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 E: Extrapolation formulas for Hill numbers of *q* = 1 and *q* ≥ 2 based on abundance data

Extrapolation for Hill numbers of order q =1

We first derive an extrapolation formula for Shannon entropy. In order for this demonstration to be self-contained, we repeat some definitions described in the main text. Let H be the true (asymptotic) Shannon entropy in the assemblage, i.e., H represents an asymptotic extrapolated value when the sample size tending to infinity. That is,

$$H = H(\infty) = -\sum_{i=1}^{S} p_i \log p_i \cdot$$

Chao et al. (2013) recently derived the following estimator of H using statistical sampling theory:

$$\hat{H} = \hat{H}(\infty) = \sum_{k=1}^{n-1} \frac{1}{k} \sum_{1 \le X_i \le n-k} \frac{X_i}{n} \frac{\binom{n-X_i}{k}}{\binom{n-1}{k}} + \frac{f_1}{n}(1-A)^{-n+1} \left\{ -\log(A) - \sum_{r=1}^{n-1} \frac{1}{r}(1-A)^r \right\}, \tag{E.1}$$

where $A = 2f_2 / [(n-1)f_1 + 2f_2]$. Let $H(n)$ be the expected entropy for a reference sample of size n,

$$H(n) = -E\left(\sum_{i=1}^{S} \frac{X_i}{n} \log \frac{X_i}{n} \right).$$

The extrapolation here is to predict the expected entropy for a sample size $n+m^*$, $H(n+m^*)$ and derive its estimator based on a reference sample. Since the entropy is a slow-varying function of sample size, it is reasonable to assume that it is linear in sample size as in the following expression:

$$H(n+m^*) = (1-w)H(n) + wH(\infty), \tag{E.2}$$

for some w, where $0 < w < 1$ and can be estimated from data. From the bias property of the entropy estimator (Basharin 1959), we have the approximation formula:

$$H(n) - H \approx -\frac{S-1}{2n}.$$

Then we can solve for the parameter w and obtain

$$w = \frac{H(n+m^*) - H(n)}{H - H(n)} = \frac{[H(n+m^*) - H] - [H(n) - H]}{H - H(n)}$$

$$\approx \frac{-1/(n+m^*) + 1/n}{1/n} = \frac{m^*}{n+m^*}.$$

Therefore, the expected entropy with sample size $n + m^*$ turns out to be:

$$H(n+m^*) = \frac{n}{n+m^*} H(n) + \frac{m^*}{n+m^*} H(\infty). \tag{E.3}$$

To find an estimator for $H(n+m^*)$, we substitute $H(\infty)$ and $H(n)$ by $\hat{H}(\infty)$ in Eq. (E.1) and $\hat{H}(n) = -\sum_{i=1}^{S} (X_i / n) \log(X_i / n)$ respectively. Then we obtain the following estimator for the expected entropy of size $n+m^*$:

$$\hat{H}(n+m^*) = \frac{n}{n+m^*} \hat{H}(n) + \frac{m^*}{n+m^*} \hat{H}(\infty). \tag{E.4}$$

For estimating the extrapolated diversity with $q = 1$, we just take the exponential function of the extrapolated entropy. That is, the proposed extrapolated estimator is

$${}^1\hat{D}(n+m^*) = \exp[\hat{H}(n+m^*)].$$

Extrapolation for Hill numbers of a general integer order $q \geq 2$

From the main text (Table 1), the extrapolation aims to predict the expected diversity of an extrapolated size $n+m^*$. That is, we want to estimate:

$${}^q D(n+m^*) = \left(\sum_{k=1}^{n+m^*} \left(\frac{k}{n+m^*} \right)^q \times E[f_k(n+m^*)] \right)^{\frac{1}{1-q}}.$$

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 $x^{(j)} = x(x-1)...(x-j+1)$. Let V_i be a binomial random variable with parameter $n+m^*$ and probability p_i. Then, we can write

$$\sum_{k=1}^{n+m^*} \left(\frac{k}{n+m^*} \right)^q \times E[f_k(n+m^*)] = \sum_{i=1}^{S} \sum_{k=1}^{n+m^*} \left(\frac{k}{n+m^*} \right)^q \binom{n+m^*}{k} p_i^k (1-p_i)^{n+m^*-k}$$

$$= \sum_{i=1}^{S} E[V_i^q / (n+m^*)^q]$$

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

$$= \sum_{i=1}^{S} \sum_{j=1}^{q} \frac{\psi(q, j)(n+m^*)^{(j)}}{(n+m^*)^q} p_i^j .$$

The last equality follows from a factorial moment property for the binomial distribution with parameter $n+m^*$ and probability p_i, i.e., $E(V_i^{(j)}) = (n+m^*)^{(j)} p_i^j$. From Good (1953), an unbiased estimator for $\sum_{i=1}^{S} p_i^j$ is $\sum_{X_i \geq j} X_i^{(j)} / n^{(j)}$, so we obtain a nearly unbiased predictor of ${}^qD(n+m^*)$ as shown in Table 1 of the main text:

$$ {}^q\hat{D}(n+m^*) = \left(\sum_{j=1}^{q} \frac{\psi(q, j)(n+m^*)^{(j)}}{(n+m^*)^q} \sum_{X_i \geq j} \frac{X_i^{(j)}}{n^{(j)}} \right)^{\frac{1}{1-q}} .$$

As m^* tends to infinity, we obtain the following nearly unbiased estimator for the asymptotic diversity ${}^qD(\infty) = (\sum_{i=}^{S} p_i^q)^{1/(1-q)}$ for $q \geq 2$:

$$ {}^q\hat{D}(\infty) = [\sum_{X_i \geq q} X_i^{(q)} / n^{(q)}]^{1/(1-q)} . \qquad \text{(E.5)}$$

This estimator can also obtained by noting that $\sum_{X_i \geq q} [X_i^{(q)} / n^{(q)}]$ is an unbiased estimator of $\sum_{i=1}^{S} p_i^q$ (Good 1953).

LITERATURE CITED

Basharin, G. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. Theory of Probability and its Applications 4:333-336.

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 review.

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