Similarity & Distance Measures

Recap: We have explored how to design and analyze the results of experiments with one or more categorical treatment factors using Analysis of Variance (One-Way ANOVA, Factorial ANOVA). We looked at experiments with one or more quantitative treatment factors and used Regression Analysis to test hypotheses (Simple Linear Regression, Multiple Regression, and Nonlinear Regression). We have analyzed the results of studies that combined categorical and quantitative experimental factors via Analysis of Covariance (ANCOVA). Something that all these designs and analyses had in common was that had only a single (univariate) response variable. The remainder of the semester will focus on design and analysis of studies with multiple (multivariate) responses.

Objective: Multivariate Statistics is a broad field; to cover it thoroughly would require far more time than we have left. The objective of the next few Modules is to introduce some multivariate methods that are frequently used in biological and ecological research.

Example: Imagine you are a limnologist looking at plankton in New Hampshire lakes and ponds. Based on plankton tows (and many hours of microscopy), you have quantified the abundance of 17 species in 30 different lakes and ponds. Your first task is to look for similarities and differences among the lakes and ponds on the basis of all 18 species. You could stare at the data for days and not really be able to see any patterns. Instead, you could do a Multivariate Cluster Analysis to generate a dendrogram (tree diagram) that helps visualize groups of lakes and ponds that are most similar and most different.

Intro: Up to this point, we have been using statistical methods to account for variability in a single dimension. For example, in the mushroom study, we were interested in determining if log type explained a significant amount of the variability in yield. The variability in yield could be visualized on the vertical axis of our plot, i.e., in one dimension. The differences between the average yield for each log type could also be visualized on the vertical axis, i.e., in one dimension. Multivariate analyses require you to start thinking multidimensionally. Let’s start with a simplified version of the NH lake and pond plankton example described above. Imagine we determine the abundance of just 2 plankton species in each of 4 ponds. We could create a 2-dimensional plot with one axis corresponding to each of the species. We could then plot a point for each pond at a position corresponding to its species abundance on each of the 2 axes. On the resulting plot, the ponds whose points are closest together in 2-dimensional space would be considered most similar. Closest together? How do you determine distances in 2-dimensional space?
Now imagine we have abundance data for a 3 plankton species. You could plot the ponds as points on a 3-dimensional plot. How do you calculate the distances between ponds in 3 dimensions? Now, try to imagine the ponds plotted in 4 dimensions, or 5 dimensions, or 18 dimensions?? How do you calculate the distance between them??

**Euclidean Distances:** In the more generalized example below, we have four "quadrats". You can think of these as the four ponds, or four forest plots, four 1-m² pvc quadrat frames you have placed at different spots on the intertidal zone along the coast, or any other sampling unit. We are labeling them as Q1, Q2, Q3, and Q4. Within each quadrat, you have quantified the abundance of species x, species y, and species z.

<table>
<thead>
<tr>
<th>Species</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>5</td>
<td>40</td>
<td>35</td>
<td>8</td>
</tr>
<tr>
<td>y</td>
<td>25</td>
<td>10</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td>z</td>
<td>25</td>
<td>5</td>
<td>7</td>
<td>27</td>
</tr>
</tbody>
</table>

If we just consider just species x, we can calculate the "Euclidean Distance" between Q1 and Q2 by subtracting the number of species x in Q1 from the number in Q2. We can plot this in one dimension, and we can write a mathematical expression for the Euclidean Distance as \((x_2-x_1)\), which in this case is \((40-5)\) or 35.

![Diagram](image)

Based on species x alone, the pairwise distance between each of the quadrats would be as follows:

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>((x_2-x_1))</td>
<td>((x_3-x_1))</td>
<td>((x_4-x_1))</td>
</tr>
<tr>
<td>35</td>
<td>0</td>
<td>((x_3-x_2))</td>
<td>((x_4-x_2))</td>
</tr>
<tr>
<td>30</td>
<td>5</td>
<td>0</td>
<td>((x_3-x_4))</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>27</td>
<td>0</td>
</tr>
</tbody>
</table>

Now let’s think about the distance between quadrats that reflects two species, species x, and species y. Do you remember the Pythagorean Theorem? Does \(c^2=a^2+b^2\) look familiar? The distance between Q1 and Q2 in the species x dimension is \((x_2-x_1)\), the distance in the species y dimension would be \((y_1-y_2)\). So, using the Pythagorean Theorem, the Euclidean Distance between Q1 and Q2 in 2-dimensions would be the square root of \((x_2-x_1)^2 + (y_1-y_2)^2\)...
... and the pairwise distances between all the quadrats in 2-dimensional space would be:

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>0</td>
<td>$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$</td>
<td>$\sqrt{(x_1-x_3)^2+(y_1-y_3)^2+(z_1-z_3)^2}$</td>
<td>$\sqrt{(x_1-x_4)^2+(y_1-y_4)^2+(z_1-z_4)^2}$</td>
</tr>
<tr>
<td>Q2</td>
<td>38.1</td>
<td>0</td>
<td>$\sqrt{(x_2-x_3)^2+(y_2-y_3)^2+(z_2-z_3)^2}$</td>
<td>$\sqrt{(x_2-x_4)^2+(y_2-y_4)^2+(z_2-z_4)^2}$</td>
</tr>
<tr>
<td>Q3</td>
<td>31.6</td>
<td>7.1</td>
<td>0</td>
<td>$\sqrt{(x_3-x_4)^2+(y_3-y_4)^2+(z_3-z_4)^2}$</td>
</tr>
<tr>
<td>Q4</td>
<td>4.2</td>
<td>34.2</td>
<td>27.9</td>
<td>0</td>
</tr>
</tbody>
</table>

So, what about the distance between quadrats in 3-dimensional space? Well, the Pythagorean Theorem works in 3-dimensions, actually it works in 4, 5, or n dimensional "hyperspace" if you can wrap your mind around that.
**Activity 1:** On the Activities 1 & 2 tab of the Similarities and Distance Workbook, complete the pairwise distance matrix using the abundance values of all 3 species. Just to make certain you are on the right track, the Euclidean Distance between Mendums Pond and Wheelwright Pond should be 6.2.

**Other Distance Metrics:** Calculating pairwise distances between sampling units is the starting point for Cluster Analysis, which is a multivariate method for visualizing multidimensional patterns in a 2-dimensional tree diagram (dendrogram). We will tackle Cluster Analysis in the next Module. While Euclidean Distances are often used in Cluster Analysis, there are other distance metrics that can be used. The remainder of the current Module will familiarize you with some of them.

**Bray-Curtis Similarity Index:** A metric that frequently used to compare sites in ecological biodiversity studies is the Bray-Curtis Similarity Index. BC values range from 0 to 1, although they are sometimes expressed as a percentage. For example, two sites that are very similar in terms of species occurrence and abundance might have a BC Similarity Index of 0.90, or 90%. Note that this is a similarity measure rather than a distance measure. It is not uncommon to see Bray-Curtis reported as a dissimilarity index, which is simply 1 minus the similarity index. So, two sites that have a BC Similarity Index of 0.90 have a BC Dissimilarity Index or 1 - 0.90 = 0.10, or 10%. One thing to keep in mind is that BC is really only valid if the sites being compared were subjected to the same sampling effort. You can't compare the number individuals in a 1 m² plot at one site and a 10 m² plot at another unless you convert the counts to a per-square-meter basis first.

The formula for a BC Similarity Index is: \( BC_{ij} = \frac{2C_{ij}}{S_i + S_j} \)

Where \( S_i \) is the total number of individuals of all the species found at one site, \( S_j \) is the total number of individuals of all the species found at the other site, and \( C_{ij} \) is the sum of the lesser of the counts for each species found at both sites.

**Activity 2:** On the Activities 1 & 2 tab of the Similarities and Distance Workbook, calculate the Bray-Curtis Similarity Index for Wheelwright and Barbados Ponds. Start in column E by using the =min() formula to find the lesser of the counts for each species in both ponds. If you sum the 3 values in column E, you will have \( C_{ij} \). If you sum column C, you will have \( S_i \), the sum of column D will give you \( S_j \). Enter the Bray-Curtis formula in E31 and you're done. (You might consider saving your workbook at this point)

**Presence/Absence Data:** In ecological studies, determining the abundance of each species present at multiple study sites can be impractical or impossible. It is sometimes more efficient to use the species presence/absence data. In the book *Ordination of Plant Communities*, David Goodall contends that “in highly heterogeneous communities, quantitative measures add little useful information over that provided by simple species lists.”
There are at least a half dozen similarity indices that can be used for multivariate analyses of binary (presence/absence) data. The ones we will look at can be expressed using the 4 values in the table below:

<table>
<thead>
<tr>
<th>Site 1</th>
<th>Site 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>a</td>
</tr>
<tr>
<td>Absent</td>
<td>b</td>
</tr>
</tbody>
</table>

Where
- **a** is the number of species that occur at both sites,
- **b** is the number of species that occur in Site 1, but not in Site 2,
- **c** is the number of species that occur in Site 2, but not in Site 1,
- **d** is the number of species that are absent from both sites.

Calculating **d** only makes sense in studies where you are calculating pairwise similarities of multiple sites (i.e. more than just 2). The six indices below are commonly used in Cluster Analysis of ecological data. They differ in the weight that they put on similarities vs differences, and the importance placed on the absence of a species from both sites. For example, in the Simple Matching Index that fact that a species is **missing** from the two sites is considered just as important as when a species **occurs** in both sites, whereas in the Positive Matching Index, only species that **occur** in both sites are considered indicative of similarity. There are no right and wrong indices, but in any given study, one index may provide clearer insights than others.

- **Positive Matching** = \( \frac{a}{a+b+c+d} \)
- **Anderberg** = \( \frac{a}{a+2(b+c)} \)
- **Jaccard** = \( \frac{a}{a+b+c} \)
- **Tanimoto** = \( \frac{a}{a+2(b+c)+d} \)
- **Simple Matching** = \( \frac{a+d}{a+b+c+d} \)
- **Sorensen's** = \( \frac{2a}{2a+b+c} \)

**Activity 3**: The Activity 3 & 4 tab of the Similarity and Distance Measurement Workbook contains presence/absence data for 17 species in 4 quadrats. The goal is to calculate the six similarity indices for each pair of quadrats. We are going to start by calculating **a**, **b**, **c**, and **d** for each pairwise comparison. Actually, **b**, **c**, and **d** are already calculated for you; you are just going to calculate **a**. The calculations use Excel "Array Formulas", which, if you haven't used them before, require some guidance. Take a look at the formula in cell I25. It calculates the number of species that occur in Quadrat 1, but not in Quadrat 2. It looks like this: \( =\text{SUM}((B5:B21=1)*(C5:C21=0)) \)
The curly brackets {} are what identifies it as an array formula. When you enter an array formula, don’t type the curly brackets. Excel will add them automatically when you finish the formula and press Shift-Ctrl-Enter (Shift-Cmd-Enter on a Mac). The other non-intuitive part of the array formula that the asterisk * means “AND” rather than multiply. So the formula above means to step through rows 5 through 21, and wherever the value in column B is 1 AND the value in C is 0, add one to the SUM.

Now it’s your turn. In cell C25, type a formula to count the species that occur in both Quadrat 1 and 2. It is going to look exactly like the formula above, except =0 will be replaced with =1. Remember, don’t type the {}. When you have typed the formula press Shift-Ctrl-Enter and the curly bracket will appear. If you did it correctly, cell C25 will show 4. If you used dollar signs correctly, you will be able to copy the formula to cells B25, D25, and E25. You will need to modify the formula for it to work in B26 through D26. The easy way to do that is copy the formula from B25 to B26, then edit the formula in B26 changing ($B$5:$B$21=1) to ($C$5:$C$21=1). You can then copy the formula in B26 to C26 through D26. Try to figure out the rest on your own, or if you can’t, watch the video demo.

When you have calculated all the pairwise values for a, four of the six pairwise similarity index tables will fill in automatically. The remaining 2 (Jaccard and Anderberg), you will need to do on your own. Again, you can watch the video demo if you need help.

Questions: In the spaces provided on the Activity 3 tab of the Similarity and Distance Measures Workbook, answer the following questions.

1. Which of the similarity indices are not influenced by species that are absent from both sites?
2. Which indices place a heavier weight on species that occur in one site but not the other?

Activity 4: Each Similarity Index has a corresponding Dissimilarity Index, which is just 1-Similarity. For example, Sorensen’s Dissimilarity Index would be:

\[
\text{Sorensen's Dissimilarity Index} = 1 - \frac{2a}{2a+b+c} = \frac{2a+b+c}{2a+b+c} - \frac{2a}{2a+b+c} = \frac{b+c}{2a+b+c}
\]

In the space provide on the Activity 4 tab of the worksheet, calculate Sorensen’s Dissimilarity Index for Q1 and Q2 using both formulas.

The Bray-Curtis Similarity Index, that you worked with in Activity 2 is actually the same as Sorensen’s Similarity Index when the quantities for each species are reduced to presence/absence (i.e. 1/0). Recall that in the formula

\[
\text{Bray-Curtis}_{ij} = \frac{2C_{ij}}{(S_i+S_j)}
\]

\(S_i\) is the total number of individuals of all the species found at Site 1, and \(S_j\) is the total number of individuals of all species at Site 2. With presence absence data, each species is recorded as either 1 (present) or 0 (absent), so adding up the 1’s for Site 1 gives us total number of species found in Site 1 regardless of the whether it occurs in site 2, so using our \(a, b, c, d\) designation, this is the equivalent of \(a+b\). Summing the 1’s for Site 2 gives us \(a+c\). So, with presence/absence data \((S_i+S_j)\) is the same as \(a+b+a+c = 2a+b+c\).
the Bray-Curtis Formula, $C_{ij}$ is the sum of the lesser of the counts for each species found at both sites. With the presence/absence data, the lesser of the counts will always be 0 for species that occur at one Site but not the other and it will only be 1 for species that occur at both sites, so, adding these up will give us the total number of species that occur at both sites, which is the same as $a$. If we substitute $a$ for $C_{ij}$ and $(2a+b+c)$ for $(S_i+S_j)$, our Bray-Curtis Similarity Index formula becomes:

$$\text{Bray-Curtis}_{ij} = \frac{2a}{2a+b+c}.$$ 

For empirical proof, use the same method that you used in Activity 2 to calculate the Similarity Index for Q1 and Q2. There is a place for you to do this at the bottom of the Activity 3 & 4 tab of the workbook.

That’s it for this Module. Save your workbook and send it to me as an email attachment.