2.1 Similarity for quantitative data matrices

Data matrix

The available biological data is assumed to consist of an array of *p* rows (species) and *n* columns (samples), whose entries are counts or densities of each species for each sample, or the total biomass of all individuals, or their percentage cover, or some other quantity of each species in each sample, which we will typically refer to as *abundance*. This includes the special case where only presence (1) or absence (0) of each species is known. For the moment nothing further is assumed about the structure of the samples. They might consist of one or more replicates (repeated samples) from a number of different sites, times or experimental treatments but this information is *not* used in the initial analysis. The strategy outlined in Chapter 1 is to *observe* any pattern of similarities and differences across the samples (i.e. let the biology 'tell its own story') and then compare this with known or *a priori* hypothesised inter-relations between the samples based on environmental or experimental factors.

Similarity coefficient

The starting point for many of the analyses that follow is the concept of *similarity* (*S*) between any pair of samples, in terms of the biological communities they contain. Inevitably, because the information for each sample is multivariate (many species), there are many ways of defining similarity, each giving different weight to different aspects of the community. For example, some definitions might concentrate on the similarity in abundance of the few commonest species whereas others pay more attention to rarer species.

The data matrix itself may first be modified; there are three main possibilities.

a) The absolute numbers (biomass/cover), i.e. the fully quantitative data observed for each species, are most commonly used. In this case, two samples are considered perfectly similar only if they contain the same species in *exactly* the same abundance.

b) The relative numbers (biomass/cover) are sometimes used, i.e. the data is *standardised* to give the percentage of total abundance (over all species) that is accounted for by each species. Thus each matrix entry is divided by its column total (and multiplied by 100) to form the new array. Such standardisation will be essential if, for example, differing *and unknown* volumes of sediment or water are sampled, so that absolute numbers of individuals are not comparable between samples. Even if sample volumes are the same (or, if different and known, abundances are adjusted to a unit sample volume, to define densities), it may still sometimes be biologically relevant to define two samples as being perfectly similar when they have the same *% composition* of species, fluctuations in total abundance being of no interest. (An example might be fish dietary data on the predated assemblage in the gut, where it is the fish doing the sampling and no control of total gut content is possible, of course.)

c) A reduction to simple presence or absence of each species may be all that is justifiable, e.g. sampling artefacts may make quantitative counts unreliable, or concepts of abundance may be difficult to define for some important faunal components.

A similarity coefficient *S* is conventionally defined to take values in the range (0, 100%), or alternatively (0, 1), with the ends of the range representing the extreme possibilities:

S = 100% (or 1) if two samples are totally similar;

S = 0 if two samples are totally dissimilar.

Dissimilarity (δ) is defined simply as 100 – *S*, the "opposite side of the coin" to similarity.

What constitutes total similarity, and particularly total dissimilarity, of two samples depends on the specific similarity coefficient adopted but there are clearly some properties that it would be desirable for a biologically-based coefficient to possess. Full discussion of these is given in Clarke, Somerfield & Chapman (2006) , e.g. most ecologists would feel that *S* should equal zero when two samples have no species in common and *S* must equal 100% if two samples have identical entries (after modification, in cases b and c above).** Such guidelines lead to a small set of coefficients termed the *Bray-Curtis family* by Clarke, Somerfield & Chapman (2006) .

Similarity matrix

Similarities are calculated between every pair of samples and it is conventional to set these *n(n* –1)/2 values out in a lower triangular matrix. This is a square array, with row and column labels being the sample numbers 1 to *n*, but it is not necessary to fill in either the diagonals (similarity of sample *j* with itself is always 100%!) or the upper right triangle (the similarity of sample *j* to sample *k* is the same as the similarity of sample *k* to sample *j*, of course).

Similarity matrices are the basis (explicitly or implicitly) of many multivariate methods, both in the representation given by a clustering or ordination analysis and in some associated statistical tests. A similarity matrix can be used to:

a) discriminate sites (or times) from each other, by noting that similarities between replicates within a site are consistently higher than similarities between replicates at different sites (ANOSIM test, Chapter 6);

b) cluster sites into groups that have similar communities, so that similarities within each group of sites are usually higher than those between groups (Clustering, Chapter 3);

c) allow a gradation of sites to be represented graphically, in the case where site A has some similarity with site B, B with C, C with D but A and C are less similar, A and D even less so etc. (Ordination, Chapter 4).

Species similarity matrix

In a complementary way, the original data matrix can be thought of as describing the pattern of occurrences of each species across the given set of samples, and a matching triangular array of similarities can be constructed between every *pair of species*. Two species are *similar* (*S´* near 100

or 1) if they have significant representation at the same set of sites, and totally *dissimilar* ($S´ = 0$) if they never co-occur. Species similarities are discussed later in this chapter, and the resulting clustering diagrams in Chapter 7 but, in most of this manual, 'similarity' refers to between-*sample* similarity.

Bray-Curtis coefficient

Of the numerous similarity measures that have been suggested over the years[¶], one has become particularly common in ecology, usually referred to as the *Bray-Curtis* coefficient, since Bray & Curtis (1957) were primarily responsible for introducing this coefficient into ecological work. The similarity between the *j*th and *k*th samples, S_{jk}, has two definitions (they are entirely equivalent, as can be seen from some simple algebra or by calculating a few examples):

$$S_{jk} = 100 \left[1 - \frac{\sum_{i=1}^{p} | y_{ij} - y_{ik} | }{\sum_{i=1}^{p} (y_{ij} + y_{ik}) } \right] = 100 \frac{\sum_{i=1}^{p} 2 \min (y_{ij}, y_{ik}) }{\sum_{i=1}^{p} (y_{ij} + y_{ik}) } \tag{2.1}$$

Here y_{ij} represents the entry in the *i*th row and *j*th column of the data matrix, i.e. the abundance for the *i*th species in the *j*th sample ($i = 1, 2, ..., p$; $j = 1, 2, ..., n$). Similarly, y_{ik} is the count for the *i*th species in the *k*th sample. |...| represents the absolute value of the difference (the sign is ignored) and min(.,.) the minimum of the two counts; the separate sums in the numerator and denominator are both over all rows (species) in the matrix.

[¶] *Legendre & Legendre (2012) , in their invaluable text on Numerical Ecology, give very many definitions of similarity, dis-similarity and distance coefficients, and PRIMER follows their suggestion of the collective term resemblance to cover any such measure and, where possible, uses their numbering system.*

Revision #41
Created 15 February 2022 08:10:34 by Arden
Updated 16 October 2024 06:08:56 by Marti