Anne Chao and Lou Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 93:2533−2547. **http://dx.doi.org/10.1890/11-1952.1**

APPENDIX B. An unbiased estimator for the expected sample coverage in coverage-based rarefaction.

Theorem B1: Under the assumption that sample species frequencies (X_1, X_2, …, X_S) obey the following multinomial model with cell total n and cell probabilities (p_1, p_2, …, p_S):

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

the minimum variance unbiased estimator of the expected coverage of a sample with size m, $E(C_m) = 1 - \sum_{i=1}^{S} p_i (1 - p_i)^m$, is

$$\hat{C}_m = 1 - \sum_{X_i \geq 1} \frac{X_i}{n} \frac{\binom{n - X_i}{m}}{\binom{n-1}{m}}, \quad m < n, \tag{B.1}$$

which is also Eq. 4b in the main text. It is readily seen that $0 \leq \hat{C}_m \leq 1$ for all m, and

$$1 - \hat{C}_m = \hat{S}_{m+1} - \hat{S}_m ,$$

where

$$\hat{S}_m = S_{obs} - \sum_{X_i \geq 1} \frac{\binom{n - X_i}{m}}{\binom{n}{m}}, \quad m < n, \tag{B.2}$$

denotes the species richness estimator in a traditional rarefaction curve for a sample of size $m < n$, and S_{obs} is the number of species observed in the reference sample of size n.

Proof: Under the multinomial assumption, the sample frequency X_i follows a binomial distribution. Then we have

$$E\left[\frac{X_i}{n}\frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}\right] = \sum_{x=1}^{n}\frac{x}{n}\frac{\binom{n-x}{m}}{\binom{n-1}{m}}\binom{n}{x}p_i^x(1-p_i)^{n-x}$$

$$= p_i(1-p_i)^m\sum_{x=1}^{n}\frac{(n-1-m)!}{(x-1)!(n-x-m)!}\,p_i^{x-1}(1-p_i)^{n-m-x}$$

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

$$= p_i(1-p_i)^m\,.$$

From the Rao-Blackwell and Lehmann-Scheffé Theorems (e.g., Casella and Berger 2002, p. 347 and p. 369), the estimator in Eq. B.1 is the unique minimum variance unbiased estimator of the expected sample coverage. Also, $1-\hat{C}_m = \hat{S}_{m+1}-\hat{S}_m$ follows directly from the following equation:

$$\frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}-\frac{\binom{n-X_i}{m+1}}{\binom{n-1}{m+1}} = \frac{X_i}{n}\frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}\,.$$

<u>Theorem B2</u>: We have the following monotonic properties:

(1) The four functions $E(S_m)$, $E(C_m)$, $\hat{S}_m$ and $\hat{C}_m$ are all non-decreasing function of sample size m;

(2) $E(S_m)$ is a non-decreasing function of $E(C_m)$;

(3) $\hat{S}_m$ is a non-decreasing function of $\hat{C}_m$.

Proof: For any $X_i \geq 1$, direct computation leads to

$$\frac{\binom{n-X_i}{m+1}}{\binom{n-1}{m+1}} \le \frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}.$$

This implies $\hat{C}_{m+1} \ge \hat{C}_m$. Since $\hat{S}_{m+1} - \hat{S}_m = 1 - \hat{C}_m \ge 0$, we obtain $\hat{S}_{m+1} \ge \hat{S}_m$. The other monotonic properties are readily seen.

LITERATURE CITED

Casella G., and R. L. Berger. 2002. Statistical Inference, Second Edition. Duxbury, Pacific Grove.