Anne Chao, T. C. Hsieh, Robin L. Chazdon, Robert K. Colwell, and Nicholas J. Gotelli. 2015. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology* 96:1189-1201.

APPENDIX B: A generalization of Good-Turing sample coverage theory.

Assume that there are S species in the assemblage and that the true species relative abundances are $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Suppose a random sample of n individuals is selected with replacement. Let X_i denote the abundance of the i-th species in the sample, $i = 1, 2, \ldots, S$. The coverage of the sample originally developed by Turing and Good (Good 1953) is expressed as

$$ {}^1C = \sum_{i \in detected} p_i = \sum_{i=1}^{S} p_i I(X_i > 0), \qquad \text{(B.1)} $$

where $I(A)$ is an indicator function that equals 1 when A is true and 0 otherwise. The leading superscript "1" in the notation ${}^1C$ signifies that this is the first-order of our generalization of sample coverage. This sample coverage can be regarded as one measure of sample completeness, giving the fraction of the true relative abundances of those species detected in the sample (or equivalently, the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample). Subtracting the sample coverage from unity gives the "coverage deficit" (Chao and Jost 2012), ${}^1C_{def}$, the proportion of the assemblage belonging to undetected species:

$$ {}^1C_{def} = 1 - {}^1C = \sum_{i=1}^{S} p_i I(X_i = 0). \qquad \text{(B.2)} $$

The second-order sample coverage is defined as the fraction of the squared true relative abundances of those species detected in the sample. It can be expressed as

$$ {}^2C = \frac{\sum_{i \in detected} p_i^2}{\sum_{i=1}^{S} p_i^2} = \frac{\sum_{i=1}^{S} p_i^2 I(X_i > 0)}{\sum_{i=1}^{S} p_i^2}. \qquad \text{(B.3)} $$

The second-order sample coverage quantifies the sample completeness for very abundant or dominant species. The expected second-order sample coverage is

$$ E({}^2C) = 1 - \frac{\sum_{i=1}^{S} p_i^2 (1 - p_i)^n}{\sum_{i=1}^{S} p_i^2}. $$

The denominator and numerator for the right-most term in the above formula can be accurately estimated. The minimum variance unbiased estimator of $\sum_{i=1}^{S} p_i^2$ is $\sum_{X_i \geq 2} X_i (X_i - 1) / [n(n-1)]$ (Good 1953). Using a similar derivation as in Chao et al. (2009),

we can obtain an estimator for $\sum_{i=1}^{S} p_i^2 (1-p_i)^n$ as $\frac{2f_2}{n(n-1)}\left[\frac{(n-2)f_2}{(n-2)f_2+3f_3}\right]^2$ if $f_2, f_3 > 0$; details are provided below for a general case. We have the following estimator for the second-order sample coverage:

$$^2\hat{C} = 1 - \frac{2f_2}{\sum_{X_i \geq 2} X_i(X_i-1)}\left[\frac{(n-2)f_2}{(n-2)f_2+3f_3}\right]^2. \qquad \text{(B.4)}$$

The deficit of the second-order sample coverage is defined as $^2C_{def} = 1 - {}^2C$, leading to the corresponding estimator:

$$^2\hat{C}_{def} = \frac{2f_2}{\sum_{X_i \geq 2} X_i(X_i-1)}\left[\frac{(n-2)f_2}{(n-2)f_2+3f_3}\right]^2. \qquad \text{(B.5)}$$

We can extend the above approach to define the *r*-th order sample coverage as

$$^rC = \frac{\sum_{i \in detected} p_i^r}{\sum_{i=1}^{S} p_i^r} = \frac{\sum_{i=1}^{S} p_i^r I(X_i > 0)}{\sum_{i=1}^{S} p_i^r}, \quad r = 1, 2, \ldots,$$

with the expected value

$$E(^rC) = 1 - \frac{\sum_{i=1}^{S} p_i^r (1-p_i)^n}{\sum_{i=1}^{S} p_i^r}.$$

We separately estimate the denominator and numerator for the right-most term in the above formula. The minimum variance unbiased estimator of $\sum_{i=1}^{S} p_i^r$ for $r \leq n$ is $\sum_{X_i \geq r} X_i^{(r)} / n^{(r)}$ (Good, 1953), where $X_i^{(r)} = X_i(X_i-1)...(X_i-r+1)$ and $n^{(r)} = n(n-1)...(n-r+1)$ denote the descending factorial. We now derive an estimator for the numerator $\sum_{i=1}^{S} p_i^r (1-p_i)^n$ (provided $f_r, f_{r+1} > 0$) following Chao et al. (2009) approach. Under the model that the species sample frequencies ($X_1, X_2, \ldots, X_S$) follow a multinomial distribution with parameters (p_1, $p_2, \ldots, p_S$), we have

$$E(f_k) = E\left(\sum_{i=1}^{S} I(X_i = k)\right) = \sum_{i=1}^{S} \frac{n^{(k)}}{k!} p_i^k (1-p_i)^{n-k}, \; k = 0, 1, \ldots, n.$$

Then we can write

$$\sum_{i=1}^{S} p_i^r (1-p_i)^n = \sum_{i=1}^{S} \frac{r!}{n^{(r)}} (1-p_i)^r \left[\frac{n^{(r)}}{r!} p_i^r (1-p_i)^{n-r}\right]$$

$$= \sum_{i=1}^{S} \frac{r!}{n^{(r)}} (1-p_i)^r P(X_i = r) = \frac{r!}{n^{(r)}} E\left\{\sum_{i=1}^{S} (1-p_i)^r I[X_i = r]\right\}.$$

Note that in the sum $\left\{\sum_{i=1}^{S}(1-p_i)^r I[X_i=r]\right\}$, only species with sample frequency r would contribute a term $(1-p_i)^r$, and other species do not contribute. Denote the mean relative abundance for all species with sample frequency r by $\overline{p}_{(r)}$. We have an approximation formula:

$$\left\{\sum_{i=1}^{S}(1-p_i)^r I[X_i=r]\right\} \approx f_r \times (1-\overline{p}_{(r)})^r,$$

implying

$$\sum_{i=1}^{S} p_i^r (1-p_i)^n \approx \frac{r!E(f_r)}{n^{(r)}}(1-\overline{p}_{(r)})^r. \qquad \text{(B.6)}$$

To estimate $\overline{p}_{(r)}$, we consider the following expected sum:

$$E\sum_{i=1}^{S}\frac{p_i}{1-p_i}I[X_i=r] = \sum_{i=1}^{S}\frac{p_i}{1-p_i}\frac{n^{(r)}}{r!}p_i^r(1-p_i)^{n-r}$$

$$= \frac{r+1}{n-r}\sum_{i=1}^{S}\frac{n^{(r+1)}}{(r+1)!}p_i^{r+1}(1-p_i)^{n-(r+1)} = \frac{r+1}{n-r}E(f_{r+1}). \qquad \text{(B.7)}$$

Applying the same approach used in the derivation of Eq. B.6, we have

$$E\sum_{i=1}^{S}\frac{p_i}{1-p_i}I[X_i=r] \approx \frac{\overline{p}_{(r)}}{1-\overline{p}_{(r)}}E(f_r). \qquad \text{(B.8)}$$

Substituting the expected frequency counts involved in Eqs. B.7 and B.8 by sample data, we obtain

$$\frac{\overline{p}_{(r)}}{1-\overline{p}_{(r)}}f_r \approx \frac{r+1}{n-r}f_{r+1}.$$

Therefore, we can solve $\overline{p}_{(r)}$ in the above equation to get

$$\overline{p}_{(r)} \approx \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}}.$$

Substituting the resulting $\overline{p}_{(r)}$ into Eq. B.6, we obtain the following estimator of $\sum_{i=1}^{S} p_i^r(1-p_i)^n$:

$$\frac{r!f_r}{n^{(r)}}\left[\frac{(n-r)f_r}{(n-r)f_r+(r+1)f_{r+1}}\right]^r.$$

These lead to the following coverage estimator:

$${}^r\hat{C}=1-\frac{r!f_r}{\sum_{X_i\geq r}X_i^{(r)}}\left[\frac{(n-r)f_r}{(n-r)f_r+(r+1)f_{r+1}}\right]^r. \qquad \text{(B.9)}$$

For $r = 1$, it reduces to Eq. 2b of the main text, and for $r = 2$, it reduces to Eq. B.4 and Eq. 3a of the main text. The corresponding estimator for the deficit of the r-th order sample coverage is

$${}^r\hat{C}_{def}=\frac{r!f_r}{\sum_{X_i\geq r}X_i^{(r)}}\left[\frac{(n-r)f_r}{(n-r)f_r+(r+1)f_{r+1}}\right]^r.$$

For $r = 1$, it reduces to Eq. 2c of the main text, and for $r = 2$, it reduces to Eq. B.5 and Eq. 3b of the main text.

LITERATURE CITED

Chao, A., R. K. Colwell, C. W. Lin, and N. J. Gotelli, N. J. 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90:1125-1133.

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

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