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 I: Hill numbers and Hurlbert's indices

Hurlbert (1971) suggested a class of species indices $\{S(m);\ m = 1, 2, \ldots\}$ as a family of useful measures. Here $S(m)$ is defined as the expected number of species in a sample of m individuals selected at random (with replacement) from an assemblage with S species and the relative abundances $\{p_1, p_2, \ldots, p_S\}$. The plot of $S(m)$ with respect to sample size m is the traditional expected species accumulation curve (SAC). Here the Hurlbert index $S(m)$ is identical to the theoretical formulas $S_{ind}(m)$ given in Eq. (B.1) for abundance data. Given a reference sample, the sample-size-based rarefaction and extrapolation formulas (Colwell et al. 2012) for species richness provide estimators of Hurlbert indices. In this Appendix, we simply use the notation $S(m)$ instead of $S_{ind}(m)$ for notational simplicity. That is, from Eq. (B.1), we have (Good 1953)

$$S(m) = \sum_{i=1}^{S} [1 - (1 - p_i)^m] = S - \sum_{i=1}^{S} (1 - p_i)^m \ . \qquad \text{(I.1)}$$

For Hurlbert indices, we have $S(1) = 1$, and as m tends to infinity, $S(m)$ tends to species richness. So species richness corresponds to the order infinity of Hurlbert's indices. We now show that Hill numbers and Hurlbert indices are mathematically equivalent in the sense that they contain the same information about biodiversity.

Since Hurlbert's indices apply to integer numbers, we only consider Hill numbers for integer order q. Here, for the first time, we show that the set of measures $\{S(m);\ m = 2, 3\ldots\}$ and the set of Hill numbers restricted to non-negative integers, i.e., $\{{}^qD; q = 0, 1, 2, \ldots\}$, contain exactly the same information, in the sense that each element of one set is a function of the other set, i.e., if we know one set, then the other set is totally known. We first prove the case of a finite order q in the following proposition.

Proposition I1: The two sets of finite elements $\{{}^2D, {}^3D, \ldots, {}^qD\}$ and $\{S(2), S(3), \ldots, S(q)\}$ for a positive integer q contain the same information.

Proof: We begin by showing for any finite $q = 2, 3, \ldots$, that any $S(q)$ can be expressed as the following function of Hill numbers $\{{}^2D, {}^3D, \ldots, {}^qD\}$. A direct expansion leads to (Leinster and Cobbold 2012)

$$S(q) = S - \sum_{i=1}^{S}(1-p_i)^q = S - \sum_{i=1}^{S}\sum_{r=0}^{q}\binom{q}{r}(-1)^r p_i^r$$

$$= S - S + \binom{q}{1}\left(\sum_{i=1}^{S} p_i^1\right) - \sum_{r=2}^{q}\binom{q}{r}(-1)^r\left(\sum_{i=1}^{S} p_i^r\right)$$

$$= q + \sum_{r=2}^{q}\binom{q}{r}(-1)^{r+1}({}^rD)^{1-r}. \qquad \text{(I.2)}$$

This formula shows that if we know the values of ${}^2D,\ {}^3D,...,{}^qD$, then we know the values of $\{S(2), S(3),..., S(q)\}$.

On the other hand, we show any Hill number of order ${}^qD$, for any fixed integer q = 2, 3, …, is a function of $\{S(2), S(3),..., S(q)\}$. This can be seen by the following identity

$$({}^qD)^{1-q} = \sum_{i=1}^{S} p_i^q = \sum_{i=1}^{S}[1-(1-p_i)]^q = \sum_{i=1}^{S}\sum_{r=0}^{q}\binom{q}{r}(-1)^r(1-p_i)^r$$

$$= S - \sum_{i=1}^{S}\binom{q}{1}(1-p_i) + \sum_{i=1}^{S}\sum_{r=2}^{q}\binom{q}{r}(-1)^r(1-p_i)^r$$

$$= S - q(S-1) + \sum_{r=2}^{q}\binom{q}{r}(-1)^r[S-S(r)]$$

$$= S - q(S-1) + S(q-1) + \sum_{r=2}^{q}\binom{q}{r}(-1)^{r+1}S(r)$$

$$= q + \sum_{r=2}^{q}\binom{q}{r}(-1)^{r+1}S(r).$$

Thus, we have

$${}^qD = \left\{q + \sum_{r=2}^{q}\binom{q}{r}(-1)^{r+1}S(r)\right\}^{1/(1-q)}$$

$$= \left\{q - \binom{q}{2}S(2) + \binom{q}{3}S(3) - ... + \binom{q}{q}(-1)^{q+1}S(q)\right\}^{1/(1-q)}. \qquad \text{(I.3)}$$

This formula shows that if we know the values of $\{S(2), S(3),..., S(q)\}$, then we know the values of ${}^2D,\ {}^3D,...,{}^qD$. Then, from Eqs. (I.2) and (I.3), we can conclude that the two finite sets $\{{}^2D,\ {}^3D,...,{}^qD\}$ and $\{S(2), S(3),..., S(q)\}$ contain the same information.

Proposition I2: The infinite sets $\{S(m);\ m = 2, 3\ldots\}$ for Hurlbert's indices and the infinite set for Hill numbers $\{{}^qD; q = 0, 1, 2,...\}$ contain the same information.

Proof: From Proposition I1, we only need to notice that ${}^0D = S(\infty)$ = species richness S, and the expression ${}^1D = \exp\{-1 + \sum_{m=2}^{\infty} S(m)/[m(m-1)]\}$ as proved by Mao (2007) using a complicated approach. Here we give a simple direct proof:

$$
\begin{aligned}
-\sum_{i=1}^{S} p_i \log p_i &= -\sum_{i=1}^{S} p_i \log[1-(1-p_i)] \\
&= \sum_{m=1}^{\infty}\sum_{i=1}^{S} \frac{p_i(1-p_i)^m}{m} \\
&= \sum_{m=1}^{\infty} \frac{S(m+1)-S(m)}{m} \\
&= \sum_{m=1}^{\infty} \frac{S(m+1)}{m} - \sum_{m=1}^{\infty} \frac{S(m)}{m} \\
&= \sum_{m=2}^{\infty} \frac{S(m)}{m-1} - \left[1+\sum_{m=2}^{\infty} \frac{S(m)}{m}\right] \\
&= -1+\sum_{m=2}^{\infty} \frac{S(m)}{m(m-1)}.
\end{aligned}
$$

Our proof implies that, if we know the SAC, then Hill numbers for non-negative integer values are all known, and vice versa. Thus, we can conclude that the two important families of diversity measures, Hurlbert's indices and Hill numbers (with non-negative integer order q), are mathematically equivalent.

The slope at the sample size m =1 of an expected SAC and rarefaction curve as a function of sample size has been noted to characterize important quantities. Given the SAC formula in Eq. (I.1), Lande (2000) noted that this slope is

$$S(2)-S(1) = 1-\sum_{i=1}^{S} p_i^{\,2} = 1-({}^{2}D)^{-1}. \qquad \text{(I.4)}$$

That is, this slope is identical to the Gini-Simpson index, which itself is a simple transformation of the Hill number for q=2. Given a reference sample of n, the traditional sample-size-based rarefaction below provides estimators of Hurlbert indices $S(m)$ for $m < n$:

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

Olszewski's (2004) found that the corresponding slope of this curve is

$$\tilde{S}_{ind}(2)-\tilde{S}_{ind}(1) = 1-\sum_{X_i\geq 2}\frac{X_i(X_i-1)}{n(n-1)} = \frac{n}{n-1}\left(1-\sum_{X_i\geq 1}\left(\frac{X_i}{n}\right)^2\right). \qquad \text{(I.5)}$$

This is Hurlbert's (1971) Probability of an Interspecific Encounter (PIE) for a sample; see Appendix J. For the coverage-based SAC, in which $S(m)$ is depicted as a function of the $C(m)$, we have similar findings. Notice that Eq. (B.8) gives the following formula of the expected coverage for a sample of size m for individual-based abundance data:

$$C(m) = 1 - \sum_{i=1}^{S} p_i (1 - p_i)^m \text{ , } m > 0. \qquad \text{(I.6)}$$

The following analytic minimum variance unbiased estimator of the expected coverage $C(m)$ is derived by Chao and Jost (2012):

$$\hat{C}(m) = 1 - \sum_{X_i \geq 1} \frac{X_i}{n} \frac{\binom{n - X_i}{m}}{\binom{n-1}{m}}, \quad m < n, \qquad \text{(I.7)}$$

where X_i is the number of individuals of species i observed in the reference sample. It then follows from Eq. (I.6) that the slope of the line connecting the origin ($S(0)$, $C(0)$) and the point ($S(1)$, $C(1)$) in a coverage-based SAC becomes

$$\frac{S(1) - S(0)}{C(1) - C(0)} = \frac{1 - 0}{[1 - \sum_{i=1}^{S} p_i (1 - p_i)] - 0} = \frac{1}{\sum_{i=1}^{S} p_i^2},$$

which is the Hill number for $q = 2$ (i.e., Simpson diversity). For the coverage-based rarefaction curve, Eq. (I.7) and the following derivation show that the estimated slope is a nearly unbiased estimator of the Simpson diversity:

$$\frac{\tilde{S}(1) - \tilde{S}(0)}{\hat{C}(1) - \hat{C}(0)} = \frac{1 - 0}{1 - [\sum_{X_i > 1} X_i (n - X_i) / n(n-1)] - 0}$$

$$= \frac{1}{[\sum_{X_i > 1} X_i (X_i - 1) / n(n-1)]}.$$

Thus, the theoretical relationship in an expected SAC is also valid for data-based estimators.

For a sample-size-based SAC, Olszewski (2004) also noticed that the slope at any sample size $m-1$, $S(m) - S(m-1)$, is the probability that the mth individual represents a species that was not found in the previous sample of size $m-1$. As proved by Chao and Jost (2012), this probability is identical to the expected *coverage deficit*:

$$S(m) - S(m-1) = 1 - C(m-1), \quad m > 0.$$

We prove below that $S(m) - S(m-1)$ turns out to be a function of Hill numbers $\{{}^2D, {}^3D, ..., {}^mD\}$. For example, the slope at the size $m = 2$ of the SAC is:

$$S(3) - S(2) = 1 - 2\sum_{i=1}^{S} p_i^2 + \sum_{i=1}^{S} p_i^3 = 1 - 2({}^2D)^{-1} + ({}^3D)^{-2}, \qquad \text{(I.8)}$$

which a function of ${}^2D$ and ${}^3D$. Generally, we can extend Lande's (2000) observation (Eq. I.4) to any sample size *m* in an expected SAC by showing that the slope of the expected SAC can be expressed as a function of Hill numbers at every point along the curve, as shown in the following proposition.

Proposition I3: The slope at any size $m \geq 2$ in an expected SAC, $S(m)-S(m-1)$, can be written as a function of Hill numbers $\{{}^2D,\ {}^3D,...,{}^mD\}$. Specifically, we have

$$S(m)-S(m-1)=1-\binom{m-1}{1}({}^2D)^{-1}+\binom{m-1}{2}({}^3D)^{-2}-\binom{m-1}{3}({}^4D)^{-3}...+(-1)^m\binom{m-1}{m-1}({}^mD)^{-(m-1)}.$$

For *m* = 2, the above reduces to Eq. (I.4), and for *m* = 3, it reduces to Eq. (I.8).

Proof: The result is seen from the following derivation:

$$\begin{aligned}
S(m)-S(m-1) &= \sum_{i=1}^{S} p_i(1-p_i)^{m-1} \\
&= \sum_{i=1}^{S} p_i\left[1+\sum_{r=1}^{m-1}\binom{m-1}{r}(-1)^r\, p_i^r\right] \\
&= 1+\sum_{i=1}^{S}\sum_{r=1}^{m-1}\binom{m-1}{r}(-1)^r\, p_i^{r+1} \\
&= 1+\sum_{r=1}^{m-1}\binom{m-1}{r}(-1)^r\left({}^{r+1}D\right)^{-r}. \\
&= 1-\binom{m-1}{1}({}^2D)^{-1}+\binom{m-1}{2}({}^3D)^{-2}-\binom{m-1}{3}({}^4D)^{-3}+...+(-1)^{m-1}\binom{m-1}{m-1}({}^mD)^{-(m-1)}.
\end{aligned}$$

For a coverage-based SAC, the slope at the coverage point $C(m-1)$, i.e., the slope of the line connecting the two points $(S(m-1), C(m-1))$ and $(S(m), C(m))$ becomes

$$\frac{S(m)-S(m-1)}{C(m)-C(m-1)}=\frac{1-C(m-1)}{C(m)-C(m-1)}.$$

Then Proposition I3 implies that this slope is a function of Hill numbers $\{{}^2D,\ {}^3D,...,{}^{m+1}D\}$.

LITERATURE CITED

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

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.

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

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577-586.

Lande, R., P. J. DeVries, and T. R. Walla. 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. Oikos 89:601-605.

Leinster, T., and C. A. Cobbold. 2012. Measuring diversity: the importance of species similarity. Ecology 93: 477-489.

Mao, C. X. 2007. Estimating species accumulation curves and diversity indices. Statistica Sinica 17:761-774.

Olszewski, T. D. 2004. A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. Oikos 104:377-387.