Anne Chao, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2013. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecology Monographs*.

Appendix C: Rarefaction and extrapolation for species richness (incidence data)

In this Appendix, we briefly review the incidence data sections of Colwell et al. (2012). We also derive new formulas for coverage-based rarefaction and extrapolation curves for incidence data.

Sample-size-based rarefaction/extrapolation

Rarefaction and extrapolation for incidence data are formulated under the model (Eqs. 2a, 2b of the main text) in which the incidence frequency counts Y_i follow a binomial distribution with T sampling units and the incidence probability is π_i for the ith species in any sampling unit, $i = 1, 2, \ldots, S$. Let $S_{sample}(t)$ be the expected number of species in a set of t sampling units randomly selected from the assemblage. If we know the true species incidence probabilities $\pi_1, \pi_2, ..., \pi_S$ of each of the S species in each sampling unit, then

$$S_{sample}(t) = \sum_{i=1}^{S}[1-(1-\pi_i)^t] = S - \sum_{i=1}^{S}(1-\pi_i)^t, \quad t \geq 1. \qquad \text{(C.1)}$$

The plot of $S_{sample}(t)$ with respect to the number of sampling units t (the sample size, in this case) is the expected species accumulation curve for incidence data. The rarefaction (interpolation) part estimates the expected species richness for a smaller number of sampling units $t < T$. Based on the incidence reference sample with frequencies Y_i, the minimum variance unbiased estimator for $S_{sample}(t)$ is (Shinozaki 1963, see Chiarucci et al. 2008 for a history of this derivation)

$$\widetilde{S}_{sample}(t) = S_{obs} - \sum_{Y_i>0}\left[\binom{T-Y_i}{t} \Big/ \binom{T}{t}\right], \; t < T. \qquad \text{(C.2)}$$

Here, $\binom{a}{b} \equiv 0$ if $a < b$. We use this conventional definition throughout this Appendix.

The extrapolation problem is to estimate the expected number of species $S_{sample}(T+t^*)$ in an augmented set of $T + t^*$ sampling units ($t^* > 0$) from the assemblage. From Eq. (C.1), we have

$$S_{sample}(T+t^*) = E(S_{obs}) + \sum_{i=1}^{S}[1-(1-\pi_i)^{t^*}](1-\pi_i)^T. \quad \text{(C.3)}$$

Chao et al. (2009, their Appendix) derived the following estimator:

$$\tilde{S}_{sample}(T+t^*) = S_{obs} + \hat{Q}_0\left[1-\left(1-\frac{Q_1}{Q_1+T\hat{Q}_0}\right)^{t^*}\right]$$

$$\approx S_{obs} + \hat{Q}_0\,[1-\exp(\frac{-t^*Q_1}{Q_1+T\hat{Q}_0})], \quad \text{(C.4)}$$

where $\hat{Q}_0$ can be any predicted value of Q_0, the number of species present in the assemblage, but not observed in any sampling unit in the reference sample. Colwell et al. (2012) suggested that $\hat{Q}_0$ can be obtained by using the Chao2 estimator (Chao 1987) or ICE (Lee and Chao 1994). The estimator from the Chao2 estimator is

$$\hat{Q}_0 = \begin{cases} [(T-1)/T]Q_1^2/(2Q_2), & if\ Q_2 > 0 \\ [(T-1)/T]Q_1(Q_1-1)/2, & if\ Q_2 = 0 \end{cases} \quad \text{(C.5)}$$

Here, we suggest using a bootstrap method to obtain unconditional variance estimators for the rarefaction estimator in Eq. (C.2) and its predictor in Eq. (C.4). See Appendix G for details. The resulting variances are then used to construct confidence intervals of the expected species richness. As with abundance data, Colwell et al. (2012) linked the rarefaction and extrapolation to form an integrated smooth curve. The corresponding confidence intervals based on a bootstrap method also join smoothly at the reference point (T, S_{obs}).

Coverage-based rarefaction/extrapolation

In this section, we derive estimators for sample coverage based on incidence data for a set of sampling units (which together form the reference sample). The sample coverage of a reference sample of T sampling units is defined as

$$C_{sample}(T) = \frac{\sum_{i=1}^{S}\pi_i I(Y_i > 0)}{\sum_{i=1}^{S}\pi_i}, \quad \text{(C.6)}$$

which represents the fraction of the total incidence probabilities of the discovered species in the reference sample. This type of sample coverage was first defined in Chao et al. (1992) for

capture-recapture data. Analogous to Eq. (B.7), a very accurate estimator of the sample coverage for reference sample size T is

$$\hat{C}_{sample}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right], \qquad \text{(C.7)}$$

where $U = \sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$ denotes the total number of incidences in the reference sample. Similarly, we can define sample coverage, $C_{sample}(t)$, for t sampling units. The expected sample coverage for t sampling units is

$$E[C_{sample}(t)] = \frac{\sum_{i=1}^{S} \pi_i[1-(1-\pi_i)^t]}{\sum_{i=1}^{S} \pi_i} = 1 - \frac{\sum_{i=1}^{S} \pi_i(1-\pi_i)^t}{\sum_{i=1}^{S} \pi_i}, \quad t \geq 1. \qquad \text{(C.8)}$$

For the rarefaction part of the curve ($t < T$), we can find the minimum variance unbiased estimator for the denominator and numerator in Eq. (C.8), respectively, but their ratio given below is only a nearly unbiased estimator of the sample coverage for t sampling units:

$$\hat{C}_{sample}(t) = 1 - \sum_{Y_i \geq 1} \frac{Y_i}{U} \frac{\binom{T-Y_i}{t}}{\binom{T-1}{t}}, \quad t < T. \qquad \text{(C.9)}$$

This equation is analogous to Eq. (B.9) for abundance data. An estimator for the expected coverage of an extrapolated sample with $T + t^*$ sampling units is

$$\hat{C}_{sample}(T + t^*) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]^{t^*+1}. \qquad \text{(C.10)}$$

As t^* tends to infinity, the extrapolated coverage estimator approaches unity, indicating complete coverage. When $t^* = 0$, Eq. (C.10) reduces to the sample coverage estimator for the reference sample as given in Eq. (C.7). As with abundance data, a smooth coverage-based interpolation and extrapolation curves with confidence intervals can be constructed for incidence data.

LITERATURE CITED

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783-791.

Chao, A., R. K. Colwell, C. W. Lin, and N. J. Gotelli. 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90:1125-1133.

Chao, A., and L. Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology, 93:2533-2547.

Chao, A., S. Lee, and S. Jeng. 1992. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. Biometrics 48:201-216.

Chiarucci, A., G. Bacaro, D. Rocchini, and L. Fattorini. 2008. Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. Community Ecology 9:121-123.

Colwell, R. K., A. Chao, N. J. Gotelli, S. Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5:3-21.

Lee, S. M., and A. Chao. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics 80:88-97.

Shinozaki, K. 1963. Note on the species area curve. Proceedings of the 10th Annual Meeting of the Ecological Society of Japan, 5.