2.3 Presence/absence data

As discussed at the beginning of this chapter, quantitative uncertainty may make it desirable to reduce the data simply to presence or absence of each species in each sample, or this may be the only feasible or cost-effective option for data collection in the first place. Alternatively, reduction to presence/absence may be thought of as the ultimate in severe transformation of counts; the data matrix (e.g. in Table 2.1a) is replaced by 1 (presence) or 0 (absence) and Bray-Curtis similarity (say) computed. This will have the effect of giving potentially equal weight to all species, whether rare or abundant (and will thus have somewhat similar effect to the Canberra coefficient, a suggestion confirmed by the comparative analysis in Chapter 16).

Many similarity coefficients have been proposed based on (0, 1) data arrays; see for example, Sneath & Sokal (1973) or Legendre & Legendre (2012) . When computing similarity between samples *j* and *k*, the two columns of data can be reduced to the following four summary statistics without any loss of relevant information:

a = the number of species which are present in both samples;

b = the number of species present in sample *j* but absent from sample *k*;

c = the number of species present in sample *k* but absent from sample *j*;

d = the number of species absent from both samples.

For example, when comparing samples 1 and 4 from Table 2.1a, these frequencies are:

		Sample 4:	**1**	**0**
Sample 1:	1		*a* = 2	*b* = 1
	0		*c* = 2	*d* = 1

In fact, because of the symmetry, coefficients must be a symmetric function of *b* and *c*, otherwise $S _ {14}$ will not equal $S _ {41}$. Also, similarity measures not affected by joint absences will not contain *d*. The following are some of the more commonly advocated coefficients.

The *simple matching* similarity between samples *j* and *k* is defined as:

$$ S _ {jk} = 100 \left[(a + d) / (a + b + c + d) \right] \tag{2.5} $$

so called because it represents the probability ($\times 100$) of a single species picked at random (from the full species list) being present in both samples or absent in both samples. Note that *S is* a function of *d* here, and thus depends on joint absences.

If the simple matching coefficient is adjusted, by first removing all species which are jointly absent from samples *j* and *k*, one obtains the *Jaccard* coefficient:

$$ S _ {jk} = 100 \left[a / (a + b + c) \right] \tag{2.6} $$

i.e. *S* is the probability ($\times 100$) that a single species picked at random (from the reduced species list) will be present in both samples.

A popular coefficient found under several names, commonly *Sørensen* or *Dice*, is

$$ S _ {jk} = 100 \left[2a / (2a + b + c) \right] \tag{2.7} $$

Note that this is identical to the Bray-Curtis coefficient when the latter is calculated on (0, 1) presence/absence data, as can be seen most clearly from the second form of equation (2.1).¶ For example, reducing Table 2.1a to (0, 1) data, and comparing samples 1 and 4 as previously, equation (2.1) gives:

$$S_{14} = 100 \left[\frac{2(0+1+1+0+0+0)}{1+2+2+1+1+0} \right] = 57.1 $$

This is clearly the same construction as substituting $a = 2$, $b = 1$, $c = 2$ into equation (2.7).

Several other coefficients have been proposed; Legendre & Legendre (2012) list at least 15, but only one further measure is given here. In the light of the earlier discussion on coefficients satisfying desirable, biologically-motivated criteria, note that there is a presence/absence form of the *Kulczynski* coefficient (2.4), a close relative of Bray-Curtis/Sørensen, namely:

$$S _ {jk} = 50 \left(\frac{a}{a+b} + \frac{a}{a+c} \right) \tag{2.8} $$

Recommendations

1. In most ecological studies, some intuitive axioms for desirable behaviour of a similarity coefficient lead to the use of the Bray-Curtis coefficient (or a closely-related measure such as Kulczynski).
2. Similarities calculated on original abundance (or biomass) values can often be over-dominated by a small number of highly abundant (or large-bodied) species, so that they fail to reflect similarity of overall community composition.
3. Some coefficients (such as Canberra and that of Gower (1971) , see later), which separately scale the contribution of each species to adjust for this, have a tendency to over-compensate, i.e. rare species, which may be arbitrarily distributed across the samples, are given equal weight to abundant ones. The same criticism applies to reduction of the data matrix to simple presence/absence of each species. In addition, the latter loses potentially valuable information about the *approximate* numbers of a species (0: absent, 1: singleton, 2: present only as a handful of individuals, 3: in modest numbers, 4: in sizeable numbers; 5: abundant; 6: highly abundant. This apparently crude scale can often be just as effective as analysing the precise counts in a multivariate analysis, which typically extracts a little information from a lot of species).
4. A balanced compromise is often to apply the Bray-Curtis similarity to counts (or biomass, area cover etc) which have been moderately, $\sqrt{}$*y*, or fairly severely transformed, log(1+*y*) or $\sqrt{} \sqrt{}$*y* (i.e. $y ^ {0.25}$). Most species then tend to contribute something to the definition of similarity, whilst the retention of some information on species numbers ensures that the more abundant species are given greater weight than the rare ones. A good way of assessing where this balance lies – how much of the matrix is being used for any specific transformation – is to view *shade plots* of the data matrix, as

seen in Figs. 7.7 to 7.10 and 9.5 and 9.6.

5. Pre-treating the data, prior to transformation, by standardisation of samples is sometimes desirable, depending on the context. This divides each count by the total abundance of all species in that sample and multiplies up by 100 to give a percent composition (or perhaps standardises by the maximum abundance). Worries that this somehow makes the species variables non-independent, since they must now add to 100, are misplaced: species variables are *always* non-independent – that is the point of multivariate analysis! *Without* sample standardisation, the Bray-Curtis coefficient will reflect both compositional differences among samples and (to a weak extent after transformation) changing total abundance at the different sites/times/treatments.[§]

[¶] *Thus the Sorensen coefficient can be obtained in two ways in the PRIMER Resemblance routine, either by taking* S8 Sorensen *in the P/A list or by transforming the data to presence/absence and selecting Bray-Curtis similarity.*

[§] *The latter is usually thought necessary, by marine benthic ecologists at least: if everything becomes half as abundant they want to know about it! However, much depends on the sampling device and the patchiness of biota; plankton ecologists usually do standardise, as will kick-samplers in freshwater, where there is much less control of 'sample volume'. Standardisation removes any contribution from totals but it does not remove the subsequent need to transform, in order to achieve a better balance of the abundant and rarer species.*

Revision #14
Created 16 February 2022 08:56:32 by Arden
Updated 27 June 2023 23:32:56 by Arden