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 D: Unveiling the species-rank incidence distribution for incidence data.

In the main text, we introduced the basic model, data format and terminology for incidence data. In order for this presentation to be self-contained, we duplicate the introductory text here; see Colwell et al. (2012) and Chao et al. (2014) for the necessary backgrounds. Assume that in the focal assemblage there are S different species indexed by 1, 2, …, S. For any sampling unit, assume that the ith species has its own unique *incidence* (or *occurrence*) probability π_i that is constant for any randomly selected sampling unit. The incidence probability π_i is the probability that species i is detected in a sampling unit. The *species incidence distribution* (SID) and the corresponding *species-rank incidence distribution* (RID) refer to the set $(\pi_1, \pi_2, \ldots, \pi_S)$ of the S species.

Assume that a set of T sampling units are randomly selected from the study area with replacement. The underlying data consist of an $S \times T$ detection/non-detection matrix $\{W_{ij}; i = 1, 2, \ldots, S, j = 1, 2, \ldots, T\}$; here $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise, $i = 1, 2, \ldots, S, j = 1, 2, \ldots, T$. Let $Y_i = \sum_{j=1}^{T} W_{ij}$ denotes the sample *species incidence frequency* of species i. Denote the *incidence frequency counts* by $(Q_0, Q_1, \ldots, Q_T)$, where Q_k is the number of species that are detected in exactly k sampling units in the data, $k = 0, 1, \ldots, T$. The unobservable zero frequency count Q_0 denotes the number of species among the S species present in the assemblage that are not detected in any of the T sampling units. Also, Q_1 represents the number of "unique" species (those that are detected in only one sampling unit), and Q_2 represents the number of "duplicate" species (those that are detected in only two sampling units). Define the *sample incidence probability* as $\hat{\pi}_i = Y_i / T$, which is the conventional plug-in estimator. The empirical RID is based on this plug-in estimator.

All derivations below are based on that Y_i, $i = 1, 2, \ldots, S$ follows a binomial distribution with the total number T and the detection probability π_i:

$$P(Y_i = y_i, i = 1, 2, \ldots, S) = \prod_{i=1}^{S} \binom{T}{y_i} \pi_i^{y_i} (1 - \pi_i)^{T - y_i} .$$

Chao et al. (1992) first defined the sample coverage for a set of T sampling units as the following expression:

$$^1C^* = \frac{\sum_{i \in detected} \pi_i}{\sum_{i=1}^{S} \pi_i} = \frac{\sum_{i=1}^{S} \pi_i I(Y_i > 0)}{\sum_{i=1}^{S} \pi_i}, \qquad \text{(D.1)}$$

where $I(A)$ is an indicator function that equals 1 when A is true and 0 otherwise. Here we use a superscript "*" to refer to the sample coverage of incidence data, and use a leading superscript "1" to signify that this is the first-order of our generalization of sample coverage. The sample coverage quantifies the fraction of the total incidence probabilities of the

discovered species in the T sampling units. It is an objective measure of sample completeness for incidence data. Subtracting the sample coverage from unity gives the fraction of the assemblage belonging to the undetected species. As with the abundance data, we denote the first-order coverage deficit by ${}^1C^*_{def}$, which can be expressed as (Chao and Jost 2012):

$$ {}^1C^*_{deficit} = 1 - {}^1C^* = \frac{\sum_{i \in undetected} \pi_i}{\sum_{i=1}^{S} \pi_i} = \frac{\sum_{i=1}^{S} \pi_i I(Y_i = 0)}{\sum_{i=1}^{S} \pi_i}. \qquad \text{(D.2)} $$

Following exactly the same approach for abundance data, we can generalize the sample coverage theory to incidence data by defining the r-th order sample coverage as

$$ {}^rC^* = \frac{\sum_{i=1}^{S} \pi_i^r I(Y_i > 0)}{\sum_{i=1}^{S} \pi_i^r}, \; r = 1, 2, \ldots \qquad \text{(D.3)} $$

with the expected value

$$ E({}^rC^*) = 1 - \frac{\sum_{i=1}^{S} \pi_i^r (1-\pi_i)^T}{\sum_{i=1}^{S} \pi_i^r}. $$

If the incidence frequency counts Q_r and Q_{r+1} are both non-zero, then we have the following estimator

$$ {}^r\hat{C}^* = 1 - \frac{r! Q_r}{\sum_{Y_i \geq r} Y_i^{(r)}} \left[\frac{(T-r)Q_r}{(T-r)Q_r + (r+1)Q_{r+1}} \right]^r, \; r = 1, 2, \ldots \qquad \text{(D.4)} $$

where $Y_i^{(r)} = Y_i(Y_i - 1)...(Y_i - r + 1)$ denotes the descending factorial. The estimator in Eq. D.4 is derived by noting that the minimum variance unbiased estimator of $\sum_{i=1}^{S} \pi_i^r$ is $\sum_{Y_i \geq r} Y_i^{(r)} / T^{(r)}$ (Good, 1953), and an estimator for $\sum_{i=1}^{S} \pi_i^r (1-\pi_i)^T$ is (if Q_r, $Q_{r+1} > 0$)

$$ \frac{r! Q_r}{T^{(r)}} \left[\frac{(T-r)Q_r}{(T-r)Q_r + (r+1)Q_{r+1}} \right]^r. $$

The derivation is parallel to that given in Appendix B for abundance data. The corresponding estimator for the deficit of the r-th order coverage is ${}^r\hat{C}^*_{def} = 1 - {}^r\hat{C}^*$, i.e.,

$$ {}^r\hat{C}^*_{def} = \frac{r! Q_r}{\sum_{Y_i \geq r} Y_i^{(r)}} \left[\frac{(T-r)Q_r}{(T-r)Q_r + (r+1)Q_{r+1}} \right]^r, \quad r = 1, 2, \ldots $$

Estimated the RID for incidence data

Similar derivation steps as those for abundance data lead to the following adjusted estimator for a detected species:

$$\tilde{\pi}_i = \frac{Y_i}{T}(1-\hat{\lambda}e^{-\hat{\theta}Y_i}), \quad Y_i > 0. \tag{D.5}$$

Since an unbiased estimator for $\sum_{i=1}^{S}\pi_i$ (the denominator of ${}^1C^*$) is $\sum_{Y_i\geq 1} Y_i/T$, and an unbiased estimator for $\sum_{i=1}^{S}\pi_i^2$ (the denominator of ${}^2C^*$) is $\sum_{Y_i\geq 2} Y_i(Y_i-1)/[T(T-1)]$, the two parameters λ and θ satisfy the following two equations:

$$\sum_{i\in \text{detected}} \pi_i \approx \sum_{Y_i\geq 1}(Y_i/T)(1-\lambda e^{-\theta Y_i}) = {}^1\hat{C}^* \times \frac{\sum_{Y_i\geq 1} Y_i}{T}; \tag{D.6}$$

$$\sum_{i\in \text{detected}} \pi_i^2 \approx \sum_{Y_i\geq 1}[(Y_i/T)(1-\lambda e^{-\theta X_i})]^2 = {}^2\hat{C}^* \times \frac{\sum_{Y_i\geq 2} Y_i(Y_i-1)}{T(T-1)}. \tag{D.7}$$

Any software can be readily used to obtain the solution $\hat{\lambda}$ and $\hat{\theta}$ in the above system of nonlinear equations. If the solution $\hat{\theta}$ is out of the range of [0, 1], we replace it by 1 so that the model reduces to the one-parameter case. A bootstrap method as the one we proposed for abundance data in the main text can be similarly applied to assess the sampling variance of the estimator $\tilde{\pi}_i$ and construct the associated confidence interval of π_i.

Based on the Chao2 lower bound or estimator (Chao, 1987), we have the following estimator for the number of the undetected species in T sampling units:

$$\hat{Q}_0 = \begin{cases} \dfrac{(T-1)}{T}\dfrac{Q_1^2}{2Q_2}, & \text{if } Q_2 > 0; \\ \dfrac{(T-1)}{T}\dfrac{Q_1(Q_1-1)}{2}, & \text{if } Q_2 = 0. \end{cases} \tag{D.8}$$

Assuming a geometric series for the incidence probabilities for the undetected species, i.e., $\pi_i = \alpha\beta^i$, $i = 1, 2, ..., \hat{Q}_0$, we obtain the following two equations in terms of two parameters α and β for the undetected species:

$$\sum_{i\in \text{undetected}} \pi_i \approx \sum_{i=1}^{\hat{Q}_0}\alpha\beta^i = {}^1\hat{C}^*_{def} \times \frac{\sum_{Y_i\geq 1} Y_i}{T}; \tag{D.9}$$

$$\sum_{i\in \text{undetected}} \pi_i^2 \approx \sum_{i=1}^{\hat{Q}_0}(\alpha\beta^i)^2 = {}^2\hat{C}^*_{def} \times \frac{\sum_{Y_i\geq 2} Y_i(Y_i-1)}{T(T-1)}. \tag{D.10}$$

We can solve $\hat{\alpha}$ and $\hat{\beta}$ by the above system of nonlinear equations. Therefore, the proposed estimated relative abundances for the undetected species are

$$\tilde{\pi}_i = \hat{\alpha}\hat{\beta}^i, i = 1, 2, \ldots, \hat{Q}_0. \tag{D.11}$$

Combining the adjustment method for detected species in Eq. D.5 and the estimated relative abundances for undetected species in Eq. D.11, we can construct a complete RAD based on T sampling units.

Soil ciliates example

The incidence frequency counts for the ciliates data described in the main text are summarized in the following table (Foissner et al. 2005):

k	1	2	3	4	5	6	7	8	9	10	12	13	14	15
Q_k	150	53	42	18	12	9	10	7	6	1	2	3	2	1

k	17	19	20	22	23	24	26	27	29	32	33	34	35	37	39
Q_k	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

The estimation procedures are generally parallel to those presented for the abundance data, as illustrated in Fig. 2 of the main text. We omit the details. The first- and second-order sample coverage estimates are respectively 88.45% and 99.38%; the corresponding coverage deficit estimates are thus respectively 11.55% and 0.62%. Also, we have $\sum_{Y_i \geq 1} Y_i / T = 1281/51 = 25.12$ (the average number of incidences per sampling unit) and $\sum_{Y_i \geq 2} Y_i (Y_i - 1) / [T(T-1)] = 6.1082$. Based on these statistics and data, the numerical solution for Eqs. D.6 and D.7 are $\hat{\lambda} = 0.3264$ and $\hat{\theta} = 0.1528$; and the numerical solution for Eqs. D.9 and D.10 are $\hat{\alpha} = 0.0139$, and $\hat{\beta} = 0.999934$; see the main text for more analysis. The estimated decay factor $\hat{\beta}$ is close to unity, so the estimated relative abundances for the 209 undetected species (Fig. 3 in the main text) differ little.

LITERATURE CITED

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783-791.

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., N. G. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species biodiversity studies. Ecological Monographs 84: 45-67.

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

Chao, A., S. Lee, and S. Jeng. 1992. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. Biometrics 48:201-216.

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.

Foissner, W., Agatha, S., and Berger, H. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha region and the Namib Desert. Denisia 5:1-1459.