Ecology, 86(5), 2005, pp. 1143–1153

ESTIMATION OF SPECIES RICHNESS: MIXTURE MODELS, THE ROLE OF RARE SPECIES, AND INFERENTIAL CHALLENGES

CHANG XUAN MAO[1,3] AND ROBERT K. COLWELL[2]

[1]*Department of Statistics, University of California, Riverside, California 92521 USA*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043 USA*

Abstract. We examine the role of rare species in the problem of estimating within-habitat species richness based on sampling data. Richness estimation can be modeled realistically for abundance-based and incidence-based data using Poisson or binomial mixtures, respectively. The problem can be reduced to estimation of the odds of the probability of a species remaining undetected in the sample or sample set. Within this rigorous statistical framework, we explore existing methods of richness estimation and assess their limitations. We do this by modeling the addition of increasing numbers of rare, undetected species to a reference assemblage, assessing the power of different methods to distinguish the modified species assemblages from the reference assemblage. (We use empirical example data sets for birds, seeds, and beetles as reference assemblages.) By considering the contributions of rare species and the role of undetected species for a fixed sampling effort, we show why the problem of richness estimation is so difficult, and we discuss what statistical models can provide.

Key words: conditional inference; mixture models; nonparametric maximum likelihood estimation; one-sided inference; richness estimation; richness extrapolation; singleton species; species accumulation curve; transient species; zero-truncated binomial; zero-truncated Poisson.

INTRODUCTION

The concept of species rarity in ecology and biogeography takes many forms, depending primarily upon spatial and temporal scale (Fisher et al. 1943, Preston 1948, Rabinowitz 1981, Gaston 1994, Magurran 2004). We will consider rarity at the habitat level, in the context of alpha diversity (Whittaker 1972) and the estimation of local species richness from sampling data.

With the exception of thorough biotic surveys in isolated, species-poor habitats, it is routine in species inventory work to find that, even after intensive sampling, some species are represented by only one or two individuals (singletons or doubletons) or are detected in only one or two samples in a replicated sample set (uniques or duplicates) (Colwell and Coddington 1994). Often, enlarging the sample (or sample set) yields additional individuals of these rare species, moving them into higher abundance or occurrence classes, but at the same time reveals additional species that now represent new singletons and doubletons or uniques and duplicates. These are the workings of Preston's demon, the moving ''veil line'' between detected and the undetected species as sample size increases (Preston 1948).

For example, in a multiyear study of the insect herbivores of a selected set of plant species in New Guinea, Novotny and Basset (2000) found that 278 of the 1050 insect species recorded (26%) were singletons, based on more than 80 000 individual insects. When a single host-plant species was considered, 45% of the leaf-chewing or sap-sucking insect species were singletons. Recent microbial inventories, made possible by molecular tools, have revealed an astonishing richness of species (however defined), with even more astonishing proportions of singletons, sometimes reaching more than 90% (Hughes et al. 2000, Falkowski and de Vargas 2004).

What do rare species mean, in inventory data? Biologists have long suspected that rare species in many assemblages are a mix of genuinely rare, ''persistent'' or ''resident'' species and ''transient'' or ''occasional'' species that may be common elsewhere (Magurran and Henderson 2003). In the most extensive local inventory of tropical ants on record, Longino et al. (2002) carried out a replicated, multimethod sampling campaign over many years at a Costa Rican rain forest site. Fig. 1 shows the results in the form of a species accumulation curve (a sample-based rarefaction curve [Gotelli and Colwell 2002]). Even when all samples are pooled, the numbers of uniques and duplicates are barely declining, as rarer and rarer occurrence classes come to light. Longino et al. (2002) considered the 51 uniques (12% of the total 437 species) individually and classified them as geographical edge species (14 species, known to be common elsewhere but rare at the site), methodological edge species (six species, probably common at the site, but not susceptible to the survey methods

Manuscript received 9 July 2004; revised and accepted 5 August 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.
[3] E-mail: cmao@stat.ucr.edu

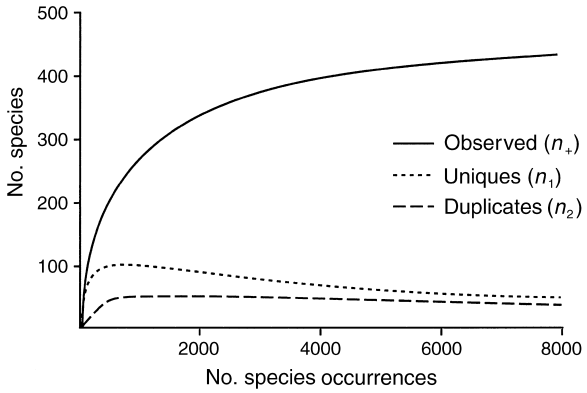

FIG. 1. Species accumulation (sample-based rarefaction) curve for rain forest ants. The lower curves show the number of species detected in only one sample (uniques) and in exactly two samples (duplicates). Note that rare species continue to accumulate even after more than 8000 samples have been pooled (Longino et al. 2002).

used), globally rare species (25 species, known elsewhere but not common anywhere), or globally unique species (known only from a single sample at the site, and nowhere else on earth).

Whatever the explanation for their rarity, the presence of substantial numbers of rare species in sampling data suggests that the inventory is incomplete—that the true species richness for the study habitat includes undetected species. What are the prospects for accounting for those undetected species statistically? Can we tell if an inventory is complete, or nearly so? In this paper, we explore these questions in the context of a statistically rigorous sampling model.

A Framework of Mixture Models

Two kinds of data

Consider an assemblage with a true richness of S species. The ith species has a true relative abundance ϕ_i for $i = 1, 2, \ldots, S$, with $\Sigma_{i=1}^{S} \phi_i = 1$. In the estimation of the species richness S, two general classes of sampling data may be distinguished: abundance-based data and incidence-based data. The simplest form of abundance-based data is a single, multispecies sample, in which the number of individuals from each species found in the sample is recorded. The number of individuals from the ith species Y_i will be treated as a Poisson random variable with a mean parameter λ_i, called the detection rate. The detection rates λ_i depend on the relative abundances ϕ_i, the probability of an individual being detected when it is present, and the sample size (the number of individuals), which, in turn, is a function of the sampling effort.

Incidence-based data, in the simplest case, consist of a set of multispecies samples (from timed observations, quadrats, traps, lures, seines, dredge hauls, mist nets, or other replicated sampling units) for which only the detection or nondetection of species in each sample is available. Let p_i be the detection probability of the ith species. Let Y_i be the number of samples in which the ith species is detected, which is a binomial random variable. We will use the detection odds $\lambda_i = p_i = (1 - p_i)$ for incidence-based data. The detection odds λ_i depend on the relative abundances ϕ_i, the probability of a species being detected when it is present, the sampling design (e.g., quadrat size), and on nonrandom, species-specific aggregation or disaggregation or individuals among samples (Colwell et al. 2004).

The Y_i will be called frequencies for both kinds of data. If $Y_i = 0$, the ith species does not appear in the sample (for abundance-based data) or sample set (for incidence-based data). (In statistical terms, an incidence-based sample set is treated as a single sample, but we will use the term ''sample set'' for such data, to conform to ecological terminology.) We will assume that if the detection is imperfect, then the effect of imperfect detection does not vary across individuals or species. Thus, for a fixed sampling effort, the detection rates/odds λ_i depend only on the relative abundances ϕ_i, in the sense that a large/small ϕ_i corresponds to a large/small λ_i. The homogeneous case means that the ϕ_i and the λ_i are identical, which is rarely, if ever, true. The heterogeneous case, in which the ϕ_i and the λ_i are allowed to vary across species, will be the focus of this article.

An empirical data set can be summarized in terms of the counts n_x, where n_x is the number of frequencies Y_i that equal x. Thus n_0 is the number of undetected species, n_1 is the number of singletons, and n_2 is the number of doubletons, etc. (For incidence-based data, the n_1 are often called uniques and the n_2 are called duplicates [Colwell and Coddington 1994]). The observed richness $n_+ = \Sigma_{x\geq 1} n_x$ is the number of species detected in the sample (for abundance-based data) or sample set (for incidence-based data), from among the S species actually present.

For example, Janzen (1973) collected an abundance-based data set on tropic beetles; Norris and Pollock (1998) analyzed an abundance-based avian dataset (1995 census data for the Wisconsin route of the North American Breeding Bird Survey); and Butler and Chazdon (1998) recorded the species of tropical plants emerging from seed-bank samples, which we treat here as a replicated, incidence-based sample set. In the beetle data, $n_+ = 78$ species were detected and the nonzero observed counts are $n_1 = 59$, $n_2 = 9$, $n_3 = 3$, $n_4 = 2$, $n_5 = 2$, $n_6 = 2$, $n_{11} = 1$; (Chao and Shen [2003] also analyzed this classic data set). In the bird data, $n_+ = 72$ species were detected and the nonzero observed counts are $n_1 = 11$, $n_2 = 12$, $n_3 = 10$, $n_4 = 6$, $n_5 = 2$, $n_6 = 5$, $n_7 = 1$, $n_8 = 3$, $n_9 = 2$, $n_{10} = 4$, $n_{12} = 1$, $n_{13} = 1$, $n_{14} = 1$, $n_{15} = 2$, $n_{16} = 1$, $n_{18} = 2$, $n_{25} = 1$, $n_{29} = 1$, $n_{30} = 1$, $n_{32} = 1$, $n_{39} = 1$, $n_{44} = 1$, $n_{53} = 1$, $n_{54} = 1$. In the seed-bank data, $n_+ = 34$ species were detected among 121 soil samples and the nonzero observed counts are $n_1 = 3$, $n_2 = 2$, $n_3 = 3$, $n_4 = 3$, $n_5 = 1$, n_6

$= 5$, $n_7 = 1$, $n_8 = 1$, $n_9 = 3$, $n_{10} = 1$, $n_{11} = 2$, $n_{13} = 1$, $n_{17} = 1$, $n_{24} = 2$, $n_{43} = 2$, $n_{47} = 1$, $n_{52} = 1$, $n_{61} = 1$.

The mixture models

In the model for abundance-based data, Y_i is a Poisson random variable with detection rate λ_i. In the incidence-based model, Y_i is a binomial random variable with detection odds λ_i. Let T be the number of samples in an incidence-based data set. A Poisson/binomial density can be written as

$$g(y; \lambda) = \begin{cases} \dfrac{\lambda^y}{y!e^\lambda} & \text{(abundance-based)} \\ \dbinom{T}{y} \dfrac{\lambda^y}{(1 + \lambda)^T} & \text{(incidence-based)} \end{cases}$$

where $y \geq 0$ is a possible value of Y_i. Note that we use the term density for the probability mass function throughout this paper. We will assume that $\lambda_1, \lambda_2, \ldots, \lambda_S$ arise as a sample of identically and independently distributed random variables from a latent distribution $H(\lambda)$. For incidence-based data, this is equivalent to the assumption that the detection probabilities p_i also arise as a random sample from a latent distribution. Such a model is called a mixture model because the unconditional distribution of the frequencies Y_i is a mixture of Poisson or binomial distributions.

In a finite mixture model, the latent distribution H is discrete over G values of parameter λ (the support points ξ_k), with mixing weights π_k, $k = 1, 2, \ldots, G$. This means that if one selects one of the λ_i randomly, then the probability that it equals ξ_k is π_k. There is another biological interpretation of a finite mixture model. An assemblage consists of G groups of species, called homogeneous groups. Within each group, the species share the same detection rate/odds, called the group detection rate/odds. The mixing weight is the group relative richness: the number of species in group k divided by S, the total species richness in the assemblage. Fig. 2 illustrates an assemblage of 20 species with four homogeneous groups.

Thus, we treat the frequencies Y_i as a sample from a Poisson or binomial mixture, which we can specify as

$$g_H(y) = \pi_1 g(y; \xi_1) + \pi_2 g(y; \xi_2) + \cdots + \pi_G g(y; \xi_G).$$

where each term in the summation represents the contribution to the mixture from one of the G homogeneous species groups. The model $g_H(y)$ includes not only detected ($Y_i > 0$), but undetected ($Y_i = 0$) species. To model real data, we need a density for only the detected species. Given n_+ observed species, the density of those $Y_i > 0$ is a zero-truncated mixture of Poisson or binomial distributions:

$$g_H(x)/[1 - g_H(0)] \qquad x \geq 1.$$

This can be written as a mixture $f_Q(x)$ of truncated Poisson/binomial densities as follows:

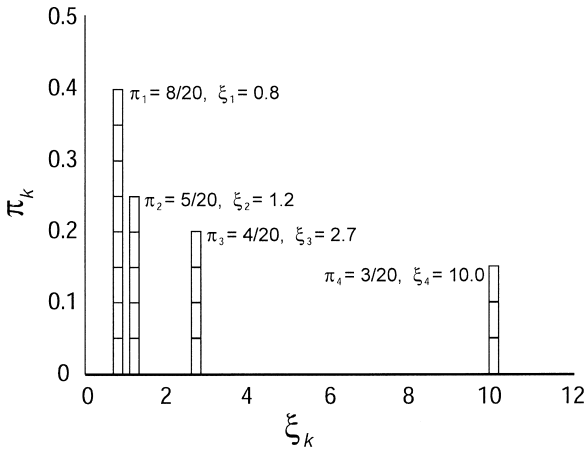

FIG. 2. Four homogeneous groups comprising an assemblage of $S = 20$ species. The group detection rates/odds ξ_k and the group relative richness π_k are indicated in the figure.

$$g_H(x)/[1 - g_H(0)] = f_Q(x) \qquad x \geq 1$$

where Q is a derived latent distribution (taking zero truncation into account) that has the same support points as the latent distribution H, but requires adjusted mixing weights ϖ_k. To be specific, we define

$$f_Q(x) = \varpi_1 f(x; \xi_1) + \varpi_2 f(x; \lambda_2) + \cdots + \varpi_G f(x; \xi_G) \qquad x \geq 1 \tag{1}$$

$$f(x; \lambda) = \frac{g(x; \lambda)}{1 - g(0; \lambda)} = \begin{cases} \dfrac{\lambda^x}{x!(e^\lambda - 1)} & \text{(abundance-based)} \\ \dbinom{T}{y} \dfrac{\lambda^x}{(1 + \lambda)^T - 1} & \text{(incidence-based)} \end{cases} \tag{2}$$

$$\varpi_k = \frac{S\pi_k[1 - g(0; \xi_k)]}{S[1 - g_H(0)]} = \frac{1 - g(0; \xi_k)}{1 - g_H(0)}\pi_k. \tag{3}$$

As a probability density, $f(x; \lambda)$ is defined only for $x \geq 1$. For notational convenience, we will use $f(0; \lambda) = g(0; \lambda) = (1 - g(0; \lambda))$, which is not a probability, but it is the odds of a species being undetected if this species has detection rate/odds λ. Note that $f(0; \lambda)$ goes to infinity as λ goes to zero at the same rate as $1/\lambda$, because it is clear that

$$B = \lim_{\lambda \to 0} \lambda f(0; \lambda) = \begin{cases} \lim_{\lambda \to 0} \dfrac{\lambda}{e^\lambda - 1} = 1 & \text{(abundance-based)} \\ \lim_{\lambda \to 0} \dfrac{\lambda}{(1 + \lambda)^T - 1} = \dfrac{1}{T} & \text{(incidence-based).} \end{cases} \tag{4}$$

This fact will be used in later discussion. As seen from the first equality in Eq. 3, the mixing weight ϖ_k is the group relative expected observed richness in the sense

TABLE 1. An example assemblage from 1995 census data for the Wisconsin route of the North American Breeding Bird Survey.

Parameter	k				
	1	2	3	4	5
ξ_k	2.1328	7.2628	13.6913	30.3418	48.7918
π_k	0.5407	0.2117	0.1362	0.0646	0.0469
ω_k	0.5717	0.1974	0.1269	0.0602	0.0437

Notes: The total species richness in this assemblage is $S = 77$. We present the support points ξ_k and mixing weights π_k of the latent distribution H and ω_k of the derived latent distribution Q. The derived latent distribution Q is the nonparametric maximum-likelihood estimate of the bird data set. H and S are calculated from Q and the observed richness.

that it is the ratio of the expected number of detected species in the kth homogeneous group over the total expected number of detected species from the assemblage. We will identify latent distribution H with the set of parameters $\xi_1, \xi_2, \ldots, \xi_G$ and $\pi_1, \pi_2, \ldots, \pi_G$ and identify the derived latent distribution Q with the set of parameters $\xi_1, \xi_2, \ldots, \xi_G$ and $\varpi_1, \varpi_2, \ldots, \varpi_G$. Table 1 is an example assemblage.

The abundance structure (for abundance-based data) or incidence structure (for incidence-based data) of an assemblage is determined by the total species richness S, the number of homogeneous groups G, the group relative richness π_k, or equivalently the group relative expected observed richness ϖ_k, and the group detection rates/odds ξ_k. Because these parameters are allowed to vary, finite mixture models can provide a good approximation to many real assemblages.

There are parametric mixture models (Fisher et al. 1947, Burnham 1972, Ord and Whitmore 1983, Sichel 1997, Dorazio and Royle 2003). For example, H can be a gamma distribution $H(\lambda; a, s)$ or a beta distribution $H(p; a, b)$ (on the detection probability p), where

$$H(\lambda; a, s) = \int_0^{\lambda} \frac{1}{s^a \Gamma(a)} z^{a-1} e^{-z/s}\, dz$$

$$H(p; a, b) = \int_0^{p} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^{a-1}(1-z)^{b-1}\, dz.$$

One can find a discrete latent distribution H that approximates a continuous gamma/beta distribution. Most importantly, the mixture $f_Q(x)$ of zero-truncated Poisson/binomial densities from the continuous distribution and that from its discrete approximation can be very close; see Fig. 3.

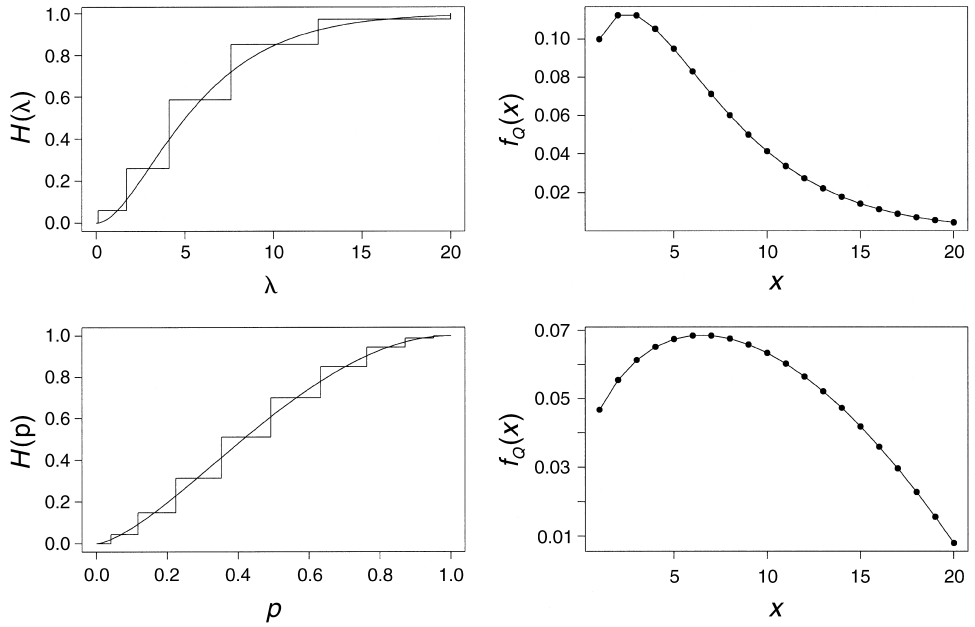

FIG. 3. A continuous latent distribution approximated by a discrete latent distribution. The top left panel presents a gamma distribution $H(\lambda; 2, 3)$ and a discrete latent distribution. The bottom left panel presents a beta distribution with $H(p; 1.5, 2)$ and a discrete latent distribution. The mixtures of the gamma distribution and its discrete approximation are presented in the top right panel. The mixtures of the beta distribution and its discrete approximation are presented in the bottom right panel. For the Poisson case, only $f_Q(x)$ with $x \leq 20$ are plotted; for the binomial case, $T = 20$.

The probability of observing a particular set of frequencies Y_i (a particular pattern of abundances or presences), called the joint density of the frequencies Y_i, is

$$\prod_{i=1}^{S} g_H(y_i) = \prod_{y\geq 0} g_H^{n_y}(y).$$

There are $S!/(\Pi_{x\geq 0} n_x!)$ ways to get the same counts n_x with $x \geq 0$ from the Y_i (note that this number includes the count for undetected species n_0). Because $n_0 + n_+ = S$, the density of the counts, as the full likelihood of the parameters S and H, becomes

$$L_1(S, H) = \frac{S!}{(S - n_+)! \prod_{x\geq 1} n_x!} g_H^{S-n_+}(0) \prod_{x\geq 1} g_H^{n_x}(x).$$

The observed counts n_x with $x \geq 1$ are called sufficient statistics because they contain all the information available for us to make statistical inference. The conditional likelihood of the observed species richness depends only on the latent distribution Q,

$$L_2(Q) = \frac{n_+!}{\prod_{x\geq 1} n_x!} \prod_{x\geq 1} f_Q^{n_x}(x) = \frac{n_+!}{\prod_{x\geq 1} n_x!} \prod_{x\geq 1} \left[\frac{g_H(x)}{1 - g_H(0)}\right]^{n_x}. \quad (6)$$

Note that $L_1(S; H) = L_2(Q)L_3(S; H)$, where $L_3(S; H)$ is the density of n_+:

$$L_3(S, H) = \frac{S!}{(S - n_+)!n_+!} g_H^{S-n_+}(0)[1 - g_H(0)]^{n_+} \quad (7)$$

which is called the marginal likelihood.

Predicting the number of undetected species

The key to estimating the true richness S from the sample data lies in modeling the undetected species, n_0. If we knew the latent distribution H, estimation of the species richness would be straightforward. The maximum likelihood estimator for S is the integer part of

$$\hat{S} = n_+ + n_+ g_H(0)/[1 - g_H(0)] \quad (8)$$

which maximizes the marginal likelihood $L_3(S; H)$ in Eq. 7 and linearly depends on H only through the odds of a species being undetected:

$$\alpha = g_H(0)/[1 - g_H(0)].$$

We can write α as $\alpha(Q)$, because it is easy to show that

$$\alpha(Q) = \sum_{k=1}^{G} \varpi_k f(0; \xi_k). \quad (9)$$

Note that $n_+\alpha$ for a fixed α predicts the number of undetected species n_0. Because α is unknown, the problem is reduced to estimation of α.

Statistical estimation methods

Various statistical methods have been applied toward estimation of the species richness S. Bunge and Fitzpatrick (1993) presented a comprehensive review; also see Colwell and Coddington (1994). Here we will summarize various methods from the point of view of mixture models, and in particular, we will include new developments in the last ten years. Each procedure can be thought to provide an estimator for α, the odds of a randomly selected species being undetected. If a procedure produces an estimator $\hat{S}$ for S directly, then we can write an estimator for α as $\hat{\alpha} = \hat{S}/n_+ - 1$.

Although Q is unknown, $f_Q(x)$ can be estimated by the sample proportion $\hat{f}(x) = n_x/n_+$, where $\hat{f}(x)$ is called the empirical density. If a parameter is a function of the $f_Q(x)$, then we can estimate it by replacing $f_Q(x)$ with $\hat{f}(x)$. To see this, consider, for example, a parameter α_{ML} and its estimator $\hat{\alpha}_{ML}$,

$$\alpha_{ML} = \alpha_{ML}(Q) = A\frac{f_Q^2(1)}{2f_Q(2)}$$

$$\hat{\alpha}_{ML} = A\frac{\hat{f}^2(1)}{2\hat{f}(2)} = A\frac{n_1^2}{2n_+n_2} \quad (10)$$

where $A = 1$ (abundance-based) or $A = 1 - 1/T$ (incidence-based).

The unfortunate fact is that the odds α cannot be written as an explicit function of the $f_Q(x)$. There are two general recipes toward estimation of α. The first, termed the approximation recipe, means that one finds some parameter $\alpha^{\#}$ that is an explicit function of the $f_Q(x)$ and is assumed to be close to α under certain situations. The estimator $\hat{\alpha}^{\#}$ will be used as an estimator for α. The parameter α_{ML} in Eq. 10 is such an example. It was shown in Mao and Lindsay (2003) and Mao (2004*a*) that for all assemblages, we have $\alpha_{ML} \leq \alpha$, with $\alpha_{ML} = \alpha$ for the homogeneous case. This means that α_{ML} is a universal lower bound for α. Chao (1984) obtained α_{ML} for the abundance-based data and Chao (1989) applied it to the incidence-based data without the factor $A = 1 - 1/T$, which is close to one for a large T. Many estimators for α that are well-known to ecologists, under the name of nonparametric estimators, in the sense that $\hat{f}(x)$ is a nonparametric estimator for $f_Q(x)$, can be thought to be based on the approximation recipe, although the original logic that leads to the development of an estimator might be something else and the development might be done in a different model (Burnham and Overton 1978, Darroch and Ratcliff 1980, Smith and van Belle 1984, Zelterman 1988, Chao and Lee 1992, Lee and Chao 1994).

The second strategy, termed the plug-in recipe, means that one finds an estimator $\hat{Q}$ for the latent distribution Q and then plugs the elements of $\hat{Q}$ (Eq. 1) into Eq. 9 to yield an estimator for α. This is also equivalent to estimating the latent distribution H. Because estimation of the latent distribution Q or H in-

volves complicated iterative algorithms, the plug-in recipe is less known and less used in ecology, although several estimation procedures have been proposed (Burnham 1972, Efron and Thisted 1976, Ord and Whitmore 1986, Mingoti and Meeden 1992, Norris and Pollock 1996, 1998, Pledger 2000, Dorazio and Royle 2003, Mao 2004*a*, *b*, Mao et al., *in press*).

Among various Q/H-estimation procedures, the ones that maximize either the full likelihood $L_1(S, H)$ in Eq. 5 or the conditional likelihood $L_2(Q)$ in Eq. 6 have been addressed thoroughly. The nonparametric maximum likelihood procedures in Norris and Pollock (1996, 1998) used $L_1(S, H)$ in Eq. 5. The nonparametric maximum likelihood procedures in Mao (2004*a*) used $L_2(Q)$ in Eq. 6. The parametric maximum likelihood procedure for incidence-based data in Dorazio and Royle (2003) used $L_1(S, H)$ in Eq. 5. The parametric maximum likelihood procedure for the abundance-based data in Efron and Thisted (1976) used a conditional likelihood similar to $L_2(Q)$ in Eq. 6.

The Role of Rare Species in Inference

Rare species are interpreted here as those with small relative abundances ϕ_i, which in turn result in small detection rates/odds λ_i. To see how the existence of rare species will affect statistical inference on the total species richness, we will consider fixing an assemblage, called a reference assemblage, and compare it with a modified assemblage, which differs from the reference assemblage by an additional group of species that have a small detection rate/odds. We will denote the reference assemblage as $C = \{S, H\}$ or $C = \{S, Q\}$ and the modified assemblage by $C^* = \{S^*, H^*\}$ or $C^* = \{S^*, Q^*\}$, where the relative species richness, the relative expected observed richness, and the detection rate/odds of the additional rare species group are denoted by π_0, ϖ_0^*, and ξ_0^*, respectively. For each species group in the reference assemblage, its group relative richness and group relative expected observed richness in the modified assemblage have been changed by the addition of the rare species group, as specified by

$$\xi_k^* = \xi_k \qquad \pi_k^* = \pi_k(1 - \pi_0^*)$$
$$\varpi_k^* = \varpi_k(1 - \varpi_0^*) \qquad k = 1, 2, \ldots, G. \quad (11)$$

Fig. 4 illustrates a reference assemblage and a modified assemblage.

The expected number of species that have frequency x from the kth homogeneous group equals the number of species in the group, $S\pi_k$, times the probability $g(x; \xi_k)$ that a species has frequency x. The expected number of species that have the frequency x from the reference assemblage C is

$$E\{n_x \mid C\} = \sum_{k=1}^{G} S\pi_k g(x; \xi_k) \qquad x \geqq 0.$$

From Eqs. 2 and 3, and the fact that $S\{1 - g_H(0)\} = E\{n_+ \mid C\}$, we can write the following:

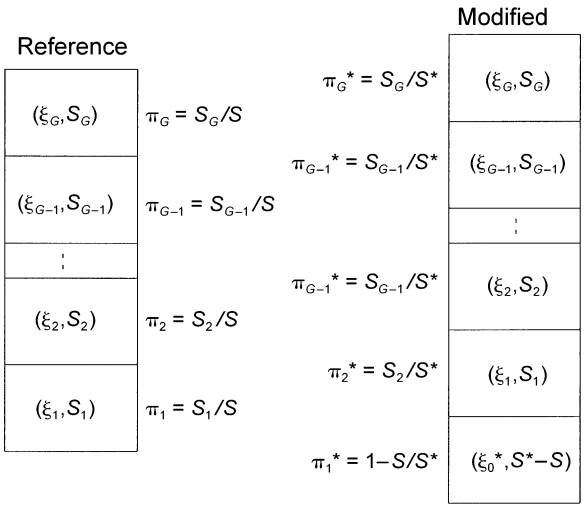

FIG. 4. A reference assemblage and a modified assemblage. Each species group is put in a small box. The modified assemblage has an additional rare species group. The group relative richness changes for a species group in the reference assemblage while the group detection rate/odds is kept the same.

$$E\{n_x \mid C\} = E\{n_+ \mid C\} \sum_{k=1}^{G} \varpi_k f(x; \xi_k) \qquad x \geq 0. \quad (12)$$

Because going from the reference assemblage C to the modified assemblage C^*, we add a group of rare species, we can write the difference as

$$E\{n_x \mid C^*\} - E\{n_x \mid C\} = (S^* - S)g(x; \xi_0^*)$$
$$= E\{n_x \mid C^*\}\varpi_0^* f(x; \xi_0^*) \qquad x \geq 0.$$

Note that ϖ_0^* is the ratio of the expected number of species being detected from the additional rare species group over the total expected number of species being detected from the modified assemblage, that is,

$$\varpi_0^* = \frac{E\{n_+ \mid C^*\} - E\{n_+ \mid C\}}{E\{n_+ \mid C^*\}} \quad (13)$$

which can be reformulated to

$$E\{n_+|C^*\} - E\{n_+|C\} = E\{n_+|C\}\varpi_0^*/\{1 - \varpi_0^*\}. \quad (14)$$

Due to Eq. 14, we can write Eq. 12 in terms of $E\{n_x \mid C\}$, a quantity determined by the reference assemblage, and ϖ_0^* and ξ_0^*, the relative expected observed richness and the detection rate/odds of the rare species group, respectively:

$$E\{n_x \mid C^*\} - E\{n_x \mid C\} = E\{n_+ \mid C\}\frac{\varpi_0^*}{1 - \varpi_0^*} f(x; \xi_0^*) \qquad (x \geq 1) \quad (15)$$

$$E\{n_0 \mid C^*\} - E\{n_0 \mid C\} = E\{n_+ \mid C\}\frac{\varpi_0^*}{1 - \varpi_0^*} f(0; \xi_0^*). \quad (16)$$

The right-hand side of each of Eqs. 14, 15, and 16

TABLE 2. Example 1 of modified assemblages.

m	2	10	50	100	1000	2000	5000	Reference
α	0.30	0.47	0.52	0.53	0.53	0.53	0.53	0.07
En_+	84.76	74.91	72.43	72.10	71.81	71.79	71.78	71.77
En_0	25.24	35.09	37.57	37.90	38.19	38.21	38.22	5.23
En_1	21.21	14.19	11.85	11.53	11.24	11.22	11.21	11.20
En_2	14.65	12.30	12.15	12.15	12.15	12.15	12.15	12.15
En_3	9.54	9.13	9.12	9.12	9.12	9.12	9.12	9.12
En_4	5.80	5.75	5.75	5.75	5.75	5.75	5.75	5.75
En_5	3.76	3.76	3.76	3.76	3.76	3.76	3.76	3.76

Notes: The modified assemblage $C^{(m)}$ is obtained by adding to the latent distribution H (the reference assemblage) in Table 1 a support point $\xi_0^{(m)} = 1/m$ with a mixing weight $\pi_0^{(m)} = 0.3$ so that $S^{(m)} = 110$. We present the odds α, the expected observed richness En, and the expected counts En_x for $x = 0, 1, 2, 3, 4, 5$.

stands for the contribution to the expected count or the expected observed species richness from the rare species group. From Eqs. 14 and 15, if the relative expected observed richness of the rare species group ϖ_0^* is small enough, then the contribution from the added rare species group to the expected observed richness or any expected observed count can be as small as possible, that is, for $x \geq 1$, when $\varpi_0^* \approx 0$,

$$0 \leq E\{n_x | C^*\} - E\{n_x | C\} \leq E\{n_+ | C^*\} - E\{n_+ | C\} \approx 0.$$

This means that, with the same sampling effort, empirically one can not tell whether the data are generated from the reference community C or from the modified community C^*. On the other hand, when both ϖ_0^* and the detection rate/odds ξ_0^* of the rare species group are small, using Eq. 4, we can write Eq. 16 as

$$E\{n_0 | C^*\} - E\{n_0 | C\} = E\{n_+ | C\}[\xi_0^* f(0; \xi_0^*)](\varpi_0^*/\xi_0^*) \approx E\{n_+ | C\}B(\varpi_0^*/\xi_0^*).$$

One can easily find ϖ_0^* and ξ_0^* such that the contribution from the rare species group to the expected unobserved count can be as large as possible, for example, using ϖ_0^* and ξ_0^* with $\varpi_0^* = \sqrt{\xi_0^*}$ and $\xi_0^* \approx 0$.

The different contributions to the expected observed counts and the expected unobserved counts from the additional rare species have profound consequences in statistical inference on the odds α and the estimation of the species richness S. From Eqs. 1, 9, and 11, we can also obtain

$$\sum_{x \geq 1} |f_{Q^*}(x) - f_Q(x)| = \varpi_0^* \sum_{x \geq 1} |f(x; \xi_0^*) - f_Q(x)| \leq 2\varpi_0^* \quad (17)$$

$$\alpha(Q^*) - \alpha(Q) = \varpi_0^* f(0; \xi_0^*) - \varpi_0^* \alpha(Q) \approx B\varpi_0^*/\xi_0^* \quad (18)$$

where the approximation in Eq. 18 holds when both ϖ_0^* and ξ_0^* are small.

Eqs. 17 and 18 suggest that the estimation procedures based on either the approximation recipe or the plug-in recipe might have trouble producing useful results. For a procedure based on the approximation recipe, for example, for α_{ML} in Eq. 10, we will have $\alpha_{ML}(Q^*) \approx \alpha_{ML}(Q)$ as $\varpi_0^* \approx 0$. If $\alpha_{ML}(Q) \approx \alpha(Q)$, then $\alpha_{ML}(Q^*) \approx \alpha(Q)$ and $\alpha_{ML}(Q^*)$ can never be close to $\alpha(Q^*)$. The same logic applies to all estimators based on the approximation recipe. There is no such parameter $\alpha^\#$ for α such that $\alpha^\#$ is close to α for all latent distributions.

Consider any procedure that estimates the latent distribution Q. Even if Q fits a data set well, then, because of Eq. 17, when ϖ_0^* is small enough, Q^* should also fit the same data as well as Q and numerically either Q^* or Q might provide a fit slightly better than the other. However, the estimators for α obtained from Q^* and Q will be dramatically different so that estimation of α by plug-in might produce extremely large values. A theoretic consequence for confidence intervals is that, for a confidence interval, for either the odds α or the species richness S, to achieve its advertised confidence level, the upper confidence limit must often infinite.

As an illustration, the assemblage in Table 1 has $S = 77$ species, taken to be the true, complete reference assemblage. We will consider a sequence of modified assemblages $C^{(m)} = \{S^{(m)}, H^{(m)}\}$ with corresponding derived latent distribution $Q^{(m)}$, that are parameterized by m. Table 2 and Fig. 5 show two sequences of modified assemblages. In Table 2, the last column (reference) shows the expected counts and observed species richness for the reference assemblage. To create the modified assemblages in Table 2, we add a fixed number of rare species (33, so that $S^{(m)} = 110$), with a fixed mixing weight $\pi_0^{(m)}$. We then vary the support point $\xi_0^{(m)}$ for the added species according to $\xi_0^{(m)} = 1/m$ so that $\varpi_0^{(m)}$ also becomes smaller as m gets larger, with m ranging from 2 to 5000, to explore the effect of adding rarer and rarer species. (The function $\xi_0^{(m)} = 1/m$ is just a convenient way to vary $\xi_0^{(m)}$ based on the index m.) When $\xi_0^{(m)}$ is large (e.g., $m = 2$, $\xi_0^{(m)} = 0.5$), the added species are reflected in the expected counts for the observed species (e.g., the expected singletons En_1 goes from 11.20 to 21.21), and thus in the expected number of observed species En_+ (which rises from 71.78 to 84.77). But as $\xi_0^{(m)}$ becomes smaller and smaller

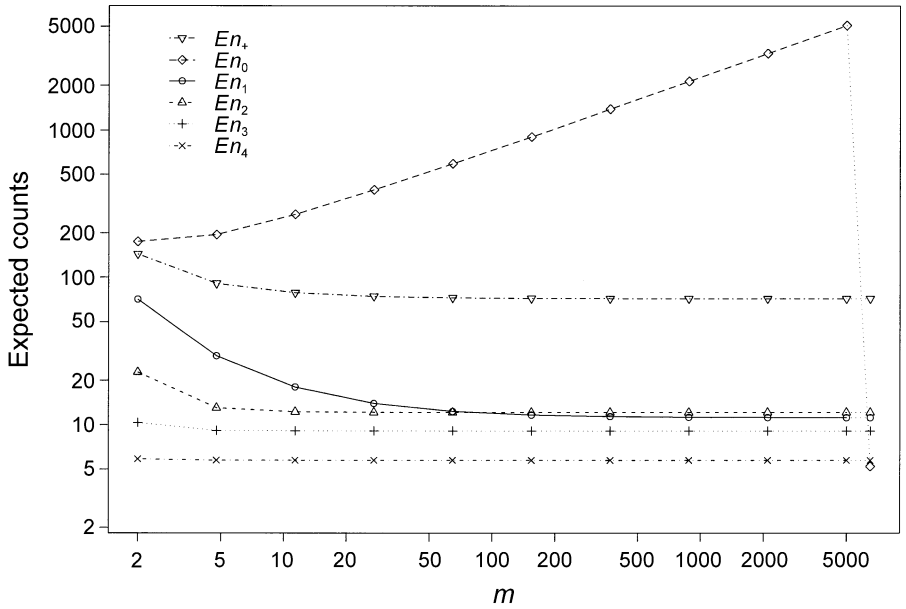

FIG. 5. Example 2 of modified assemblages. Reference (C) and modified ($C^{(m)}$) assemblages with a variable number of added rare species, variable mixing weight, and variable support point for the rare species group. The modified assemblage $C^{(m)}$ is obtained by adding to the derived latent distribution Q in Table 1 a support point $\xi_0^{(m)} = m^{-3/2}$ with a mixing weight $\varpi_0^{(m)} = m^{-1}$. The expected observed richness $E\{n_+|C^{(m)}\}$ and the expected counts $E\{n_x|C^{(m)}\}$ for $x = 0, 1, 2, \ldots, 5$ are plotted. The points not on the curves are for the reference assemblage, but each is linked to a curve by dotted lines. The ordinates are logarithmically scaled.

(index m becomes larger and larger), the expected observed counts (and thus the expected observed richness) for the modified assemblages converge on the expected observed counts for the reference assemblage. Meanwhile, the number of undetected species En_0 rises asymptotically to a constant, as reflected in the odds against detection $\alpha(Q^{(m)})$, which approaches an asymptote of 0.53.

This unsettling result is further illustrated in Fig. 5, where we vary both $\xi_0^{(m)}$ (by the arbitrary function $\xi_0^{(m)} = m^{-3/2}$) and mixing weight $\pi_0^{(m)}$ (indirectly by the arbitrary function $\varpi_{(m)0} = 1/m$), so that the true richness of the modified assemblages $S^{(m)}$ is not constant, but increases with m. The same qualitative pattern seen in Table 2 emerges. The expected observed counts and the expected observed richness for the modified assemblages converge on the expected observed counts and the expected observed richness for the reference assemblage. Meanwhile the number of undetected species and the odds against detection continue to rise (instead of approaching an asymptote), because we have allowed the true richness of the modified assemblages to increase with index m.

Table 3 presents α_{ML} in Eq. 10 calculated in the reference assemblage and modified assemblages that have been considered in Table 2 and Fig. 5. While α_{ML} is close to the true odds α in the reference assemblage, there is a substantial difference between α_{ML} and α in the modified assemblages. If a data set is generated from the modified assemblage $C^{(m)}$, using α_{ML} will yield a large bias. We can conclude that all nonparametric estimators can have a large bias in the presence of rare species because of Eqs. 17 and 18. The confidence intervals based on the asymptotic normality of these

TABLE 3. The approximation parameter α_{ML} in Eq. 10 is compared with α in the reference assemblage in Table 1 and modified assemblages in Table 2 (first pair of rows) and Fig. 5 (second pair of rows).

	m							
Parameter	2	10	50	100	1000	2000	5000	Reference
Tables 1 and 2								
α	0.30	0.47	0.52	0.53	0.53	0.53	0.53	0.07
α_{ML}	0.18	0.11	0.08	0.08	0.07	0.07	0.07	0.07
Fig. 5								
α	1.22	3.18	7.13	10.07	31.70	44.79	70.78	0.07
α_{ML}	0.77	0.19	0.09	0.08	0.07	0.07	0.07	0.07

TABLE 4. The two estimates $\hat{Q}$ and $\hat{Q}_{\#}$ with $\rho_k = \xi_k/(1 + \xi_k)$ (seed bank data).

	k					
Parameter	1	2	3	4	5	6
$\hat{Q}$						
ρ_k	0	0.0292	0.0680	0.1786	0.3827	0.4885
ω_k	0.0545	0.3179	0.3940	0.0866	0.1138	0.0332
$\hat{Q}_{\#}$						
ρ_k	⋯	0.0197	0.0636	0.1774	0.3827	0.4885
ω_k	⋯	0.2889	0.4752	0.0889	0.1138	0.0332

nonparametric estimators usually can not achieve their advertised confidence level.

Consider calculating the nonparametric maximum likelihood estimators (NPMLE) for the latent distribution Q in the seed-bank data and the beetles data (Lindsay 1983*a*, *b*). The NPMLE $\hat{Q}$ for Q that maximizes the conditional likelihood $L_2(Q)$ in Eq. 6 in the seed-bank data is presented in Table 4, where we use the detection probability instead of the detection odds, and the group detection probability is denoted by $\rho_k = \xi_k/(1 + \xi_k)$. Note that $\hat{Q}$ has a support point at zero, which means that an assemblage with infinitely many rare species whose relative abundances are extremely small is the most plausible one to generate such data. Another likelihood-based estimate $\hat{Q}_{\#}$ is also presented in Table 4, which is obtained by the EM algorithm starting from the distribution that eliminates the zero support point in $\hat{Q}$. Both estimates fit the data very well. While the log-maximized likelihoods $\log L_2(\hat{Q}_{\#}) = -113.55$ and $\log L_2(\hat{Q}) = -113.24$ are close to one another, the estimates $\alpha(\hat{Q}) = \infty$ and $\alpha(\hat{Q}_{\#}) = 0.03$ are dramatically different: the former suggests there are many undetected species while the latter suggests there are few. The observed seed-bank dataset has relatively few rare species (e.g., only three singletons and two doubletons), suggesting that few species remain undetected, as indicated by $\hat{Q}_{\#}$. But the NPMLE $\hat{Q}$ suggests that there might be many rare species, because the likelihood can be increased slightly by allowing many additional rare species, as in $\hat{Q}$. An estimator for the odds α calculated from an NPMLE might be severely biased.

For some data sets, it is possible to have an NPMLE for Q without a zero or tiny support point so that an estimate for the odds α seems acceptable. However, when one tries to construct confidence intervals for α or S by means of bootstrap, there will usually be some bootstrap resamples that yield an NPMLE with a tiny or zero support point, which makes the upper confidence limit extremely large. For example, the NPMLE $\hat{Q}$ for Q in the beetle data is presented in Table 5, yielding an estimate $\alpha(\hat{Q}) = 3.29$. We take 200 bootstrap resamples from the estimated conditional likelihood $L_2(\hat{Q})$. There are 14 resamples that produce the NPMLE for Q with the smallest support point less than 1×10^{-5}, among which one has a zero support point. The 95% conditional confidence interval for α is $(1.71, 1 \times 10^{7})$. We remark that the numeric results of bootstrap might vary across different runs if one wishes to run the algorithm again because they are random. Clearly the upper confidence limit is useless. Because the NPMLE for Q based on bootstrap resamples are different in magnitude, a smoothed density of the random variable $\log \alpha(\hat{Q})$ rather than $\alpha(\hat{Q})$ constructed from the 199 resample estimates for α is presented in Fig. 6. Note that the density of $\log \alpha(\hat{Q})$ is still highly skewed and has a long right tail, which means that there is always a small but not negligible probability of obtaining a large value of $\log \alpha(\hat{Q})$ or equivalently a large value of $\alpha(\hat{Q})$. Note that because $\hat{S} = n_+(1 + \alpha(\hat{Q}))$ is simply a linear function in $\alpha(\hat{Q})$, the conclusions also apply if we replace $\alpha(\hat{Q})$ with $\hat{S}$.

TABLE 5. The estimate $\hat{Q}$ (beetle data).

	k		
Parameter	1	2	3
ξ_k	0.2252	3.4105	9.7892
ω_k	0.8288	0.1553	0.0155

CONCLUSION

The problem of estimating the species richness S has long been a challenge to both statisticians and ecologists. We have attempted to provide an overview of the

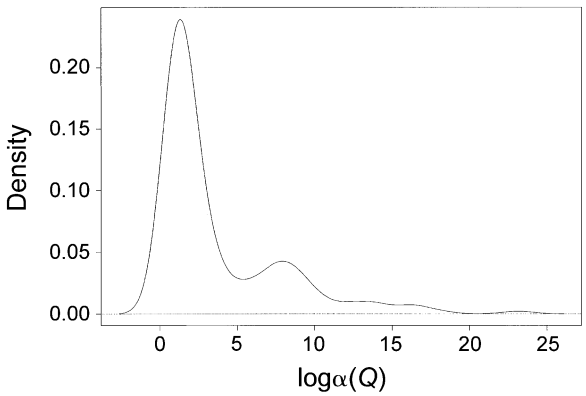

FIG. 6. The estimated density of $\log \alpha(\hat{Q})$ based on the resample.

challenges inherent in estimating richness when many rare species are known or suspected to be present. Our conclusions may appear pessimistic, but we consider it is important to explore the fundamental issues that reflect on claims for estimators and confidence intervals, to caution against overoptimistic conclusions from parametric models, and warn against the blind applications of statistical methods, although such a warning has been issued. For example, statistician I. J. Good, one of the first to tackle this problem (Good 1953), later stated: ''I don't believe it is usually possible to estimate the number of species . . . but only an appropriate lower bound to that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species.'' (Bunge and Fitzpatrick 1993).

Although from typical empirical data, one cannot exclude the existence of many rare species statistically, it is nonetheless possible to infer, with mixture models, how many species should exist at a particular confidence level, that is, to compute with rigor a lower bound for the species richness S, other than the observed richness n_+. Besides mixture models, all existing methods for estimation of S can be treated as improvements upon n_+, the number of observed species, which is a clearly negatively biased estimator for true species richness S.

Using mixture models, it is also possible to estimate rigorously, with confidence intervals, the expected increment in richness that would result from increasing observed samples to two or three times their actual size, without attempting to find a true asymptote (Colwell et al. 2004, Mao et al., *in press*; see also Shen et al. 2003, for a different approach; in fact, it is not at all clear that a true asymptote even exists, for some taxa in some habitats, as discussed in the *Introduction*.) Such an extrapolation, for many purposes in ecology and conservation biology, will often represent a most useful and welcome savings in time and resources. In short, we do not counsel despair, but rather realistic expectations and a cautious approach to inference.

Special Feature

Acknowledgments

We are grateful to A. Ellison for the invitation to prepare this paper and the constructive comments from the two referees. This work was supported by US-NSF grant DEB-0072702 to R. K. Colwell.

Literature Cited

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

Burnham, K. P. 1972. Estimation of population size in multiple capture–recapture studies when capture probabilities vary among animals. Dissertation. Oregon State University, Corvallis, Oregon, USA.

Burnham, K. P., and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika **65**:625–634.

Butler, B. J., and R. L. Chazdon. 1998. Species richness, spatial variation, and abundance of the soil seed bank of a secondary tropical rain forest. Biotropica **30**:214–222.

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

Chao, A. 1989. Estimating population size for sparse data in capture–recapture experiments. Biometrics **45**:427–438.

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

Chao, A., and T. J. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics **10**: 429–443.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B **345**:101–118.

Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology **85**:2717–2727.

Darroch, J. N., and D. Ratcliff. 1980. A note on capture-recapture estimation. Biometrics **36**:149–153.

Dorazio, R. M., and J. A. Royle. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics **59**:351–364.

Efron, B., and R. Thisted. 1976. Estimating the number of unseen species: how many words did Shakespeare know? Biometrika **63**:435–447.

Falkowski, P. G., and C. de Vargas. 2004. Shotgun sequencing in the sea: a blast from the past? Science **304**(5667): 58–60.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 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.

Gaston, K. E. 1994. Rarity. Chapman and Hall, London, UK.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika **40**: 237–264.

Gotelli, N., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters **4**:379–391.

Hughes, J. B., J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan. 2000. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and Environmental Microbiology **67**:4399–4406.

Janzen, D. H. 1973. Sweep samples of tropic foliage insects: description of study sites, with data on species abundances and size distributions. Ecology **54**:659–686.

Lee, S. M., and A. Chao. 1994. Estimating population size via sample coverage for closed capture–recapture models. Biometrics **50**:88–97.

Lindsay, B. G. 1983*a*. The geometry of mixture likelihoods: a general theory. The Annals of Statistics **11**:86–94.

Lindsay, B. G. 1983*b*. The geometry of mixture likelihoods. Part II: the exponential family. The Annals of Statistics **11**: 783–792.

Longino, J., R. K. Colwell, and J. A. Coddington. 2002. The ant fauna of a tropical rainforest: estimating species richness three different ways. Ecology **83**:689–702.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, London, UK.

Magurran, A. E., and P. A. Henderson. 2003. Explaining the excess of rare species in natural species abundance distribution. Nature **422**:714–718.

Mao, C. X. 2004*a*. Estimating the unknown sample size from truncated binomial or geometric mixtures. Technical report. Department of Statistics, University of California, Riverside, California, USA.

Mao, C. X. 2004*b*. Predicting the conditional probability of discovering a new class. Journal of American Statistical Association **99**:1108–1118.

Mao, C. X., R. K. Colwell, and J. Chang. *In press*. Estimating the species accumulation curve using mixtures. Biometrics.
Mao, C. X., and B. G. Lindsay. 2003. Estimating the number of classes using Poisson mixtures. Technical report. Department of Statistics, University of California, Riverside, California, USA.
Mingoti, S. A., and G. Meeden. 1992. Estimating the total number of distinct species using presence and absence data. Biometrics **48**:863–875.
Norris, J. L. I., and K. H. Pollock. 1996. Nonparametric MLE under two closed capture–recapture models with heterogeneity. Biometrics **52**:639–649.
Norris, J. L. I., and K. H. Pollock. 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environmental and Ecological Statistics **5**:391–402.
Novotny, V., and Y. Basset. 2000. Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. Oikos **89**:564–572.
Ord, J. K., and G. Whitmore. 1986. The Poisson-inverse Gaussian distribution as a model for species abundance. Communications in Statistics, Theory and Methods **15**: 853–871.
Pledger, S. 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. Biometrics **56**:434–442.
Preston, F. W. 1948. The commonness, and rarity, of species. Ecology **29**:254–283.
Rabinowitz, D. 1981. Seven forms of rarity. Pages 205–217 *in* H. Singe, editor. The biological aspects of rare plant conservation. Wiley, Chichester, UK.
Shen, T. J., A. Chao, and C. F. Lin. 2003. Predicting the number of new species in future taxonomic sampling. Ecology **84**:798–804.
Sichel, H. S. 1997. Modeling species-abundance frequencies and species individual functions with the generalized inverse Gaussian-Poisson distribution. South African Statistical Journal **31**:13–37.
Smith, E. P., and G. van Belle. 1984. Nonparametric estimation of species richness. Biometrics **40**:119–129.
Whittaker, R. H. 1972. Evolution and measurement of species diversity. Taxon **21**:213–251.
Zelterman, D. 1988. Robust estimation in truncated discrete distributions with application to capture–recapture experiments. Journal of Statistical Planning and Inference **18**: 225–237.