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 K: PIE and rarefaction

In the literature, the measure "Probability of an Interspecific Encounter (PIE)" (Hurlbert 1971) has been widely applied to quantify biodiversity. Some authors refer to PIE to an assemblage-level parameter whereas others consider it a data-level statistic. To clarify these, we first distinguish two levels: assemblage-level PIE (in terms of assemblage parameters) and data-level PIE (in terms of sample data).

(1) *Assemblage-level PIE*

In Hulbert's (1971) original definition, he defined PIE as the proportion of potential inter-individual encounters that are interspecific (as opposed to intraspecific), assuming every individual in the entire assemblage can encounter all other individuals. Let N_i be the true number of individuals for species i in the entire assemblage, and $\sum_{i=1}^{S} N_i = N$. He gave a formula for PIE as

$$PIE = \frac{N}{N-1}\left(1-\sum_{i=1}^{S}\left(\frac{N_i}{N}\right)^2\right) = 1-\sum_{i=1}^{S}\frac{N_i(N_i-1)}{N(N-1)} . \qquad \text{(K.1)}$$

This PIE is a parameter that needs to be estimated from data. Here PIE is the probability that two individuals belong to different species if these two individuals are taken from the *assemblage without replacement* (individulas cannot be repeated). Again, note here "without replacement" refers to an interpretaion of the parameter *PIE*, not to how data are collected in sampling schemes. Then later Hurlbert defined a related parameter "the complement of the Simpson index" or "the Gini-Simpson index" as

$$H_{GS} = 1-\sum_{i=1}^{S}\left(\frac{N_i}{N}\right)^2 = 1-\sum_{i=1}^{S} p_i^2 , \qquad \text{(K.2)}$$

where $p_i = N_i / N$ denotes the relative abundance of the ith species. This is the probability that two individuals belong to different species if these two individuals are taken from the assemblage *with replacement* (here individuals can be repeated). Note here "with replacement" refers to an interpretation of the parameter H_{GS}, not to the sampling scheme under which data are collected.

(2) ***Data-level PIE***

We now discuss the sampling schemes under which data are collected. We distinguish two types of sampling schemes:

(2a) If n individuals are taken *with replacement* from the assemblage, and let X_i be the sample fequency of the ith species. Then X_i is a binomial distribution. In this case, an unbaised estimator for the Gini-Simpson index is

$$\hat{H}_{GS} = \frac{n}{n-1}\left(1 - \sum_{X_i \ge 1}\left(\frac{X_i}{n}\right)^2\right) = 1 - \sum_{X_i \ge 1}\frac{X_i(X_i - 1)}{n(n-1)} \equiv P\hat{I}E_w .$$

This estimator $P\hat{I}E_w$ refers to a data-level statistic (the subindex w denotes sampling with replacement). It is denoted by $P\hat{I}E_w$ because it has a similar probability interpretation as PIE defined in Eq. (K.1). $P\hat{I}E_w$ can be interpreted as the probability that two individuals belong to different species if these two individuals are taken *without replacement* from the *sample data*. But here data are taken from the entire assemblage *with* replacement. Statistical estimation theory implies that $P\hat{I}E_w$ is an unbiased estimator for the Gini-Simpson index:

$$E(P\hat{I}E_w) = H_{GS} . \qquad \text{(K.3)}$$

In this sampling scheme, we do not have an unbiased estimator for the assemblage-level PIE defined in Eq. (K.1).

(2b) If n individuals are taken *without* replacement from the assemblage, and let X_i be the sample fequency of the ith species. Then X_i is a hypergeometric distribution. In this case, an unbaised estimator for PIE is

$$P\hat{I}E_{wor} \equiv \frac{n}{n-1}\left(1 - \sum_{X_i \ge 1}\left(\frac{X_i}{n}\right)^2\right) = 1 - \sum_{X_i \ge 1}\frac{X_i(X_i - 1)}{n(n-1)} .$$

(The subindex *wor* denotes sampling without replacement). That is, we have

$$E(P\hat{I}E_{wor}) = PIE . \qquad \text{(K.4)}$$

Note that at the data-level, although the two statistics $P\hat{I}E_w$ and $P\hat{I}E_{wor}$ have the same formulas as a function of sample frequencies, the data are collected from different sampling schemes.

PIE and the rarefaction curves

(a) If n individuals are taken *with replacement* from the assemblage, the expected SAC at the size m is (see Eq. (B.1) in Appendix B)

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

The slope of the expceted SAC at the base is calculated as

$$S(2) - S(1) = 1 - \sum_{i=1}^{S} p_i{}^2 = H_{GS} \ .$$

In this case, the traditional rarefaction for sample size m has the form: (see Eq. (B.2) in Appendix B)

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

The slope of this rarefaction curve at the base is

$$\tilde{S}_{ind}(2) - \tilde{S}_{ind}(1) = = 1 - \sum_{X_i \ge 1} \frac{X_i (X_i - 1)}{n(n-1)} = P\hat{I}E_w \, .$$

It then follows from Eq. (K.3) that the slope at base in the traditional rarefaction curve is an unbaised estimator for the Gini-Simpson index (which is the slope at the base of the expected SAC when sampling is conducted with replacement).

(b) If n individuals are taken *without replacement* from the assemblage, the expected SAC at the size m is

$$S_{wor}(m) = S - \sum_{i=1}^{S} \binom{N - N_i}{m} \Big/ \binom{N}{m}.$$

(The subindex *wor* denotes sampling without replacement). Then the slope of the expected SAC at the base is

$$S_{wor}(2) - S_{wor}(1) = 1 - \sum_{i=1}^{S} \frac{N_i (N_i - 1)}{N(N-1)} = PIE \ .$$

In this case, the rarefaction has exactly the same form as in the case of sampling with replacement (but X_i is a hypergeometric distribution instead of a binomial distribution):

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

The slope of the rarefaction curve at the base is

$$\tilde{S}_{wor}(2) - \tilde{S}_{wor}(1) = 1 - \sum_{X_i \geq 1} \frac{X_i(X_i - 1)}{n(n-1)} = P\hat{I}E_{wor}\,.$$

It then follows from Eq. (K.4) that the slope at base in the rarefaction curve is an unbaised estimator for the assemblage PIE (which is the slope at the base of the expected SAC obtained when sampling is conducted without replacement).

LITERATURE CITED

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