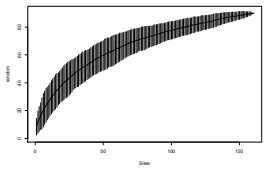
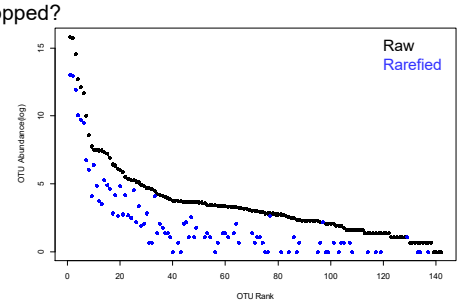
```
Treatment       n richness
  Naturals     74      127
  Pool exposed 80       87
  Pools         6       32

Treatment       n    Shannon
  Naturals     74 1.13720178
  Pool exposed 80 1.17596629
  Pools         6 0.82936825

Treatment       n  Jevenness
  Naturals     74 0.23475596
  Pool exposed 80 0.26332075
  Pools         6 0.23930509

Mean Brav Curtis 0.4614
```

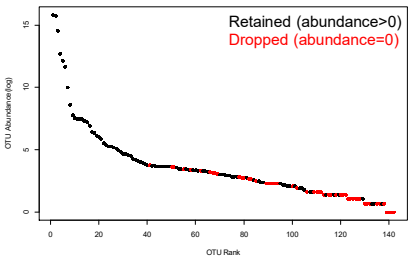


```
Treatment       n richness
  Naturals     74       73
  Pool exposed 80       60
  Pools         6       20

Treatment       n    Shannon
  Naturals     74 1.19500029
  Pool exposed 80 1.23457736
  Pools         6 0.84805103

Treatment       n  Jevenness
  Naturals     74 0.27852502
  Pool exposed 80 0.30153236
  Pools         6 0.28308639

Mean Bray Curtis 0.2650
```



- Everyone had slightly different results…rrarefy will produce different results each time.
- Richness declined by 30-40%, which species were dropped?

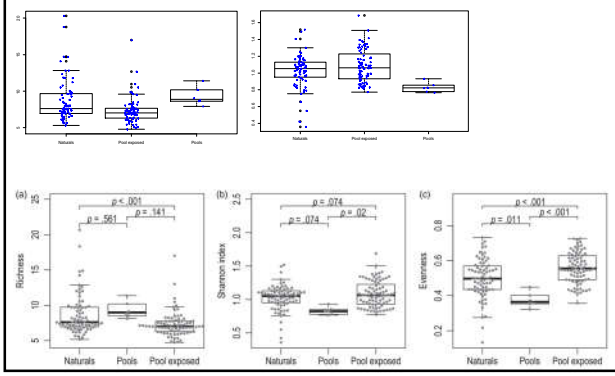




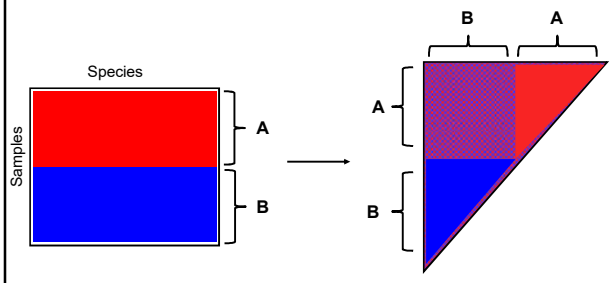
- Bray Curtis dissimilarity dropped from 0.46 to 0.27. Most of the increase in similarity came from controlling abundance of common species, not the loss of rare species.
- Use a qualitative metric to look at changes in species composition
- Mean steinhaus dissimilarity: raw 0.653 vs. 0.506 rarefied

- What if we retain only the 60 most abundant OTU (dropping 82 rare OTU)?
  - Mean Bray Curtis
    - Raw 0.4614
    - Rarefied 0.2649
  - Mean Steinhaus
    - Raw 0.6070
    - Rarefied 0.4774
- So, eliminating almost all of the rare species did not change Bray Curtis at all. Steinhaus similarity increased, and the difference between raw and rarefied decreased.

- Doing 1000 rarefactions of the data as described…results are essentially identical to what is in the paper.



Distance matrix analyses



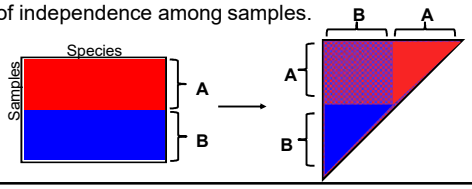
What can you do with a large triangular similarity matrix?

Recall that each element in the triangular matrix is a similarity measure between two objects (samples in this case).

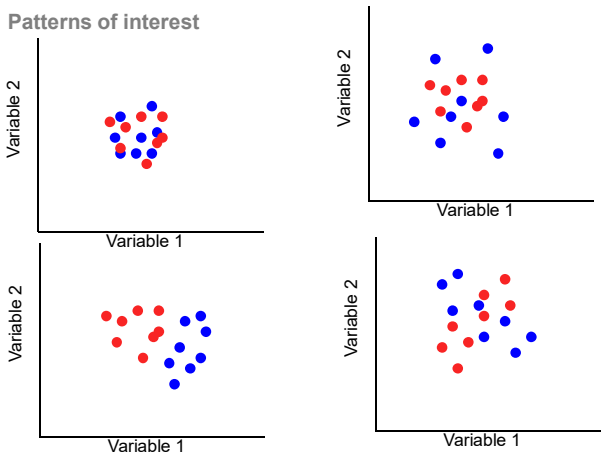
What if your hypothesis is that two groups of samples are different?

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



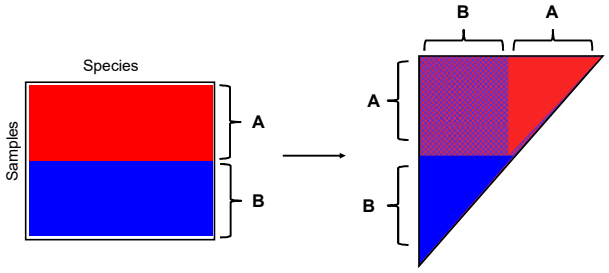
Three techniques to test for these patterns in a similarity matrix.

- **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.



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

$$R = \frac{(B - W)}{\frac{N(N-1)}{4}}$$

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

Clarke, K. R. (1993). Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18, 117-143.

	10	9	8	7	6	5	4	3	2	1
1	0.5	0.6	0.4	0.5	0.6	0.8	0.7	0.7	0.8	
2	0.6	0.4	0.5	0.6	0.7	0.7	0.8	0.9		
3	0.4	0.5	0.5	0.6	0.7	0.8	0.9			
4	0.5	0.3	0.4	0.5	0.6	0.9				
5	0.5	0.6	0.6	0.5	0.7					
6	0.8	0.7	0.9	0.7						
7	0.9	0.8	0.7							
8	0.7	0.8								
9	0.9									
10										

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

	10	9	8	7	6	5	4	3	2	1
1	32	24	41	32	24	7	14	14	7	
2	24	41	32	24	14	14	7	1		
3	41	32	32	24	14	7	1			
4	32	45	41	32	24	1				
5	32	24	24	32	14					
6	7	14	1	14						
7	1	7	14							
8	14	7								
9	1									
10										

Rank similarities.

	10	9	8	7	6	5	4	3	2	1
1	0.5	0.6	0.4	0.5	0.6	0.8	0.7	0.7	0.8	
2	0.8	0.4	0.8	0.6	0.7	0.7	0.4	0.9		
3	0.4	0.7	0.6	0.6	0.7	0.8	0.7			
4	0.5	0.3	0.4	0.5	0.6	0.9				
5	0.5	0.6	0.6	0.5	0.7					
6	0.8	0.5	0.9	0.7						
7	0.6	0.8	0.7							
8	0.7	0.8								
9	0.8									
10										

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

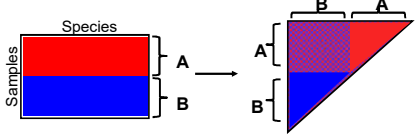
	10	9	8	7	6	5	4	3	2	1
1	39	31	43	36	25	11	18	20	5	
2	9	40	6	30	22	21	41	1		
3	42	16	28	24	14	4	19			
4	38	45	44	34	29	3				
5	35	27	26	33	13					
6	10	37	2	17						
7	32	7	23							
8	15	8								
9	12									
10										

Mean rank within 1-5 and 6-10 = 15.3
Mean similarity between = 29.6

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

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

site	year	creek	Species1	Species2	Species3	Species4	Species5	Species6	Species7	Species8	Species9	Species10
A_07_1	2007	A	2	87	0	16	20	3	1	0	9	21
A_07_2	2007	A	2	82	0	2	31	8	2	0	6	39
A_08_1	2008	A	1	19	0	9	12	18	100	0	8	14
A_08_2	2008	A	1	17	0	12	15	16	85	0	7	31
A_09_1	2009	A	2	44	0	13	36	24	5	0	8	1
A_09_2	2009	A	4	30	0	45	25	36	9	0	8	0
B_07_1	2007	B	1	38	2	11	8	20	1	2	1	20
B_07_2	2007	B	0	43	1	13	25	13	3	1	0	33
B_08_1	2008	B	1	7	1	5	0	1	96	1	4	11
B_08_2	2008	B	1	5	1	19	5	12	19	0	1	11

- Load the data as you normally would.
- You will need to separate the community data from the variables describing your groups.

```
sample_data<-read.csv(file="sample2.csv", row.names=1, header=TRUE)
year<-as.factor(sample_data$year)
creek<-sample_data$creek
community<-subset(sample_data,select=-c(year,creek))
```

- This code yields four objects
  - Sample_data: everything in the original data file
  - year and creek: variables with year and creek factors
  - Community: community matrix without factors

ANOSIM (vegan package)

- R Code:
  - `community_dist<-vegdist(community_proportion, method='bray')`
  - `year_ano<-anosim(community_dist,year, permutations=5000)`
  - `summary(year_ano)`
  - `plot(year_ano)`
  - `#plot a frequency distribution ofthe 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=)
  - Parallel – specify how many cores to use in parallel
- Output
  - Observed R, vector of permuted Rs, significance of observed R

ANOSIM

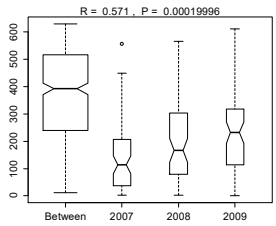
- Can use any similarity measure on virtually any data
- Converts to rank similarity (information loss)
- Two way designs can be tested (see Primer ver. 6)
- Use of ranks means differences in the amount of variability (dispersion) not detected
- Pairwise tests often not built in (but see Primer)
- Often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

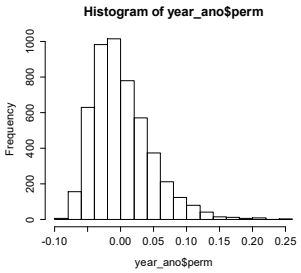
ANOSIM output

Dissimilarity: bray

ANOSIM statistic R: 0.4576
Significance: 0.001

Based on 999 permutations





Mean of rank similarities within each group and between all groups.

```
tapply(year_ano$dis.rank,year_ano$class.vec,mean)
```

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 Δ = weighted average within group similarity
- Permutations – randomize matrix, recalculate Δ
- Significance of Δ assessed by distribution of permuted Δ scores.
- Code

```
community_dist<-vegdist(community_proportion, method='bray')
year_mrpp<-mrpp(community_dist,year,permutations=5000)
year_mrpp
```

- Options
  - As in ANOSIM - permutations, strata
  - Weights – three methods for weighting sample size for Δ calculation

J. Van Sickle 1997. Using mean similarity dendrograms to evaluate classifications. *Journal of Agricultural, Biological, and Environmental Statistics* 2:370-388.

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 Δ
  - Vector of permuted Δ and expected Δ
  - Significance of observed Δ
  - 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
  - Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46.
  - McArdle, B.H. and M.J. Anderson. 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, **82**: 290–297.
- 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
  - Excoffier, L., P.E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 131(2):479-491.

Multivariate Analysis of Variance Using Distance Matrices

- 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?

Multivariate Analysis of Variance Using Distance Matrices

- Should be applicable any time ANOSIM or MRPP can be used.
- Should be more robust than ANOSIM or MRPP
- Higher level ANOVA designs, including interactions
- Code:

```
permanova<-adonis(community_proportion ~ creek*year,
method="bray", permutations=10000)
print(permanova)
```

- 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 ~ X1
        - Linear regression
      - Y ~ X1+X2
        - Multiple regression without interaction
      - Y~ X1*X2
        - Multiple regression with interaction
      - Y~F1
        - 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

Multivariate Analysis of Variance Using Distance Matrices

- Output
  - Standard ANOVA table with % variance accounted for by each variable (factor) and the residuals (error)
  - Observed and permuted pseudo-F
  - Species coefficients for each level of each factor

```
Number of permutations: 10000

Terms added sequentially (first to last)

           Df SumsOfSqs MeanSqs F.Model      R2    Pr(>F)
creek       5    0.4985 0.09969  1.9898 0.13887    0.0304 *
year        2    1.6113 0.80567 16.0804 0.44891 9.999e-05 ***
creek:year 10    0.5778 0.05778  1.1533 0.16098    0.3073
Residuals  18    0.9018 0.05010         0.25125
Total      35    3.5895                 1.00000
---
```

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

Excoffier, L., P.E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 131(2):479-491.

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 netowork, 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
- Papers
  - Caspers, B.A., F.C. Schroeder, S. Franke, W. Streich and C.C. Voigt. 2009. Odour-based species recognition in two sympatric species of sac-winged bats (*Saccopteryx bilineata, S. laptura*): combining chemical analyses, behavioral observations and odour preference tests. *Behavior Ecology and Sociobiology* **63**: 741-749.
  - Robidoux, M., P. Giorgio, and A. Derry. 2015. Effects of humic stress on the zooplankton from clear and DOC-rich lakes. Freshwater biology **60**: 1263-1278.
- Text: Chapter 3, information on similarity measures
- For information on formulas in R, review chapter 11 -help.start()

Assignment

- New dataset (spaeth.csv) – note that the factors you will use are in the first few columns.
- 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 also include an interaction term
  - Test for year and season differences while only doing randomizations within creek