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 D. Theoretical proof of the proposed unbiased algorithm in coverage-based rarefaction.

Theorem D1: Consider the following resampling procedure: take m +1 individuals *without* replacement from a given reference sample of size n, and record the number of singletons in this new sample. Then given the species sample frequencies $\{X_1, X_2, ..., X_S\}$ in the reference sample, one minus the average proportion of singletons in a resample of size m +1 is an unbiased estimator of $\hat{C}_m$ ($m < n$) under this sampling probability model. This estimator is the unique minimum variance unbiased estimator based on resample frequencies.

Proof: If we select m+1 individuals *without* replacement from the given data $\{X_1, X_2, ..., X_S\}$, then the species frequencies $\{Z_1, Z_2, ..., Z_S\}$ in the resample of size m+1 follows a generalized hyper-geometric distribution:

$$P(Z_1 = z_1, ..., Z_S = z_S) = \binom{X_1}{z_1}\binom{X_2}{z_2}...\binom{X_S}{z_S} / \binom{n}{m+1}. \qquad \text{(D.1)}$$

The marginal distribution of the frequency Z_i follows a hyper-geometric distribution with probability function:

$$P(Z_i = z_i) = \binom{X_i}{z_i}\binom{n - X_i}{m+1-z_i} / \binom{n}{m+1}.$$

The expected number of singletons is then

$$\sum_{i=1}^{S} P(Z_i = 1) = \sum_{i=1}^{S} \binom{X_i}{1}\binom{n - X_i}{m} / \binom{n}{m+1}$$

Then the proportion of singletons in a resample size of m+1 is equal to

$$\frac{1}{m+1}\sum_{i=1}^{S} P(Z_i=1)=\frac{1}{m+1}\sum_{i=1}^{S}\binom{X_i}{1}\binom{n-X_i}{m}\Big/\binom{n}{m+1}=\sum_{X_i\geq 1}\frac{X_i}{n}\frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}.$$

It follows from Theorem B1 that $E[1-\frac{1}{m+1}\sum_{i=1}^{S} I(Z_i=1)]=\hat{C}_m$, where I(.) is an indicator function, and the expectation here is taken with respect to the probability model in Eq. D.1. This proves that our proposed algorithm is unbiased. From the Rao-Blackwell and Lehmann-Scheffé Theorems (e.g., Casella and Berger 2002, p. 347 and p. 369), our estimator is the unique minimum variance unbiased estimator of the sample coverage under the sampling model in Eq. D.1.

LITERATURE CITED

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