ENVIRONMETRICS
Environmetrics 2006; **17**: 129–138
Published online 6 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.755

An empirical method for inferring species richness from samples

Paul A. Murtaugh*,† and David S. Birkes

Department of Statistics, Oregon State University, Corvallis, OR 97331, U.S.A.

SUMMARY

We introduce an empirical method of estimating the number of species in a community based on a random sample. The numbers of sampled individuals of different species are modeled as a multinomial random vector, with cell probabilities estimated by the relative abundances of species in the sample and, for hypothetical species missing from the sample, by linear extrapolation from the abundance of the rarest observed species. Inference is then based on likelihoods derived from the multinomial distribution, conditioning on a range of possible values of the true richness in the community. The method is shown to work well in simulations based on a variety of real data sets.

KEY WORDS: confidence intervals; diversity; multinomial; resampling; taxon richness

1. INTRODUCTION

Patterns of the diversity and relative abundance of species in biological communities have long been of interest to ecologists (for example, see Ludwig and Reynolds, 1988; Ricklefs and Schluter, 1994; and Hubbell, 2001). The simplest summary of diversity is species richness—the total number of species present in a community. Since the likelihood of inclusion of rare species increases with the size of a sample, much attention has focused on the estimation of richness from the numbers of species present in samples. This problem is germane to many practical issues, e.g. the assessment of biodiversity in a consistent way across many different biological systems (Larsen and Herlihy, 1998). How do we compare the species richness of two communities that have been sampled with different intensities?

There is a vast statistical literature on the estimation of the number of classes in a partitioned population, but the problem is a difficult one (see reviews by Bunge and Fitzpatrick, 1993; and Chao, 2003). Approaches that have been developed include extrapolation of the relationship between number of species detected and the spatial extent and/or intensity of sampling (Pielou, 1975; Colwell and Coddington, 1994; Flather, 1996); parametric modeling of species' relative abundances (Fisher *et al.*, 1943; Bulmer, 1974); a variety of non-parametric techniques (Burnham and Overton, 1979; Smith and van Belle, 1984; Chao, 1987; Chao and Lee, 1992); and Bayesian approaches that place priors on

*Correspondence to: Paul A. Murtaugh, Department of Statistics, Oregon State University, Corvallis, OR 97331, U.S.A.
†E-mail: murtaugh@stat.oregonstate.edu;

Contract/grant sponsor: U.S. Environmental Protection Agency; contract/grant number: CR82-9096-01.

Received 17 January 2005
Accepted 20 June 2005

the number and relative abundances of species (Solow, 1994; Christen and Nakamura, 2000; Rodrigues *et al*., 2001). Some of these methods are included in a web-based software package that is widely used by ecologists (Colwell, 2004).

Here we introduce an empirical method of estimating species richness in which the numbers of sampled individuals of different species are modeled as a multinomial random vector, with cell probabilities estimated by the relative abundances of species in the sample and, for hypothetical species missing from the sample, by linear extrapolation from the abundance of the rarest observed species. Point estimates and confidence intervals for richness are based on likelihoods derived from the multinomial distribution, conditioning on a range of possible values of the true richness in the community. We evaluate the methodology by applying it to data sets obtained by resampling from real biological communities.

It should be noted that real data sets rarely focus exclusively on biological species. Groupings of individuals are inevitably influenced by collection methods, the ease of identifying different organisms, the investigator's interests and experience, and many other factors. How organisms are grouped has obvious consequences for numerical summaries and generalizations about patterns of diversity in natural communities (for example, see Hurlbert, 1971; and Paine, 1988). We use the shorthand 'species richness' with the understanding that the 'species' may actually represent a variety of other possible groupings.

2. METHODOLOGY

Suppose we obtain n individuals in a random sample from a community of S species, or in a set of samples that can justifiably be pooled. Let S_n denote the number of species represented in the sample ($S_n \leq S$), and let $n_1, n_2, \ldots, n_{s_n}$ denote the numbers of individuals of these species in the sample, in ascending order, where s_n is a particular value of S_n. Our goal is to find the value(s) of S that are most consistent with the observation of s_n species and $n_1, \ldots, n_{s_n}$ individuals in our sample.

2.1. Relative abundances of species

We estimate the relative abundances of species in hypothetical communities of S species, where S is varied from s_n upwards. If $S = s_n$, we can approximate the relative abundance of species j in the community as $p_j = n_j/n$, for $j = 1, \ldots, s_n$. If $S > s_n$, obtain initial estimates of the relative abundances of the species occurring in the sample as $p_j = n_j/n$, $j = S - s_n + 1, \ldots, S$. Assume that the $S - s_n$ species missing from the sample are in fact the rarest in the community, and that their relative abundances can be estimated by linear extrapolation as follows. For $j = 1, \ldots, S - s_n$, set

$$p_j = \frac{j}{S - s_n + 1}(0.5 p_{S-s_n+1}) \tag{1}$$

That is, on a plot of the p_js vs. abundance rank, we read the values of $p_1, \ldots, p_{S-s_n}$ from a line connecting the points (0,0) and $(S - s_n + 1,\ 0.5 p_{S-s_n+1})$; see Figure 1. We then standardize the ps so that they sum to one,

$$\hat{p}_j = \frac{p_j}{\sum_{i=1}^{S} p_i} \tag{2}$$

for $j = 1, \ldots, S$.

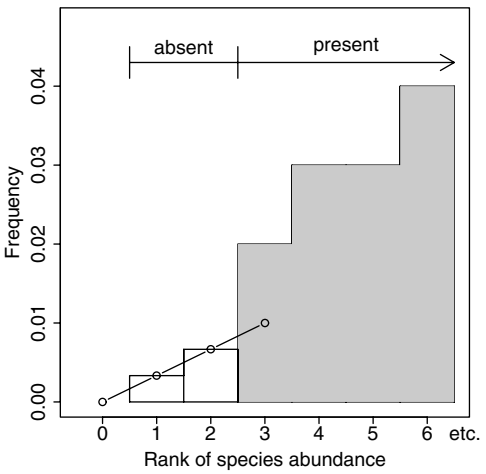

Figure 1. Relative abundances of species in a hypothetical community, showing linear extrapolation for two species assumed to be present in the community but missing from the sample (cf. Equation 1)

This procedure for imputing relative abundances of unobserved species (Equation 1) is motivated not by a particular parametric model, but rather by its robust performance in simulations and applications to real data. We explored many other models of the relative abundances of 'missing' species, as is discussed later, but none performed as well as the above approach.

2.2. *The likelihood*

Given $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_S)$ calculated with (2), and conditioning on the total sample size n, the numbers of sampled individuals from the S species have a multinomial distribution:

$$(n_1, n_2, \ldots, n_S) \sim \text{Multinomial}(n; \hat{\boldsymbol{p}}) \tag{3}$$

In order to find the values of S most consistent with our data, we evaluate the likelihoods

$$\text{L}(S; s_n, \hat{\boldsymbol{p}}) \equiv \text{P}(S_n = s_n | S, \hat{\boldsymbol{p}}) \tag{4}$$

for $S = s_n, s_n + 1, \ldots$. In spite of the simplifying assumption used to derive $\hat{\boldsymbol{p}}$ in the preceding section, it is not always true that the species missing from the sample will be those that are rarest in the community. Consequently, to evaluate the likelihood in (4), one must enumerate all $\binom{S}{S-s_n}$ ways of getting $S - s_n$ zero counts and compute the associated probabilities. These computations become impracticable as S gets large, so we resorted to a Monte Carlo approach: a large number of multinomial random vectors (usually 1000) was generated from (3), and the proportion of vectors having $S - s_n$ zero counts was used as an approximation for the likelihood in (4).

2.3. *Estimation*

For a particular sample, likelihoods are calculated as in (4) for $S = s_n,\ s_n + 1, \ldots$. Our point estimate of S is the value for which the likelihood is maximized.

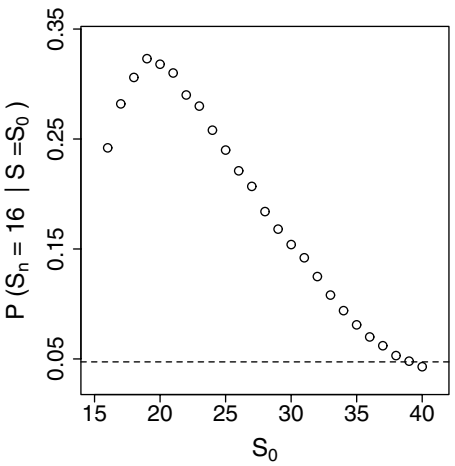

Figure 2. For a hypothetical sample having 16 species represented by 1, 1, 2, 2, 3, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 25 individuals, respectively, these are values of the likelihood in (4) for different values of S. The horizontal line is the demarcation between likelihoods for Ss that are included in a 95% confidence interval and those for Ss not in the interval—see (6). Here $\hat{S} = 19$, with a 95% confidence interval of 16 to 39

The function $\mathrm{L}(S;\ s_n, \hat{\boldsymbol{p}})$ is similar to a profile likelihood function, except that $\hat{\boldsymbol{p}}$ is an empirical estimate, rather than a maximum likelihood estimate (MLE), and the dimensionality of $\hat{\boldsymbol{p}}$ changes with S. If $\mathrm{L}(S;\ s_n,\ \hat{\boldsymbol{p}})$ were a true profile likelihood, then the statistic

$$W = 2\left[\log \mathrm{L}(\hat{S}; s_n, \hat{\boldsymbol{p}}) - \log \mathrm{L}(S_0; s_n,\ \hat{\boldsymbol{p}})\right] \tag{5}$$

where $\hat{S}$ is the MLE of S, would have an asymptotic χ^2_1 distribution when $S = S_0$ (for example, see Azzalini, 1996). Equation (5) implies that an approximate $100(1-\alpha)\%$ confidence interval for S is

$$\left\{S_0\ :\ \mathrm{L}(S_0; s_n,\ \hat{\boldsymbol{p}})\ >\ \mathrm{L}\,(\hat{S}; s_n,\ \hat{\boldsymbol{p}})\exp\left(-0.5\chi^2_{1,1-\alpha}\right)\right\} \tag{6}$$

where $\chi^2_{1,\,1-\alpha}$ is the $100(1-\alpha)$ percentile of a χ^2 distribution with 1 degree of freedom. Figure 2 is a graphical illustration of the steps in this estimation procedure.

We implemented our methodology in functions written in R (R Development Core Team, 2004), which are available on request.

2.4. A note on a similar method

Emigh (1983) developed an exact expression for the probability of observing a certain number of classes from a multinomial distribution, which we approximate by simulation (4). As explained earlier, the exact approach requires the enumeration of $\binom{S}{S-s_n}$ subsets for each value of s_n. For the data sets used here, this is not feasible. For example, our largest data set includes 118 species, and $\binom{118}{59}$ is 2.4×10^{34}.

Emigh (1983) used the probability distribution of the number of observed classes to estimate the actual number of classes in a way that is superficially similar to our approach. For different subsets of the s_n observed species—the most common species, the two most common species, etc.—he finds the maximum likelihood estimate of S and then assigns equal probability to the unobserved species.

Table 1. Sources of data used in the figures. S is the number of taxa ('species') represented in pooled samples consisting of N organisms. The data sets that are numbered are those shown in Figure 3; results for the non-numbered data sets are included in Figures 4–6. A complete listing of the species counts is available on request

No.	Description and source
1	Macroinvertebrates in a stream (CAS01) in the Oregon Cascades (Li *et al*., 2001): $S = 75$, $N = 2577$
–	Macroinvertebrates in stream CAS05: $S = 82$, $N = 13,478$
–	Macroinvertebrates in stream CAS06: $S = 68$, $N = 1517$
–	Macroinvertebrates in stream CAS08: $S = 66$, $N = 3727$
2	Macroinvertebrates in a stream (VAL04) in the Willamette Valley, Oregon (Li *et al*., 2001): $S = 57$, $N = 4372$
–	Macroinvertebrates in stream VAL02: $S = 36$, $N = 3123$
3	Zooplankton in Cornell Experimental Pond # 124 (Murtaugh, 1989): $S = 12$, $N = 1330$
–	Zooplankton in Cornell Experimental Pond # 127: $S = 11$, $N = 2563$
–	Zooplankton in Lake Washington, Seattle, on 3 November 1978 (Murtaugh, unpublished): $S = 11, N = 1850$
4	Trees in a submontane forest, Mt. Kinasalapi, Kitangland, Mindanao (Pipoly and Madulid, 1998): $S = 44$, $N = 1062$
5	Trees in the Nanjenshan forest, Taiwan (Sun *et al*., 1998): $S = 118$, $N = 37,609$
6	Marine benthos collected off San Diego, EPA Environmental Monitoring and Assessment Program regional survey (K. Ritter, personal communication): $S = 82$, $N = 7745$
–	Marine benthos collected off Channel Island: $S = 63$, $N = 3456$

(This assumption about the relative abundances of missing species was also made, in an entirely different context, by Solow and Polasky, 1999.) A goodness-of-fit test compares this vector of relative abundances to the vector of observed counts, and the smallest S for which the goodness-of-fit P-value exceeds 0.05 is chosen as the best estimate of the true number of species in the population.

Applied to samples from real data sets, as described below, Emigh's method consistently gave estimates of S that were just two to three species greater than the number observed in the sample, s_n. This led to extreme underestimation of the actual numbers of species in most of the data sets.

3. RESULTS

For comparison with our method, we calculate values of the Abundance-based Coverage Estimator (ACE) of Chao and Lee (1992) and Chao *et al.* (1993); another non-parametric estimator, usually referred to as 'Chao 1' (for instance, see Chazdon *et al*., 1998), developed by Chao (1984), along with an algorithm for obtaining confidence intervals (Chao, 1987); and a first-order jackknife estimator, 'Jack 1', developed by Burnham and Overton (1979). We applied all of these methods to 13 data sets resampled from real biological communities (Table 1). As noted earlier, the 'species' in these data sets are actually groupings of varying taxonomic complexity. For each community, samples collected at roughly the same time and place were pooled, and the resulting large number of individuals was assumed to include all of the species present in the community. Subsamples of various sizes, repeatedly drawn *with replacement* from the pooled data, were then used to generate estimates of species richness.

Sampling with replacement means we are doing simulations based on the empirical distribution of individuals among species. Since none of these data sets is a complete census of an actual community,

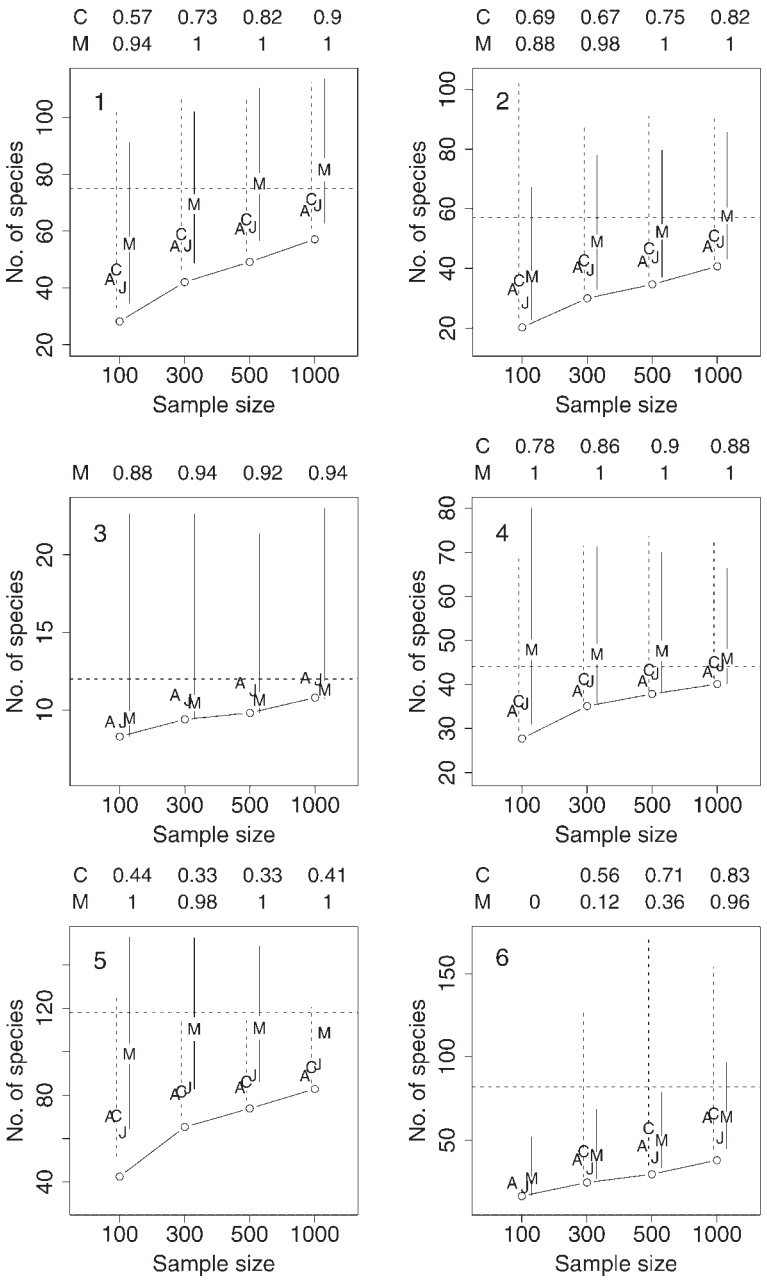

Figure 3. Estimates of species richness in repeated samples from six of the data sets described in Table 1 (identified in upper left-hand corner of each plot). 'True' richnesses are shown by the horizontal lines. The circles show the mean observed richnesses in the repeated samples. The letters show mean predictions from 50 re-samplings using our method (M), and 1000 re-samplings using the first-order jackknife (J), ACE (A) and Chao 1 (C) estimators. Vertical lines connect mean upper and lower limits of 95 per cent confidence intervals from methods C and M. If results are not shown for method C, more than 10 per cent of the re-samplings lead to undefined estimates. The numbers running across the top of each plot are the proportions of intervals that include the true richness

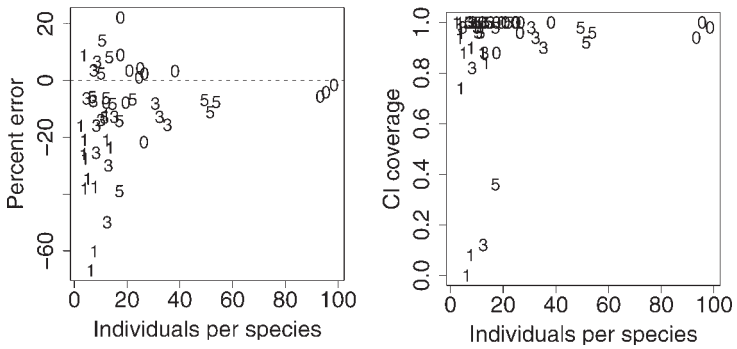

Figure 4. Relationships of prediction accuracy (upper plot) and confidence interval coverage (lower plot) to the average number of individuals per species in random samples from the 13 data sets in Table 1. 'Percent error' is 100 × (predicted number of species minus actual number of species)/(actual number). Each point represents a set of 50 re-samplings for a particular data set and sample size ('1' = 100, '3' = 300, '5' = 500, and '0' = 1000 individuals per sample)

the total species richness in each set probably underestimates the true species richness in the system from which the data were gathered. The larger the *N*s in Table 1, the more closely the empirical distribution will match the true distribution of individuals among species in the community.

The accuracy of our method's predictions and the coverage of the associated confidence intervals are quite good for most of the empirical distributions—six examples are shown in Figure 3. All of the methods tend to underestimate the true richness, although their performance improves as sample size increases.

Communities with many rare species provide the greatest challenge for all of these methods, as shown for data set 6 in Figure 3. Thirty-five percent of the species in this data set are singletons (i.e. species represented by just one individual), the highest value for the data sets shown in Table 1. In communities with a few very common species and many rare ones, only very large samples are likely to provide an adequate representation of species richness.

As shown in Figure 4, the total number of individuals in the sample divided by the number of species represented may be a useful metric for judging how well our method is likely to work in specific examples. For samples with at least 30 individuals per species, our estimates are usually within 20 per cent of the true values, and the coverage of our confidence intervals is usually within a few percent of the nominal level.

Figures 5 and 6 summarize comparisons of our method with the methods included in Figure 3. Our estimates are generally more accurate and less variable than those provided by the Jack 1 and ACE methods (Figure 5). As shown in Figure 6, our method is more accurate than the Chao 1 method in all but five data set/sample size combinations, and the coverage of our confidence intervals is much closer to the nominal level. In fact, none of the Chao 1 intervals achieves the nominal 95 per cent coverage.

A possible weakness of our approach is the 'ad hoc' nature of our imputation of the relative abundances of unobserved species (1), although other authors have also used arbitrary imputation methods (Emigh, 1983; and Solow and Polasky, 1999). We explored many other models of the relative abundances of missing species, including ones based on the 'broken-stick' distribution (MacArthur, 1960; Pielou, 1969); exponential and other non-linear extrapolations from the abundance of the rarest observed species; various rank-size relations encompassed by Zipf's Law (Read, 1988); and choices of multipliers other than 0.5 in (1). None surpassed the overall accuracy of the approach outlined above, and most resulted in confidence intervals with much less than the nominal coverage.

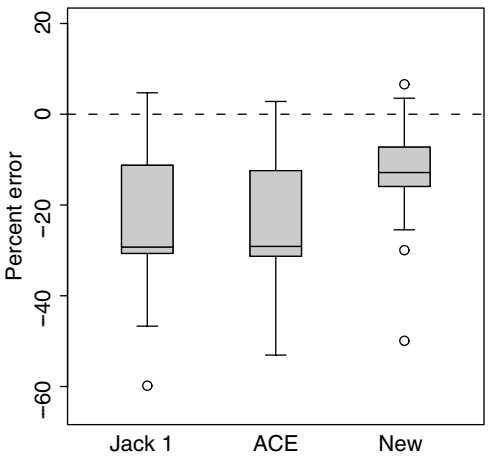

Figure 5. Boxplots of prediction accuracy (see legend for Figure 4) for three different methods applied to the 13 data sets in Table 1, for samples of size 300 only. The jackknife and ACE results are based on 1000 re-samplings; the results for our method are based on 50 re-samplings

The estimation method presented here appears to work well for data modeled after a variety of real communities. Its performance generally surpasses those of several competing estimators that are currently in use. Like the other methods, our method usually underestimates the 'true' richness but becomes increasingly accurate as the sample size increases. The associated confidence intervals usually have at least the nominal coverage for sufficiently large samples.

Because the accuracy of richness estimates depends strongly on the size of the sample relative to the (unknown) number of species present, it may be ill-advised to compare communities based on samples in which a fixed number of individuals is counted, as argued by Cao, *et al*. (2002). It may be more sensible to sample (or count) until some threshold number of individuals per sampled species is achieved, or until some other 'stopping' criterion is met (Christen and Nakamura, 2003).

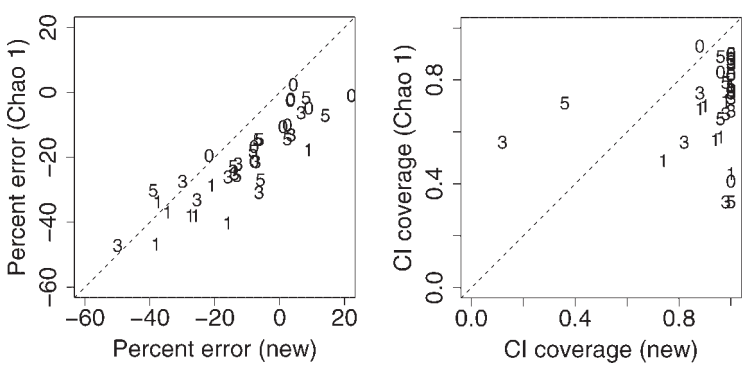

Figure 6. Comparisons of prediction accuracy (upper plot) and confidence interval coverage (lower plot) between the Chao 1 method and our method, for the 10 data sets for which the Chao 1 estimator could be calculated for at least 90 per cent of re-samplings. The labeling of points corresponds to sample size, as explained in the legend to Figure 4

It would be of obvious interest to identify features of a sample—whether number of individuals per species, evenness of the distribution of relative abundances, or other attributes—that are markers of communities for which richness will be difficult to project. Such features would be signals that richness is likely to be underestimated, and they might even provide an empirical basis for adjustments that reduce the bias of the estimator.

ACKNOWLEDGEMENTS

We thank A. Herlihy for graciously providing data on Oregon stream invertebrates; K. Ritter for providing data from the EPA surveys; A. Chao and D.P. Larsen for helpful guidance in reviewing existing methods and literature; and an anonymous reviewer for suggesting the profile-likelihood analogy.

The research described in this article has been funded by the U.S. Environmental Protection Agency through the STAR Cooperative Agreement CR82-9096-01 National Research Program on Design-Based/Model-Assisted Survey Methodology for Aquatic Resources at Oregon State University. It has not been subjected to the Agency's review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

Azzalini A. 1996. *Statistical Inference Based on the Likelihood*. Chapman & Hall: London.

Bulmer MG. 1974. On fitting the Poisson lognormal distribution to species abundance data. *Biometrics* **30**: 101–110.

Bunge J, Fitzpatrick M. 1993. Estimating the number of species: a review. *Journal of the American Statistical Association* **88**: 364–373.

Burnham KP, Overton WS. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60**: 927–936.

Cao Y, Williams DD, Larsen DP. 2002. Comparison of ecological communities: the problem of sample representativeness. *Ecology* **72**: 41–56.

Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**: 265–270.

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

Chao A. 2003. Species richness estimation. In *Encyclopedia of Statistical Sciences* (2nd edn), Balakrishnan N, Read CB, Vidakovic B (eds). Wiley: New York.

Chao A, Lee S-M. 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**: 210–217.

Chao A, Ma M-C, Yang MCK. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika* **80**: 193–201.

Chazdon RL, Colwell RK, Denslow JS, Guariguata, MR. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica. In *Forest Biodiversity Research, Monitoring and Modeling*, Dallmeier F, Comiskey JA (eds). Parthenon Publishing: Paris; 285–309.

Christen JA, Nakamura M. 2000. On the analysis of accumulation curves. *Biometrics* **56**: 748–754.

Christen, JA, Nakamura M. 2003. Sequential stopping rules for species accumulation. *Journal of Agricultural, Biological, and Environmental Statistics* **8**: 184–195.

Colwell RK. 2004. Estimates: statistical estimation of species richness and shared species from samples. Version 7. User's Guide and application published at: http://purl.oclc.org/estimates.

Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society (Series B)* **345**: 101–118.

Emigh TH. 1983. On the number of observed classes from a multinomial distribution. *Biometrics* **39**: 485–491.

Fisher RA, Corbet AS, Williams CB. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**: 42–58.

Flather CH. 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography* **23**: 155–168.

Hubbell SP. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press: Princeton.

Hurlbert SH. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**: 577–586.

Larsen DP, Herlihy AT. 1998. The dilemma of sampling streams for macroinvertebrate richness. *Journal of the North American Benthological Society* **17**: 359–366.
Li J, Herlihy A, Gerth W, Kaufmann P, Gregory S, Urquhart S, Larsen DP. 2001. Variability in stream macroinvertebrates at multiple spatial scales. *Freshwater Biology* **46**: 87–97.
Ludwig JA, Reynolds JF. 1988. *Statistical Ecology: A Primer on Methods and Computing*. John Wiley & Sons: New York.
MacArthur RH. 1960. On the relative abundance of species. *American Naturalist* **94**: 25–36.
Murtaugh PA. 1989. Size and species composition of zooplankton in experimental ponds with and without fishes. *Journal of Freshwater Ecology* **5**: 27–38.
Paine RT. 1988. Food webs: road maps of interactions or grist for theoretical development? *Ecology* **69**: 1648–1654.
Pielou EC. 1969. *An Introduction to Mathematical Ecology.* Wiley-Interscience: New York.
Pielou EC. 1975. *Ecological Diversity.* Wiley: New York.
Pipoly JJ, III, Madulid DA. 1998. Composition, structure and species richness of a submontane moist forest on Mt. Kinasalapi, Mindanao, Phillippines. In *Forest Biodiversity Research, Monitoring and Modeling*, Dallmeier F, Comiskey JA (eds). Parthenon Publishing: Paris; 591–600.
R Development Core Team. 2004. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.
Read CB. 1988. Zipf's Law. In *Encyclopedia of Statistical Sciences* (Vol. 9), Kotz S, Johnson NL, Read CB (eds). Wiley: New York; 674–676.
Ricklefs RE, Schluter D (eds). 1994. *Species Diversity in Ecological Communities: Historical and Geographical Perspectives*. University of Chicago Press: Chicago.
Rodrigues J, Milan LA, Leite JG. 2001. Hierarchical Bayesian estimation for the number of species. *Biometrical Journal* **43**: 737–746.
Smith EP, van Belle G. 1984. Nonparametric estimation of species richness. *Biometrics* **40**: 119–129.
Solow AR. 1994. On the Bayesian estimation of the number of species in a community. *Ecology* **75**: 2139–2142.
Solow AR, Polasky S. 1999. A quick estimator for taxonomic surveys. *Ecology* **80**: 2799–2803.
Sun I-F, Hsieh C-F, Hubbell SP. 1998. Structure and species composition of a sub-tropical rain forest in southern Taiwan on a wind-stress gradient. In *Forest Biodiversity Research, Monitoring and Modeling*, Dallmeier F, Comiskey JA (eds). Parthenon Publishing: Paris; 563–590.

