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 D: Proof details for some formulas (Eqs. 5, 8, 9b, 11a and 11b of the main text) and a replication principle

Proposition D1 (Eqs. 5 and 8 in the main text): Assume that sample species frequencies (X_1, X_2, …, X_S) obey this multinomial model with cell total n and cell probabilities (p_1, p_2, …, p_S):

$$P(X_1 = x_1,..., X_S = x_S) = \frac{n!}{x_1!... x_S!} p_1^{x_1} p_2^{x_2} ... p_S^{x_S} .$$

(a) The frequency counts expected in a sample of size m consist of the frequency counts $\{E[f_k(m)]; k = 1, \ldots, m\}$ and

$$E[\, f_k(m)] = \sum_{i=1}^{S} \binom{m}{k} p_i^k (1 - p_i)^{m-k} \text{ , } k = 0, 1, \ldots, m.$$

(b) Let $S_{ind}(m)$ be the expected species richness in a sample of size m; see Eq. (B.1) of Appendix B. Then the species richness based on the expected frequency counts $\{E[f_k(m)]; k = 1, \ldots, m\}$ is identical to $S_{ind}(m)$ for any m. That is,

$$\sum_{k=1}^{m} E[\, f_k(m)] = S_{ind}(m) .$$

This implies that the species richness based on the expected frequency counts in a sample of size m is identical to the expected species richness in a sample of the same size.

(c) The minimum variance unbiased estimator of the expected frequency count $E[\, f_k(m)]$ is

$$\hat{f}_k(m) = \sum_{X_i \geq k} \frac{\binom{X_i}{k}\binom{n - X_i}{m - k}}{\binom{n}{m}}, \quad m < n, \; k \geq 1.$$

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

Proof:

(a) Assume that the *i*th species are represented by Z_i (m) individuals in a sample of size m. Under the multinomial model, the variable $Z_i(m)$ follows a binomial distribution with parameter m and probability p_i. Then $f_k(m) = \sum_{i=1}^{S} I[Z_i(m) = 1]$, which implies

$$E\,[\,f_k(m)] = \sum_{i=1}^{S} P[Z_i(m) = k] = \sum_{i=1}^{S} \binom{m}{k} p_i^k (1-p_i)^{m-k}\ ,\ m < n.$$

(b) The conclusion in (b) follows from the following identity:

$$\sum_{k=1}^{m} E[\,f_k(m)] = \sum_{i=1}^{S} \sum_{k=1}^{m} \binom{m}{k} p_i^k (1-p_i)^{m-k} = \sum_{i=1}^{S} [1-(1-p_i)^m] = S_{ind}(m).$$

(c) Under the multinomial assumption, the sample frequency X_i in the reference sample follows a binomial distribution with parameter n and probability p_i. Then we have

$$E[\,\hat{f}_k(m)] = E \sum_{X_i \geq k} \frac{\binom{X_i}{k}\binom{n-X_i}{m-k}}{\binom{n}{m}}$$

$$= \sum_{i=1}^{S} \sum_{x=k}^{n-m+k} \frac{\binom{x}{k}\binom{n-x}{m}}{\binom{n}{m}} \binom{n}{x} p_i^x (1-p_i)^{n-x}$$

$$= \sum_{i=1}^{S} \binom{m}{k} p_i^k (1-p_i)^{m-k} \sum_{x=k}^{n-m+k} \binom{n-m}{x-k} p_i^{x-k} (1-p_i)^{n-m-(x-k)}$$

$$= \sum_{i=1}^{S} \binom{m}{k} p_i^k (1-p_i)^{m-k} \sum_{y=0}^{n-m} \binom{n-m}{y} p_i^y (1-p_i)^{n-m-y}$$

$$= \sum_{i=1}^{S} \binom{m}{k} p_i^k (1-p_i)^{m-k}.$$

From the Rao-Blackwell and Lehmann-Scheffé Theorems (e.g., Casella and Berger 2002, p. 347 and p. 369), the estimator is the unique minimum variance unbiased estimator of the expected abundance frequency.

Proposition D2 (Eq. 9b in the main text): The estimated species richness ${}^0\hat{D}(m)$ based on the observed species frequency counts $\hat{f}_k(m)$, $k = 1, \ldots, m$ for a rarefied sample of size m, is identical to the traditional abundance-based rarefaction function. That is, we have

$$ {}^0\hat{D}(m) = \sum_{k=1}^{m} \hat{f}_k(m) \; = \tilde{S}_{ind}(m), $$

where $\tilde{S}_{ind}(m)$ is the traditional rarefaction formula given in Eq. (B.2),

$$ \tilde{S}_{ind}(m) = S_{obs} - \sum_{X_i>0} \left[\binom{n-X_i}{m} \Big/ \binom{n}{m} \right], \; m < n. $$

This concludes that the species richness based on the estimated frequency counts in a rarefied sample of size m is identical to the estimated species richness in a rarefied sample of the same size. This implies the data-based version of the theoretical relationship in (b) of Proposition D1 is valid for rarefied samples.

Proof: The result follows from the following derivations:

$$ \sum_{k=1}^{m} \hat{f}_k(m) = \sum_{k=1}^{m} \sum_{X_i \geq k} \frac{\binom{X_i}{k}\binom{n-X_i}{m-k}}{\binom{n}{m}} = \sum_{X_i \geq 1} \sum_{k \neq 0} \frac{\binom{X_i}{k}\binom{n-X_i}{m-k}}{\binom{n}{m}} $$

$$ = \sum_{X_i \geq 1} \left\{ 1 - \binom{n-X_i}{m} \Big/ \binom{n}{m} \right\} = S_{obs} - \sum_{X_i>0} \left[\binom{n-X_i}{m} \Big/ \binom{n}{m} \right]. $$

Proposition D3 (Eqs. 11a and 11b of the main text): Under the model assumption in Proposition D1, we have the following conclusions.

(a) For $m \geq 1$, we have

$$ {}^2D(m) = \frac{1}{\sum_{k=1}^{m} \left(\frac{k}{m}\right)^2 \times E[\, f_k(m)]} = \frac{1}{\frac{1}{m} + \frac{m-1}{m} \sum_{i=1}^{S} p_i^2}. $$

(b) The two estimators for ${}^2\hat{D}(m)$ in Eq. 11b of the main text are identical. That is,

$$^2\hat{D}(m) = \frac{1}{\sum_{k=1}^{m}\left(\frac{k}{m}\right)^2 \hat{f}_k(m)} = \frac{1}{\frac{1}{m} + \frac{m-1}{m}\sum_{i=1}^{S}\frac{X_i(X_i-1)}{n(n-1)}} \ .$$

Proof:

(a) Let $Z_i(m)$ be a binomial distribution with parameter m and probability p_i, as we defined in the proof of Proposition D1. Then we can express

$$\begin{aligned}
\sum_{k=1}^{m}\left(\frac{k}{m}\right)^2 \times E[f_k(m)] &= \sum_{i=1}^{S} E\left(\frac{Z_i(m)}{m}\right)^2 \\
&= \sum_{i=1}^{S}\left(\mathrm{var}\left(\frac{Z_i(m)}{m}\right) + \left[E\left(\frac{Z_i(m)}{m}\right)\right]^2\right) \\
&= \sum_{i=1}^{S} \frac{mp_i(1-p_i)}{m^2} + \sum_{i=1}^{S} p_i^2 = \frac{1}{m} + \frac{m-1}{m}\sum_{i=1}^{S} p_i^2 .
\end{aligned}$$

(b) Note that we have

$$\sum_{k=1}^{m}\left(\frac{k}{m}\right)^2 \hat{f}_k(m) = \sum_{k=1}^{m}\left(\frac{k}{m}\right)^2 \sum_{X_i \geq k} \frac{\binom{X_i}{k}\binom{n-X_i}{m-k}}{\binom{n}{m}}$$

$$= \sum_{X_i \geq 1} E\left(\frac{R_i}{m}\right)^2 ,$$

where R_i is a hypergeometric random variable with the probability density function:

$$P(R_i = k) = \frac{\binom{X_i}{k}\binom{n-X_i}{m-k}}{\binom{n}{m}} .$$

Using the expectation and variance of a hypergeometric distribution, we obtain

$$\sum_{X_i \geq 1} E\left(\frac{R_i}{m}\right)^2 = \sum_{X_i \geq 1}\left(\mathrm{var}\left(\frac{R_i}{m}\right) + \left[E\left(\frac{R_i}{m}\right)\right]^2\right)$$

$$= \sum_{X_i \geq 1} \left(\frac{1}{m} \frac{X_i}{n} \frac{(n - X_i)}{n} \frac{(n-m)}{n-1} + \left(\frac{X_i}{n} \right)^2 \right)$$

$$= \frac{1}{m} + \frac{m-1}{m} \sum_{X_i \geq 1} \frac{X_i (X_i - 1)}{n(n-1)}.$$

A replication principle and its generalization

Proposition D4: A replication principle for the model of abundance data. Assume that Assemblage 2 consists of K replicates of Assemblage 1. Each replicate has the same number of species and the same species abundances as Assemblage 1, but with completely different, unique species in each replicate. A sample of m individuals is taken from Assemblage 1. Then the sample size needed in Assemblage 2 to attain the same expected sample coverage is approximately Km, 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.

Proof: Without loss of generality, we prove the theorem for $K = 2$ (doubling property). Assume that there are S species in Assemblage 1, with species relative abundances or species probabilities (p_1, p_2, …, p_S). Since Assemblage 2 is a doubled replicate of Assemblage 1, in Assemblage 2 there are $2S$ species with relative abundances ($p_1/2$, $p_1/2$, $p_2/2$, $p_2/2$, …, $p_S/2$, $p_S/2$). As proved by Chao and Jost (2012, their Appendix A), if a sample of m individuals is taken from Assemblage 1, then the sample size needed in Assemblage 2 to attain the same expected sample coverage is approximately Km. In this case, the expected diversity of order q in Assemblage 1 based on Eq. 6 of the main text is

$$ {}^{q}D_1(m) = \left(\sum_{k=1}^{m} \left(\frac{k}{m} \right)^q \times E[\, f_{k,1}(m)] \right)^{\frac{1}{1-q}}, \; m \geq 1, \; q \neq 1. \qquad \text{(D.1)}$$

Here the abundance frequency count $f_{k,1}(m)$ is the number of species represented by exactly k individuals in a sample of size m in Assemblage 1. From Eq. 5 of the main text, when m is large enough we have for $k = 0, 1, \ldots, m$

$$E\,[\, f_{k,1}(m)] = \sum_{i=1}^{S} \binom{m}{k} p_i^k (1 - p_i)^{m-k} \approx \sum_{i=1}^{S} \frac{(mp_i)^k}{k!} e^{-mp_i}. \qquad \text{(D.2)}$$

The approximation is satisfactory in the following sense: either the both sides of (D.2) are negligible, or the relative error with respective to the right hand side of (D.2) tends to zero; see Harris (1959, his Appendix A) for proof details. For Assemblage 2, the expected diversity of order q with a sample of $2m$ is

$$ {}^{q}D_2(m) = \left(\sum_{k=1}^{2m} \left(\frac{k}{2m} \right)^q \times E[\, f_{k,2}(2m)] \right)^{\frac{1}{1-q}}, \; m \geq 1, \; q \neq 1. \qquad \text{(D.3)} $$

Here the abundance frequency count $f_{k,2}(2m)$ is the number of species represented by exactly k individuals in a sample of size $2m$ in Assemblage 2. Again, it follows from Harris (1959) that for $k = 0, 1, \ldots, 2m$

$$ E\,[\, f_{k,2}(2m)] = 2\sum_{i=1}^{S} \binom{2m}{k} \left(\frac{p_i}{2} \right)^k \left(1 - \frac{p_i}{2} \right)^{2m-k} \approx 2\sum_{i=1}^{S} \frac{(mp_i)^k}{k!} e^{-mp_i} \; . \qquad \text{(D.4)} $$

The above approximation is in the same sense as that in Eq. (D.2). This shows the following relationship for any k:

$$ E\,[\, f_{k,2}(2m)] \approx 2E\,[\, f_{k,1}(m)]. \qquad \text{(D.5)} $$

Note that in Assemblage 1, no species probability in the abundance set ($p_1, p_2, \ldots, p_S$) of Assemblage 1 is greater than unity, so any species probability in the abundance set ($p_1/2, p_1/2, p_2/2, p_2/2, \ldots, p_S/2, p_S/2$) of Assemblage 2 is not greater than 1/2. Note that for any $k > m$ in Eq. (D.3), we have $k/(2m) > 1/2$. So the limit in the summation is only from $k = 1$ to $k = m$. It then follows from (D.3) and (D.5) that we have

$$ {}^{q}D_2(m) = \left(\sum_{k=1}^{2m} \left(\frac{k}{2m} \right)^q \times E[\, f_{k,2}(2m)] \right)^{\frac{1}{1-q}} \approx \left(\sum_{k=1}^{m} \left(\frac{k}{2m} \right)^q \times 2E[\, f_{k,1}(m)] \right)^{\frac{1}{1-q}} $$

$$ = \left(\sum_{k=1}^{m} 2^{1-q} \left(\frac{k}{m} \right)^q \times E[\, f_{k,1}(m)] \right)^{\frac{1}{1-q}} = 2[{}^{q}D_1(m)], $$

which shows that the expected diversity of Assemblage 2 for a sample of size $2m$ is approximately double that of Assemblage 1 for a size of m, if both samples are standardized to the same degree of completeness. A similar proof can be made for $q = 1$ and any value of k.

A critical step in the proof is the approximation formula $E\,[\, f_{k,2}(2m)] \approx 2E\,[\, f_{k,1}(m)]$ derived in Eq. (D.5). A simple explanation can be seen from the perspective of species frequencies. Note that the expected species frequencies for the abundance set ($p_1, p_2, \ldots, p_S$) of Assemblage 1, based on a sample of size m, are ($mp_1, mp_2, \ldots, mp_S$). The expected species frequencies for the abundance set ($p_1/2, p_1/2, p_2/2, p_2/2, \ldots, p_S/2, p_S/2$) of Assemblage 2, based on a sample of size $2m$, are ($mp_1, mp_1, mp_2, mp_2, \ldots, mp_S, mp_S$), which is a duplication of the frequency set of

Assemblage 1. This intuitively validates Eq. (D.5) and explains why the sample size must be doubled in Assemblage 2 in order to standardize sample coverage. It also explains why the expected diversity for standardized sample size does not obey the replication principle.

Following the proof of Chao and Jost (2012, their Appendix A), we can obtain a generalization of the replication principle by means of the following proposition.

Proposition D5: A generalization of the replication principle discussed in Proposition D4. 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.

LITERATURE CITED

Casella, G. and R. L. Berger. 2002. Statistical inference, second Edition. Duxbury, Pacific Grove, California, USA.

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

Harris, B. 1959. Determining bounds on integrals with applications to cataloging problems. Annals of Mathematical Statistics 30:521-548.