The Well-Tempered Assemblage: Reducing Bias in the Estimation of Species Rank Abundance Distributions

N. J. Gotelli [a], A. Chao [b], R. K. Colwell [c], R. Chazdon [c] and T. Hsieh [b]

[a] Department of Biology,
University of Vermont,
Burlington, U.S.A.
Nicholas.Gotelli@uvm.edu

[b] Institute of Statistics,
National Tsing Hua University,
Hsin-Chu, Taiwan

[c] Department of Ecology & Evolutionary Biology,
The University of Connecticut,
Storrs, U.S.A.

Abstract: Most plant and animal assemblages are characterized by a few common species and many uncommon or rare species. Understanding the mechanisms shaping the species abundance distribution (SAD) has long been a major research focus in ecology. Beginning with seminal work by R.A. Fisher in the 1940s, ecologists have fit simple statistical models such as the geometric series, log normal, and exponential series to species abundance data. This distribution-fitting approach is based on the use of the simple "plug-in" estimator $\hat{p}_i = n_i / N$, where $\hat{p}_i$ is the estimated relative frequency of species i , n_i is the number of individuals observed of species i, and N is the number of individuals in the sample. However, with incomplete sampling and undetected species, $\hat{p}_i$ is a biased estimator of the true relative frequency of the species in the sample, and the degree of bias increases with the relative rarity of each species. Using the concept of sample coverage and the theory of frequency estimation by I.J. Good and A. Turing, we estimated the true species abundance distribution (SAD) based on a random sample of individuals. We separately estimated relative frequencies for the set of species detected in the sample and for the set of species undetected in the sample. We then combined the two parts to obtain an estimated SAD in which the relative frequency for each species has been tuned or adjusted to minimize the bias inherent in the traditional plug-in estimator. To examine the performance of the tuned estimators, we created artificial data sets by randomly sampling from common statistical distributions, and by randomly sampling from large empirical distributions based on very thorough field censuses. With sufficient sample size or sample coverage, the tuned estimators closely matched the true SADs. These more accurate estimators of relative frequency should aid ecologists in understanding and modeling SADs.