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 H: Rarefaction and extrapolation of Hill numbers for incidence data

All our derivations for the results in Table 2 of the main text for incidence data are generally parallel to those for abundance data, but some modifications are required. In this Appendix, we only sketch the necessary modifications. For incidence data, our model is based on the following binomial product model for the observed incidence-based species frequencies $(Y_1, Y_2, \ldots, Y_S)$: (see Eq. 2b in the main text)

$$P(Y_i = y_i, i = 1, 2, \ldots, S) = \prod_{i=1}^{S} \binom{T}{y_i} \pi_i^{y_i} (1-\pi_i)^{T-y_i} .$$

As defined in the main text, the Hill numbers ${}^q\varDelta(t)$ for the model of incidence data are expressed as

$${}^q\varDelta = \left(\sum_{i=1}^{S} \left[\frac{\pi_i}{\sum_{j=1}^{S} \pi_j} \right]^q \right)^{1/(1-q)} , \ q \geq 0, \ q \neq 1. \qquad \text{(H.1)}$$

These are the Hill numbers based on the relative incidences $\pi_i / \sum_{j=1}^{S} \pi_j$, i = 1, 2,…, S.

For any sample of size t, define the incidence frequency count $Q_k(t)$ as the number of species detected in exactly k sampling units. For the reference sample of size T, we just use Q_k for notational simplicity, i.e., $Q_k = Q_k\,(T)$. The expected value of $Q_k(t)$ can be expressed as:

$$E\,[Q_k(t)] = \sum_{i=1}^{S} \binom{t}{k} \pi_i^k (1-\pi_i)^{t-k} , \ k = 0, 1, \ldots, t. \qquad \text{(H.2)}$$

In particular, $E[Q_0(t)] = \sum_{i=1}^{S} (1-\pi_i)^t$ is the expected number of undetected species in t samples. In the reference sample, we define $U = \sum_{Y_i > 0} Y_i \ = \sum_{j=1}^{T} jQ_j$ as the total number of incidences in the T samples. Here U is an observable variable in the reference sample. Define U_t as the expected total number of incidences for t sampling units, and we have:

$$U_t = \sum_{j=1}^{t} jE[Q_j(t)] = t\sum_{i=1}^{S} \pi_j , \ t \geq 1. \qquad \text{(H.3)}$$

Here, U_t is an unobservable parameter and must be estimated from the reference sample.

As discussed in the main text for abundance data, we define the expected diversity ${}^{q}\Delta(t)$ for t sampling units as the Hill numbers based on the expected incidence frequency counts which are formed by averaging out incidence counts for t sampling units. Suppose a random sample of t sampling units are taken from the entire assemblage, then we obtain a set of incidence frequency counts for this sample, $\{Q_k(t); k = 1, \ldots, t\}$. After an infinite number of such samples have been taken, the average of $Q_k(t)$ for each k =1, 2, …, t tends to $E[Q_k(t)]$ derived in Eq. (H.2). The frequency counts expected in t sampling units consists of the frequency counts $\{E[Q_k(t)]; k = 1, \ldots, t\}$ with the expected total incidences $U_t = \sum_{j=1}^{t} jE[Q_j(t)]$. Note that, for a set of t sampling units, the relative incidences of species are simply $1/U_t$ (there are $E[Q_1(t)]$ such species), $2/U_t$ (there are $E[Q_2(t)]$ such species), …, t/U_t (there are $E[Q_t(t)]$ such species). Thus we can obtain the expected diversity ${}^{q}\Delta(t)$ for t sampling units (t can be any positive integer, not necessarily restricted to $t < T$) as

$$ {}^{q}\Delta(t) = \left(\sum_{k=1}^{t} \left(\frac{k}{U_t} \right)^{q} \times E[Q_k(t)] \right)^{\frac{1}{1-q}} = \left(\sum_{k=1}^{t} \left(\frac{k}{\sum_{j=1}^{t} jE[Q_j(t)]} \right)^{q} \times E[Q_k(t)] \right)^{\frac{1}{1-q}}. \qquad \text{(H.4)} $$

Rarefaction refers to the case where $t < T$ whereas extrapolation refers to the case $t > T$. Throughout the paper and appendices, the theoretical formulas for rarefaction and extrapolation of Hill numbers for the model of incidence data refer to Eq. (H.4). All theoretical formulas for q = 0, 1, and 2, and in general for order $q > 2$, are provided in Table 2 of the main text (the first column). For finding estimators in the rarefaction part, we just replace the parameter U_t in Eq. (H.4) by an estimator (see below) and the expected incidence counts $E[Q_k(t)]$ by their estimators given in the following proposition.

<u>Proposition H1</u>: Under a binomial product model (Eq. 2b in the main text), the minimum variance unbiased estimator for the expected incidence frequency count $E[Q_k(t)]$ is

$$ \hat{Q}_k(t) = \sum_{Y_i \geq k} \frac{\binom{Y_i}{k}\binom{T-Y_i}{t-k}}{\binom{T}{t}}, \quad t < T, \ t \geq 1. \qquad \text{(H.5)} $$

Here, $\binom{a}{b} \equiv 0$ if $a < b$. We use this conventional definition throughout this Appendix. The proof is parallel to that in Proposition D1 of Appendix D and is thus omitted.

Rarefaction/Extrapolation for Species Richness (q = 0)

A similar proof as in Proposition D2 (Appendix D) shows that the rarefaction estimator ${}^0\hat{\Delta}(t)$ is identical to the traditional sample-based rarefaction function. That is,

$$ {}^0\hat{\Delta}(t) = \sum_{k=1}^{t} \hat{Q}_k(t) \ = \tilde{S}_{sample}(t) = S_{obs} - \sum_{Y_i>0} \left[\binom{T-Y_i}{t} \Big/ \binom{T}{t} \right]. $$

Since our estimator $\hat{Q}_k(t)$ for the incidence frequency counts are valid only for $t < T$, they can be used only for rarefaction, but not for extrapolation. The extrapolation estimator for species richness for the expected number of species ${}^0D(T+t^*)$ in a sample of size $t = T + t^*, (t^* > 0)$ is shown in Eq. (C.4) and also in Table 2 of the main text.

Rarefaction/Extrapolation for Shannon diversity (q = 1)

The theoretical formula of diversity of order q =1 for a rarefied sample of size t is

$$ {}^1\Delta(t) = \exp\left(\sum_{k=1}^{t} \left(-\frac{k}{U_t} \log \frac{k}{U_t} \right) \times E[Q_k(t)] \right). \qquad \text{(H.6)} $$

Here we need an estimator for $U_t = \sum_{j=1}^{t} jE[Q_j(t)] = t \sum_{i=1}^{S} \pi_i$. Since $E(U) = T \sum_{i=1}^{S} \pi_j$, an unbiased estimator of the total incidence probabilities, $\sum_{i=1}^{S} \pi_j$, is U/T . This implies from Eq. (H.3) that an unbiased estimate of U_t for any t is $\hat{U}_t = tU/T$. Replacing U_t by $\hat{U}_t$ and $E[Q_k(t)]$ by $\hat{Q}_k(t)$, given in Eq. (H.5), we obtain the rarefaction estimator:

$$ {}^1\hat{\Delta}(t) = \exp\left(\sum_{k=1}^{t} \left(-\frac{k}{\hat{U}_t} \log \frac{k}{\hat{U}_t} \right) \times \hat{Q}_k(t) \right), \ t < T. $$

For the extrapolation of Hill number of q = 1, there is no unbiased estimator $\hat{Q}_k(t)$ as $t > T$. we adopt an approach similar to the one used for abundance data. For incidence data, define $H = H(\infty)$ as the true entropy in the assemblage:

$$ H = H(\infty) = -\sum_{i=1}^{S} \left(\pi_i \Big/ \sum_{j=1}^{S} \pi_j \right) \log\left(\pi_i \Big/ \sum_{j=1}^{S} \pi_j \right). $$

Also, define $H(T)$ as the expected entropy for the reference sample of size T, i.e.,

$$ H(T) = -E\left[\sum_{i=1}^{S} (Y_i/U) \log(Y_i/U) \right]. $$

Chao et al. (2013) recently obtained a nearly unbiased estimator for the entropy under the model of incidence data:

$$\hat{H}_{sample} = \frac{T}{U}\hat{H}_0 + \log\frac{U}{T} \ , \tag{H.7}$$

where

$$\hat{H}_0 = \sum_{k=1}^{T-1}\frac{1}{k}\sum_{1\le Y_i \le T-k}\frac{Y_i}{T}\frac{\binom{T-Y_i}{k}}{\binom{T-1}{k}} + \frac{Q_1}{T}(1-A)^{-T+1}[-\log A - \sum_{r=1}^{T-1}\frac{1}{r}(1-A)^r],$$

and $A = 2Q_2 / [(T-1)Q_1 + 2Q_2]$. Thus, an estimator of the Shannon diversity is ${}^1\hat{\Delta} = \exp(\hat{H}_{sample})$.

After some expansions, we can obtain the following two approximation formulas:

$$H(T) - H \approx -\frac{S - \sum_{i=1}^{S}\pi_i}{2U} \ ,$$

$$H(T+t^*) - H \approx -\frac{S - \sum_{i=1}^{S}\pi_i}{2U(T+t^*)/T} .$$

As with the abundance data, we assume that there is a linear relationship in the theoretical entropy function:

$$H(T+t^*) \approx (1-w)H(T) + wH(\infty).$$

We can then solve for the parameter *w* to obtain

$$w = \frac{H(T+t^*) - H(T)}{H(\infty) - H(T)} \approx \frac{t^*}{T+t^*} .$$

To find an estimator for $H(T+t^*)$, we substitute $H(\infty)$ and $H(T)$ by $\hat{H}_{sample}$ given in Eq. (H.7) and $\hat{H}(T) = -\sum_{i=1}^{S}(Y_i/U)\log(Y_i/U)$, respectively. Then, we obtain the following estimator for the expected entropy of size T+t^*:

$$\hat{H}(T+t^*) = \frac{T}{T+t^*}\hat{H}(T) + \frac{t^*}{T+t^*}\hat{H}_{sample} \ . \tag{H.8}$$

As the augmented sample size t^* tends to infinity, the extrapolated formula (H.8) tends to the entropy estimator $\hat{H}_{sample}$ in Eq. (H.7). For estimating the extrapolated diversity of q = 1, we just take the exponential function of the extrapolated entropy,

$$ {}^{1}\hat{\Delta}(T+t^*) = \exp[\hat{H}(T+t^*)]. \qquad \text{(H.9)} $$

Rarefaction/Extrapolation for Simpson diversity (q = 2)

For q = 2, the theoretical formula for any sample size t is

$$ {}^{2}\Delta(t) = \frac{1}{\sum_{k=1}^{t}\left(\frac{k}{U_t}\right)^2 \times E[Q_k(t)]},\ t \geq 1. \qquad \text{(H.10)} $$

Replacing U_t by $\hat{U}_t$ and $E[Q_k(t)]$ by $\hat{Q}_k(t)$, we obtain our proposed rarefaction estimator:

$$ {}^{2}\hat{\Delta}(t) = \frac{1}{\sum_{k=1}^{t}\left(\frac{k}{\hat{U}_t}\right)^2 \times \hat{Q}_k(t)} = \frac{1}{\frac{1}{t}\times\frac{1}{U/T} + \frac{t-1}{t}\sum_{Y_i>0}\frac{Y_i(Y_i-1)}{U^2(1-1/T)}}. \qquad \text{(H.11)} $$

Applying our general formula (H.10) to an augmented sample size of T + t^*, we obtain the following expected diversity of order 2:

$$ {}^{2}\Delta(T+t^*) = \frac{1}{\sum_{k=1}^{T+t^*}\left(\frac{k}{U_{T+t^*}}\right)^2 \times E[Q_k(T+t^*)]}. $$

The denominator in the above formula can be simplified to

$$ \begin{aligned} &\sum_{k=1}^{T+t^*}\left(\frac{k}{U_{T+t^*}}\right)^2 \times E[Q_k(T+t^*)] \\ &= \sum_{i=1}^{S}\sum_{k=1}^{T+t^*}\left(\frac{k}{U_{T+t^*}}\right)^2 \binom{T+t^*}{k}\pi_i^k(1-\pi_i)^{T+t^*-k} \\ &= \left(\frac{1}{U_{T+t^*}}\right)^2 \sum_{i=1}^{S}[(T+t^*)\pi_i + (T+t^*)(T+t^*-1)\pi_i^2]. \end{aligned} $$

In the above formula, we substitute $\sum_{i=1}^{S}\pi_j$, $\sum_{i=1}^{S}\pi_i^2$ and U_{T+t*} respectively, by their unbiased estimators U/T , $\sum_{i=1}^{S}Y_i(Y_i-1)/[T(T-1)]$, and $\hat{U}_{T+t*}=(T+t^*)U/T$, and obtain a nearly unbiased estimator for the extrapolated diversity ${}^2D(T+t^*)$:

$$ {}^2\hat{\varDelta}(T+t^*)=\frac{1}{\dfrac{1}{T+t^*}\times\dfrac{1}{U/T}+\dfrac{T+t^*-1}{T+t^*}\sum\limits_{Y_i>0}\dfrac{Y_i(Y_i-1)}{U^2(1-1/T)}}. \qquad \text{(H.12)} $$

As t^* tends to infinity, we obtain the following nearly unbiased estimator for the asymptotic diversity:

$$ {}^2\hat{\varDelta}(\infty)=\frac{(1-1/T)U^2}{\sum\limits_{Y_i>1}Y_i(Y_i-1)}. \qquad \text{(H.13)} $$

Rarefaction/Extrapolation for Hill number of order q> 2

In the theoretical formula ${}^q\varDelta(t)$ given in Eq. (H.4) of any order q, we can replace U_t by $\hat{U}_t$ and $E[Q_k(t)]$ by $\hat{Q}_k(t)$ to obtain our proposed rarefaction estimator:

$$ {}^q\hat{\varDelta}(t)=\left(\sum_{k=1}^{t}\left(\frac{k}{\hat{U}_t}\right)^q\times\hat{Q}_k(t)\right)^{\frac{1}{1-q}}. $$

For extrapolation, we let $\psi(q,j)$ be the Stirling number of the second kind defined by the coefficient in the expansion $x^q=\sum_{j=1}^{q}\psi(q,j)\,x^{(j)}$, where $x^{(j)}=x(x-1)...(x-j+1)$ denotes the falling factorial function. Also, let V_i be a binomial random variable with parameter $T+t^*$ and probability π_i. Then,

$$ \sum_{k=1}^{T+t^*}\left(\frac{k}{U_{T+t*}}\right)^q\times E[Q_k(T+t^*)] $$

$$ =\sum_{i=1}^{S}\sum_{k=1}^{T+t^*}\left(\frac{k}{U_{T+t*}}\right)^q\binom{T+t^*}{k}\pi_i^k(1-\pi_i)^{T+t^*-k} $$

$$ =\left(\frac{1}{U_{T+t*}}\right)^q\sum_{i=1}^{S}\;EV_i^q $$

$$= \frac{1}{(U_{T+t*})^q} \sum_{i=1}^{S} \sum_{j=1}^{q} \psi(q, j) E[V_i^{(j)}]$$

$$= \sum_{i=1}^{S} \sum_{j=1}^{q} \frac{\psi(q, j)(T+t^*)^{(j)}}{(U_{T+t^*})^q} \pi_i^j .$$

The last equality follows from a moment property of a binomial distribution with parameter T+t^* and probability π_i: $E(V_i^{(j)}) = (T+t^*)^{(j)} \pi_i^j$. Replacing U_{T+t*} and $\sum_{i=1}^{S} \pi_i^j$, respectively, with their unbiased estimators $\hat{U}_{T+t*} = U(1+t^*/T)$ and $\sum_{Y_i \geq j} Y_i^{(j)} / T^{(j)}$ (Good 1953), we obtain the proposed nearly unbiased predictor for $q > 2$ as shown in Table 2 of the main text:

$${}^q\hat{\Delta}\,(T+t^*) = \left(\frac{1}{(U/T)^q} \sum_{j=1}^{q} \frac{\psi(q, j)(T+t^*)^{(j)}}{(T+t^*)^q} \sum_{Y_i \geq j} \frac{Y_i^{(j)}}{T^{(j)}} \right)^{\frac{1}{1-q}} .$$

As t^* tends to infinity, the nearly unbiased estimator for the asymptotic diversity ${}^q\Delta(\infty) = [\sum_{i=1}^{S} \pi_i^q / (\sum_{j=1}^{S} \pi_j)^q]^{1/(1-q)}$ for $q \geq 2$ is:

$${}^q\hat{\Delta}(\infty) = \left(\frac{1}{(U/T)^q} \sum_{Y_i \geq q} \frac{Y_i^{(q)}}{T^{(q)}} \right)^{1/(1-q)} . \qquad \text{(H.14)}$$

This estimator can also obtained by noting that $\sum_{Y_i \geq q} [Y_i^{(q)} / T^{(q)}]$ is an unbiased estimator of $\sum_{i=1}^{S} \pi_i^q$ (Good 1953) and U/T is an unbiased estimator for $\sum_{i=1}^{S} \pi_i$.

A replication principle and its generalization for the model of incidence data

Proposition H2: A replication principle for the model of incidence data. Assume Assemblage 2 consists of K replicates of Assemblage 1. Each replicate has the same number of species and the same species incidence probabilities as Assemblage 1, but with completely different, unique species in each replicate. A sample of t sampling units is taken from Assemblage 1. Then number of sampling units needed in Assemblage 2 to attain the same expected sample coverage is approximately Kt, and the expected diversity of any order $q \geq 0$ in Assemblage 2 for the sample with standardized coverage is approximately K times of that in Assemblage 1.

Proposition H3: A generalization of the replication principle discussed in Proposition H2. If Assemblage 2 is unambiguously K times more diverse than Assemblage 1 (i.e., for all $q \geq 0$, Hill number of order q of Assemblage 2 is K times that of Assemblage 1), then in the coverage-based standardization, the expected diversity of any order $q \geq 0$ in Assemblage 2 is approximately K times of that in Assemblage 1.

The proof for these two propositions is generally parallel to that for abundance data (Propositions D4 and D5 in Appendix D) and thus is omitted.

LITERATURE CITED

Chao, A., Y. T. Wang and L. Jost. 2013. Entropy and species accumulation curve: a nearly unbiased entropy estimator via discovery rates of new species. Under revision, Methods in Ecology and Evolution.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237-264.