Week 13: Multivariate statistics (Part 2)

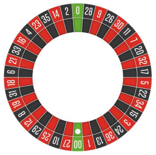
ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

1

Statistics vignette



- Imagine that you're a gambler w/ $10 at a casino
- You play a game where:

50%: +$1

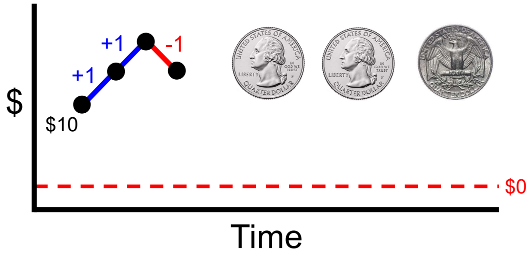
50%: -$1

- Play until (1) you go broke, (2) casino goes broke, or (3) you leave

2

Random walk

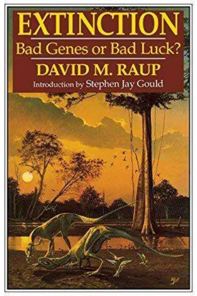
If you don't leave, which is more likely to happen first: (1) you go broke, or (2) casino goes broke?

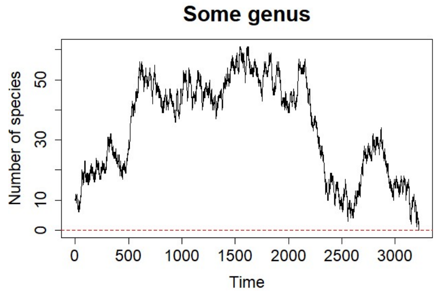


3

Gambler's Ruin

- Because of the lower absorbing boundary, all clades (e.g., genera) will eventually go extinct





4

Lecture outline

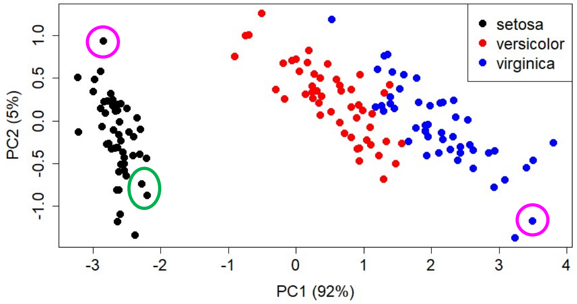
- Distance matrices
  - Euclidean distance
- Ordination continued
  - Principal coordinates analysis
  - Non-metric multidimensional scaling
- Classification
  - Cluster analysis
- Primer on multivariate regression

Distance matrices

Euclidean distance

One goal of PCA

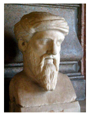
- Plot MV data in fewer dimensions, while preserving distances between observations



Similar SL, SW, PL, PW

Different SL, SW, PL, PW

What distance?

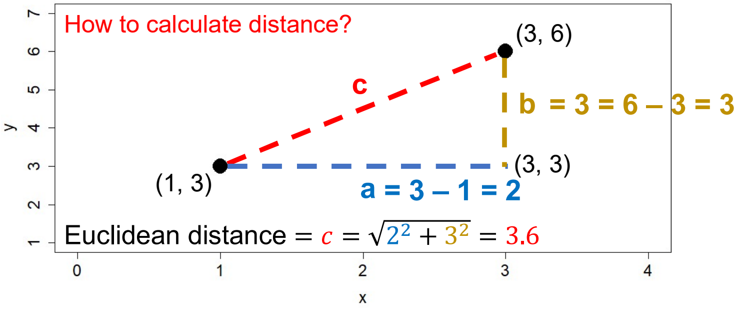


Pythagorean Theorem

$$a^2 + b^2 = c^2$$

$$c = \sqrt{a^2 + b^2}$$

- For PCA, **Euclidean distance**
- In two dimensions:



Euclidean distance $= c = \sqrt{2^2 + 3^2} = 3.6$

Euclidean distance

- Measures the distance between two points in ***any*** number of dimensions (n)

$$d_{i,j} = \sqrt{\sum_{k=1}^{n} \underbrace{(y_{i,k} - y_{j,k})}_{\text{Difference in x-, y-, z-, etc. dimension } (k)}^2}$$

c(1, 2, 3, 4)
c(3, 4, 5, 6) ⇒ c(2^2, 2^2, 2^2, 2^2) ⇒ $\sqrt{16}$ ⇒ 4

Euclidean distance

```
    Sepal.L  Sepal.W  Petal.L  Petal.W
1       5.1      3.5      1.4      0.2
2       4.9      3.0      1.4      0.2
```

$$\sqrt{(5.1-4.9)^2 + (3.5-3.0)^2 + (1.4-1.4)^2 + (0.2-0.2)^2}$$

Euclidean distance = 0.54

Distance matrices

- E.g., Euclidean dist. of first four plants in `iris`

```
     1     2     3     4
1 0.00  0.54  0.51  0.65
2 0.54  0.00  0.30  0.33
3 0.51  0.30  0.00  0.24
4 0.65  0.33  0.24  0.00
```

- Cell in i^{th} row and j^{th} column is distance between i^{th} and j^{th} observations
- Diagonal elements always zero (e.g., comparing observation to itself)
- Upper triangle is identical to lower triangle

Euclidean distance

```
    Sepal.L  Sepal.W  Petal.L  Petal.W
1       5.1      3.5      1.4      0.2
2       4.9      3.0      1.4      0.2
```

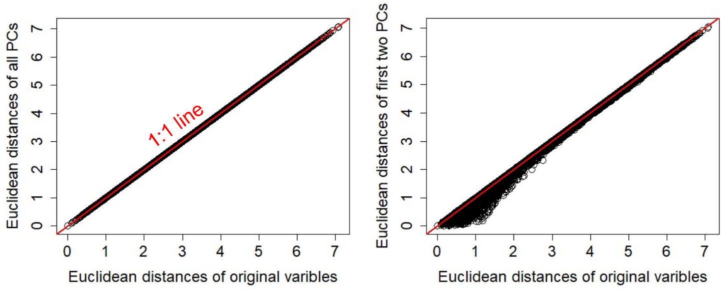
Euclidean distance = 0.54

```
        PC1      PC2      PC3      PC4
1     -2.68    -0.32     0.03      0.0
2     -2.71     0.18     0.21      0.1
```

Euclidean distance = 0.54

Exactly the same!

Shepard diagram


Euclidean distances of all PCs
0
1
2
3
4
5
6
7
1:1 line
0
1
2
3
4
5
6
7
Euclidean distances of original varibles
Euclidean distances of first two PCs
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
7
Euclidean distances of original varibles

Rigid rotation

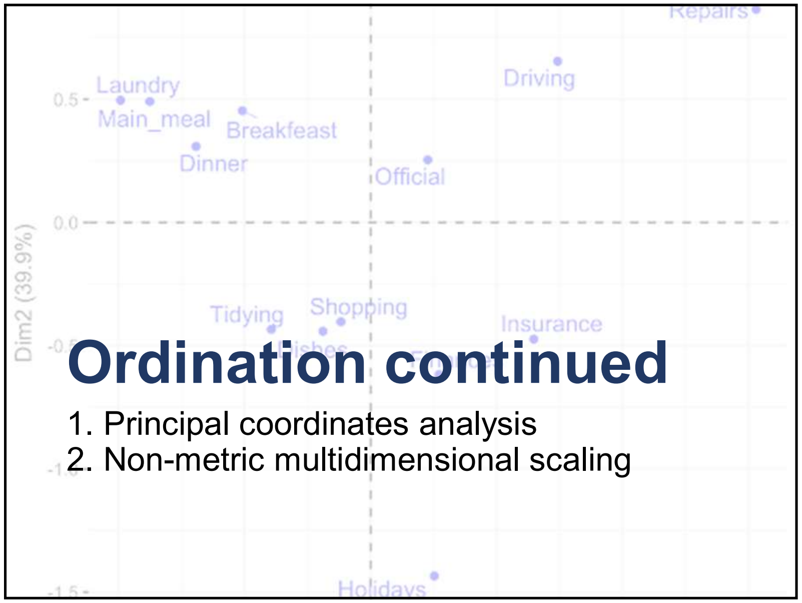
- PCA simply rotates the cloud of data points, **<u>preserving the distances between points,</u>** to align with axes of greatest variation (PCs)
- Euclidean distances will be identical only when **<u>all</u>** PCs are considered
- If data are scaled first, PCA preserves distances between scaled versions of data points

Different kinds of distances

- Euclidean distances work best with continuous data
- <u>Categorical data (e.g., comparing species abundances across sites)</u>
  - Chi-square
  - Bray-Curtis
- <u>Binary outcomes (e.g., stone tool type presence/absence across sites)</u>
  - Jaccard
  - Sørensen
- <u>Mixed continuous & categorical</u>
  - Gower

Gotelli & Ellison 2018, pg. 404 (there are *many* more)

Questions?



Ordination continued

1. Principal coordinates analysis
2. Non-metric multidimensional scaling

Goals of ordination (review)

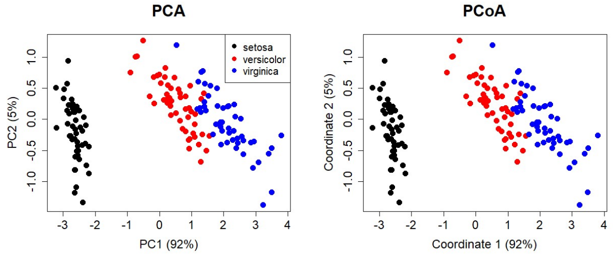
- Distill MV data into fewer variables/dimensions that can be plotted or analyzed
- Maintain (as best as possible) original distances between points
- Create new variables/axes, where the first one is fit through greatest variation, second one is fit through greatest residual variation, etc.
- New variables/axes are perpendicular and independent to each other

Principal coordinates analysis (PCoA)

- AKA **metric multidimensional scaling**
- Generalized ordination technique done on **any** distance matrix (e.g., chi-square, Jaccard)
- Specifically, does a **singular value decomposition** on the distance matrix (so slightly different from PCA)
- Unlike PCA, does not combine variables into super-variables (i.e., PCs); just plots MV data on smaller number of axes using distance matrix (i.e., PCA w/o the loadings)

Comparing PCA & PCoA

- PCoA done on Euclidean distance matrix produces the **same exact** scatter plot as PCA
- Use PCA instead of PCoA if you have continuous variables (loadings are interesting!)



Another PCoA example

- Hair & eye color of statistics students (`HairEyeColor`)

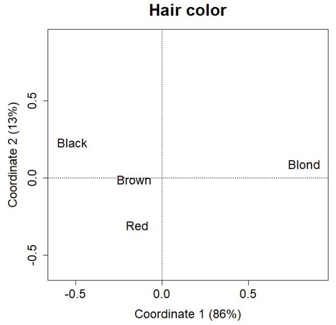
```
         Eye
Hair     Brown Blue Hazel Green
  Black     68   20    15     5
  Brown    119   84    54    29
  Red       26   17    14    14
  Blond      7   94    10    16
```

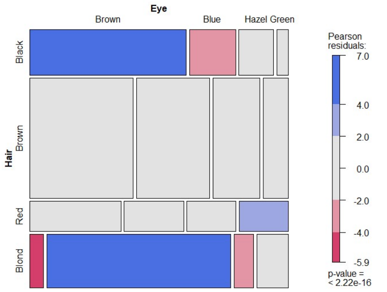
- Euclidean distance not appropriate for categorical variables, so use chi-square dist.

21

Hair color

- Points closer together have more similar eye color proportions





22

Or eye color

- Because PCoA doesn't use correlated variables for collapsing, can plot either rows (hair color) or columns (eye color)
- i.e., distance matrix can be calculated using either rows or columns

Hair color

```
      Black Brown Red
Brown   0.4
Red     0.7   0.3
Blond   1.3   1.0 1.0
```

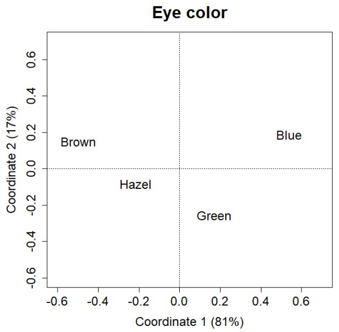
Eye color

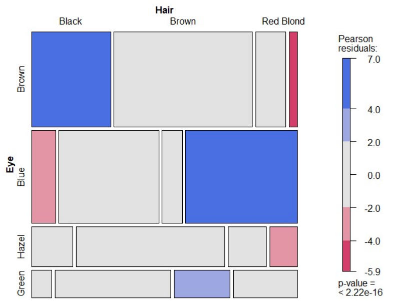
```
      Brown Blue Hazel
Blue    1.0
Hazel   0.4  0.8
Green   0.8  0.6   0.5
```

23

Eye color

- Points closer together have more similar hair color proportions

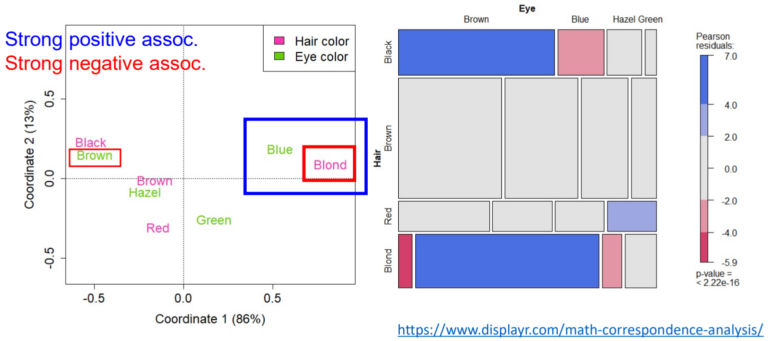




24

Can put on same plot!

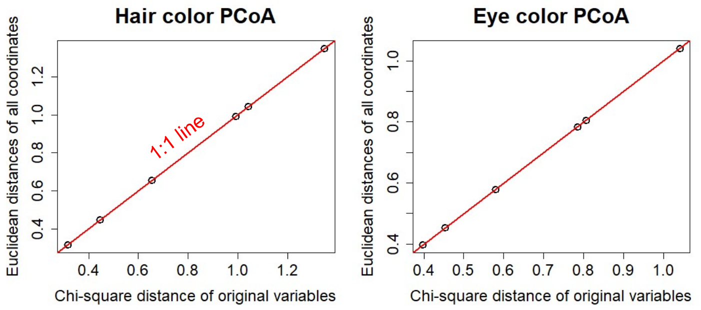
- PCoA done on chi-square distance matrix is **correspondence analysis** if, e.g., hair weighted by # individuals in each color


Strong positive assoc.
Strong negative assoc.

https://www.displayr.com/math-correspondence-analysis/

Shepard diagram

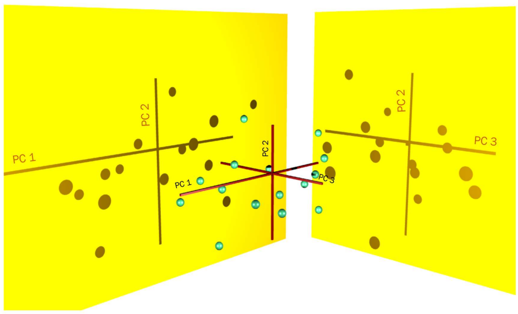
- Chi-square distances can be 100% preserved in Euclidean space



Questions?

What ordination does

- Projects MV data cloud onto 2D plane
- Only ever see a few dimensions of the MV data



Non-metric multidimensional scaling (NMDS)

- Instead of plotting a few out of many dimensions, NMDS forces MV data cloud into only a few axes (i.e., 2 or 3)
- Computer iteratively finds the solution (explained on next slide)
- **NO** singular value decomposition, so axes cannot be interpreted as % variance explained and are not fit through greatest variation
- Works with **any** distance matrix (like w/ PCoA)

29

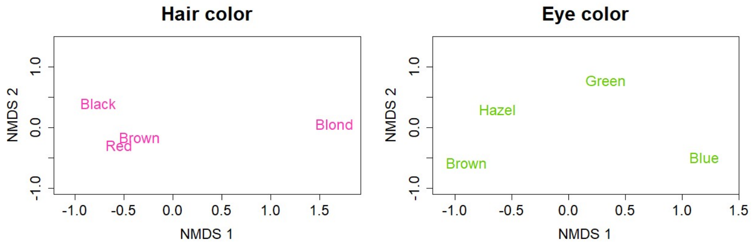
How NMDS works

1. Generate a distance matrix of your data
2. Pick number of dimensions to plot (e.g., 2)
3. Place observations' points on plot, using some starting configuration (can choose this)
4. Quantify how well Euclidean distances between NMDS points correspond w/ distance matrix (cf. Shepard plot)
5. Adjust points to increase correspondence
6. Repeat Steps 4 & 5 until correspondence is maximized

30

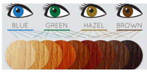
NMDS plots

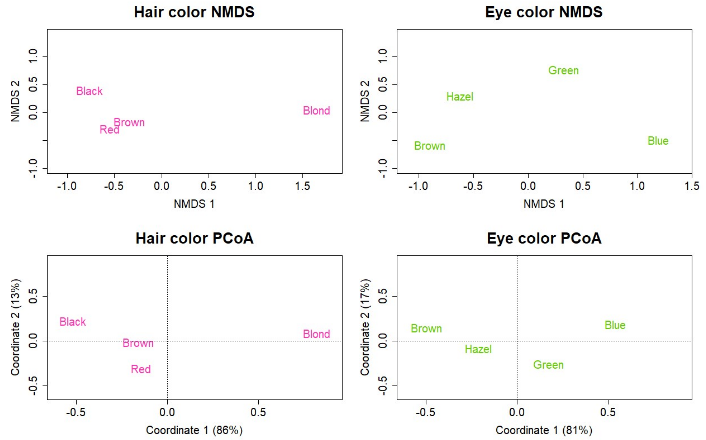
- Done using chi-square distance matrix
- Points that are closer are more similar
- Orientation relative to axes is arbitrary (why it's difficult to plot hair & eye color on same plot)



31

Compare NMDS & PCoA

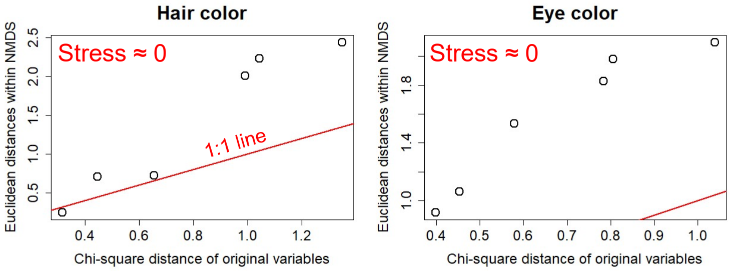




32

Shepard diagram

- Rank order, not exact distances, is preserved
- **Stress** is a measure of how much distances don't correspond (general rule: below 0.2 is OK)



Questions?

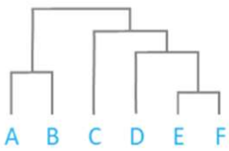
Classification

Cluster analysis

What is classification?

- Placing objects w/ more similar measurements into the same group
- Most popular method is a hierarchical one, called **cluster analysis**

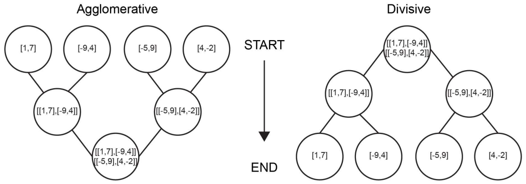
Dendrogram



"A" is most similar to which data point?
(interpret like a phylogenetic tree)

Two main kinds

1. **Agglomerative clustering (“bottom-up”)**
   - Starts w/ separate observations and successively groups them into larger clusters
2. **Divisive clustering (“top-down”)**
   - Starts w/ one cluster and splits data into smaller clusters until each observation is its own cluster


Agglomerative

Divisive

START

END

[1,7]

[-9,4]

[-5,9]

[4,-2]

[[1,7],[-9,4]]

[[-5,9],[4,-2]]

[[1,7],[-9,4]]
[[-5,9],[4,-2]]

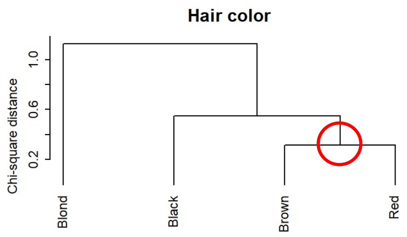
Two main kinds

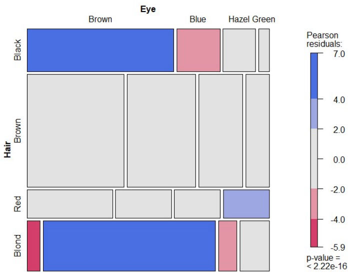
1. **Agglomerative clustering (“bottom-up”)**
2. **Divisive clustering (“top-down”)**

- Done using **any** distance matrix
- Different algorithms use different rules for how clusters should be made
- Cluster analysis creates a tree-diagram known as a **dendrogram**
- This is an exploratory/visualization method

Agglomerative

- Most popular algorithm in anthropology is **UPGMA (unweighted pair group method with arithmetic mean)**

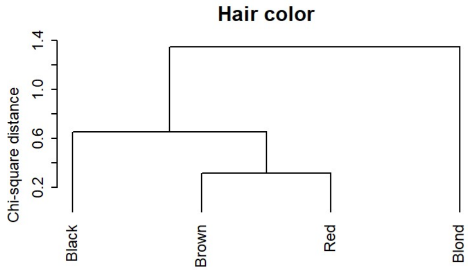




χ^2 dist. = 0.32 → same dist. btw Brown & Red

Divisive

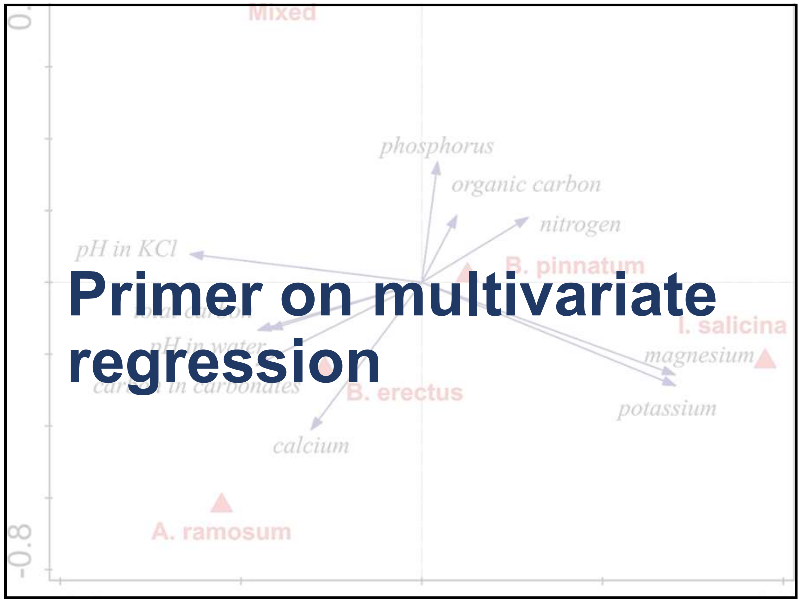
- Not as common as agglomerative clustering
- Dendrograms are like a mobile: you can rotate clusters around stems w/o changing results



Non-hierarchical clustering

1. ***K*-means clustering**
   - Groups observations into *k* clusters (*k* is determined *a priori*)
   - Observations within a cluster are more similar than between clusters
2. **Discriminant analysis**
   - Classifies observations into pre-defined groups
   - Like PCA, axes are linear combination of variables, where estimated coefficients (eigenvectors) maximize differences btw groups
   - Need to cross-validate model
   - Superseded by other methods (i.e., machine learning)

Questions?



Primer on multivariate regression

Multivariate regression

- Regression of two or more DVs as a function of one or more IVs

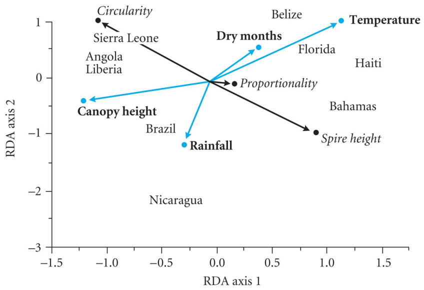
1. **Redundancy analysis (RDA)**
   - Uses a PCA to analyze variables
   - Continuous DVs
2. **Canonical correspondence analysis (CCA)**
   - Uses correspondence analysis to analyze variables
   - Frequencies of categorical DVs
   - Won't go over this

Redundancy analysis

- In essence, calculates a PCA of DVs & estimates coefficients for each IV for each PC
- Uses Monte Carlo methods to test if there's a significant relationship between IVs and DVs
- I have never seen this used in anthropology

RDA example (from textbook)

- **<u>DV</u>**: three variables of snail shell shape from nine countries
- **<u>IV</u>**: four environmental variables



Questions?

Summary

- Distance matrices quantify how (dis)similar variables are
  - There are many distance metrics out there
- PCoA ordinates data using **<u>any</u>** distance matrix
  - Euclidean distance matrix → PCA
  - Chi-square distance matrix → correspondence analysis
- Cluster analysis uses **<u>any</u>** distance matrix to group similar variables together in dendrogram
- RDA is a regression with multiple continuous DVs