Contents

2

1 Estimation of Species Richness and Shared Species Richness

Anne Chao and Chun-Huo Chiu

1.1 Introduction

Species richness (i.e., the number of species) is the simplest and the most intuitive measure for characterizing the diversity of an assemblage in biological sciences. In biogeographic studies, species range maps and local and regional floras and faunas generally provide only species presence-absence information for each locality. For these studies, species richness thus becomes the only measure that can be used to quantify diversity. Even when species abundances are available, in conservation analyses the actual species count in an area is often the most relevant diversity measure. Species richness is a special case of a continuum of diversity measures that vary in their sensitivity to species abundances; it is the diversity of order 0 [53], completely insensitive to abundances.

However, the compilation of complete species inventories often requires extraordinary efforts and is an almost unattainable goal in practical applications. There are undetected species in almost every taxonomic survey or species inventory. The observed number of species in samples inevitably underestimates the true species richness (observed plus undetected). Both theoretical and practical considerations show that if there are many rare species in a hyper-diverse assemblage, then it becomes statistically difficult to obtain an accurate point estimate of species richness from small samples. The estimation of species richness from samples has been extensively discussed in the literature. The topic is important for comparing assemblages in conservation and management of biodiversity, for assessing the effects of human disturbance on biodiversity, and for making environmental policy decisions. Various approaches have been proposed and compared. See [8][15][36][50][52][67] and [68] for reviews of species richness estimation. See also [2] and [96] for various sampling aspects and relevant methodologies.

In addition to estimating the species richness of assemblages of plants or animals, the techniques discussed here have a wide range of applications in various other disciplines, as will be outlined in Section 10.

Under any sampling scheme, the number of observed species in samples usually increases with the number of individuals encountered, the number of samples collected, or the area sampled. To control

for this dependence when comparing the sample richnesses of different assemblages, ecologists often standardize sampling efforts [51][52], or standardize sample completeness [1][63]. By using individual-based or sample-based data, rarefaction is a traditional method to down-sample the larger samples until they are the same size (the same number of individuals or the same number of sampling units) as the smallest sample [58][85], and then the richnesses of these samples with equal sampling efforts are compared. A main disadvantage of rarefaction is the loss of information involved in when reducing the sizes of some of the samples. Extrapolation solves this problem. Based on an estimate of asymptotic species richness, extrapolation predicts richnesses of hypothetical samples larger than the actual samples, instead of smaller than the actual samples as in rarefaction [27][35][49][89] linked rarefaction and extrapolation and showed that extrapolation can be reliable to about twice the actual sample sizes. This allows for rigorous statistical comparison not only for rarefied but also for extrapolated species richness values.

Compared with estimating species richness in a single assemblage, the estimation of species richness shared by multiple assemblages has received relatively little attention. The number of shared species among assemblages can be used to describe assemblage overlap and forms a basis to construct various types of beta diversity and similarity/dissimilarity measures [36][62][64][67]. These measures are often used to quantify spatial or temporal variation in species composition, to set conservation priorities or to evaluate regional conservation plans. The application of these measures will shed light on the ecological factors that cause differentiation between assemblages, and on the factors which control stochastic genetic differentiation between incompletely-isolated subpopulations of a species. These topics are among the most fundamental questions in community ecology and in evolutionary theory.

1.2 Traditional Curve-Fitting Approaches

Traditional curve-fitting approaches to estimating species richness include the following:

(1) Using parametric curves to extrapolate a species-accumulation or species-area curve to predict its asymptote, which is used as an estimate of species richness. This approach has a long history and various curves have been presented [45]; a summary with discussion is provided in [15]. Among the proposed asymptotic functions are the negative exponential function, Weibull model, logistic equation, and the Michaelis-Menten equation. This approach does not directly use information on the frequencies of common and rare species, but simply forecasts the shape of the rising curve.

(2) Fitting a truncated parametric distribution or functional form to the observed species abundances to obtain an estimate of species richness. The earliest approach was proposed by Preston [82], who fitted a truncated log-normal curve to the (properly grouped) frequencies and used the integrated value of the fitted curve over the real line as an estimate of the total number of species. Other truncated distributions (e.g., negative binomial, geometric, Zipf-Mandelbrot, logarithmic; see [67]) can also be applied. Although this approach uses information on the frequencies of common and rare species, it simply fits a curve to the observed frequency data.

The curve-fitting approaches generally do not provide the variances of the resulting estimates without further assump-

tions on sampling theory. The advantages and disadvantages of the curve-fitting approaches are shared by other sampling-theory-based parametric approaches; see later discussion.

1.3 Sampling-Theory-Based Approaches

The work by Fisher, Corbet, and Williams [44] provided the mathematical foundation for statistical sampling-theory-based approaches to estimate species richness. Since then, a large body of literature discussing models and estimation under various sampling plans has been published and applied to many research fields. Below, we review this sampling-theory approach separately for two type of sampling data: individual-based (abundance) data and sample-based (incidence) data. For each type of data, different sampling models and estimators of species richness are reviewed. Because sampling variation is an inevitable component of biological surveys, it is equally important to assess the variance (or standard error) of an estimator and provide a confidence interval that will reflect sampling uncertainty.

Two Types of Data Our notation and terminology generally follow [35]. Consider an assemblage consisting of N total individuals, each belonging to one of S distinct species. Let N_i (true species absolute abundance) be the number of individuals of the ith species in the assemblage, $i = 1, 2, \ldots, S$, $N_i > 0$, and $N = \sum_{i=1}^{S} N_i$. The relative abundance p_i of species i is N_i/N, so that $\sum_{i=1}^{S} p_i = 1$. Here N, S, N_i, and p_i represent the true underlying assemblage size, species richness of the assemblage, the abundance and the relative abundance of the ith species. These parameters are unknown but can be estimated based on a random sample from the assemblage. We distinguish between two sampling data structures.

(1) Individual-based (Abundance) Data In many biological studies (e.g., bird, insect, mammal and plant), it is often the case that one individual is observed or encountered at a time and classified as to species identity. Assume that a random sample of n individuals is taken from the assemblage and a total of S_{obs} species are observed. This basic sample is referred to as a reference sample. This type of data can be obtained by two different sampling schemes. (a) Discrete-type sampling in which sampling unit is an individual. For example, we sample 100 individuals in a study area. Here sample size n is fixed by design and each species can be represented by at most n individuals. (b) Continuous-type sampling in which sampling efforts are measured in a continuous scale such as time, area or water volume. For example, we sample 30 ha or 100 hours in a study area. Here the number of observed individuals in this sampling protocol is a random variable and each species can be represented by many individuals without a limit.

Let X_i (sample species frequency) be the number of individuals of the ith species which are observed in the sample, $i = 1, 2, \ldots, S$. Only those species with $X_i > 0$ are observable in the sample, and $\sum_{i=1}^{S} X_i = n$ (only species with $X_i > 0$ contribute to the sum). Let f_k (abundance frequency counts), $k = 0, 1, \ldots, n$, be the number of species represented by exactly k individuals in the reference sample. Thus, we have $n = \sum_{i=1}^{S} X_i = \sum_{k \geq 1} k f_k$, and $S_{obs} = \sum_{k \geq 1} f_k$. In particular, f_1 is the number of species represented by exactly one individual ("singletons") in the reference sample, and f_2 is the number of species represented by exactly two individuals ("doubletons"). Also, f_0 denotes the number of undetected species in the ref-

erence sample. Here "undetected species" means species that are present in the assemblage of N individuals and S species, but were not detected in the reference sample of n individuals and S_{obs} species. Because $S = S_{obs} + f_0$, species richness estimation is equivalent to the inference about the number of undetected species f_0.

(2) Sample-based (Incidence) Data. In many biodiversity studies, the sampling unit is not an individual, but a trap, net, quadrat, plot, or timed survey. It is these sampling units, and not the individual organisms, that are actually sampled randomly and independently. Quadrat sampling provides an example in which the study area is divided into a number of quadrats with approximately the same area, and a sample of quadrats is randomly selected for survey. There are other examples: similar sampling is conducted by several investigators, or trapping records are collected over multiple occasions. Counting the exact number of individuals for each species appearing within each sampling unit may often become impossible for micro-organisms, invertebrates or plants. In most cases, only their incidence (presence or absence) can be recorded. Estimation is based on a set of sampling units in which the incidence of each species is recorded in each sampling unit instead of its abundance. We use the general term "sampling unit" in what follows to refer to a quadrat, occasion, site, transect line, team, occasion, a period of fixed time, a fixed number of traps, or an investigator etc. There is a natural time ordering among temporally replicated samples, but such an ordering does not exist in most of the other types of sampling schemes.

The reference sample for incidence data consists of a set of T sampling units. The presence or absence of each species within each sampling unit is recorded, to form a species-by-sampling-unit incidence matrix (W_{ij}) with S rows and T columns. The value of the element W_{ij} of this matrix is unity if species i is present in the jth community, and zero if it is absent. The row sum of the incidence matrix, $Y_i = \sum_{i=1}^{T} W_{ij}$ denotes the incidence-based frequency of species i, for $i = 1$ to S. Here, Y_i is analogous to X_i in the individual-based abundance vector. Species present in the assemblage but not detected in any sampling unit have $Y_i = 0$. The total number of species observed in the reference sample is S_{obs} (only species with $Y_i > 0$ contribute to S_{obs}).

For most applications, the sufficient statistics from the species-by-sampling-unit incidence matrix are the incidence frequency counts $(Q_1, Q_2, \ldots, Q_T)$ where Q_k denotes the number of species that are detected in exactly k sampling units, $k = 0, 1, \ldots, T$. That is, Q_k is the number of species each represented exactly $Y_i = k$ times in the incidence matrix sample. For the incidence matrix, $\sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$, and $S_{obs} = \sum_{k=1}^{T} Q_k$. Thus, based on the terminology used in Colwell and Coddington (1994), Q_1 represents the number of "unique" species (those that are each detected in only one sampling unit) and Q_2 represents the number of "duplicate" species (those that are each detected in exactly two sampling units). The zero frequency count Q_0 denotes the number of species among the S species in the assemblage that are not detected in any of the T sampling units. Since $S = S_{obs} + Q_0$, species richness estimation is equivalent to the inference about the number of undetected species Q_0.

1.4 Species Richness Estimation (Abundance Data)

Suppose n individuals (with n fixed in advance) are independently observed from the study site. The commonly used model

is the multinomial model. In this discrete-type sampling, it is assumed that the sampling procedure itself does not substantially alter the species relative abundances $(p_1, p_2, \ldots, p_S)$. This assumption is fulfilled if individuals are sampled with replacement so that any individual can be repeatedly observed. If sampling is done without replacement, so that any individual can only be observed at most once in the sample, then a hypergeometric model is more appropriate as will be discussed below. In practice the two probability models differ little when the biological populations being sampled are sufficiently large and sample size is small relative to population size.

Generally, species detection probability or rate in any observation is a combination of species abundance and individual detectability, which is determined by many possible factors (such as individual movement patterns, color, size, and vocalizations). If we assume that all individuals have the same detectability, then the observed species frequencies $(X_1, X_2, \ldots, X_S)$ for given S and $(p_1, p_2, \ldots, p_S)$ follow a multinomial distribution (The undetected species, i.e., $X_i = 0$, do not contribute to this likelihood.)

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_S = x_S) = \frac{n!}{x_1! \ldots x_S!} p_1^{x_1} p_2^{x_2} \ldots p_S^{x_S}. \quad (1)$$

In this special case, the species detection rate for the ith species is simply the true relative abundance $p_i = N_i/N$. Individuals classified to the same species are often indistinguishable and any individual may be observed repeatedly, but the number of individuals represented by any species is at most n which is fixed by design. A more general model assumes that the detectability of any individual of the ith species is $\theta_i > 0$, which varies with species. Under this general model, the species detection rate for the ith species in any individual becomes $\psi_i = p_i\theta_i / \sum_{k=1}^{S} p_k\theta_k$, $i = 1, 2, \ldots, S$. Thus, a more general setting is the following model that allows heterogeneous individual detectability:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_S = x_S) = \frac{n!}{x_1! \ldots x_S!} \psi_1^{x_1} \psi_2^{x_2} \ldots \psi_S^{x_S}. \quad (2)$$

Here, the sample frequency X_i of species i is a binomial distribution with the detection probability ψ_i being a normalized product of species relative abundance p_i and individual detectability θ_i.

Assume that the assemblage is surveyed by a continuous-type sampling efforts and that the total amount efforts is increased from 0 to A units. Since the number of observed individuals of any species has no upper limit, a common approach is based on the Poisson model which can take value from 0 to infinity. This approach can be traced back to Fisher et al. [44], who assumed that individuals of the ith species arrive a sample according to a Poisson process with a mean species occurrence or detection rate $A\lambda_i$, here λ_i represents the mean rate per unit of effort. In some applications, the exact arrival times for each individual are available, but in most biological samplings, only the frequencies of discovered species are recorded, and these frequencies would be sufficient for estimating species richness [74]. In this sampling scheme, the sample size n (the number of individuals observed in the experiment) is a random variable and n can be any positive integer. The probability distribution for the observed frequencies is a product-Poisson distribution:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_S = x_S) = \prod_{i=1}^{S} (A\lambda_i)^{x_i} \frac{exp(-A\lambda_i)}{x_i!}. \quad (3)$$

(The undetected species, i.e., $X_i = 0$, do not contribute to this likelihood.)

Although n is a random variable, we can consider the conditional distribution of the

frequencies $(X_i, X_2, \ldots, X_S)$ given n (here $n = \sum_{k=1}^{S} X_k$). The conditional distribution is a multinomial distribution with cell total n and cell probabilities $\lambda_i / \sum_{k=1}^{S} \lambda_k$, $i = 1, 2, \ldots, S$. In other words, inference under a multinomial model can be regarded as a conditional procedure under a product-Poisson model. If we assume that the Poisson rate λ_i is proportional to the product of species relative abundance p_i and individual detectability θ_i, then this conditional multinomial distribution is identical to the model in Equation (2). This is also the reason that many estimators are shared by both the product-Poisson model under continuous-effort sampling schemes and a multinomial model under discrete-effort sampling schemes. Coleman's [33] area-based model is basically a special case of the product-Poisson model. Coleman considered that the reference sample is obtained by a survey in a specified site of area A. Within this site, the ith species occurs at a species-specific mean rate $A\lambda_i$ and the probability distribution is same as that in Equation (3).

Parametric or Likelihood-based Models Fisher et al. [44] adopted a parametric approach in their pioneering work on species richness estimation. In this approach, one assumes a parametric distribution $f(\lambda; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of parameters, for the species detection rates $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ in the product-Poisson model (3) or for the relative abundances $(p_1, p_2, \ldots, p_S)$ in the multinomial model (1). Most parametric approaches are based on the product-Poisson model (3) under a continuous sampling framework. When S is large and unknown, it is statistically difficult to deal with inference problems with so many parameters such as $(\lambda_1, \lambda_2, \ldots, \lambda_S)$. Assuming a parametric distribution for $(\lambda_1, \lambda_2, \ldots, \lambda_S)$, we see that the whole inference problem is reduced to the estimation of S and θ, so that usual inference procedures can be applied. This is a main advantage of parametric approaches.

If the distribution $f(\lambda; \boldsymbol{\theta})$ is a degenerate distribution with all probabilities at a fixed point λ, then this reduces to a homogeneous model (with equal detection rates for all species) with $\lambda_1 = \lambda_2 = \ldots = \lambda_S = \lambda$. Although this homogeneous model is rarely valid in practice, it provides a starting framework for species richness estimation and has been discussed extensively in the literature [15]. An approximate maximum likelihood estimator (MLE) under the homogeneous model is the solution $\hat{S}$ of the following equation:

$$S_{obs} = \hat{S}[1 - exp(-n/\hat{S})], \qquad (4a)$$

with an asymptotic variance estimator for the solution $\hat{S}$

$$\hat{var}(\hat{S}) \approx \hat{S}/[exp(n/\hat{S}) - (n/\hat{S}) - 1].$$

A highly efficient estimator which can be regarded as a coverage-based estimator for the special case of a homogeneous model is [39]

$$\hat{S}_0 = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}}, \qquad (4b)$$

where $S_{abun} = \sum_{i>\kappa} f_i$, and $S_{rare} = \sum_{i=1}^{\kappa} f_i$. Here $\hat{C}_{rare} = 1 - f_i / \sum_{i=1}^{\kappa} i f_i$ is an estimate of "sample coverage" for rare species group (see "Non-parametric Approaches" for details). The value κ is a cut-off point which separates the observed frequencies into "abundant" and "rare" species groups. The choice of κ will be discussed below. Confidence interval of species richness based on the MLE or $\hat{S}_0$ can be constructed by using an asymptotic variance and a log-transformation so that the lower bound of the interval is not less than S_{obs} [13][29].

To formulate the parametric theory under a general distribution $f(\lambda; \boldsymbol{\theta})$, we first construct the likelihood function of S and

$\boldsymbol{\theta}$ based on both observed and undetected species. For any mixing density $f(\lambda;\boldsymbol{\theta})$, define $p_{\boldsymbol{\theta}}(k)$, $k = 0, 1, \ldots$ as the probability that any species is observed k times in the sample, then from Equation (3) we have

$$p_{\boldsymbol{\theta}}(k) = \int_0^\infty (A\lambda)^k \frac{exp(-A\lambda)}{k!} f(\lambda;\boldsymbol{\theta}) d\lambda, \tag{5a}$$

and $E(f_k) = S p_{\boldsymbol{\theta}}(k)$. Consider that each species can be classified into any of the following disjoint categories: undetected, detected once, detected twice,. . .etc. Then the likelihood function for S and $\boldsymbol{\theta}$ from all species can be written as

$$L(S,\boldsymbol{\theta}) = \frac{S!}{(S-S_{obs})!\prod_{k\geq 1} f_k!} \times [p_{\boldsymbol{\theta}}(0)]^{S-S_{obs}} \prod_{k\geq 1} [p_{\boldsymbol{\theta}}(k)]^{f_k}. \tag{5b}$$

The species richness estimation thus reduces to an inference with parameters S and θ, and traditional estimation procedures can be applied. For example, the unconditional maximum likelihood estimator (UMLE) and its asymptotic variance are obtained based on the above full likelihood (5b). A conditional (on S_{obs}) maximum likelihood estimator (CMLE) is often more convenient to obtain as follows. Note that likelihood (5b) can be factored as $L(S,\boldsymbol{\theta}) = L_b(S,\boldsymbol{\theta}) L_c(\boldsymbol{\theta})$,

$$L_b(S,\boldsymbol{\theta}) = \frac{S!}{(S-S_{obs})!S_{obs}!}[1-p_{\boldsymbol{\theta}}(0)]^{S_{obs}} \times [p_{\boldsymbol{\theta}}(k)]^{N-S_{obs}}, \tag{5c}$$

and

$$L_c(\boldsymbol{\theta}) = \frac{S_{obs}!}{\prod_{k\geq 1} f_k!} \prod_{k\geq 1} \left[\frac{p_{\boldsymbol{\theta}}(k)}{1-p_{\boldsymbol{\theta}}(0)}\right]^{f_k}. \tag{5d}$$

where $L_b(S,\boldsymbol{\theta})$ is a likelihood with respect to S_{obs}, a binomial (S, $1 - p_{\boldsymbol{\theta}}(0)$), and $L_c(\boldsymbol{\theta})$ is a multinomial likelihood with respect to $\{f_k; k \geq 1\}$ with cell total S_{obs} and zero-truncated cell probabilities $p_{\boldsymbol{\theta}}(k)/[1 - p_{\boldsymbol{\theta}}(0)]$, $k \geq 1$ The first likelihood $L_b(S;\boldsymbol{\theta})$ results in the CMLE $\hat{S}_{CMLE} = S_{obs}/[1 - p_{\boldsymbol{\theta}}(0)]$, where $\hat{\boldsymbol{\theta}}$ maximizes the second likelihood $L_c(\boldsymbol{\theta})$ [84]. Both types of MLE's can also be regarded as empirical Bayes estimators if we think of the mixing distribution as a prior having unknown parameters that must be estimated.

Fisher et al. [44] adopted a two-parameter gamma distribution with $\boldsymbol{\theta} = (\tau, \beta)$ and density $f(\lambda;\tau,\beta) = \beta^{-\tau}\lambda^{\tau-1}exp(-\lambda/\beta)/\Gamma(\tau)$, i.e., the gamma-Poisson or gamma-mixed Poisson model. Since the squared coefficient of variation of this gamma distribution is $1/\tau$, the parameter τ measures inversely the degree of heterogeneity among species detection rates. The $p_{\boldsymbol{\theta}}(k)$, or equivalently $E(f_k)$, $k = 0, 1, 2,\ldots$ correspond to individual terms of a negative-binomial distribution.

$$p_{\boldsymbol{\theta}}(k) = \frac{\Gamma(k+\tau)}{\Gamma(k+1)\Gamma(\tau)}\left(\frac{\beta}{1+\beta}\right)^k \left(\frac{1}{1+\beta}\right)^\tau, \quad k = 0, 1, 2, \ldots$$

In the special case of $\tau = 1$ (i.e., $f(\lambda;\boldsymbol{\theta})$ is an exponential distribution), the model is equivalent to a broken-stick model ([79], p. 285). In this case, the $p_{\boldsymbol{\theta}}(k)$, $k = 0, 1, 2, \ldots$, correspond to the terms of a geometric distribution. Define $x = \beta/(1 + \beta)$ and $\alpha = S/[-log(1 - x)]$ (Fisher's alpha). Given $k > 0$, as τ tends to 0 (i.e., the degree of heterogeneity among species detection rates tends to infinity), we have $p_{\boldsymbol{\theta}}(k) \to x^k/\{k[-log(1 - x)]\}$, or equivalently, $E(f_k) \to \alpha x^k/k$, the well-known Fisher's log-series model. But this model does not yield an estimate of species richness ([79], p. 274). Fisher's alpha has been used as an informative diversity measure. Based on sample data, Fisher's α and the parameter x in the log-series model can be solved from the two equations: $S_{obs} = -\alpha \log(1-x)$ and $n = \alpha x/(1-x)$. Thus two data sets with the same sample size and the

observed numbers of species would result in the same Fisher's α estimate. That is, Fisher's α completely ignores the species sample frequencies.

Other parametric models include the log-normal [7], inverse-Gaussian [77], and generalized inverse-Gaussian [91]. The chief weakness of these methods is that simulations show that they work well only when the correct form of the species detection rates is already known (e.g., [76]), but this is never the case for empirical data. Another weakness is that extensive iterative procedures are required to find the UMLE and CMLE, and in some cases the iterative steps fail to converge to a value and thus the UMLE or CMLE may not be obtainable. Chao and Bunge [16] proposed an explicit estimator under the gamma-Poisson model:

$$\hat{S} = S_{abun} + \sum_{i=2}^{\kappa} f_i / \left[1 - \frac{f_1 \sum_{i=1}^{\kappa} i^2 f_i}{(\sum_{i=1}^{\kappa} i f_i)^2}\right], \quad (6)$$

where S_{abun} and the cut-off point κ are defined as in (4b). A variance estimator is obtained by a standard asymptotic method; see [16].

By assigning various priors for parameters (S, α, β) in a gamma-Poisson model, a fully Bayesian hierarchical approach was proposed in [83]. Complicated calculations are handled by computer-intensive algorithms through the use of Gibbs sampling, a Markov Chain Monte Carlo method. The reader is referred to the above reference and [3] for previous work in the Bayesian direction.

One can also assume a parametric distribution for the relative abundances $(p_1, p_2, \ldots, p_S)$ in the multinomial model (1). For example, to characterize the theoretical patterns, a functional form is selected to model the relative abundances $(p_1, p_2, \ldots, p_S)$. The most popular functional forms include the geometric $p_i \propto \alpha(1-\alpha)^{i-1}$ and the Zipf-Mandelbrot law $p_i \propto (1+\alpha)^{-\theta}$, where α and θ are parameters. Although these types of models can produce species richness estimates [8], they are mainly useful for describing the features of abundant species especially for applications in linguistics. Moreover, simulation studies have shown that the estimators derived from these models do not perform satisfactorily [9]. A random-parameter model assuming that $(p_1, p_2, \ldots, p_S)$ follows a Dirichlet distribution leads to a broken stick model as described earlier.

A difficulty shared by the curve-fitting and parametric approaches lies in the selection of a parametric function or distribution. Two models with different parametric functions or distributions may fit the data equally well, but they yield widely different estimates. Also, a parametric model which gives a good fit to the data does not necessarily result in a satisfactory species richness estimate. These approaches do not work well in comparisons with empirical or simulated data sets; see [52] and [50]. Another drawback is that the parametric approaches are difficult to generalize to deal with shared species richness in multiple assemblages.

Non-parameteric Approaches The above concerns have led to non-parametric approaches, which avoid making assumptions about species detection rates or species abundance distributions. We mainly focus on the Chao1 and ACE-type estimators and briefly review the jackknife and non-parametric maximum likelihood estimator (NPMLE) approaches (see [15] for other methods):

Estimator by Chao [12] If there are many undetectable or "invisible" species in a hyper-diverse assemblage, then it will be impossible to obtain a good estimate of species richness. Therefore, a reliable lower bound for species richness is often of

more practical use than an imprecise point estimate. Based on the concept that rare species carry the most information about the number of undetected species [47], the Chao1 estimator uses only the numbers of singletons and doubletons (and the observed richness) to estimate the number of undetected species and obtain a lower bound for species richness [12].

The Chao1 lower bound was originally derived under the multinomial model given in Equation (1) in which each species' detection probability is just its relative abundance, and all individuals of all species have the same detectabilities. It is actually also valid under the more general model given in Equation (2) in which individuals' detectabilities can vary from species to species. We review how the Chao1 estimator is derived from this latter model, under which the distribution of sample frequency X_i is a binomial distribution with a species-specific probability $\psi_i = p_i\theta_i / \sum_{k=1}^{S} p_k\theta_k$. Then for $k = 0, 1, \ldots$

$$P(X_i = k) = \binom{n}{k}\psi_i^k(1-\psi_i)^{n-k}. \quad (7a)$$

This gives

$$\begin{aligned} E(f_k) &= \sum_{i=1}^{S} P(X_i = k) \\ &= \sum_{i=1}^{S} \binom{n}{k}\psi_i^k(1-\psi_i)^{n-k}. \quad (7b) \end{aligned}$$

In particular, the expected number of undetected species, singletons and doubletons are respectively:

$$E(f_0) = \sum_{i=1}^{S}(1-\psi_i)^n, \quad (8a)$$

$$E(f_1) = \sum_{i=1}^{S} n\psi_i(1-\psi_i)^{n-1}, \quad (8b)$$

$$E(f_2) = \sum_{i=1}^{S} \binom{n}{2}\psi_i^2(1-\psi_i)^{n-2}. \quad (8c)$$

The Cauchy-Schwarz inequality leads to

$$\begin{aligned} &\left[\sum_{i=1}^{S}(1-\psi_i)^n\right]\left[\sum_{i=1}^{S}\psi_i^2(1-\psi_i)^{n-2}\right] \\ &\geq \left[\sum_{i=1}^{S}\psi_i(1-\psi_i)^{n-1}\right]^2. \quad (9) \end{aligned}$$

Combining (8a), (8b), (8c) and (9), we then obtain a theoretical lower bound for $E[f_0]$:

$$E(f_0) \geq \frac{(n-1)}{n}\frac{[E(f_1)]^2}{2E(f_2)}.$$

The term $(n-1)/n$ in the above can be dropped if n is large enough. Replacing the expected values by the observed data, we have the following estimator which is referred to as the Chao1 estimator in ecological literature (e.g., [36]):

$$\begin{aligned} \hat{S}_{Chao1} &= S_{obs} + f_1^2/(2f_2), \quad \text{if } f_2 > 0 \\ &= S_{obs} + f_1(f_1-1)/2, \quad \text{if } f_2 = 0 \end{aligned} \quad (10)$$

with an associated variance estimator (if $f_2 > 0$) [13]:

$$\begin{aligned} &\hat{var}(\hat{S}_{Chao1}) \\ &= f_2\left[\frac{1}{2}(\frac{f_1}{f_2})^2 + (\frac{f_1}{f_2})^3 + \frac{1}{4}(\frac{f_1}{f_2})^2\right]. \quad (11a) \end{aligned}$$

If $f_2 = 0$, the variance formula (11a) becomes:

$$\begin{aligned} \hat{var}(\hat{S}_{Chao1}) &= \frac{f_1(f_1-1)}{2} \\ &+ \frac{f_1(2f_1-1)^2}{4} - \frac{f_1^4}{4\hat{S}_{Chao1}}. \quad (11b) \end{aligned}$$

Under many classes of species detection rates, the lower bound is very sharp if the reference sample size is large enough[17]. This justifies the use of the Chao1 estimator as a valid point estimator of species richness for large n. A confidence interval, which indicates the possible range of the

true species richness, is constructed by using the variance estimator (11a) and (11b) and a log-transformation [13].

The Chao1 estimator may be useful for data sets in which it is too time-consuming to count the frequencies of all abundance classes, but it is relatively easy to count just the number of singleton and doubleton species. The Chao1 estimator is intuitive and very easy to calculate, and often perform just as well as more complex asymptotic estimators. The Chao1 estimator is featured in several computer software packages (see Section 11).

Recently, the Chao1 estimator is extended to deal with data for sampling without replacement [24], in which sampling units cannot be repeatedly observed. For example, the model in Equation (1) becomes a generalized hypergeometric distribution with a known population size N,

$$P(X_i = x_i, i = 1, 2, \ldots, S) = \binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_s}{x_S}\Big/\binom{N}{n}.$$

The Chao1 estimator under this model is generalized to

$$\hat{S}_{Chao1.wor} = S_{obs} + \frac{f_1^2}{\frac{n}{n-1}2f_2 + \frac{q}{1-q}f_1}, \tag{12}$$

where the subscript "*wor*" refers to "without replacement", and $q = n/N$ denotes the known sampling ratio (the ratio of sample size to the population size). A similar generalization can be made for the model (2) and the resulting estimator is identical to the estimator (12). When only a small portion of individuals are taken from the entire universe of N individuals in the assemblage, so that the sample fraction q approaches zero, the lower bound approaches the Chao1 estimator. On the other hand, when q approaches 1 so that all individuals are observed, $q/(1-q)$ approaches infinity and our lower bound reduces to the number of observed species, which equals the true parameter when all individuals have been observed.

The ACE [23][26][21] The concept of "sample coverage" was originally developed for cryptographic analyses during World War II by the founder of modern computer science, Alan Turing, and by his colleague I. J. Good [47][48]. Under the multinomial model given in Equation (1), the "coverage" of a sample is mathematically defined as $C = \sum_{i=1}^{S} p_i I(X_i > 0)$, where $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. That is, the sample coverage represents the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample. It is a reliable measure of the degree of sample completeness. Subtracting the sample coverage from unity gives the probability that a new, previously-unsampled species would be found if the sample were enlarged by one individual. This concept has played an essential role in species richness estimation [15][23].

Contrary to most people's intuition, Good and Turing discovered that sample coverage can be very accurately and efficiently estimated using only information contained in the sample itself, as long as the sample size is reasonably large. A robust estimate of the coverage of a sample of size n is simply $1 - f_1/n$, where f_1 is the number of singletons as defined earlier [47][48]. Based on the concept of sample coverage, the Abundance-based Coverage Estimator (ACE) was developed by Chao and Lee [23] and Chao et al. [21] under the model (1) and a similar ACE was derived by Chao et al. [26] under the product-Poisson model (3). The two models yield almost identical ACE estimates due to the close relationship between the product-Poisson model and the multino-

mial model.

Here we present the ACE under the multinomial model (1). It is assumed in this approach that the species detection probabilities are fully characterized by their mean $\overline{p} = 1/S$ and CV (coefficient of variation). The squared CV, γ^2, is defined as $\gamma^2 = [S^{-1}\sum_{i=1}^{S}(p_i - \overline{p})^2]/\overline{p}^2$. The CV parameter is used to characterize the degree of heterogeneity. The larger the CV, the greater the degree of heterogeneity among species detection probabilities. The CV vanishes if and only if the community is homogeneous.

To apply the concept of sample coverage to species richness estimation, we need to specify a cut-off value κ which separates frequencies into "rare" and "abundant" groups. The reasons for such a separation are mainly because (i) only rare species carry information about undetected species; and (ii) the parameter CV is very difficult to estimate for a highly heterogeneous assemblage, restricting inference to the "rare" species gives a more stable estimate of CV. The cut-off $\kappa = 10$ works well with many empirical data sets. However, for highly heterogeneous assemblages so that the data exhibit a long tail of frequencies, the cut-off $\kappa = 10$ may be too low. In this case, we suggest the use of a data-dependent cut-off $\kappa = n/S_{obs}$, which is the species average frequency. Thus, our recommendation can be summarized as in one rule as $\kappa = max(10, n/S_{obs})$. As defined in (4b), $S_{abun} = \sum_{i>\kappa} f_i$ denotes the total number of observed species in the abundant species group, and the number of observed species in the rare species group is $S_{rare} = \sum_{i=1}^{\kappa} f_i$. The Good-Turing coverage estimate for the rare species group becomes $\hat{C}_{rare} = 1 - f_1/\sum_{i\le\kappa} if_i$. The basic idea in the ACE is to account for the heterogeneity by adjusting the estimator $S_{abun} + S_{rare}/\hat{C}_{rare}$ (an estimator under a homogeneous model; see Equation 4b). The ACE is expressed as:

$$\hat{S}_{ACE} = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\hat{\gamma}^2_{rare}, \tag{13a}$$

where $\hat{\gamma}^2_{rare}$ is the square of the estimated CV and

$$\hat{\gamma}^2_{rare} = max\{\frac{S_{rare}}{\hat{C}_{rare}}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{\sum_{i=1}^{\kappa} if_i} \times \frac{1}{(\sum_{i=1}^{\kappa} if_i - 1)} - 1, 0\}. \tag{13b}$$

From the last term in the ACE, it is then readily seen that if heterogeneity exists, then the estimator which ignores the heterogeneity would have negative bias and the magnitude of the negative bias is proportional to the magnitude of heterogeneity. For highly heterogeneous assemblages, $\hat{\gamma}_{rare}$ in (13b) generally underestimates. A modified CV estimate was derived in [23] and the resulting estimator called ACE-1 has the following form:

$$\hat{S}_{ACE\text{-}1} = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\tilde{\gamma}^2_{rare}, \tag{14a}$$

where

$$\hat{\gamma}^2_{rare} = max\{\hat{S}_{ACE}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{\sum_{i=1}^{\kappa} if_i} \times \frac{1}{(\sum_{i=1}^{\kappa} if_i - 1)} - 1, 0\}. \tag{14b}$$

Approximate variance for the ACE and ACE-1 can be obtained based on a standard asymptotic approach [23][26]. The variance estimators are then used to construct confidence intervals of species richness via a log-transformation. Although the above presentation for ACE and ACE-1 is under the model in Equation (1), its validity under the general model in Equation (2) and the product-Poisson model in Equation (3) can be shown by parallel arguments.

The Jackknife Estimator [10] Jackknife techniques were developed as a general method to reduce the bias of a biased estimator. Here the biased estimator is the number of species observed in the sample. The basic idea with the jth order jackknife method is to consider sub-data by successively deleting j individuals from the data. The first-order jackknife turns out to be

$$\begin{aligned}\hat{S}_{jk1} &= S_{obs} + \frac{n-1}{n} f_1 \\ &\approx S_{obs} + f_1. \end{aligned} \tag{15a}$$

That is, only the number of singletons is used to estimate the number of undetected species. The second-order jackknife estimator, which uses singletons and doubletons, has the form:

$$\begin{aligned}\hat{S}_{jk2} &= S_{obs} + \frac{2n-3}{n} f_1 - \frac{(n-2)^2}{n(n-1)} f_2 \\ &\approx S_{obs} + 2f_1 - f_2. \end{aligned} \tag{15b}$$

Higher-order jackknife estimators are available. All estimators can be expressed as linear combinations of frequencies and thus variances and confidence intervals can be obtained [10].

Non-parametric MLE [75][94] A mixed Poisson model with a non-parametric mixing distribution F is considered in this approach. Replacing $p_\theta(k)$ in (5a) by $p_F(k) = \int [e^{-A\lambda}(A\lambda)^k/k!] dF(\lambda)$ for $k = 0, 1, \ldots$, we see the likelihood can be expressed as a function of S and the distribution F. This approach need not specify a parametric form for F, but to evaluate the NPMLE of F. The NPMLE turns out to be a finite mixture of point masses. This is equivalent to dividing the species detection rates into several classes, with the rates in each class being identical. Complex computation procedures (EM algorithm, penalized likelihood approach, cross-validation, or bootstrap method) are involved to obtain point and interval estimators. See [4] for real data analysis and [94] for a recent update along this direction.

Example 1 (Simulated Plant Abundance Data) An ideal test for comparison of various estimators is to examine their performance on real data sets with known parameters. Although such data sets exist for incidence data (see Example 2 below), proper abundance data sets with known species richnesses are not available to us. We therefore investigate their performance and behavior for simulated data generated from an assemblage with known species richness. We treated the 150-year field observations [73] for endangered and rare vascular plant species in the central portion of the southern Appalachian region as the true entire assemblage. The species-abundance distribution for this survey is reproduced in Table 1; a total of 188 species were recorded out of 1008 individuals. The assemblage from which data were generated thus includes 188 species, and there are 61 species with a relative abundance of 1/1008, 35 species with an abundance of 2/1008,... etc. The CV value for this distribution is 1.56, which indicates a relatively high degree of heterogeneity among species abundances.

We conducted a simulation experiment by selecting individuals with replacement from this species abundance distribution. We consider the multinomial model given in Equation (1): all individuals are assumed to have the same detectabilities so that the detection probability of any species is equal to its relative abundance. In our experiment, 100 simulated samples of size $n = 500$ were generated from the assumed assemblage. The expected number of species observed in a sample of size 500 is about 131. For each simulated sample, we used the SPADE program to obtain species richness estimates and their estimated s.e's computed from an asymp-

Table 1: Abundance frequency counts of the extant rare vascular plant species (188 species, 1008 individuals) in the southern Appalachians [73].

i	1	2	3	4	5	6	7	8	9	10	11	12	13
f_i	61	35	18	12	15	4	8	4	5	5	1	2	1

i	14	15	16	19	20	22	29	32	40	43	48	67
f_i	2	3	2	1	2	1	1	1	1	1	1	1

totic method. In Table 2, we present the average estimates and their average s.e.'s over 100 simulated samples for each of the following estimators: two estimators under a homogeneous model, Chao1 estimator, ACE, ACE-1, first-order and second-order jackknife estimators, and three parametric estimators (UMLE, CMLE and Chao and Bunge 2002 estimator) under a gamma-Poisson model. Since all species frequencies in any generated samples are less than 10, a cut off =10 implies that we used all sample frequencies to compute ACE and ACE-1. For each estimator, the resulting 100 estimates were then used to compute the sample s.e. and sample root mean squared error (RMSE), where RMSE is the square root of mean squared error (MSE, the sum of variance and squared bias). The sample s.e. quantifies the variation of an estimator and thus is a measure of precision. The sample RMSE quantifies the distance between an estimator and the true parameter, and thus is a measure of accuracy. The SPADE program also outputs a 95% confidence interval for each generated data set. Out of the resulting 100 intervals, the percentage that the 100 intervals cover the true value is also shown for each estimator in the last column of Table 2.

A satisfactory estimator should have small bias, small RMSE, and the coverage probability of the associated confidence interval should be close to the nominal confidence level (95% in our case). The two estimates under a homogeneous model (150.6 and 134.6), which ignore the heterogeneity among species abundances, have severe negative biases, leading to large values of RMSE. In contrast, all three parametric estimates (Chao and Bunge 2002 estimator, UMLE and CMLE) under a gamma-Poisson model exhibit positive biases. Their low precision causes low accuracy and overly conservative confidence intervals.

The Chao1 estimator (171.2 with s.e. 15.38) is slightly negatively biased (bias is about 7.7%) and behaves as a tight lower bound of species richness. The ACE (174.1 with s.e. 13.15) with a cut-off point of 10 exhibits similar behavior as the Chao1 estimator. Since the heterogeneity among species abundances is relatively high, it is expected that the ACE-1 with the same cut-off point is preferable to the ACE. Table 2 shows that the ACE-1 (187.2) has the lowest bias among the estimators considered in Table 2. For these data, the two jackknife estimators also perform satisfactorily as compared with the ACE and ACE-1. Although the first-order jackknife exhibits the smallest RMSE, the coverage of its confidence interval is much lower than the anticipated nominal value of 95%. The ACE-1 performs best in terms of bias and interval coverage probability (96%). For the Chao1, ACE and ACE-1, the average s.e.'s based on an asymptotic method (in column 4) are generally close to the sample s.e̊values (in column 5), implying the asymptotic method produces satisfactory

Table 2: Comparison of various species richness estimates based on 100 simulation samples of size 500 from the Appalachian plant abundance frequency distribution. True species richness = 188. All estimates are based on the SPADE output.

Estimator/Model	Equation number in text	Average estimate	Average s.e.	Sample s.e.	Sample RMSE	95% C.I. coverage
Homogeneous Model	(4b)	150.6	6.07	8.04	38.21	0
Homogeneous (MLE)	(4a)	134.6	1.93	6.12	53.77	0
Chao1 (Chao, 1984)	(10)	171.2	15.38	15.50	22.79	0.84
ACE (Chao & Lee, 1992)	(13a)	174.1	13.15	12.74	18.82	0.86
ACE-1 (Chao & Lee, 1992)	(14a)	187.2	19.26	17.63	17.56	0.96
1st order jackknife	(15a)	177.2	9.58	9.82	14.59	0.88
2nd order jackknife	(15b)	195.4	16.54	15.94	17.49	0.85
Gamma-Poisson (Chao & Bunge, 2002)	(6)	205.1	35.29	33.35	37.34	0.97
Gamma-Poisson-UMLE	(5b)	207.0	43.30	41.03	45.05	1.00
Gamma-Poisson-CMLE	(5c)	208.7	47.51	32.72	39.56	0.99

estimated s.e.s for these data.

1.5 Species Richness Estimation (Incidence Data)

For sample-based data, the reference sample consists of a species-by-sampling-unit incidence matrix. This $S \times T$ matrix is similar to a capture-recapture matrix for estimating the size of an animal population. There is a simple analogy between species richness estimation for a multiple-species assemblage and population size estimation for a single species. An "individual" animal in capture-recapture studies corresponds to a "species" in the richness estimation. The estimating target in the former is population size and in the latter is species richness. Also, the capture probability in a capture-recapture experiment corresponds to species detection probability, which is defined as the chance of encountering at least one individual of a given species. Therefore, the estimation techniques in the capture-recapture technique can be directly applied to estimate species richness. The major difference is that in population studies individuals are often not distinguishable from each other, thus animals are often captured and tagged or marked in order to have individual capture records, but in species richness estimation species are easily classified from sighting. Comprehensive reviews of methodology and applications are provided by [2][87][88], and short overviews specifically on population size estimation are given in [14][19] and [20].

Most developments in incidence data are based on a sequence of useful models proposed by Pollock [81] for analyzing capture-recapture data. We formulate these models in terms of species richness estimation. Three sources of variations in species detection probability are considered: (i) model $\boldsymbol{M}_t$, which allows species detection probabilities to vary by time or sampling unit; (ii) model $\boldsymbol{M}_b$, which allows behavioral responses to previous detection records; and (iii) model $\boldsymbol{M}_h$, which allows heterogeneous detection probabilities among species. Various combinations

of the above three variations (i.e., models $\boldsymbol{M}_{tb}$, $\boldsymbol{M}_{th}$, $\boldsymbol{M}_{bh}$ and $\boldsymbol{M}_{tbh}$) and the model $\boldsymbol{M}_0$, in which no variation exists, are also considered. A large number of statistical estimation methods have been proposed in the capture-recapture literature. These estimators rely on many different approaches: the maximum likelihood, the jackknife method, the bootstrap method, log-linear or generalized log-linear models, Bayesian methods, mixture models, sample coverage procedures, and martingale estimating functions [15][87][88]. Some of the models have been used [5][11][96] to estimate species richness.

Models with behavioral response (i.e., models $\boldsymbol{M}_b$, $\boldsymbol{M}_{bh}$, $\boldsymbol{M}_{tb}$ and $\boldsymbol{M}_{tbh}$) allow the detection probability of any species to depend on whether the observer has already recorded it in "previous" sampling units. Thus ordering is implicitly involved in these four models. Almost all estimation procedures derived under these models depend on the ordering of the samples. These models are thus useful only for temporally replicated samples, especially when the sampling is conducted by a single investigator or when only data on the accumulation of previously undiscovered species are used. Therefore, models $\boldsymbol{M}_t$, $\boldsymbol{M}_h$ and $\boldsymbol{M}_{th}$ are more potentially useful for species estimation. Since heterogeneity is expected in natural communities, this leaves models $\boldsymbol{M}_h$ and $\boldsymbol{M}_{th}$.

A multiplicative form of model $\boldsymbol{M}_{th}$ assumes that the detection probability P_{ij}, the probability of detecting the ith species in the jth sampling unit, has the form $P_{ij} = \pi_i e_j$, $0 < \pi_i e_j < 1$; here the parameters $\{e_1, e_2, \ldots, e_T\}$, $\{\pi_1, \pi_2, \ldots, \pi_S\}$, are used, respectively, to denote the unknown sampling-unit effects and heterogeneity pattern. The latter is mostly determined by species abundance structure and individual detectabilities whereas the former is closely related to sampling efforts, quadrat area, sampling method, landscape and other environmental variables associated with each sampling unit. When the sampling-unit effects can be assumed to be identical (e.g, equal-size quadrats, equal-effort sampling with similar protocols), this model reduces to model $\boldsymbol{M}_h$, i.e., $P_{ij} = \pi_i$. Here $\sum_{i=1}^{S} \pi_i$ may be greater than 1. (For example, the detection probability of the first species might be 0.6 and for the second species 0.7). Model $\boldsymbol{M}_h$ assumes that the ith species has its own unique detection probability π_i that remains constant over sampling units. That is, each element W_{ij} in the incidence matrix is a Bernoulli random variable (since $W_{ij} = 0$ or $W_{ij} = 1$), with probability π_i that $W_{ij} = 1$ and probability $1-\pi_i$ that $W_{ij} = 0$. The probability distribution for the incidence matrix is

$$\begin{aligned} P(W_{ij} = w_{ij}; i = 1, 2, \ldots, S; j = 1, 2, \ldots, T) \\ = \prod_{j=1}^{T} \prod_{i=1}^{S} \pi_i^{w_{ij}} (1 - \pi_i)^{1 - w_{ij}} \\ = \prod_{i=1}^{S} \pi_i^{y_i} (1 - \pi_i)^{T - y_i}. \end{aligned} \tag{16a}$$

The row sums $(Y_1, Y_2, \ldots, Y_S)$ are thus the sufficient statistics under model $\boldsymbol{M}_h$, and our analysis is based on the incidence frequency counts Q_k defined from $(Y_1, Y_2, \ldots, Y_S)$. From (16a), the model is also equivalent to a product-binomial model for the observed species frequencies:

$$\begin{aligned} P(Y_i = y_i; i = 1, 2, \ldots, S) \\ = \prod_{i=1}^{S} \binom{T}{y_i} \pi_i^{y_i} (1 - \pi_i)^{T - y_i}. \end{aligned} \tag{16b}$$

Quadrat sampling has been widely used to estimate abundance of plants and other sessile organisms. The models for quadrat sampling can be formulated as a special type of model $\boldsymbol{M}_h$ with π_i specifically in terms of species occupancy rate and species detectability in each quadrat. Suppose that the region under investigation is divided into T^* disjoint quadrats of the same area,

a sample of T quadrats are randomly selected. Then each selected quadrat is surveyed and the presence or absence of any species for each of these T quadrats is recorded. We assume that T is relatively small to T^* so that sampling with replacement and the sampling without replacement differ little. Let M_i be the unknown number of occupied quadrats by the ith species, $i = 1, 2, \ldots, S$. Assume that in each sampling unit, the conditional probability of detecting species i in any selected quadrat (given species i is present) is $0 < \alpha_i \leq 1$. That is, any selected sampling unit need not be completely censused. The model assumes that out of these M_i quadrats, species i can only be detected in U_i quadrats. Here U_i is also unknown and $M_i \geq U_i \geq 1$. (For any species with $U_i = 0$, there is no chance to detect this species in any sample, so it should be excluded in the estimating target.) Thus, U_i is a truncated binomial distribution with probability

$$P(U_i = u) = \binom{M_i}{u} \frac{\alpha_i^u (1-\alpha_i)^{M_i - u}}{1 - (1-\alpha_i)^{M_i}}, \quad u = 1, 2, \ldots, M_i.$$

In the other $T^* - U_i$ quadrats, either species i is absent or it is present but cannot be detected. Here we may assume any types of distributions for species detection probability α_i (e.q., constant, uniform distribution and beta distributions). The sample frequencies $(Y_1, Y_2, \ldots, Y_S)$ given $U_i = u_i$, $i = 1, 2, \ldots, S$, follow a product-binomial distribution:

$$P(Y_i = y_i; i = 1, 2, \ldots, S) = \prod_{i=1}^{S} \binom{T}{y_i} (\frac{u_i}{T^*})^{y_i} (1 - \frac{u_i}{T^*})^{T - y_i}, \quad 1 \leq u_i \leq M_i. \tag{16c}$$

This is a special case of model $\boldsymbol{M}_h$ in which the detection probability for species i in any sampling unit is $\pi_i = U_i/T^*$, which has an approximately mean value $M_i \alpha_i / T^*$, a product of occupancy rate M_i/T^* and the detectability α_i.

Parametric or likelihood-based approach As in the abundance model, further assumptions about the detection probabilities $\pi_1, \pi_2, \ldots, \pi_S$ under model $\boldsymbol{M}_h$ can be made so that the number of parameters can be reduced and the inference procedures can be simplified. A common parametric model is the beta-binomial model, where the detection probabilities are assumed to be a random sample from a beta distribution [11][40] with density $f(\pi; \alpha, \beta) = \Gamma(\alpha + \beta)\pi^{\alpha-1}(1 - \pi)^{\beta-1}/[\Gamma(\alpha)\Gamma(\beta)]$. The likelihood is similar to that in Equation (5b), with $p_{\boldsymbol{\theta}}(k)$ replaced by the following:

$$\int_0^1 \binom{T}{k} \pi^k (1-\pi)^{T-k} f(\pi; \alpha, \beta) d\pi = \binom{T}{k} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta+T-k)}{\Gamma(\alpha+\beta+T)}, \quad k = 0, 1, 2, \ldots \tag{17}$$

Then based on (5b) and (5c), the UMLE and CMLE can be obtained and all estimation procedures are similar to those discussed for the abundance data. Numerical iterations are required to obtain estimates, which may not be obtainable due to failure of convergences.

There are other parametric assumptions. For example, Pledger [80] assumed a latent class or a finite mixture model. That is, there are several groups of species and in each group the detection rates are assumed to be homogeneous within each class. Coull and Agresti [38] used a normal distribution to model the effects $\{l_1, l_2, \ldots, l_S\}$, where $l_i = logit(\pi_i) = log[\pi_i/(1 - \pi_i)]$. These approaches work well only when the specified parametric models are the true models. For example, in the latent class model, it works well only when there are actually contain groups of individuals that are

thought to have different detection rates of capture. Standard inference estimation procedures involving numerical iterations are then applied to obtain species richness estimates and the associated confidence intervals.

Non-Parametric Approaches

The Chao2 Estimator [13] Most of the non-parametric estimators were originally derived for capture-recapture experiments [14]. Under model $\boldsymbol{M}_h$, it follows from Equation (16b) that each species incidence frequency is a binomial distribution. Using similar steps as we used in our derivation of the Chao1 estimator for the abundance model, we obtain the following Chao2 lower bound for species richness under the model $\boldsymbol{M}_h$: [13]

$$\begin{aligned}\hat{S}_{Chao2} &= S_{obs} + \frac{T-1}{T}\frac{Q_1^2}{2Q_2}, \\ &\qquad \text{if } Q_2 > 0 \\ &= S_{obs} + \frac{T-1}{T}\frac{Q_1(Q_1-1)}{2}, \\ &\qquad \text{if } Q_2 = 0 \end{aligned} \tag{18a}$$

Unlike the Chao1 estimator, here the factor $(T-1)/T$ cannot be neglected because in sample-based data, T may not be sufficiently large. When $Q_2 > 0$, a variance estimator for the Chao2 lower bound is:

$$\begin{aligned}&\hat{var}(\hat{S}_{Chao2}) \\ &= Q_2\left[\frac{A}{2}(\frac{Q_1}{Q_2})^2 + A^2(\frac{Q_1}{Q_2})^3 + \frac{1}{4}A^2(\frac{Q_1}{Q_2})^2\right],\end{aligned} \tag{18b}$$

where $A = (T-1)/T$. When $Q_2 = 0$, the variance is modified to [29]:

$$\begin{aligned}\hat{var}(\hat{S}_{Chao2}) &= \frac{AQ_1(Q_1-1)}{2} \\ &\quad + \frac{A^2Q_1(2Q_1-1)^2}{4} \\ &\quad - \frac{A^2Q_1^4}{4\hat{S}_{Chao2}}.\end{aligned}$$

The Chao2 estimator is also valid under the model in Equation (16c). This implies that as long as quadrats are randomly selected, the use of the Chao2 estimator is justified in quadrat sampling even if species are spatially aggregated in the study area.

In some surveys, a sample of T sampling units is randomly selected without replacement from a total of T^*units so that any unit cannot be repeatedly selected. Chao and Lin [24] modify the model in Equation (16c) by assuming that the sample frequencies $(Y_1, Y_2, \ldots, Y_S)$ given $U_i = u_i$ follow a product-hypergeometric distribution:

$$\begin{aligned}&P(Y_i = y_i, i = 1, 2, \ldots, S) \\ &\quad = \prod_{i=1}^{S}\binom{u_i}{y_i}\binom{T^*-u_i}{T-y_i}\Big/\binom{T^*}{T}, \\ &\qquad 1 \le u_i \le M_i.\end{aligned}$$

That is, $(Y_1, Y_2, \ldots, Y_S)$ are independent but non-identically distributed random variables and each follows a hypergeometric distribution. They extended the Chao2 estimator to the following estimator

$$\hat{S}_{Chao2.wor} = S_{obs} + \frac{Q_1^2}{\frac{T}{T-1}2Q_2 + \frac{q}{1-q}Q_1},$$

where again the subscript "*wor*" refers to "without replacement", and $q = T/T^*$ denotes the known sampling ratio. When the sample fraction q approaches zero, the lower bound approaches the Chao2 estimator given in (18a). When q approaches 1 so that all individuals are observed, $q/(1-q)$ approaches infinity and the lower bound tends to the number of observed species, which equals the true parameter when all sampling units have been selected.

The ICE [65] Parallel to the ACE, there is a corresponding Incidence-based Coverage Estimator (ICE) for incidence data under model $\boldsymbol{M}_h$. The definition of "sample coverage" is modified to $C = \sum_{i=1}^{S}\pi_i I(Y_i > 0)/\sum_{i=1}^{S}\pi_i$, which can be

very accurately estimated from data. As with the ACE, we first select a cut-off point κ that partitions the data into an infrequent species group (incidence frequency not larger than κ) and a frequent species group (incidence frequency larger than κ). Denote the number of species in the frequent group by $S_{freq} = \sum_{i>\kappa} Q_i$ and the number of species in the infrequent group by $S_{infreq} = \sum_{i=1}^{\kappa} Q_i$. The estimated sample coverage for the infrequent group is $\hat{C}_{infreq} = 1 - Q_1/\sum_{i=1}^{\kappa} iQ_i$. Let the number of sampling units that include at least one infrequent species be T_{infreq}. Then the ICE is expressed as

$$\hat{S}_{ICE} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}}\hat{\gamma}^2_{infreq}, \tag{19a}$$

where $\hat{\gamma}^2_{infreq}$ is the estimate of the squared CV of the species detection probabilities $\pi_1, \pi_2, \ldots, \pi_S$:

$$\hat{\gamma}^2_{infreq} = max\{\frac{S_{infreq}}{\hat{C}_{infreq}}\frac{T_{infreq}}{(T_{infreq}-1)} \times \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{(\sum_{i=1}^{\kappa} iQ_i)(\sum_{i=1}^{\kappa} iQ_i - 1)} - 1, 0\}. \tag{19b}$$

Here, we also recommend using $\kappa = max(10, n/S_{obs})$ as in the ACE. If CV approaches zero, then the ICE is reduced to the following estimator for a homogeneous model (i.e., $\pi_1 = \pi_2 = \ldots = \pi_S$)

$$\hat{S}_0 = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}}, \tag{20}$$

which is similar to that in Equation (4b). For highly-heterogeneous cases, a modified ICE-1 is suggested:

$$\hat{S}_{ICE\text{-}1} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}}\tilde{\gamma}^2_{infreq}, \tag{21}$$

where

$$\tilde{\gamma}^2_{infreq} = max\{\hat{S}_{ICE}\frac{T_{infreq}}{(T_{infreq}-1)} \times \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{(\sum_{i=1}^{\kappa} iQ_i)(\sum_{i=1}^{\kappa} iQ_i - 1)} - 1, 0\}.$$

Under model $\boldsymbol{M}_h$, confidence intervals of species richness associated with ICE and ICE-1 can be obtained based on an analytic asymptotic variances [65].

The Jackknife Estimator [10] For incidence data, the first-order jackknife for T sampling units is:

$$\hat{S}_{jk1} = S_{obs} + \frac{T-1}{T}Q_1, \tag{22a}$$

and the second-order jackknife is:

$$\hat{S}_{jk2} = S_{obs} + \frac{2T-3}{T}Q_1 - \frac{(T-2)^2}{T(T-1)}Q_2. \tag{22b}$$

All jackknife estimators can be expressed as linear combinations of incidence frequency counts and thus approximate variances and confidence intervals can be obtained [10].

Non-parametric MLE [69][71] The NPMLE approach can be similarly applied to incidence data. A mixed binomial model with a non-parametric mixing distribution F is considered in this approach. Substituting $p_F(k) = \int_0^1 \binom{T}{k}\pi^k(1-\pi)^{T-k}dF(\pi)$ for $k = 0, 1, \ldots$, into Equation (5b), we obtain the likelihood in terms of S and the distribution F. As in the abundance data case, the NPMLE of F is a finite mixture of point masses and numerical procedures are needed to obtain the NPMLE. See [69][71] for details.

Example 2 (Real Data from a Cottontail Population with known parameter) As described earlier, the estimation of species richness for incidence data

Table 3: Comparison of various population size estimates based on the SPADE output for cottontail rabbit capture-recapture data [41]. True population size = 135.

Estimator/Model	Eq. in text	Estimate	s.e.	95% confidence interval
Homogeneous Model	(20)	109	10.6	(93.9, 136.9)
Chao2 (Chao, 1987)	(18a)	130.6	22.8	(100.9, 195.6)
ICE	(19a)	133.0	8.3	(118.9, 151.7)
ICE-1	(21)	153.9	15.2	(129.4, 189.8)
1st order jackknife	(22a)	116.6	8.9	(102.6, 138.0)
2nd order jackknife	(22b)	141.4	14.9	(118.2, 177.6)
Beta-binomial-CMLE	(17), (5c)	320.6	752.4	(88.1, 5004.8)
Beta-binomial-UMLE	(17), (5b)	*	*	(*, *)

$*$ iterative steps do not converge

is equivalent to the estimation of population size in a capture-recapture experiment. In the context of population size estimation, there were some well known real capture-recapture data sets collected from populations with known sizes. We use the cottontail capture-recapture data provided in [41] for illustration. Edwards and Eberhardt conducted a live-trapping study on a confined cottontail population of known size. In their study, 135 cottontail rabbits were penned in a 4-acre rabbit-proof enclosure. Live trapping was conducted for 18 consecutive nights. To distinguish individuals, a unique tag was attached to each individual so that the capture history of each individual captured in the experiment was known. A total of 142 captures were recorded and there were $S_{obs} = 76$ distinct rabbits. For these data, the incidence frequency counts (Q_1 to Q_7) were 43, 16, 8, 6, 0, 2, 1. In Table 3, we show various population size estimates, their estimated s.e.'s and 95 confidence intervals based on the output from the SPADE program.

The CV estimate computed from Equation (13b) is 0.654, indicating the existence of significant heterogeneity among individuals' capture probabilities. Any estimate under a homogeneous model (equal-capture probabilities) would have a severe negative bias. In Table 3, the homogeneous model using Equation (20) gives an estimate of 109 which is far from the true population size of 135. The UMLE under a beta-binomial model is not obtainable due to the failure of convergence of the iterative steps in computation. The corresponding CMLE exhibits a large positive bias and a large variation. These render the two parametric estimates useless. The Chao2 estimate gives a very sharp lower bound 130.6 (s.e. 22.8) with a 95% confidence interval of (100.9, 195.6). The ICE gives an estimate of 133 with a 95% confidence interval of (118.9, 151.7). The ICE has the lowest bias among those considered in Table 3. The first-order jackknife estimator severely underestimates whereas the second-order jackknife estimator slightly overestimates, with a wider 95% confidence interval (118.2, 177.6) than that based on the ICE.

1.6 Rarefaction and Extrapolation (Abundance Data)

Discrete-Type Sampling The classic rarefaction model refers to the multinomial

model in Equation (1) based on species relative abundances. The model can be extended to the more general model in Equation (2) based on species detection probabilities. If we knew the true species detection probabilities $\psi_1, \psi_2, \ldots, \psi_S$ of each of the S species, we could compute the expected number of species $S_{ind}(m)$ in a random subset of m individuals from the reference sample ($m < n$), using the following function:

$$S_{ind}(m) = S - \sum_{i=1}^{S}(1-\psi_i)^m. \qquad (23a)$$

However, we need to estimate it based on a sample of n individuals with observed species frequencies X_i for species i. It follows from [92] that the minimum variance unbiased estimator (MVUE) for $S_{ind}(m)$ is

$$\tilde{S}_{ind}(m) = S_{obs} - \sum_{X_i>1} \binom{n-X_i}{m} \Big/ \binom{n}{m}. \qquad (23b)$$

There are two kinds of variance associated with the estimator $\tilde{S}_{ind}(m)$. A "conditional" (on reference sample) variance only focuses on the variation of species richness in the sampling procedure of the given reference sample. That is, if the sampling for the reference sample were stopped at the size m, this is what the variation about the species richness estimate would be. Thus the conditional variance tends to zero when the rarefied sample size approaches the size of the reference sample because species richness of sample size of n is fixed. An "unconditional" variance gives the variation of species richness if another new sample of size m is taken from the entire assemblage. Therefore, the unconditional variance does not vanish when sample size tends to n. In most applications, unconditional variance is more useful as the inference is often not restricted to the reference sample. An asymptotic unconditional variance for $\tilde{S}_{ind}(m)$ was derived in [35]:

$$\sigma^2_{ind}(m) = \sum_{k=1}^{n}(1-\alpha_{km})^2 f_k - \tilde{S}^2_{ind}(m)/S_{est}. \qquad (24)$$

where S_{est} is an estimate of species richness (Chao1 or ACE), $\alpha_{km} = (n-k)!(n-m)!/[n!(n-k-m)!]$ for $k \le n-m$, and $\alpha_{km} = 0$ otherwise. This variance works well if sample size is sufficiently large, but it generally overestimates for small sample sizes. Thus the confidence intervals constructed by using this asymptotic variance are conservative. When sample size is not large, a bootstrap variance is suggested [35]. Details are given in the User's Guide of the iNEXT program (see Section 11).

The extrapolation problem is to estimate the expected number of species $S_{ind}(n+m^*)$ in an augmented sample of $n+m^*$ individuals from the assemblage ($m^* > 0$). The theoretical formula, for a given S_{obs}, can be expressed as

$$S_{ind}(n+m^*) = S_{obs} - \sum_{i=1}^{S}\left[1-(1-\psi_i)^{m^*}\right](1-\psi_i)^n. \qquad (25a)$$

Note that when m^* tends to infinity, $E[S_{ind}(n+m^*)]$ tends to species richness. Based only on the reference sample, with observed species frequencies X_i and their frequency counts f_i, Shen et al. [89] derived the following useful predictor:

$$\begin{aligned}\tilde{S}_{ind}(n+m^*) &= S_{obs} - \hat{f}_0\left[1-(1-\frac{f_1}{n\hat{f}_0})^{m^*}\right] \\ &\approx S_{obs} + \hat{f}_0\left[1-exp(-\frac{m^*}{n}\frac{f_1}{\hat{f}_0})\right]. \qquad (25b)\end{aligned}$$

They also derived an asymptotic unconditional variance estimator for $\tilde{S}_{ind}(n+m^*)$.

A bootstrap variance estimator is used instead in the iNEXT program.

In the prediction formula (25b), we must determine $\hat{f}_0$, an estimator for f_0 (the number of undetected species). We suggest that $\hat{f}_0$ can be obtained by using either the Chao1 estimator ($\hat{f}_0 = \hat{S}_{Chao1} - S_{obs}$) or the ACE estimator ($\hat{f}_0 = \hat{S}_{ACE} - S_{obs}$). When m^* tends to infinity, the extrapolated estimator approaches $S_{obs} + \hat{f}_0$, implying the selected species richness estimator is exactly the asymptotic value of our extrapolation formula. We would suggest that the extrapolation is reliable, in terms of the bias with respect to the true value $S_{ind}(n + m^*)$, at most only up to a tripling of the reference sample size, or more conservatively, a doubling of reference sample size. The precision of the extrapolation generally depends on the amount of data in the reference sample. Sparse data would lead to very large variances especially for long-range forecast. On the other hand, abundant data would result in precise estimates for all extrapolated values and the selected species richness estimator, but the bias may become substantial for long-range forecast. Similar conclusions apply to all extrapolations in the following sections.

Continuous-Type Sampling Under the area-based product-Poisson model (Coleman rarefaction) given in Equation (3), the rarefaction is to estimate the expected number of species $S_{area}(a)$ in a random sub-area of size a within the reference area of size A ($a < A$). If we knew the true Poisson occurrence rates $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ of each of the S species in the assemblage, we could compute

$$S_{area}(a) = S - \sum_{i=1}^{S} exp(-a\lambda_i). \quad (26a)$$

Based on species abundances X_i in the reference sample, Coleman [33] obtained the following estimator

$$\tilde{S}_{area}(a) = S_{obs} - \sum_{X_i>0} \left(1 - \frac{a}{A}\right)^{X_i}. \quad (26b)$$

This estimator is the MVUE for $S_{area}(a)$ from basic statistical theory.

The extrapolation is to estimate the expected number of species $S_{area}(A + a^*)$ in an augmented area $A + a^*$ ($a^* > 0$). Given S_{obs}, the theoretical formula can be expressed as

$$S_{area}(A + a^*) = S_{obs} - \sum_{i=1}^{S} [1 - exp(-a^*\lambda_i)]\, exp(-A\lambda_i). \quad (27a)$$

Working from species abundances X_i in the reference sample, Chao and Shen [27] proposed an estimator for $S_{area}(A + a^*)$,

$$\tilde{S}_{area}(A + a^*) = S_{obs} - \hat{f}_0 \left[1 - exp(-\frac{a^*}{A}\frac{f_1}{\hat{f}_0})\right], \quad (27b)$$

where $\hat{f}_0 = \hat{S}_{Chao1} - S_{obs}$ or $\hat{f}_0 = \hat{S}_{ACE} - S_{obs}$. They also derived an asymptotic variance estimator for $\tilde{S}_{area}(A + a^*)$. As in the individual-based case, we can use an asymptotic unconditional variance or a bootstrap method to assess the variance for the estimator (26b) and the predictor (27b).

Example 3 (Beetle Abundance Data) We used two beetle data sets provided in [60][61] to illustrate how to apply individual-based rarefaction and extrapolation to the same reference sample. The purpose is to compare beetle species richness between Osa second-growth site and Osa old-growth site. This example was also discussed in [35]. The data are duplicated in Table 4. The sample sizes (number

of individual beetles) for the Osa second-growth is much larger than the size for the Osa old-growth site (976 vs. 237 individuals, see Figure 1). From the unstandardized raw data (the reference samples), one might conclude that the second-growth site has more beetle species than the old-growth site (140 vs. 112) (Figure 1, solid points). Using Equation (23b), we plot for each site the individual-based rarefaction $\tilde{S}_{ind}(m)$ for m ranges from 1 to the size of a reference sample; see the two solid lines in Figure 1. When the sample size in the second-growth site is rarefied down to 237 individuals to match the size of the old-growth sample, the ordering of the two sites is reversed. The interpolated species richness for 237 individuals in the second-growth site is only 70, considerably less than the old-growth site, with 112 species.

Applying the extrapolation formula (Equation 25b with $\hat{f}_0 = \hat{S}_{Chao1} - S_{obs}$) to the Janzen data set to increase the sample size in each site yields the extrapolated curves (broken line curves) for each site in Figure 1. For each of the extrapolation curves, we plot $\tilde{S}_{ind}(n + m^*)$, where $n + m^*$ ranges from reference sample size 0 to 1300. Even though the mathematical derivations for interpolation and extrapolation are fundamentally different, the interpolation and extrapolation curves join smoothly at the single data point of the reference sample. The confidence intervals for both interpolation and extrapolation curves based on bootstrap s.e.'s are also linked smoothly.

For both samples, the interpolated and extrapolated richness, and their 95% confidence interval, increased with sample size. For extrapolation, the variances associated with the extrapolated values are relatively small up to a doubling of the reference sample, signifying quite accurate extrapolation in this range. For the Osa old-growth site, the extrapolation is extended to five times of the original sample size in order to compare with the Osa second-growth curve. This long-range extrapolation inevitably yields very wide confidence intervals due to relatively sparse data. For the Osa the second-growth site with more abundant data, the extrapolation is extended to only less than double the reference sample size, yielding a quite accurate extrapolated estimate with a narrow confidence interval.

Based on Figure 1, even though the Osa old-growth site extrapolation for large sample sizes exhibits high variance, the old-growth and second-growth confidence intervals do not overlap up to size of 1300, except initially for very small sample sizes, and the two 95% confidence intervals based on a bootstrap method do not overlap for any sample size considered. This implies that beetle species richness for any sample size is significantly greater in the old-growth site than that in the second-growth site for sample size up to 1300 individuals.

When the augmented size tends to infinity in the extrapolation, each curve approaches the Chao1 estimator. In the second-growth site, the Chao1 estimate is 284.1 (s.e. 50.5) with a 95% confidence interval of (213.9, 420.9). The corresponding Chao1 estimate for the old-growth site is 464.8 (s.e. 136.8) with a 95% confidence interval of (281.4, 846.9). Although the two intervals overlap considerably due to a large variance associated with the estimate of the old-growth site, the Chao1 estimates imply that the species in the old-growth assemblage is richer. A similar conclusion is also valid if the ACE, ACE-1, and second-order jackknife are used in the extrapolation formula. (Only the first-order jackknife implies a reversed ordering.)

The traditional rarefaction downsamples to standardize sample size. A recent new proposal by [1] and [63] is to standardize sample coverage. A given sample size may be sufficient to find all species in one assemblage, but may find only a small percentage

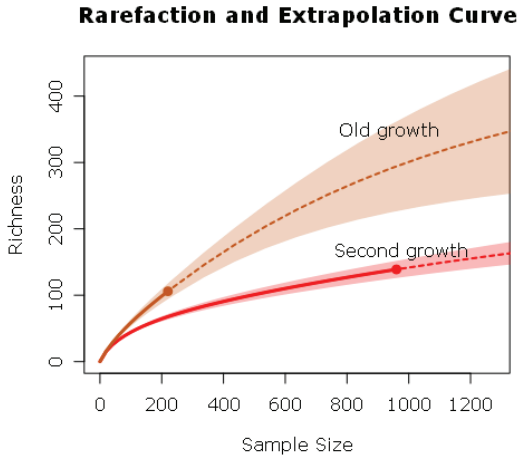

Figure 1: Individual-based interpolation (solid lines) and extrapolation (broken lines) from two reference samples (filled circles) of beetles from Costa Rica Osa old-growth site and Osa second-growth site [60] [61]. The two curves and their associated 95% bootstrap confidence intervals are shown up to a sample size of 1300. This curve is part of the output from program iNEXT (see Section 11). This figure is slightly different from that in [35] because different methods are used in constructing confidence intervals.

Table 4: The species abundance frequency counts for beetles from two sites on the Osa Peninsula in southwestern Costa Rica [60] [61].

Osa second-growth site: $S_{obs} = 140$, $n = 976$

i	1	2	3	4	5	6	7	8	9	10	11	12	14
f_i	70	17	4	5	5	5	5	3	1	2	3	2	2

i	17	19	20	21	24	26	40	57	60	64	71	77
f_i	1	2	3	1	1	1	1	2	1	1	1	1

Osa old-growth site: $S_{obs} = 112$, $n = 237$

i	1	2	3	4	5	6	7	8	14	42
f_i	84	10	4	3	5	1	2	1	1	1

of species in another assemblage that contains many rare species. Subsamples standardized by size may therefore represent different degrees of completeness. As described earlier, "sample coverage" characterizes the degree of completeness of a given sample. For example, the sample coverage estimates show that these two samples have very different degrees of completeness. In the second-growth site, the number of singletons was $f_1 = 70$, implying that the sample completeness (i.e., coverage estimate) is $1 - f_1/n = 1 - 70/976 = 93\%$. In the old-growth site, there were $f_1 = 84$ singletons, thus the sample completeness is $1 - 84/237 = 65\%$. In other words, for the old-growth site, the sample only covers 65% of the population, whereas the sample in the second-growth site covers 93%. Rarefaction to a target level of sample coverage allows informative comparison of equally-complete samples from multiple assemblages. See the two above-mentioned papers and [22] for analytic formulas and details.

1.7 Rarefaction and Extrapolation (Incidence Data)

Sample-based rarefaction is formulated under model $\boldsymbol{M}_h$ (Equation 16b) in which the incidence frequency counts Y_i follows a binomial distribution with T sampling units and detection probability π_i for the ith species in any sampling unit, or under a quadrat sampling (Equation 16c), with model $\boldsymbol{M}_h$ and $\pi_i = u_i/T^*$. Let $S_{sample}(t)$ be the expected number of species in a set of t sampling units randomly selected from the assemblage. If we knew the true species detection probabilities $\pi_1, \pi_2, \ldots, \pi_S$ of each of the S species in each sampling unit, we could compute

$$S_{sample}(t) = S - \sum_{i=1}^{S} (1 - \pi_i)^t. \quad (28a)$$

Based on the incidence reference sample with frequencies Y_i, the MVUE for $S_{sample}(t)$ is

$$\tilde{S}_{sample}(t) = S_{obs} - \sum_{Y_i > 0} \frac{\binom{T-Y_i}{t}}{\binom{T}{t}}. \quad (28b)$$

This analytic formula was first derived by Shinozaki [90] and rediscovered multiple times [32]. Colwell et al. ([37], their Equation 6) developed an asymptotic estimator for the unconditional variance in terms of the frequency counts Q_k, similar to Equation (24).

The extrapolation problem is to estimate the expected number of species $S_{sample}(T + t^*)$ in an augmented set of $T + t^*$ sampling units ($t^* > 0$) from the assemblage. For a given S_{obs}, the theoretical formula can be written as

$$S_{sample}(T + t^*) = S_{obs} + \sum_{i=1}^{S} \left[1 - (1 - \pi_i)^{t^*}\right] (1 - \pi_i)^T. \quad (29a)$$

Chao et al. ([18], their Appendix) obtained an estimator

$$\tilde{S}_{sample}(T + t^*) = S_{obs} - \hat{Q}_0 \left[1 - (1 - \frac{Q_1}{Q_1 + T\hat{f}_0})^{t^*}\right] \approx S_{obs} + \hat{Q}_0 \left[1 - exp(-\frac{tQ_1}{Q_1 + T\hat{Q}_0})\right], \quad (29b)$$

where $\hat{Q}_0$ can be obtained by using either the Chao2 estimator ($\hat{Q}_0 = \hat{S}_{Chao2} - S_{obs}$) or the ICE estimator ($\hat{Q}_0 = \hat{S}_{ICE} - S_{obs}$). In the program iNEXT, a bootstrap variance estimator is adopted for the estimator in (28b) and (29b).

Example 4 (Soil Ciliates Incidence Data) This data set includes soil ciliate species presence or absence data for a total of 51 soil samples from three areas (Southern Namib Desert, Central Namib Desert and Atosha Pan) of Namibia, Africa. The numbers of soil samples are respectively 15, 17 and 19 in the three areas. The incidence frequency counts are shown in Table 5 (original data are provided in [46] p.58-63). Detailed sampling locations, procedures and species identification were described in [46]. A total of 331 species were recorded in the data. The Atosha Pan has the highest observed species richness ($S_{obs} = 234$), the Central area has the lowest observed species richness ($S_{obs} = 136$) and the Southern area is in between ($S_{obs} = 154$).

As in the beetle data, the rarefaction and extrapolation curves (Equations 28b and 29b) are linked as shown in Figure 2. For sample-based rarefaction, we plot $\tilde{S}_{sample}(t)$ (solid line curves in Figure 2) for t ranges from 1 to the size of a reference sample. When all sample sizes of the three areas are standardized to the smallest size 15, the ordering of the three areas is the same as that based on the observed richness values.

Applying the formula (29b) with $\hat{Q}_0 = \hat{S}_{Chao2} - S_{obs}$ to the three samples to increase the soil sample size in each area yields the extrapolated curves (broken line curves) for each area in Figure 2. For each of the extrapolation curves, we plot $\tilde{S}_{sample}(T + t^*)$, where $T + t^*$ ranges from the reference sample size to a size of 38. As in the abundance data, we have a smooth curve for the interpolation and extrapolation curves that are linked at the single data point of the reference sample. Similar smooth curves occur for the confidence intervals. Since data are abundant, all the extrapolated values have acceptable uncertainty with relatively narrow confidence intervals. Except for very small sample sizes, the three bootstrap intervals do not overlap up to 38 soil samples. Therefore, the ordering for ciliate species richness for any sample size would remain the same as that for the observed richness values up to 38 soil samples.

When the augmented size tends to infinity in the extrapolation, each curve approaches the Chao2 estimator. In the Southern Namib Desert, the Chao2 estimate is 270.3 (s.e. 34.9) with a 95% confidence interval of (219.4, 360.8); in the Central Namib Desert, the Chao2 estimate is 216 (s.e. 26.1) with a 95% confidence interval of (178.9, 285.1); in the Etosha Pan, the estimate is 402.2 (s.e. 41.4) with a 95% confidence interval of (338.5, 504.7). All estimates indicate that there are still a substantial fraction of undetected species in the current data. The interval for the Etosha Pan does not overlap with that for the Central Namib Desert, but the intervals for the other two pairs of areas do overlap. These estimates imply the same ordering as that from the observed species richness. The same ordering is shown based on the ICE, ICE-1, and the first two orders of jackknife estimators.

1.8 Shared Species Richness Estimation (Abundance Data)

As indicated by Colwell and Coddington [36], the problem of estimating the true number of species shared by two (or more) sites or biotas based on sample data presents a difficult but important challenge. The first statistical estimator of shared species was developed by Chao et al. [21] who developed a generalization of ACE- and ICE-type shared species richness estimators for two assemblages. However, these estimators cannot be easily extended to the general case when there are more than two assemblages. Pan et al. [78] proposed the Chao1-type and Chao2-type

Table 5: Incidence frequency counts for ciliates in three regions of Namibia Desert (Original data are given in [46, p. 58-63]).

Southern Namib Desert ($T = 15$, $S_{obs} = 154$)

i	1	2	3	4	5	6	7	8	9	10	11	12	13
Q_i	85	29	14	9	5	1	1	2	2	1	2	2	1

Central Namib Desert ($T = 17$, $S_{obs} = 136$)

i	1	2	3	4	5	6	7	8	9	11	12	15	16
Q_i	69	28	13	4	3	7	1	2	1	1	1	3	3

Etosha Pan ($T = 19$, $S_{obs} = 234$)

i	1	2	3	4	5	6	7	8	9	10	11	12	14
Q_i	125	44	26	14	6	5	4	3	2	2	1	1	1

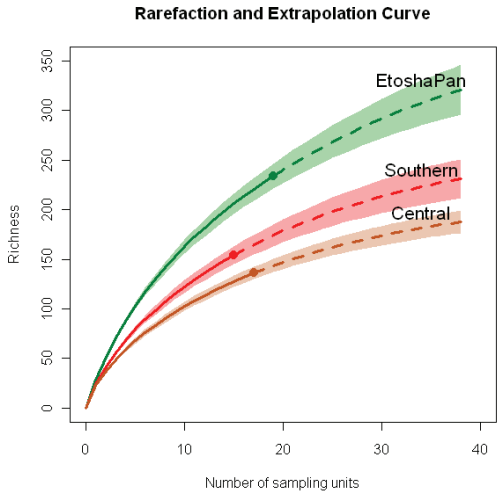

Figure 2: Sample-based interpolation (solid lines) and extrapolation (broken lines) from three reference samples (filled circles) of ciliate species from the Southern Namib Desert, Central Namib Desert and Etosha Pan of Namibia, Africa [46]. The three curves and their associated 95% bootstrap confidence intervals are shown up to 38 soil samples. This curve is part of the output from the iNEXT program.

lower bounds for shared species richness and provided a unified approach to construct lower bounds for more than two assemblages. We will mainly focus on the Chao1-type and Chao2-type shared species richness estimators.

Assume that there are S_1 species in Assemblage I and there are S_2 species in Assemblage II. Under the multinomial model in Equation (2), the species detection probabilities in Assemblages I and II are denoted by $(\psi_{11}, \psi_{21}, \ldots, \psi_{S_1,1})$ and $(\psi_{12}, \psi_{22}, \ldots, \psi_{S_2,2})$, respectively. $\sum_{i=1}^{S_1} \psi_{i1} = \sum_{i=1}^{S_2} \psi_{i2} = 1$. Let the number of shared species be S_{12}. Without loss of generality, we assume that the first S_{12} species are the shared species. Two random samples (Sample I with size n_1 and Sample II with size n_2) are taken from Assemblages I and II, respectively. Denote the observed species frequencies in the two samples by $(X_{11}, X_{21}, \ldots, X_{S_1,1})$ and $(X_{12}, X_{22}, \ldots, X_{S_2,2})$ respectively. Assume that D_{12} shared species are observed. Let f_{jk} denote the number of shared species that are observed j times in Sample I and k times in Sample II. In particular, f_{11} denotes the number of shared species that are singletons in both samples, and f_{00} denotes the number of shared species that are missed in both samples. Also, f_{+0} denotes the number of shared species that are observed in Sample I but not observed in Sample II, and a similar interpretation for f_{0+}.

The Chao1-shared Lower Bound [78]
Like the Chao1 species richness estimator, the Chao1-shared estimator was derived [78] as a lower bound by using the Cauchy-Schwarz inequality based on information from rare shared species. The basic idea, like the Chao1 estimator, is that abundant shared species carry negligible information about the undetected shared species. Only rare shared species carry nearly all such information. Since $S_{12} = D_{12} + f_{+0} + f_{0+} + f_{00}$ and only D_{12} is observable, our approach is to find a lower bound for each of the expected values of the other three terms, i.e., $E(f_{+0})$, $E(f_{0+})$ and $E(f_{00})$. Assuming a multinomial model (Equation 2) for each of the two sets of frequencies, we have

$$E(f_{00}) = \sum_{i=1}^{S_{12}} (1-\psi_{i1})^{n_1} (1-\psi_{i2})^{n_2},$$

$$E(f_{+0}) = \sum_{i=1}^{S_{12}} [1-(1-\psi_{i1})^{n_1}](1-\psi_{i2})^{n_2},$$

$$E(f_{0+}) = \sum_{i=1}^{S_{12}} (1-\psi_{i1})^{n_1} [1-(1-\psi_{i2})^{n_2}].$$

To obtain a lower bound for $E(f_{+0})$, note that

$$E(f_{+1}) = \sum_{i=1}^{S_{12}} [1-(1-\psi_{i1})^{n_1}] \times n_2 \psi_{i2} (1-\psi_{i2})^{n_2-1},$$

$$E(f_{+2}) = \sum_{i=1}^{S_{12}} [1-(1-\psi_{i1})^{n_1}] \times \binom{n_2}{2} \psi_{i2}^2 (1-\psi_{i2})^{n_2-2}.$$

The following Cauchy-Schwarz inequality

$$\left[\sum_{i=1}^{S_{12}} [1-(1-\psi_{i1})^{n_1}](1-\psi_{i2})^{n_2}\right] \times \left[\sum_{i=1}^{S_{12}} [1-(1-\psi_{i1})^{n_1}]\psi_{i2}^2 (1-\psi_{i2})^{n_2-2}\right] \geq \left[\sum_{i=1}^{S_{12}} [1-(1-\psi_{i1})^{n_1}]\psi_{i2}(1-\psi_{i2})^{n_2-1}\right]^2$$

leads to

$$E(f_{+0}) \geq \frac{(n_2-1)}{n_2} \frac{[E(f_{+1})]^2}{2E(f_{+2})}.$$

Similarly, a lower bound for $E(f_{0+})$ is

$$E(f_{0+}) \geq \frac{(n_1-1)}{n_1} \frac{[E(f_{1+})]^2}{2E(f_{2+})}.$$

A lower bound for $E(f_{00})$ is obtained by noting

$$E(f_{11}) = \sum_{i=1}^{S_{12}} n_1\psi_{i1}(1-\psi_{i1})^{n_1} \times n_2\psi_{i2}(1-\psi_{i2})^{n_2-1},$$

$$E(f_{22}) = \sum_{i=1}^{S_{12}} \binom{n_1}{2}\psi_{i1}^2(1-\psi_{i1})^{n_1-2} \times \binom{n_2}{2}\psi_{i2}^2(1-\psi_{i2})^{n_2-2}.$$

Again, a similar Cauchy-Schwarz inequality

$$\left[\sum_{i=1}^{S_{12}}(1-\psi_{i1})^{n_1}(1-\psi_{i2})^{n_2}\right] \times \left[\sum_{i=1}^{S_{12}}\psi_{i1}^2(1-\psi_{i1})^{n_1-2}\psi_{i2}^2(1-\psi_{i2})^{n_2-2}\right] \geq \left[\sum_{i=1}^{S_{12}}\psi_{i1}(1-\psi_{i1})^{n_1-1}\psi_{i2}(1-\psi_{i2})^{n_2-1}\right]^2$$

gives

$$E(f_{00}) \geq \frac{(n_1-1)}{n_1}\frac{(n_2-1)}{n_2}\frac{[E(f_{11})]^2}{4E(f_{22})}.$$

Combining the above three lower bounds, we obtain a lower bound for the shared species richness:

$$\hat{S}_{12} = D_{12} + k_1\frac{f_{1+}^2}{2f_{2+}} + k_2\frac{f_{+1}^2}{2f_{+2}} + k_1k_2\frac{f_{11}^2}{4f_{22}}, \quad (30a)$$

where $k_i = (n_i-1)/n_i$. This estimator is referred to as the Chao1-shared estimator because it can be regarded as an extension of the Chao1 estimator to the case of two assemblages. In Equation (30a), note that only those shared species that are singletons or doubletons in at least one of the samples are used to infer the undetected shared species richness. In many cases, the sample sizes n_1 and n_2 are large for abundance data so the terms $(n_1-1)/n_1$ and $(n_2-1)/n_2$ can be dropped in the above formula. If at least one variable in the set $\{f_{2+}, f_{+2}, f_{22}\}$ is zero, then a modified estimator is

$$\tilde{S}_{12} = D_{12} + k_1\frac{f_{1+}(f_{1+}-1)}{2(f_{2+}+1)} + k_2\frac{f_{+1}(f_{+1}-1)}{2(f_{+2}+1)} + k_1k_2\frac{f_{11}(f_{11}-1)}{4(f_{22}+1)}. \quad (30b)$$

Note that non-shared species play no role in Equations (30a) and (30b), although any observed non-shared species could actually be a shared species. Because the proposed estimator can be regarded as a function of the statistics $(D_{12}, f_{11}, f_{22}, f_{1+}, f_{2+}, f_{+1}, f_{+2})$, a variance estimator can be obtained by using a standard asymptotic approach under a multinomial model, and this can be used to construct a confidence interval for the true parameter.

When there are more than two assemblages, a "shared" species is defined as that the species belongs to all assemblages. For example, in the case of three assemblages, assume that there are S_{123} species shared by three assemblages and a random sample is taken from each of the three assemblages. With self-explanatory notation generalization, a lower bound for shared species richness is

$$\hat{S}_{123} = D_{123} + k_1\frac{f_{1++}^2}{2f_{2++}} + k_2\frac{f_{+1+}^2}{2f_{+2+}} + k_3\frac{f_{++1}^2}{2f_{++2}} + k_1k_2\frac{f_{11+}^2}{4f_{22+}} + k_1k_3\frac{f_{1+1}^2}{4f_{2+2}} + k_2k_3\frac{f_{+11}^2}{4f_{+22}} + k_1k_2k_3\frac{f_{111}^2}{8f_{222}}. \quad (31)$$

The Chao1-shared estimators and their asymptotic variances as well as confidence intervals for shared species richness are

featured in the SPADE program up to five assemblages. Chao and Lin [24] formulated the corresponding models and derived shared species richness estimators under sampling without replacement.

Example 5 (Tropic Tree Species Abundance Data) Pan et al. [78] conducted simulation studies to examine the performance of the shared species richness estimators derived in Equations (30a), (30b) and (31). Here we apply those estimators to tree species data collected in 2000 by Chazdon and colleagues on different tree sizes as measured by diameter at breast height (DBH) from the plots in the Lindero Sur (LUSR) old-growth forest (> 200 years) of La Selva, Costa Rica. The complete data were tabulated in [28]. We focus on two sizes: trees (≥ 25 cm in DBH) and saplings (1-5 cm in DBH). All trees and saplings were marked and measured for diameter within a 1 ha plot. There were 508 individual saplings of 101 species, and 119 individual trees of 37 species. There were 29 shared species in data ($D_{12} = 29$.) If we regard the data as a reference sample from the entire LSUR forest, then we can link the rarefaction (Equation 23b) and extrapolation curves using the Chao1 estimate in the prediction formula (Equation 25b). Beyond the base point, the linked curve for saplings is always above the curve for trees. Except for initial sample sizes, the two confidence intervals for any sample sizes do not overlap, indicating the two expected species richnesses are significantly different. When the sample size tends to infinity, the extrapolated curve for saplings tends to the Chao1 estimate (119.1 with s.e. 8.81) with a 95% confidence interval of (108.3, 145.7). The extrapolated curve for trees tends to 58.5 (s.e. 12.46) with a 95% confidence interval of (44.5, 98.7). The two intervals do not intersect even in this limiting case.

We can infer species richness shared by the two different size assemblages. Using our notation, the data give the following information from the observed rare shared species: $f_{11} = 3$, $f_{12} = 1$, $f_{21} = 4$, $f_{22} = 3$, $f_{1+} = 6$, $f_{2+} = 7$, $f_{+1} = 14$, $f_{+2} = 9$, where the sapling and tree assemblages are referred to as Assemblage I and Assemblage II, respectively. It follows from Equation (30a) that the Chao1-shared species richness estimator is $\hat{S}_{12} = 43.1$ with an asymptotic s.e. of 8.95. A 95% confidence interval using a log-transformation is (33.5, 73.1). The analysis suggests that at least one-third of the shared species were not detected in the samples. In addition to shared species richness, ecologists often also compute some other measures to quantify the compositional similarity/dissimilarity between assemblages; see [64] for details.

1.9 Shared Species Richness Estimation (Incidence Data)

The Chao2-shared Lower Bound The Chao1-shared estimator developed for abundance data can be directly adapted to deal with the incidence case. All notation and model formulation are similar to those for abundance data. When there are two assemblages, assume Sample I consists of data for T_1 sampling units randomly taken from Assemblage I and Sample II consists of data for T_2 sampling units randomly selected from Assemblage II. In each selected sampling unit, only presence/absence data are recorded. Assume model $\boldsymbol{M}_h$ for each assemblage (see Equations 16b and 16c), and assume the two sets of probabilities $(\pi_{11}, \pi_{21}, \ldots, \pi_{S_1,1})$ and $(\pi_{12}, \pi_{22}, \ldots, \pi_{S_2,2})$ represent species detection probabilities in Assemblages I and II, respectively.

Let Y_{i1} and Y_{i2} denote the number of sampling units containing the ith species in Sample I and II, respectively. Let Q_{jk}

denote the number of shared species that are detected in j sampling units in Sample I and k units in Sample II. Similarly, we can define the statistics Q_{+k} and Q_{j+}. Pan et al. [78] showed that the lower bound and the modified version for the number of shared species based on incidence counts have the same forms as in Equations (30a) and (30b) except that the samples sizes n_1 and n_2 should be respectively replaced by T_1 and T_2, and the abundance counts $(f_{11}, f_{22}, f_{1+}, f_{2+}, f_{+1}, f_{+2})$ replaced by the incidence counts $(Q_{11}, Q_{22}, Q_{1+}, Q_{2+}, Q_{+1}, Q_{+2})$. These are the observed shared statistics from shared species that are "uniques" or "duplicates" in at least one of the two samples. The resulting shared estimator is referred to as the Chao2-shared estimator. See [78] for asymptotic variance and extension to the general case. All the shared species richness estimators are featured in the SPADE program.

Example 6 (Soil Ciliate Incidence Data) We used the soil ciliate species data to illustrate the shared species richness estimation for two-assemblage and three-assemblage cases. The data given in Table 5 are not sufficient to obtain shared species richness estimators. Species identities or pair frequency counts are required to compute the observed shared richness D_{12} and the statistics $(Q_{11}, Q_{22}, Q_{1+}, Q_{2+}, Q_{+1}, Q_{+2})$. The species identities details are provided in [46].

For the two-assemblage case, we use the Southern Namib Desert and Etosha Pan for illustration. If we regard the two areas as Assemblage I and Assemblage II, respectively, then the observed shared richness between the two areas is $D_{12} = 97$ species. From the data, we have the following observed shared information from the infrequent shared species: $Q_{11} = 24$, $Q_{22} = 7$, $Q_{1+} = 42$, $Q_{2+} = 20$, $Q_{+1} = 36$, and $Q_{+2} = 19$. These statistics carry almost all information about the number of undetected shared species and are sufficient to obtain the Chao2-shared species richness estimate. We obtain $\hat{S}_{12} = 188.7$ (s.e. 33.56) with a 95% confidence interval of (142.7, 280.7). For these two areas, we can conclude that at least half of the shared species was not detected. Similar inference can be made for the other two pairs of areas. For the three-assemblage case, there were 65 species shared by all three areas in data. The shared species richness estimator using a similar equation as that in (31) gives an estimate of 125.7 (s.e. 30.12) with a 95% confidence interval of (89.2, 217.3). This estimate indicates that there is still a substantial fraction of undetected species shared by the three assemblages. Our approach reveals the extent of underestimation and provides helpful information for understanding community overlap of micro-organisms.

1.10 Applications

In the following, we list some application areas along with specific goals in each:

- Population biology: estimation of the size (i.e., the total number of individuals) of biological populations based on traditional capture-recapture data [14][96] or DNA-based genetic-tagging data [66][72].
- Genetics: estimation of the number of genes or alleles based on sample frequency counts [57].
- Genomics: estimation of the richness of operational taxonomic units (OTUs) or the number of unique, nonredundant gene sequences based on microbial or other samples [55][70][86].
- Medical science and epidemiology: estimation of the number of different cases for a specific disease by merging

several incomplete lists of individuals [14][56].

- Environmental science: estimation of the number of organic pollutants that were discharged to a water environment using survey data in the study area [59].
- Software reliability: estimation of the number of undiscovered bugs in a piece of software when data in debugging processes are available [6].
- Numismatics: estimation the number of die types for ancient coins found in a hoard [54].
- Archaeology: estimation the richness of paradigmatic classes of stone tools based on sample data from archaeological sites [43].
- Linguistics: estimation the size of vocabulary for an author based on his/her known writings [42].

1.11 Software

We only list free software that can be downloaded from the Internet.

- EstimateS: developed by Colwell [34] for computing a variety of biodiversity functions, estimators (including estimators for species richness and shared species richness), and indices based on biotic sampling data. http://purl.oclc.org/estimates.
- SPADE (Species Prediction And Diversity Estimation): developed by Chao and Shen [29]. SPADE features estimation of species richness, shared species richness, diversity indices, similarity and dissimilarity measures based on biotic or genetic data. http://chao.stat.nthu.edu.tw/software CE.html
- CARE-2 (CApture-REcapture): developed by Chao and Yang [31] for estimating population size based on capture-recapture data. Covariates models and analyses are also featured. http://chao.stat.nthu.edu.tw/software CE.html.
- iNEXT (interpolation-extrapolation): an R package developed by Chao et al. [25] specifically for computing and plotting rarefaction and extrapolation curves as shown in Figures 1 and 2 of this article. http://chao.stat.nthu.edu.tw/software CE.html.
- WS2m: software developed by Turner et al. [93] for the measurement and analysis of species diversity. http://eebweb.arizona.edu/diversity/.
- SPECIES: an R package for species richness estimation developed by Wang [95]. The package is available from the Comprehensive R Archive http://CRAN.R-project.org/package=SPECIES.

Acknowledgments. The work was supported by the National Science Council of Taiwan under Contract Number: 97-2118-M007-MY3. The authors sincerely thank Lou Jost for editing and providing very helpful comments and suggestions.

References

1. Alroy, J. (2010). The shifting balance of diversity among major marine animal groups. *Science*, **329**, 1191–1194.
2. Amstrup, S. C., McDonald, T. L. and Manly, B. F. J. (2005). *Handbook of Capture-Recapture Analysis*. Princeton University Press, Princeton, USA.
3. Barger, K. and Bunge, J. (2010). Objective Bayesian Estimation for the Number of Species, *Bayesian Analysis*, **5**, 765–786.

4. Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J. Roy. Stat. Soc. C.-App.*, **54**, 721–737.
5. Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. (1998). Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, **79**, 1018–1028.
6. Briand, L. C., El Emam, K., Freimut, B. G., and Laitenberger, O. (2000). A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Trans. Software Engrg.*, **26**, 518–540.
7. Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species abundance data. *Biometrics*, **30**, 101–110.
8. Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Ass.*, **88**, 364–373.
9. Bunge, J., Fitzpatrick. M., and Handley, J. (1995). Comparison of three estimators of the number of species. *J. Appl. Stat.*, **22**, 45–59.
10. Burnham, K. P., and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, **65**, 625–633.
11. Burnham, K. P. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, **60**, 927–936.
12. Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.*, **11**, 265–270.
13. Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.
14. Chao, A. (2001). An overview of closed capture-recapture models. *J. Agric. Bio. Environ. Stat.*, **6**, 158–175.
15. Chao, A. (2005). Species estimation and applications. In N. Balakrishnan, C. Read, B, and B. Vidakovic (Eds.), *Encyclopedia of Statistical Sciences* (pp. 7907–7916). New York: Wiley.
16. Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, **58**, 531–539.
17. Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, **62**, 361–371.
18. Chao, A., Colwell, R. K., Lin, C. W. and Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, **90**, 1125–1133.
19. Chao, A., and Huggins, R. M. (2005). Modern closed population Capture-Recapture Models. In B. Manly, T. McDonald and S. Amstrup (Eds.). *The Handbook of Capture-Recapture Analysis* (pp. 58–87): Princeton University Press.
20. Chao, A. and Huggins, R. M. (2005). Classical closed population models. In B. Manly, T. McDonald and S. Amstrup (Eds.). *The Handbook of Capture-Recapture Analysis* (pp. 22–35): Princeton University Press.
21. Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000). Estimating the number of shared species in two communities. *Statist. Sinica*, **10**, 227–246.
22. Chao, A. and Jost. L. (2012). Coverage-based rarefaction: standardizing samples by completeness rather than by sample size. Manuscript.
23. Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Ass.*, **87**, 210–217.
24. Chao, A., and Lin, C.-W. (2012). Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. To appear in *Biometrics*.
25. Chao, A., Lin, Shang-Yi, and Hsieh, T. C. (2012). User's Guide for Program iNEXT (interpolation-extrapolation).

Available at: http://chao.stat.nthu.edu.tw/.

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

27. Chao, A., and Shen, T.-J. (2004). Nonparametric prediction in species sampling. *J. Agric. Bio. Environ. Stat.*, **9**, 253–269.

28. Chao, A., Shen, T.-J., and Hwang, W.-H. (2006). Application of Laplace's boundary-mode approximations to estimate species and shared species richness. *Aust. N. Z. J. Stat.*, **48**, 117–128.

29. Chao, A. and Shen, T. J. (2010). User's Guide for Program SPADE (Species Prediction And Diversity Estimation). Available at: http://chao.stat.nthu.edu.tw/.

30. Chao, A., Tsay, P. K., Lin, S.-H., Shau, W.- Y. and Chao, D.- Y. (2001). Tutorial in biostatistics: the applications of capture-recapture models to epidemiological data. *Stat. Med.*, **20**, 3123–3157.

31. Chao, A. and Yang, H. C. (2006). User's Guide for Program CARE-2 (Capture-Recapture, second part). Available at: http://chao.stat.nthu.edu.tw/.

32. Chiarucci, A., Bacaro, G., Rocchini, D., and Fattorini, L. (2008). Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Community Ecol.*, **9**,121–123.

33. Coleman, B. D. (1981). On random placement and species-area relations. *Math. Biosci.*, **54**, 191–215.

34. Colwell, R. K. (2011). Estimates: Statistical Estimation of Species Richness and Shard Species from Samples. Version 9. User's guide and application published at http://purl.oclc.org/estimates.

35. Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., et al. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblage. *J. Plant Ecol.*, **5**, 3–21.

36. Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. Royal Soc., London, Series B*, **345**, 101–118.

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

38. Coull, B. A., and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, **55**, 294–301.

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

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

41. Edwards, W. R. and Eberhardt, L. (1967). Estimating cottontail abundance from livetrapping data. *J. Wildl. Manag.*, **31**, 87–96.

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

43. Eren, M. I., Chao, A., Hwaung, W.-H. and Colwell, R. K. (2012). Estimating the richness of a population when the maximum number of classes is fixed: a nonparametric solution to an archaeological problem. Under revision, *PLoS One*.

44. Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, **12**, 42–58.

45. Flather, C. H. (1996). Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *J. Biogeogr.*, **23**, 155–168.

46. Foissner, W., Agatha, S., and Berger, H. (2002). Soil Ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa),

with emphasis on two contrasting environments, the Etosha Region and the Namib Desert. *Denisia*, **5**, 1–1063.

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

48. Good, I. J. (2000). Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma.*J. Statist. Comput. Simul.*, **66**, 101–111.

49. Good, I. J. and Toulmin, G. (1956). The number of new species and the increase of population coverage when a sample Is increased. *Biometrika*, **43**, 45–63.

50. Gotelli, N. J. and Chao, A. (2012). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. Levin, S. A. (Ed).*The Encyclopedia of Biodiversity*, 2nd Edition, Elsevier, New York.

51. Gotelli, N. J., and Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, **4**, 379–391.

52. Gotelli, N. J., and Colwell., R. K. (2011). Estimating species richness. In *Biological Diversity: Frontiers in Measurement and Assessment* pp. 39–54 (eds A. Magurran and B. McGill) Oxford: Oxford University Press.

53. Hill, M. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, **54**, 427–432.

54. Holst, L. (1981). Some asymptotic results for incomplete multinomial or Poisson samples. *Scand. J. Statist.*, **8**, 243–246.

55. Hong, S. H., Bunge, J., Jeon, S. O., and Epstein, S. S. (2006). Predicting microbial species richness. *Proc. Natl. Acad. Sci.*, **103**, 117–122.

56. Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epid. Reviews*, **17**, 243–264.

57. Huang, S. P. and Weir, B. S. (2001). Estimating the total number of alleles using a sample coverage method. *Genetics*, **159**, 1365–1373.

58. Hurlbert, S. H. (1971). The Nonconcept of species diversity: A critique and alternative parameters. *Ecology*, **52**, 577–586.

59. Janardan, K. G. and Schaeffer, D. J. (1981). Methods for estimating the number of identifiable organic pollutants in the aquatic environment. *Water Resources Res.*, **17**, 243–249.

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

61. Janzen, D. H. (1973). Sweep samples of tropical foliage insects: Effects of seasons, vegetation types, elevation, time of day, and insularity. *Ecology*, **54**, 687–708.

62. Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.

63. Jost, L. (2010). The relation between evenness and diversity. *Diversity*, **2**, 207–232.

64. Jost, L., Chao, A. and Chazdon, R. L. (2011). Compositional similarity and beta diversity. In *Biological Diversity: Frontiers in Measurement and Assessment* pp. 66–84 (eds A. Magurran and B. McGill) Oxford: Oxford University Press.

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

66. Lukacs, P. M. and Burnham, K. P. (2005). Review of capture-recapture methods applicable to noninvasive genetic sampling. *Mol. Ecol.*, **14**, 3909–3919.

67. Magurran, A. E. (2004). *Measuring Biological Diversity* , Oxford: Blackwell.

68. Magurran, A. E. and McGill, B. J. (2011). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford: Oxford University Press.

69. Mao C. X. and Colwell, R. K. (2005). Estimation of species richness: mixture models, the role of rare species, and inferential challenges. *Ecology*, **86**, 1143–1153.

70. Mao C. X. , and Lindsay BG (2007). Estimating the number of classes. *Ann. Stat.*, **35**, 917–930.

71. Mao, C. X. and You, N. (2009). On comparison of mixture models for closed population capture-recapture studies. *Biometrics*, **65**, 547–553.

72. Miller, C. R., Joyce, P. and Waits, L. P. (2005). A new method for estimating the size of small populations from genetic mark-recapture data. *Mol. Ecol.*, **14**, 1991–2005.

73. Miller, R. I., and R. G. Wiegert. (1989). Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, **70**, 16–22.

74. Nayak, T. K. (1991). Estimating the number of component processes of a superimposed process. *Biometrika*, **78**, 75–81.

75. Norris III, J. L. and Pollock, K. H. (1998). Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environ. Ecol. Statist.*, **5**, 391–402.

76. O'hara, R. B. (2005). Species richness estimators: how many species can dance on the head of a pin? *J. Anim. Ecol.*, **74**, 375–386.

77. Ord, J. K. and Whitmore, G. A. (1986). The Poisson-inverse Gaussian distribution as a model for species abundance. *Commun. Statist.-Theory Methods*, **15**, 853–871.

78. Pan, H. Y., Chao, A., and Foissner, W. (2009). A non-parametric lower bound for the number of species shared by multiple communities. *J. Agric. Bio. Environ. Stat.*, **14**, 452–468.

79. Pielou, E. C. (1977). *Mathematical Ecology.* Wiley, New York.

80. Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, **56**, 434–442.

81. Pollock, K. H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *J. Amer. Statist. Ass.*, **86**, 225–238.

82. Preston, F. W. (1948). The commonness and rarity of species. *Ecology*, **29**, 254–283.

83. Rodrigues J., Milan L. A., and Leite, J. G. (2001). Hierarchical Bayesian estimation for the number of species. *Biometrical J.*, **43**, 737–746.

84. Sanathanan, L. (1977). Estimating the size of a truncated sample. *J. Amer. Statist. Ass.*, **72**, 669–672.

85. Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *Am. Naturalist*, **102**, 243–282.

86. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M. and Hollister, E. B. (2009). Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

87. Schwarz, C. J. and Seber, G. A. F. (1999). A review of estimating animal abundance III. *Stat. Sci.*, **14**, 427–456.

88. Seber, G. A. F. (1982). *The Estimation of Animal Abundance (2nd Edition)*, Griffin, London.

89. Shen, T.-J., Chao, A., and Lin, J.-F. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology*, **84**, 798–804.

90. Shinozaki, K. (1963). Notes on the species-area curve, 10th Annual Meeting of the Ecological Society of Japan (Abstract). p.5.

91. Sichel, H. S. (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. *S. Afri. Statist. J.*, **31**, 13–37.

92. Smith, W. and Grassle, J. F. (1977). Sampling properties of a family of diversity measures. *Biometrics*, **33**, 283–292.

93. Turner, W., Leitner, W., and Rosenzweig, M. (2001). WS2m: software for the measurement and analysis of species diversity. http://eebweb.arizona.edu/diversity/.

94. Wang J.P. (2010). Estimating the species richness by a Poisson-compound gamma model. *Biometrika*, **97**, 727–740.

95. Wang, J.P. (2011). SPECIES: An R package for species richness estimation. *J. Stat. Software*, **40**, 1–15.

96. Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and Management of Animal Populations.* Academic Press, San Diego, CA.

Anne Chao and Chun-Huo Chiu
Institute of Statistics
National Tsing Hua University
Hsin-Chu, Taiwan
chao@stat.nthu.edu.tw