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 G: A bootstrap method to construct an unconditional variance estimator for any interpolated or extrapolated estimator

Abundance data

Under the multinomial model given in Eq. 1 of the main text, we suggest the use of a bootstrap method to approximate the variance of any interpolated or extrapolated diversity estimator developed in this paper. After a variance estimator is obtained, the variance can then be applied to construct a confidence interval of the expected diversity. Here we use the interpolated species richness estimator ${}^q\hat{D}(m)$ given in Table 1 of the main text as an example. Parallel steps can be formulated for any other estimators.

In the bootstrap procedure, we first need to construct the "*bootstrap assemblage*" which mimics the true entire assemblage. We first determine the true species richness in this bootstrap assemblage. As in the main text, define the *abundance frequency count* f_k as the number of species each represented by exactly k individuals in the reference sample. Thus, f_1 denotes the number of singletons and f_2 denotes the number of doubletons in the sample. Let $\hat{f}_0$ be any proper estimator of the number of undetected species. Using the Chao1 estimator (Chao 1984), we have

$$\hat{f}_0 = \begin{cases} [(n-1)/n] f_1^2/(2f_2), & if \;\; f_2 > 0 \\ [(n-1)/n] f_1(f_1-1)/2, & if \;\; f_2 = 0. \end{cases}$$

Since the number of species in the "bootstrap assemblage" must be an integer, we define $\hat{f}_0^*$ as the smallest integer which is greater than or equal to $\hat{f}_0$. Thus, there are S_{obs} + $\hat{f}_0^*$ species in the bootstrap assemblage. Although the Chao1 estimator is theoretically a lower bound, simulations have suggested that it can be used to estimate the variance of any estimator. This is mainly because very rare species that are not counted in the lower bound have almost negligible effect on variance.

Next we determine the true relative abundances for those species in the bootstrap assemblage. For the S_{obs} species that are observed in the reference sample, assume that the ith species is represented by $X_i > 0$ individuals. The sample relative abundance X_i/n on average overestimates the true relative abundance p_i. This is seen from the following conditional expectation:

$$E\left(\frac{X_i}{n}\middle| X_i > 0\right) = \frac{p_i}{1-(1-p_i)^n} > p_i\,.$$

Thus, we need to tune or adjust the sample frequency X_i /n. Chao and Jost (2012) showed that a very accurate sample coverage estimator for the reference sample based on individual-based abundance data is

$$\hat{C}_{ind}(n) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]\text{ , if } f_2 > 0.$$

When f_2 = 0, a modified formula is

$$\hat{C}_{ind}(n) = 1 - \frac{f_1}{n}\left[\frac{(n-1)(f_1-1)}{(n-1)(f_1-1)+2}\right]\text{ , if } f_2 = 0.$$

Using the concept of sample coverage, Chao et al. (2013, in preparation) derive that the tuned relative abundance in the bootstrap assemblage for the *i*th species is:

$$\hat{p}_i = \frac{X_i}{n}\left[1 - \hat{\lambda}\left(1 - \frac{X_i}{n}\right)^n\right],$$

where

$$\hat{\lambda} = \frac{1-\hat{C}_{ind}(n)}{\sum_{X_i \ge 1} \frac{X_i}{n}\left(1-\frac{X_i}{n}\right)^n}\,.$$

For the remaining $\hat{f}_0^*$ species in the bootstrap assemblage (i.e., those species that were not detected in the sample but exist in the bootstrap assemblage), we assume they all have the same probability $[1-\hat{C}_{ind}(n)]/\ \hat{f}_0^*$. This assumption may look to be restrictive, but the effect on the resulting variance estimator is small.

After the bootstrap assemblage is determined, a random sample of *m* individuals is then generated with replacement. Then a bootstrap estimate ${}^q\hat{D}(m)$ is calculated for the generated sample, i.e., all statistics in our estimators are replaced by those computed from the generated data. Replicate the procedure *B* times and obtain *B* bootstrap estimates (*B* = 200 in our examples). Some preliminary simulations suggested that in our examples a replication size of 200 is sufficient to obtain stable variance estimates and confidence intervals. The bootstrap variance estimator of the estimator ${}^q\hat{D}(m)$ is the sample variance of these *B* estimates. The resulting bootstrap *s.e.* of ${}^{q}\hat{D}(m)$ is then used to construct a 95% confidence interval ${}^{q}\hat{D}(m)$ ± 1.96 *s.e.*[${}^{q}\hat{D}(m)$] for the expected diversity of order q in a sample of size m. Similar procedures can be used to derive variance estimators for any other estimator and the associated confidence intervals.

Incidence data

Consider the independent Bernoulli model in Eq. 2a of the main text. We assume that each species may or may not be detected in each of T independent sampling units (quadrats, plots, traps, microbial culture plates, etc.). We assume in the reference sample that species i is detected in Y_i samples, for i = 1, 2, …, S. Define the *incidence frequency count* Q_k as the number of species that are detected in exactly $Y_i = k$ samples, k = 0, 1, …, T. The independent Bernoulli model assumes that the incidence probability of species i in any sample is π_i. Thus Y_i is a binominal random variable with parameters (T, π_i), as shown in Eq. 2b of the main text. Here we describe the bootstrap assemblage for each sampling unit as all procedures are parallel to those for abundance data. Let $\hat{Q}_0$ be any proper estimator of the number of undetected species. Using the Chao2 estimator (Chao 1987) for species richness, we have

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

As with the abundance data, we define $\hat{Q}_0^*$ as the smallest integer which is greater than or equal to $\hat{Q}_0$. This assures that the number of species in the bootstrap assemblage, $S_{obs} + \hat{Q}_0^*$, is an integer. The species incidence probabilities for the S_{obs} observed species in this bootstrap assemblage are estimated by

$$\hat{\pi}_i = \frac{Y_i}{T}\left[1 - \hat{\tau}\left(1 - \frac{Y_i}{T}\right)^T\right],$$

where

$$\hat{\tau} = \frac{\frac{U}{T}[1 - \hat{C}_{sample}(T)]}{\sum_{Y_i \geq 1} \frac{Y_i}{T}\left(1 - \frac{Y_i}{T}\right)^T}.$$

Here $U = \sum_{i=1}^{S} Y_i$ denotes the total number of incidences in the reference sample, $\hat{C}_{sample}(T)$ denotes the sample coverage estimate for the reference sample (see Eq. (C.7)),

$$\hat{C}_{sample}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right], \text{ if } Q_2 > 0,$$

and

$$\hat{C}_{sample}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)(Q_1-1)}{(T-1)(Q_1-1) + 2}\right], \text{ if } Q_2 = 0.$$

For the remaining $\hat{Q}_0^*$ species that are in the bootstrap assemblage but were not detected in the sample, we assume they all have the same incidence probabilities $(U/T)[1-\hat{C}_{sample}(T)]/\hat{Q}_0^*$. Given the bootstrap assemblage, then a Bernoulli random variable W_{ij} in the independent Bernoulli product model (Eq. 2a of the main text) can be generated, and thus an incidence data matrix is obtained. Other procedures follow those for the abundance data as described above.

LITERATURE CITED

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

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

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