Appendix D. Chao's Abundance-based Jaccard and Sorensen Similarity Indexes and Their Estimators

Chao's Abundance-based Jaccard and Sørensen Indexes

Let the probabilities of species discovery Assemblages 1 and 2 be denoted respectively by $(p_1, p_2, ..., p_{S_1})$ and $(\pi_1, \pi_2, ..., \pi_{S_2})$, where $p_i > 0$, $\pi_i > 0$ and $\sum_{i=1}^{S_1} p_i = \sum_{i=1}^{S_2} \pi_i = 1$. Without loss of generality, we assume the first S_{12} species are shared species, that is, the shared species are indexed by 1, 2,..., S_{12}. In Assemblage 1, let *U* denote the total relative abundances of individuals belonging to the *shared* species, $U = p_1 + p_2 + ... + p_{S_{12}}$. Likewise in Assemblage 2, let *V* denote the total relative abundances of individuals belonging to *shared* species, $V = \pi_1 + \pi_2 + ... + \pi_{S_{12}}$. We obtain the following abundance-based indices in terms of *U* and *V*:

$$J_{abd} = \frac{UV}{U + V - UV} \text{ and}$$

$$L_{abd} = \frac{2UV}{U + V}.$$

(In EstimateS output, J_{abd} is called the "Chao-Jaccard-Raw Abundance-based" estimator and L_{abd} is called the "Chao-Sørensen-Raw Abundance-based" estimator.)

Estimators for the Indexes Based on Abundance Data

A random sample of *n* individuals (Sample 1) is taken from Assemblage 1 and a random sample of *m* individuals (Sample 2) is taken from Assemblage 2. Denote the species frequencies in the *samples* by $(X_1, X_2, ..., X_{S_1})$ and $(Y_1, Y_2, ..., Y_{S_2})$, respectively. (Note that if a species is missing from a sample, X_i or Y_i will equal zero.) Thus the pair of frequencies for the S_{12} species truly shared by the two *assemblages* are $(X_1, Y_1)(X_2, Y_2)...(X_{S_{12}}, Y_{S_{12}})$. Assume that D_{12} of the S_{12} shared species available are actually observed in both samples, and their frequencies are the first D_{12} pairs. Thus, an additional $S_{12} - D_{12}$ species are shared by the two assemblages, but absent from one or both of the samples.

To incorporate the effect of unseen shared species, we use the frequencies of *observed* rare, shared species to estimate an appropriate adjustment term for *U* and *V* to account for *unseen* shared species. We first define the indicator function *I*(*expression*) such that $I = 1$ if *expression* is true, and $I = 0$ if *expression* is false. Let $f_{1+} = \sum_{i=1}^{D_{12}} I[X_i = 1, Y_i \geq 1]$ be the observed number of *shared* species that are singletons ($X_i = 1$) in Sample 1 (these species must be present in Sample 2, but may have any abundance). Now, let f_{2+} be the observed number of *shared* species that are doubletons ($X_i = 2$) in Sample 1. Similarly, we define f_{+1} and f_{+2} to be the observed number of shared species that are, respectively, singletons ($Y_i = 1$) and doubletons ($Y_i = 2$) in Sample 2.

Then the estimator for *U* is

$$\hat{U} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(m-1)}{m} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \frac{X_i}{n} I(Y_i = 1).$$

Notice that the first term in the right hand side of this equation denotes the observed total of frequencies associated with the observed shared species; the second term accounts for the estimated effect of unseen shared species. Similarly, we have

$$\hat{V} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(n-1)}{n} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \frac{Y_i}{m} I(X_i = 1).$$

When $f_{+2} = 0$ or $f_{2+} = 0$, replace f_{+2} and f_{2+} in the denominators by $f_{+2} + 1$ or $f_{2+} + 1$, respectively. If the value of $\hat{U}$ or $\hat{V}$ is greater than 1 (which rarely happens), then it is replaced by 1. Our proposed abundance-based Jaccard and Sørensen estimators are

$$\hat{J}_{abd} = \frac{\hat{U}\hat{V}}{\hat{U} + \hat{V} - \hat{U}\hat{V}} \text{ and}$$

$$\hat{L}_{abd} = \frac{2\hat{U}\hat{V}}{\hat{U} + \hat{V}}$$

(In EstimateS output, $\hat{J}_{abd}$ is called the "Chao-Jaccard-Est Abundance-based" estimator and $\hat{L}_{abd}$ is called the "Chao-Sørensen-Est Abundance-based" estimator.)

Estimators for the Indexes Based on Replicated Incidence Data

Suppose we take a set of *w* replicated incidence samples from Assemblage *X* and a set of *z* replicated incidence samples from Assemblage *Y*. For both sets of samples *combined*, there are *S* species. The number of samples in which a species is found in Assemblage *X* or *Y* is the *frequency* for that species in that sample set. The frequencies for species *i* are thus defined as

$$X_i = \sum_{j=1}^{w} x_{ij} \text{ and } Y_i = \sum_{j=1}^{z} y_{ij},$$

where x_{ij} and y_{ij} represent the presence (1) or absence (0) of species *i* in sample *j*. Note that X_i or Y_i will be zero for some species, unless all species are shared and observed.

For replicated incidence data, f_{1+} is the number of observed shared species that occur in exactly one sample ($X_i = 1$) in *X* and f_{2+} is the number of observed shared species that occur in exactly two samples ($X_i = 2$) in *X*; f_{+1} and f_{+2} are the corresponding numbers for sample matrix *Y*. Define the sum of the incidence frequencies for the matrices as

$$n = \sum_{i=1}^{S} X_i \text{ and } m = \sum_{i=1}^{S} Y_i .$$

Then the proposed estimators are

$$\hat{U}_{inc} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(z-1)}{z} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \left[\frac{X_i}{n} I(Y_i = 1) \right]$$

and

$$\hat{V}_{inc} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(w-1)}{w} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \left[\frac{Y_i}{m} I(X_i = 1) \right]$$

(The same modifications described for the abundance-based equations may be applied here if $f_{+2} = 0$ or $f_{2+} = 0$.) Thus, our proposed incidence-based Jaccard and Sørensen estimators are

$$\hat{J}_{inc} = \frac{\hat{U}_{inc}\hat{V}_{inc}}{\hat{U}_{inc} + \hat{V}_{inc} - \hat{U}_{inc}\hat{V}_{inc}} \text{ and}$$

$$\hat{L}_{inc} = \frac{2\hat{U}_{inc}\hat{V}_{inc}}{\hat{U}_{inc} + \hat{V}_{inc}}$$

(In EstimateS output, $\hat{J}_{inc}$ is called the "Chao-Jaccard-Est Incidence -based" estimator and $\hat{L}_{inc}$ is called the "Chao-Sørensen-Est Incidence-based" estimator.)

Variance Estimators

EstimateS uses a bootstrap method (Chao et al. in press) to obtain a variance estimator for the Chao's abundance-based Jaccard and Sørensen similarity index estimators. Assume that, in the data, there are a total of $D = D_1 + D_2 - D_{12}$ pairs of abundance values (see Section 3.2 for notation). These *D* species include observed shared species (for which both abundances are non-zero) and observed unique species (for which one of the abundances is zero).

The method is as follows: (a) Resample D pairs, *with* replacement, from the collection of D pairs. (b) For the Chao Jaccard or Chao Sørensen estimator, calculate an adjusted estimate (called a bootstrap estimate) based on the resampling data. (c) Repeat the procedure in (a) B times and obtain B bootstrap estimates. The bootstrap variance estimator of the adjusted estimator is the sample variance of these B estimates.

(In EstimateS output, "Chao-Jaccard-EstSD" and "Chao-Sørensen-EstSD" are the bootstrap standard deviations based on this method.)

March 8, 2005