Chao, A. and Chiu, C. H. (2016). Species richness: estimation and comparison. Wiley StatsRef: Statistics Reference Online. 1-26.

Species Richness: Estimation and Comparison

Anne Chao and Chun-Huo Chiu

INSTITUTE OF STATISTICS, NATIONAL TSING HUA UNIVERSITY, HSIN-CHU, TAIWAN

Keywords: abundance data, extrapolation, incidence data, interpolation, rarefaction, sample coverage, species richness, standardization

Abstract: On the basis of the sampling data from an assemblage, estimation of species richness (observed plus undetected) is statistically difficult especially for highly-diverse assemblages with many rare species. Simple counts of species richness in samples typically underestimate and strongly depend on sampling effort and sample completeness. There are two approaches to infer species richness and make fair comparisons among multiple assemblages based on possibly unequal-sampling effort and incomplete samples that miss many species. (1) An asymptotic approach: this approach compares the estimated asymptotes of species accumulation curves. It is based on statistical sampling-theory methods of estimating species richness. Both parametric and nonparametric methods are reviewed. We focus on the nonparametric estimators which are universally valid for all species abundance distributions. (2) A non-asymptotic approach: this approach compares the estimated species richnesses of standardized samples with a common finite sample size or sample completeness. It is based on the seamless sample-size- and coverage-based rarefaction and extrapolation sampling curves. This approach aims to compare species richness estimates for equally-large or equally-complete samples. These two approaches allow researchers to efficiently use all data to make robust and detailed inferences about species richness. Two R packages (SpadeR and iNEXT) are applied to rainforest tree data for illustration.

Species richness (i.e., the number of species) is the simplest, most intuitive and most frequently used measure for characterizing the diversity of an assemblage (*see* **Diversity measures**). Species richness possesses intuitive mathematical properties, and features prominently in foundational models of community ecology. 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. In this article, we focus on the estimation and comparison of species richness based on sampling data from each of the focal assemblages. The topic is important for understanding the causes and processes of biodiversity, for assessing the effects of human disturbance on biodiversity, and for making environmental policy decisions. See Refs 1–7 for a background and developments and applications on this topic.

In nearly all biodiversity studies, however, the compilation of complete species census and inventories often requires extraordinary efforts and is an almost unattainable goal in practical applications. There are undiscovered species in almost every taxonomic survey or species inventory. Consequently, the simple count of species (empirical or observed richness) in a sample underestimates the true species richness (observed plus undetected), with the magnitude of the negative bias possibly substantial. In addition, empirical richness strongly depends on sampling effort and thus also depends on sample completeness. Generally, there are two approaches (an asymptotic approach via species richness estimation and a non-asymptotic approach via rarefaction and extrapolation) to infer species richness and make fair comparisons among multiple assemblages based on possibly unequal-sampling effort and incomplete samples that miss many species.

First, the asymptotic approach aims to estimate the asymptote of a species accumulation curve. Then the estimated asymptote is used as a species richness estimate which can be compared across assemblages. This approach is based on statistical sampling-theory methods of estimating species richness. Both parametric and nonparametric sampling-theory-based estimation methods are reviewed. We focus on the nonparametric estimators which are universally valid for all species abundance distributions.

Species richness estimation based on sampling data has a long history in various disciplines. In general contexts, "species" can be defined in a broad sense: they may be biological species, individuals of a target population, patients/cases in epidemiology and medical sciences, bugs in software programs, words in a book, genes or alleles in genetic code, or other discrete entities. Thus, the topic of species richness estimation and comparison has had a wide range of applications not only in biological sciences but also in many other disciplines. This cross-discipline topic has been extensively discussed in the literature[1,5,7].

We briefly review in Section 2 the traditional non-sampling-based methods and their drawbacks. The 1943 paper by **Fisher**, Corbet and Williams[8] provided the mathematical foundation on statistical sampling approaches to estimate species richness. Since then, an enormous number of models and methods based on statistical sampling theory have been proposed in the literature to estimate species richness. In general, there are two frameworks: parametric and nonparametric.

Their relative merits and all details are reviewed in Section 2 after an introduction of two types of data (abundance data and incidence data) is presented.

Second, the non-asymptotic approach aims to compare species richness estimates for equally-large or equally-complete samples. It compares the estimated species richness of standardized samples with a common finite sample size or sample completeness (as measured by sample coverage; see later text for details). The earliest development of standardizing sample size for abundance data by rarefaction was proposed in a series of papers by Sanders and his followers[9-13], but see Chiarucci *et al*.[14] for a historical review. Ecologists typically use rarefaction to down-sample the larger samples until they are the same size as the smallest sample, and then compare the richnesses of these equally-large samples, but this implies that some data from the larger samples are thrown away. To avoid discarding data, Colwell *et al*.[15] proposed using a sample-size-based rarefaction (interpolation) and extrapolation (prediction) sampling curve that can be rarefied to smaller sample sizes or extrapolated to larger sample sizes. Chao and Jost[16] proposed standardizing samples by a given degree of sample completeness rather than size. The authors developed a coverage-based rarefaction and extrapolation methodology.

We review in Section 3 the sample-size- and coverage-based integration of rarefaction and extrapolation sampling curves of species richness. These two types of rarefaction and extrapolation represent a unified standardization method for quantifying and comparing species richness across multiple assemblages. When the sample is nearly complete (i.e., sample size is sufficiently large and sample completeness approaches be unity), the estimates of the non-asymptotic approach tend to those in the corresponding asymptotic approach.

In Section 4, we use real data to demonstrate the application of the R package SpadeR (Species-richness Prediction And Diversity Estimation in R) to obtain nonparametric species richness estimates. We also illustrate the use of the R package iNEXT (iNterpolation/EXTrapolation) to obtain the sample-size- and coverage-based integrated rarefaction and extrapolation sampling curves. These methods allow researchers to efficiently use all available data to make more robust and detailed inferences about species richness of the sampled assemblages, and also to make objective comparisons of species richness across assemblages.

1 Two Types of Data

We generally follow the notation and terminology used in Colwell *et al.*[15] and Chao et *al.*[17] 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

$N = \sum_{i=1}^{S} N_i$ be the total population size. The relative abundance of species i is $p_i = N_i/N$, so that $\sum_{i=1}^{S} p_i = 1$. Here N, S, N_i, and p_i represent the true but unknown underlying parameters of the assemblage. We distinguish between two sampling data structures.

1.1 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 its 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 observed sample is referred to as a *reference sample*. This type of data can be obtained by two different sampling schemes. (i) Discrete-type sampling in which the sampling unit is an individual. For example, we sample a fixed number of n individuals in a study area. Here sample size n is fixed by design and each species can be represented by at most n individuals. (ii) Continuous-type sampling in which sampling efforts are measured in a continuous scale such as time, area or water volume. For example, we sample a fixed area or a fixed amount of time in a study site. 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 $\sum_{i=1}^{S} X_i = n$. Let f_k (*abundance frequency counts*), $k = 0, 1, ..., n$, 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 reference 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.

1.2 Incidence Data

In some surveys, the sampling unit is 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 variations: 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 (detection or non-detection) 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, fixed period of time, fixed number of traps, investigator, and so on.

The *reference sample for incidence data* consists of a set of T sampling units. The detection or non-detection 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 detected in the jth sampling unit, and zero if it is not detected. The row sum of the incidence matrix $Y_i = \sum_{j=1}^{T} W_{ij}$ 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 frequency 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 incidence matrix are the *incidence-based frequency counts* $(Q_1, Q_2, ..., Q_T)$, 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. Here 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. Let U be the total number of incidences in the matrix. We have $U = \sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$. Since $S = S_{obs} + Q_0$, species richness estimation is equivalent to the inference about the number of undetected species Q_0.

2 Asymptotic Approach: Species Richness Estimation

The earliest attempts to study communities started with finding the relationship between species richness and the area that the survey covered. A classic species-area or *species accumulation curve* (or collector's curve, species-cover curve) is a plot of the accumulated number of species found with respect to the number of units of effort expended. The effort may correspond to either a continuous type (area, trap-time, volumes) or a discrete-type (such as individuals, sampling occasions, quadrats, number of nets). This curve, as a function of effort, monotonically increases and typically approaches an asymptote, which is the total number of species. An asymptotic approach refers to the estimation of the asymptote of a species accumulation curve.

The traditional curve-fitting approach uses parametric curves to fit a species-accumulation or species-area curve to predict its asymptote, which is used as an estimate of species richness. Among the proposed asymptotic functions are the negative exponential function, the **Weibull** function, the **logistic** function, and the **Michaelis-Menten equation**[1,18]. Although intuitive, this approach does not directly use information on the frequencies of common and rare species, but rather only uses presence data to forecast the shape and asymptote of the rising curve.

Another type of curve-fitting approach involves fitting a parametric species abundance distribution or functional form to the observed species frequencies to obtain an estimate of species richness. The earliest such approach was proposed by Preston[19], who fitted a log-normal curve to the (properly grouped) observed frequencies in order to estimate the portion of the assemblage below a lower limit of observed abundance that he called the "veil line." Then the integrated value of the fitted curve over the real line can be used as an estimate of species richness. Other zero-truncated distributions (e.g., negative binomial, geometric, Zipf-Mandelbrot, logarithmic) can also be applied[2]. Although this approach uses information on the frequencies of common and rare species, it simply fits a curve to the observed frequency data.

A major problem with the curve-fitting approaches is that they are not based on any statistical sampling model, so the variances of the resulting asymptotes cannot be evaluated without imposing further assumptions. Thus, rigorous and statistical comparisons of estimators among assemblages cannot be made. Another problem is that several different functional forms may fit the same data set equally well, yet yield drastically different estimates of the asymptote, implying theoretical difficulties for the selection of a proper distribution or functional form. Therefore, we mainly focus on the sampling-theory-based methodologies which are presented below separately for abundance data and incidence data.

2.1 Species Richness Estimation for Abundance Data

We first discuss the discrete-type sampling in which the sampling unit is an

individual; see Section 1.1 for notation and terminology. Suppose n individuals (with n fixed in advance) are independently observed from the study site. A commonly used model is the multinomial model. That is, the observed species frequencies $(X_1, X_2, \ldots, X_S)$ for given S and the relative abundances $(p_1, p_2, \ldots, p_S)$ follow a **multinomial distribution**:

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

Here the species detection probability for the ith species is simply its relative abundance. All our inference procedures are derived from the following marginal distribution of the sample frequency X_i:

$$P(X_i = x_i) = \binom{n}{x_i} p_i^{X_i} (1 - p_i)^{n - x_i} . \quad (1b)$$

A more general model allows the detectability of individuals to vary with species. The detectability of individuals is determined by many possible factors such as individual movement patterns, color, size, and vocalizations. This general model assumes that the detectability of any individual of the ith species is $\theta_i > 0$, which varies with species. It also assumes that the species detection probability of the ith species is proportional to the product of this species abundance N_i and the detectability θ_i of any individual of the same species. Under this general model, the species detection probability for the ith species in any observation becomes $\psi_i = N_i\theta_i / \sum_{k=1}^{S} N_k\theta_k = p_i\theta_i / \sum_{k=1}^{S} p_k\theta_k$, $i = 1, 2, \ldots, S$. Thus, a more general setting is the following model that allows for heterogeneous individual detectability:

$$P(X_1 = x_1, \ldots, X_S = x_S) = \frac{n!}{x_1! \ldots x_S!} \psi_1^{x_1} \psi_2^{x_2} \ldots \psi_S^{x_S} . \quad (1c)$$

where the detection probability ψ_i is the normalized product of species abundance and individual detectability. Under the special case that all individuals have the same detectability, model (1c) reduces to model (1a). The two models, (1a) and (1c), are identical in structure, implying that the two models are equivalent in the sense that all inference procedures are the same.

In discrete-type sampling, it is assumed that the sampling procedure itself does not substantially alter the species detection probabilities $(\psi_1, \psi_2, \ldots, \psi_S)$. This assumption is fulfilled if individuals are sampled with replacement so that any individual can be repeatedly observed. If sampling is done without replacement, meaning that any individual can only be observed at most once in the sample, then a hypergeometric model is more appropriate as will be discussed in Section 2.1.2. 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.

Next, consider the continuous-type sampling scheme. Assume that the

assemblage is surveyed through continuous-type sampling efforts and that the total amount of 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 values from 0 to infinity. This approach can be traced back to Fisher *et al.*[8], who assumed that individuals of the ith species arrive in a sample according to a Poisson process with mean species occurrence or detection rate $A\lambda_i$, where λ_i represents the mean rate per unit of effort.

In some continuous-time sampling data, the exact arrival times for each individual are available, but in most biological surveys, only the frequencies of discovered species are recorded, and these frequencies would be sufficient for estimating species richness. 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:

$$P(X_1 = x_1,..., X_S = x_S) = \prod_{i=1}^{S} (A\lambda_i)^{x_i} \frac{\exp(-A\lambda_i)}{x_i!}. \qquad (2)$$

Although n is a random variable, we can consider the conditional distribution of the frequencies $(X_1, X_2,..., X_S)$ given $n = \sum_{k=1}^{S} X_k$. 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, …, 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 abundance N_i and individual detectability θ_i, then this conditional multinomial distribution is identical to the model given in Equation (1c). 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 area-based model[20] is basically a special case of the product-Poisson model. Coleman considered that the observed 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 the same as that given in Equation (2).

2.1.1 Parametric Models

Fisher *et al.*[8] adopted a parametric approach in their pioneering work on species richness estimation. In this approach, one assumes a parametric distribution $f(\lambda;\theta)$, where θ denotes a vector of parameters for the species detection rates $(\lambda_1, \lambda_2,..., \lambda_S)$ in the product-Poisson model (2) or for the species (relative) abundances in the multinomial model. Most parametric approaches are

based on the product-Poisson model under a continuous-type sampling-effort framework. When S is large, the large number of parameters $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ makes inference problems statistically difficult to deal with. 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 conventional inference procedures can be applied. This is a major advantage of parametric models.

If the distribution $f(\lambda;\theta)$ is a degenerate distribution with all probabilities at a fixed point λ, then this reduces to a homogeneous model (i.e., equal detection rates for all species) with $\lambda_1 = \lambda_2 = \ldots = \lambda_S \equiv \lambda$. 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[3]. An approximate maximum likelihood estimator (MLE) under the homogeneous model is the solution $\hat{S}$ of the following equation:

$$S_{obs} = \hat{S}[1 - \exp(-n/\hat{S})], \tag{3}$$

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

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

To formulate the parametric theory under a general distribution $f(\lambda;\theta)$, we first construct the likelihood function of S and θ based on both observed and undetected species. For any mixing density $f(\lambda;\theta)$, define $p_\theta(k)$, $k = 0, 1, \ldots$ as the probability that any species is observed k times in the sample. Then from Equation (2) we have

$$p_\theta(k) = \int_0^\infty (A\lambda)^k \frac{\exp(-A\lambda)}{k!} f(\lambda;\theta)d\lambda, \; k = 0, 1, \ldots. \tag{4}$$

and $E(f_k) = S\,p_\theta(k)$. 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 θ from all species can be written as

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

On the basis of the above likelihood, 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 (5a). A conditional (on S_{obs}) maximum likelihood estimator (CMLE) is often more convenient to obtain as follows.

Note that likelihood (5a) can be factored as $L(S,\theta) = L_b(S,\theta)L_c(\theta)$, where

$$L_b(S,\theta) = \frac{S!}{(S-S_{obs})!S_{obs}!}\left(1-p_\theta(0)\right)^{S_{obs}}\left(p_\theta(0)\right)^{N-S_{obs}}, \qquad (5b)$$

and

$$L_c(\theta) = \frac{S_{obs}!}{\prod_{k\geq 1} f_k!}\prod_{k\geq 1}\left(\frac{p_\theta(k)}{1-p_\theta(0)}\right)^{f_k}. \qquad (5c)$$

Here $L_b(S,\theta)$ is a likelihood for a binomial (S, $1-p_\theta(0)$), which is the distribution of S_{obs}; $L_c(\theta)$ is a multinomial likelihood with respect to $\{f_k; k\geq 1\}$ with cell total S_{obs} and zero-truncated cell probabilities $p_\theta(k)/[1-p_\theta(0)]$, $k\geq 1$. The first likelihood $L_b(S,\theta)$ results in the CMLE[21] $\hat{S}_{CMLE} = S_{obs}/[1-p_{\hat{\theta}}(0)]$, where $\hat{\theta}$ maximizes the second likelihood $L_c(\theta)$. 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. Extensive iterative procedures are often required to find the UMLE and CMLE, and in some cases the iterative steps fail to converge properly and thus the UMLE or CMLE may not be obtainable.

Fisher *et al.*[8] adopted a two-parameter gamma distribution with $\theta = (\tau, \beta)$ 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 $p_\theta(k)$, or equivalently $E(f_k)$, $k = 0$, 1, 2, ..., correspond to individual terms of a negative-binomial distribution.

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

In the special case of $\tau = 1$ (i.e., $f(\lambda;\theta)$ is an exponential distribution), the model is equivalent to a broken-stick model[22]. In this case, the $p_\theta(k)$, $k = 0$, 1, 2, ..., correspond to the terms of a geometric distribution.

Fisher *et al.*[8] considered the extreme case in which τ tends to 0, i.e., the degree of heterogeneity among species detection rates tends to infinity. In this extreme case, $p_\theta(k) \to x^k/\{k[-\log(1-x)]\}$, where $x = \beta/(1+\beta)$. This implies the well-known Fisher's log-series model: $E(f_k) = S\,p_\theta(k) \to \alpha\,x^k/k$, where $\alpha = S/[-\log(1-x)]$. However, this model does not yield an estimate of species richness[22]. On the basis of the sampling data, Fisher's α and the parameter x in the log-series model are simply solved from the two equations in terms of sample size and observed numbers of species: $S_{obs} = -\alpha\log(1-x)$ and $n = \alpha\,x/(1-x)$. Thus, two data sets with the same sample size and observed numbers of species would result in the same value of Fisher's α. Fisher's α completely ignores the sample

species frequencies, although it has been used as a diversity measure in the literature.

Other parametric models for $f(\lambda;\theta)$ include the log-normal[23], inverse-Gaussian[24], and generalized inverse-Gaussian[25]. 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[26], but this is rarely the case for empirical data.

One can also assume a parametric distribution for the species relative abundances $(p_1, p_2, \ldots, p_S)$ in the multinomial model (1a) to characterize the theoretical patterns. The most popular functional forms include the geometric $p_i \propto \alpha(1-\alpha)^{i-1}$ and the Zipf-Mandelbrot law $p_i \propto (i+\alpha)^{-\theta}$, where α and θ are parameters. Although these types of models can produce species richness estimates[7], 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[27].

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. In addition, these approaches do not perform well in comparisons with empirical or simulated data sets[5,6,13]. Most importantly, when there are multiple assemblages, the parametric approach does not permit meaningful comparisons of assemblages with different distribution functions. For example, a log-normal assemblage cannot be compared to an assemblage whose species-rank distribution follows a geometric series. A practical problem, as noted earlier for obtaining CMLE and UMLE, is that in some cases the iterative steps fail to converge properly and thus, species richness estimates may not be obtainable. A related issue is that it is almost impossible to generalize the parametric approach to incorporate species' evolutionary histories or functional differences among species based on species traits[28].

2.1.2 Non-parametric Models

The non-parametric approach, which makes no assumptions about the mathematical form of the underlying distributions of species abundance or species detection rates, avoids the above-mentioned drawbacks and is more robust in applications. In the following, we review three types of analytic nonparametric estimators which are universally valid for all species abundance distributions and allow for comparison among multiple assemblages. An intuitive and basic concept in non-parametric species richness estimation is that abundant species (which are certain to be detected in samples) contain almost no information about the undetected species richness, whereas rare species (which are likely to be either

undetected or infrequently detected) contain almost all the information about the undetected species richness. Therefore, most nonparametric estimators of the number of undetected species are based on the lower-order frequency counts, especially the numbers of singletons and doubletons for abundance data.

Chao1-type estimators

When there are many undetectable or "invisible" species in a highly-diverse assemblage, it is statistically impossible to obtain a good estimate of species richness. Therefore, an accurate lower bound for species richness is often of more practical use than an imprecise point estimate. Chao[29,30] derived a lower bound of undetected species richness in terms of the numbers of singletons and doubletons; the corresponding lower bound of species richness given below is referred to as the *Chao1 estimator*[1]:

$$\hat{S}_{Chao1} = \begin{cases} S_{obs} + [(n-1)/n][f_1^2/(2f_2)], & \text{if } f_2 > 0 \\ S_{obs} + [(n-1)/n] f_1(f_1-1)/2, & \text{if } f_2 = 0 \end{cases} \tag{6a}$$

$$\approx \begin{cases} S_{obs} + f_1^2/(2f_2), & \text{if } f_2 > 0 \\ S_{obs} + f_1(f_1-1)/2, & \text{if } f_2 = 0 \end{cases}$$

The above estimator is valid under both the multinomial and product-Poisson models discussed in Section 2.1. A simple analytic variance estimator (if $f_2 > 0$)[30] is:

$$\hat{\text{var}}(\hat{S}_{Chao1}) = f_2\left[\frac{k}{2}\left(\frac{f_1}{f_2}\right)^2 + k^2\left(\frac{f_1}{f_2}\right)^3 + \frac{1}{4}k^2\left(\frac{f_1}{f_2}\right)^4\right],$$

where $k = 1 - 1/n$. If $f_2 = 0$, the above variance formula is modified to:

$$\hat{\text{var}}(\hat{S}_{Chao1}) = \frac{kf_1(f_1-1)}{2} + \frac{k^2 f_1(2f_1-1)^2}{4} - \frac{k^2 f_1^4}{4\hat{S}_{Chao1}}.$$

A confidence interval of species richness based on the Chao1 estimator can be constructed using an asymptotic variance and a log-transformation[30,31] so that the lower bound of the interval is not less than S_{obs}. In the special case of homogeneous case (i.e., all species detection probabilities or rates are equal), a bias-corrected estimator (referred to as *Chao1-bc* estimator) is

$$\hat{S}_{Chao1-bc} = S_{obs} + [(n-1)/n] f_1(f_1-1)/[2(f_2+1)]. \tag{6b}$$

Although the Chao1 estimator is derived as a lower bound of species richness, it generally works satisfactorily as a point estimator when an undetected species in the sample has approximately the same chance of being detected as a singleton[32]. This condition is satisfied if the sample size is very large or the *rare* species are nearly homogeneous in terms of detection probabilities; in the latter case, other species could be highly heterogeneous.

An improved lower bound, which makes use of the additional information of tripletons and quadrupletons to estimate undetected species richness, was recently derived by Chiu *et al*.[33] The corresponding lower bound of species richness is referred to as *iChao1 estimator* (here the sub-index *i* stands for "improved"):

$$\hat{S}_{iChao1} = \hat{S}_{Chao1} + \frac{(n-3)}{n}\frac{f_3}{4f_4} \times \max\left(f_1 - \frac{(n-3)}{(n-1)}\frac{f_2 f_3}{2f_4},\ 0\right) \qquad (6c)$$

$$\approx \hat{S}_{Chao1} + \frac{f_3}{4f_4} \times \max\left(f_1 - \frac{f_2 f_3}{2f_4},\ 0\right).$$

They also provided an analytic variance estimator to construct the associated confidence intervals.

Chao and Lin[34] extended the Chao1 estimator to deal with data based on sampling without replacement, i.e., sampling units cannot be repeatedly observed. The model in Equation (1a) is thus modified to the following generalized hypergeometric distribution with population size *N*,

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

The Chao1 estimator under this model is generalized to

$$\hat{S}_{Chao1.wor} = S_{obs} + \frac{f_1^2}{\frac{n}{n-1}2f_2 + \frac{q}{1-q}f_1}, \qquad (6d)$$

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 or the proportion of sampled area). 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 all individuals are observed, so that *q* approaches 1, $q/(1-q)$ approaches infinity and our lower bound reduces to the number of observed species, which equals the true parameter in this special case.

Coverage-based estimators (ACE-type estimators)

The *ACE* (Abundance-based Coverage Estimator) of species richness is based on the concept of "*sample coverage*" (or simply "*coverage*"), which was originally developed for cryptographic analyses during World War II by the founder of modern computer science, Alan Turing, and his colleague I. J. Good[35,36,37]. Under Model (1a), the coverage of a sample is interpreted as the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample. Mathematically, for an observed sample of size *n* with species frequencies (X_1, X_2, …, X_S), the sample coverage can be expressed as

$C = \sum_{i=1}^{S} p_i I(X_i > 0)$, where $I(A)$ is the indicator function, i.e., $I(A) = 1$ if the event A occurs, and 0 otherwise. See Section 3.1.2 for more details.

Sample coverage is an objective measure of the degree of sample completeness and can be very accurately and efficiently estimated using only information contained in the reference sample itself, as long as the sample size is reasonably large, as shown by Turing [36,37]. His estimator is surprisingly elegantly simple: it is just the complement of the proportion of singletons. Turing's sample coverage estimator is very efficient[38] and has found wide applications in various research fields. Chao and Lee[39] applied it to develop the ACE approach. See Chao and Jost[16] for a review on other applications.

The ACE model assumes that the species relative abundances (p_1, p_2, …, p_S) are fully characterized by their mean $\overline{p} = 1/S$ and CV (coefficient of variation), where 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 among species abundances. The larger the CV is, the greater will be the degree of heterogeneity. The CV vanishes if and only if all species have the same abundances (i.e., the assemblage is homogeneous).

To apply the concept of sample coverage to species richness estimation, a cut-off value κ is needed to separate species frequencies into "rare" (frequency $\leq \kappa$) and "abundant" (frequency > κ) groups. The cut-off κ =10 works well for many empirical data sets. For highly heterogeneous communities such as bacterial or microbial sequencing data, an alternative choice is κ = max (10, n/S_{obs})[5]. The reason for a cut-off point is that abundant species carry almost no information about undetected species. In addition, the parameter CV is statistically hard to estimate when there are very abundant species; restriction to the rare species group helps reduce the magnitude of CV so that a more accurate estimate of the CV can be obtained.

Let the total number of observed species in the abundant species group be $S_{abun} = \sum_{i>\kappa} f_i$ and the number of observed species in the rare species group be $S_{rare} = \sum_{i=1}^{\kappa} f_i$. Because detected rare species contain nearly all the information about the undetected species, the ACE approach estimates the number of undetected species using information from the rare species group. Let $n_{rare} = \sum_{i=1}^{\kappa} i f_i$ be the sample size for the rare species group. Turing's coverage estimate for this group is $\hat{C}_{rare} = 1 - f_1/n_{rare}$, which measures the sample completeness of the subsample restricted to rare species. In the special case of homogeneous abundances for rare species (CV = 0), the coverage-based estimator[40] is:

$$\hat{S}_0 = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}}. \qquad (7a)$$

The basic idea of the ACE[3,39] is to adjust the estimator in (7a) by accounting for heterogeneity. The resulting ACE is expressed as

$$\hat{S}_{ACE} = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\hat{\gamma}^2_{rare}, \tag{7b}$$

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

$$\hat{\gamma}^2_{rare} = \max\left\{\frac{S_{rare}}{\hat{C}_{rare}}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{(\sum_{i=1}^{\kappa} if_i)(\sum_{i=1}^{\kappa} if_i - 1)} - 1,\ 0\right\}. \tag{7c}$$

For species-rich and highly heterogeneous assemblages (e.g., species richness > 1000 and estimated CV for the whole data > 2), the estimator $\hat{\gamma}_{rare}$ in Equation (7c) and the resulting ACE generally underestimate. In such cases, a modified estimator, ACE-1, was derived[39]. An approximate variance for the ACE and ACE-1 can be obtained using standard statistical approximation theory.

Jackknife estimators

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 behind the *j*-th order jackknife method is to consider sub-data by successively deleting *j* individuals from the data. Despite the fact that Cormack[41] implied the jackknife method does not have a theoretical basis for bias reduction of species richness estimation, the first two orders of jackknife estimators are widely used in various fields. The first-order jackknife is expressed as

$$\hat{S}_{jk1} = S_{obs} + \frac{n-1}{n} f_1 \approx S_{obs} + f_1. \tag{8a}$$

This estimator implies that the number of undetected species is approximately the same as the number of singletons. The second-order jackknife estimator has the form:

$$\hat{S}_{jk2} = S_{obs} + \frac{2n-3}{n} f_1 - \frac{(n-2)^2}{n(n-1)} f_2 \approx S_{obs} + 2f_1 - f_2\ . \tag{8b}$$

This estimator implies that the number of undetected species is approximately the same as the difference between $2f_1$ and f_2. 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[42,43].

Extensive simulations conducted by Chiu *et al.*[33] based on various species abundance models revealed that the two jackknife estimators typically underestimate when the sample size is relatively small, but exceed the true species richness and overestimate at larger sample sizes. Thus, there is a limited range of

sample sizes (near crossing points) where jackknife estimators are close to the true species richness. This is likely the reason why many studies found the jackknife estimators to have a relatively good performance. However, the theoretical behaviour is not predictable because the narrow range of good performance changes with each model. Outside this range, the two jackknife estimators may have appreciable biases. The jackknife estimators often exhibit counter-intuitive patterns: their bias, accuracy and coverage probability regularly do not improve as sample size increases, whereas the other non-parametric estimators presented in this section always improve.

2.2 Species Richness Estimation for Incidence Data

As indicated in Section 1.2, the reference sample for incidence data consists of a species-by-sampling-unit incidence matrix [W_{ij}]. Each element in the matrix corresponds to either detection or non-detection of a species. Most statistical estimation methods reviewed below for incidence data were originally developed for estimating population sizes in the context of **capture-recapture** research. In typical capture-recapture experiments, data consist of an individual-by-trapping-sample matrix with the elements of the matrix corresponding to either the capture or non-capture of an individual. Thus, there is a simple analogy between the incidence data in species richness estimation for a multiple-species assemblage and the capture-recapture data in population size estimation for a single species. An "individual" animal in capture-recapture studies corresponds to a "species" in species richness estimation. The estimating target in the former is population size whereas in the latter it is species richness. Consequently, the estimation techniques in the capture-recapture models 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, while in species richness estimation, species are easily classified from sighting. See Refs 44–46 for comprehensive reviews of methodology and applications, and see Refs 47–49 for short overviews specifically on population size estimation.

Following Colwell *et al*.[15], we adopt a *product-Bernoulli model*, which assumes that the *i*th species has its own unique *detection or incidence probability* π_i that is constant for any randomly selected sampling unit. The incidence probability π_i is the probability that species i is detected in a sampling unit. This model, in the context of capture-recapture research, is referred to as Model $\mathbf{M}_{\mathrm{h}}$, where the sub-index "h" denotes heterogeneity among individual capture probabilities[47]. Under the product-Bernoulli model, each element W_{ij} in the incidence matrix is a Bernoulli random variable with probability π_i that W_{ij}=1 and

probability $1-\pi_i$ that W_{ij}=0. The probability distribution for the incidence matrix can be expressed as

$$P(W_{ij}=w_{ij})=\prod_{i=1}^{S}\prod_{j=1}^{T}\pi_i^{w_{ij}}(1-\pi_i)^{1-w_{ij}}=\prod_{i=1}^{S}\pi_i^{y_i}(1-\pi_i)^{T-y_i}\ . \qquad (9a)$$

The marginal distribution for the incidence-based frequency Y_i for the i-th species follows a binomial distribution:

$$P(Y_i=y_i)=\binom{T}{y_i}\pi_i^{y_i}(1-\pi_i)^{T-y_i}\ . \qquad (9b)$$

Because of the resemblance between Models (1b) and (9b), the inference procedures for abundance and incidence data are generally parallel: the incidence-based frequency Y_i, the incidence probability π_i, and the number of sampling units T are analogous respectively to X_i, p_i and sample size n in the abundance data. As a result, the incidence-based frequencies counts (Q_1, Q_2, …, Q_T) are analogous to the frequency counts (f_1, f_2, …, f_n) in the abundance data.

A more general capture-recapture model is called Model $\mathbf{M}_{\mathrm{ht}}$ where the sub-index "h" denotes heterogeneity among individual capture probabilities and the sub-index "t" denotes time-varying effects[47]. Model $\mathbf{M}_{\mathrm{ht}}$ can also be adapted to estimate species richness for incidence data; it assumes that the detection probability of the i-th species in the j-th sampling unit is the product of the "heterogeneity" effect π_i and the sampling-unit effect v_j. Here, as in Model $\mathbf{M}_{\mathrm{h}}$, the heterogeneity effect means that each species has its own unique incidence rate π_i; the sampling unit effect v_j is closely related to some known and unknown factors for the j-th sampling unit, possible examples of which include sampling method/efforts, quadrat area, weather variability, sampler's capability and other environmental variables associated with each sampling occasion.

Since there are many factors that may be involved in the sampling-unit effects, the effects $\{v_1, v_2,..., v_T\}$ are usually modeled as random variables taken from an unknown probability density function $h(v)$. For given sampling-unit effects $\{v_1, v_2,..., v_T\}$, the probability distribution of the incidence matrix is an extension of Equation (9a):

$$P\{W_{ij}=w_{ij}, i=1,...,S, j=1,...,T \mid v_1, v_2,..., v_T\}=\prod_{i=1}^{S}\prod_{j=1}^{T}(\pi_i v_j)^{w_{ij}}(1-\pi_i v_j)^{1-w_{ij}}\ . \qquad (10a)$$

Integrating out all possible values of $\{v_1, v_2,..., v_T\}$, we obtain the following binomial model for the incidence-based frequency Y_i:

$$P(Y_i=y_i)=\binom{T}{y_i}\left[\pi_i\int vh(v)dv\right]^{y_i}\left[1-\pi_i\int vh(v)dv\right]^{T-y_i}\equiv\binom{T}{y_i}\mu_i^{y_i}(1-\mu_i)^{T-y_i}\ , \qquad (10b)$$

where $\mu_i = \pi_i \int vh(v)dv$. That is, the frequency Y_i under Model $\mathbf{M}_{ht}$ is a binomial random variable with detection probability μ_i. When there are no sampling-unit effects such that sample effects can be assumed to be identical to unity (e.g., equal-size quadrats, equal-effort sampling with similar protocols), Model $\mathbf{M}_{ht}$ reduces to Model $\mathbf{M}_h$. Note that the distributions in Models (9b) and (10b) are identical in structure, implying that the two models are equivalent in the sense that all inference procedures are the same. Without loss of generality, we only consider Model $\mathbf{M}_h$ and all estimators reviewed in the following two sections all based on the distribution given in Equation (9b).

2.2.1 Parametric Models

A commonly used parametric approach is the **beta-binomial** model, where the detection rate π_i in model (9b) is assumed to be a random sample from a **beta distribution**[43,50]. The likelihood is similar to that in Equation (4) with $P_\theta(k)$ replaced by a beta-binomial form. Therefore, the maximum likelihood or empirical Bayes estimation procedures can be similarly obtained. Then, based on Equations (5a) and (5b), 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 alternative parametric assumptions. Instead of using a multiplicative of heterogeneity effect π_i and sampling-unit effect v_j as in Model $\mathbf{M}_{ht}$, Huggins[51] proposed a logistic model which can be expressed as $\pi_i v_j /(1+\pi_i v_j)$; this is also known as the Rasch model in educational statistics. There are several approaches to this model including the log-linear approach, mixture models and latent class models[47]. The relevant covariates or auxiliary variables can be easily incorporated to explain heterogeneity effects in analysis.

As with the parametric models for abundance data, these approaches work well only when the specified parametric models are the true models. When this is in fact the case, standard inference estimation procedures involving numerical iterations can be applied to obtain species richness estimates and the associated confidence intervals. For example, in the latent class model, it works well only when the studied population actually contains groups of individuals that are thought to have different detection rates. However, such an assumption is not directly testable based on empirical data. Other advantage and drawbacks for the above parametric models are similar to those discussed for abundance data in Section 2.1.1.

2.2.2 Non-parametric Models

As with the abundance model, a major advantage of the non-parametric models is that they are valid for all types of distributions for the detection rates $\{\pi_1, \pi_2, ..., \pi_S\}$ in model (9b). Corresponding to the Chao1-type and the ACE-type, we have for incidence data the Chao2-type and ICE-type. The estimation procedures are generally parallel simply by replacing the sample size n and the capture frequency counts $(f_1, f_2, \ldots, f_n)$ in abundance data with the number of trapping samples T and the incidence-based frequency counts $(Q_1, Q_2, \ldots, Q_T)$, respectively.

Chao2-type estimators

For incidence data, the corresponding estimator of species richness is called the *Chao2 estimator*, the formula of which is [30]:

$$\hat{S}_{Chao2} = \begin{cases} S_{obs} + [(T-1)/T]Q_1^2/(2Q_2), & if\ Q_2 > 0 \\ S_{obs} + [(T-1)/T]Q_1(Q_1-1)/2, & if\ Q_2 = 0 \end{cases} \qquad (11a)$$

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

$$\hat{\text{var}}(\hat{S}_{Chao2}) = Q_2\left[\frac{A}{2}\left(\frac{Q_1}{Q_2}\right)^2 + A^2\left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{4}A^2\left(\frac{Q_1}{Q_2}\right)^4\right], \qquad (11b)$$

where $A = (T-1)/T$. When $Q_2 = 0$, the variance is modified to[31]:

$$\hat{\text{var}}(\hat{S}_{Chao2}) = \frac{AQ_1(Q_1-1)}{2} + \frac{A^2Q_1(2Q_1-1)^2}{4} - \frac{A^2Q_1^4}{4\hat{S}_{Chao2}}.$$

Chiu *et al.*[33] derived the corresponding *iChao2 estimator*:

$$\hat{S}_{iChao2} = \hat{S}_{Chao2} + \frac{(T-3)}{4T}\frac{Q_3}{Q_4} \times \max\left(Q_1 - \frac{(T-3)}{2(T-1)}\frac{Q_2Q_3}{Q_4},\ 0\right). \qquad (11c)$$

The variance formulas for the above-mentioned estimators can be evaluated using standard statistical approximation methods. Chao and Lin[34] generalized the Chao2 estimator to data based on sampling without replacement. The resulting estimator is similar to Equation (6d) but with (f_1, f_2) and n from the latter being replaced respectively by (Q_1, Q_2) and T.

Incidence Coverage-based estimators (ICE-type estimators)

Parallel to the ACE, there is a corresponding Incidence-based Coverage Estimator (*ICE*) for incidence data under the model given in Equation (9b). However, the definition of "sample coverage" for incidence data is slightly different: it is defined as the fraction of the total incidence probabilities of the detected species in the reference sample, or, in mathematical terms $C = \sum_{i=1}^{S}\pi_i I(Y_i > 0)/\sum_{i=1}^{S}\pi_i$. This type of sample coverage was first defined and estimated in Chao *et al.*[52] for capture-recapture data.

As with ACE, a cut-off point κ is first selected to partition the data into an infrequent species group (incidence frequency not larger than κ) and a frequent species group (incidence frequency larger than κ). The cut-off κ =10 is recommended. 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 $\hat{C}_{infreq} = 1 - Q_1 / \sum_{i=1}^{\kappa} iQ_i$. In the special case that detection probability is homogeneous (i.e., $\pi_1 = \pi_2 = ... = \pi_S$) for the infrequent group, the coverage-based estimator is

$$\hat{S}_0 = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}}, \tag{12a}$$

which is similar to that in Equation (7a).

The basic idea of the ICE[52,53] is to adjust the estimator in (12a) by accounting for heterogeneity among the detection probabilities. Let the number of sampling units that include at least one infrequent species be T_{infreq}. Then the ICE is expressed as

$$\hat{S}_{ICE} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}} \hat{\gamma}^2_{infreq}, \tag{12b}$$

where $\hat{\gamma}^2_{infreq}$ is the estimate of the squared CV of the species detection probabilities $(\pi_1, \pi_2, ..., \pi_S)$ in the infrequent species group,

$$\hat{\gamma}^2_{infreq} = \max\left\{ \frac{S_{infreq}}{\hat{C}_{infreq}} \frac{T_{infreq}}{(T_{infreq}-1)} \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{\left(\sum_{i=1}^{\kappa} iQ_i\right)\left(\sum_{i=1}^{\kappa} iQ_i - 1\right)} - 1,\ 0 \right\}. \tag{12c}$$

A similar ICE-1 estimator for species-rich and highly heterogeneous assemblages can also be obtained[54].

Jackknife estimators

For incidence data, the first- and second-order jackknife estimators were originally developed by Burham and Overton[42] in the context of capture-recapture studies. The formulas for the first two orders of jackknife are obtained by replacing (f_1, f_2) with (Q_1, Q_2), and replacing n with T in Equations (8a) and (8b). An approximate variance estimator was also given in Burnham and Overton[42].

3. Non-asymptotic Approach: rarefaction and Extrapolation

When there are multiple assemblages, the sample-size- and coverage-based integration of rarefaction and extrapolation represent a unified standardization method from which fair and meaningful comparisons of species richness can be made across assemblages. This method aims to compare the non-asymptotic portion of species accumulation curves. We review below the method separately for abundance and incidence data.

3.1 Rarefaction/Extrapolation for Abundance Data

3.1.1 Sample-size-based Rarefaction and Extrapolation

Under the multinomial model in Equation (1c), let $S(m)$ denote the number of species in a hypothetical random sample of m individuals for any $m = 1, 2, \ldots$from the assemblage. If we knew the true species detection probabilities $\psi_1,\psi_2,...,\psi_S$ of the S species, we could compute the expected value of $S(m)$ via the following function:

$$E[S(m)] = S - \sum_{i=1}^{S}(1-\psi_i)^m, \ m = 1, 2,.... \tag{13a}$$

On the basis of an observed sample of size n ("reference sample") of S_{obs} species, a sample-size-based rarefaction and extrapolation curve represents an estimator of the expected species accumulation curve which depicts $E[S(m)]$ with respect to the sample size m. All samples are standardized by estimating the expected species richness for a common sample size; this sample size can be smaller than the reference sample (traditional rarefaction) or larger than the reference sample (extrapolation). An unbiased estimator[10,55] for the expected species richness in a rarefied sample of size m, $m < n$, is

$$\hat{S}(m) = S_{obs} - \sum_{X_i>0}\binom{n-X_i}{m}\Big/\binom{n}{m}, \ m < n. \tag{13b}$$

Colwell *et al*.[15] and Chao and Jost[16] followed the approach of Shen *et al*. [56] and derived the following species richness estimator for the expected number of species in an extrapolated sample of size $n+m^*$:

$$\hat{S}(n+m^*) = S_{obs} + \hat{f}_0\left[1-\left(1-\frac{f_1}{n\hat{f}_0+f_1}\right)^{m^*}\right], \quad m^* > 0, \tag{13c}$$

where $\hat{f}_0$ is the estimated zero-frequency count based on the Chao1 estimator (Eq. 6a), and f_1 denotes the number of singletons. For a short-range prediction (e.g., m^* is much lower than n), the prediction formula is approximately $\hat{S}(n+m^*) \approx S_{obs} + (f_1/n)m^*$, which is independent of the choice of $\hat{f}_0$. This implies that the extrapolation formula in Eq. (13c) is very robust and reliable even though the

undetected species richness estimator $\hat{f}_0$ is a lower bound. Previous experiences by Colwell *et al.*[15] suggested that the prediction size can be extrapolated at most to double the observed sample size. Chao *et al*.[17] proposed a **bootstrap** method to obtain the variance estimators for the estimators $\hat{S}(m)$ and $\hat{S}(n+m^*)$ as well as to construct the associated confidence intervals.

The integrated sample-size-based sampling curve includes a rarefaction part (which plots $\hat{S}(m)$ as a function of $m < n$), and an extrapolation part (which plots $\hat{S}(n+m^*)$ as a function of $n+m^*$), which join smoothly at the reference point (n, S_{obs}). The confidence intervals for the two parts based on the bootstrap method also join smoothly.

3.1.2 Coverage-based Rarefaction and Extrapolation

The concept of "coverage" (as an objective measure of sample completeness) has been widely applied in many research fields. Chao and Jost[16] proposed standardizing samples by matching their sample coverage based on rarefaction or extrapolation to a target level of sample coverage. This allows fair comparison of equally-complete samples (i.e., equal fraction of population individuals). The coverage-based rarefaction and extrapolation curve represents an estimator of the species accumulation curve, the latter of which plots $E[S(m)]$ as a function of sample completeness.

As indicated in Section 2.1.2, Turing's definition of the sample coverage C of an observed sample of size n with species frequencies ($X_1, X_2, \ldots, X_S$) is expressed as $C \equiv C(n) = \sum_{i=1}^{S} p_i I(X_i > 0) = 1 - \sum_{i=1}^{S} p_i I(X_i = 0)$, where $I(A)$ is the indicator function, i.e., $I(A) = 1$ if the event A occurs, and 0 otherwise. Here we affix the sample size n to the notation C to facilitate rarefaction and extrapolation with various sample sizes. Generally, for any sample size $m = 1, 2, \ldots$, let $C(m)$ denote the expected sample coverage of a hypothetical sample of size m. The expected value of $C(m)$ is a function of m, and can be expressed as

$$E[C(m)] = 1 - \sum_{i=1}^{S} \psi_i (1-\psi_i)^m \,, \; m = 1, 2, \ldots \qquad (14a)$$

To construct the coverage-based rarefaction and extrapolation curve, we need to estimate $E[C(m)]$ separately for three cases: (i) the observed or reference sample (i.e., $m = n$, a case on which Turing and Good focused in their cryptanalysis); (ii) a rarefied sample for $m < n$; and (iii) an extrapolated sample for $m = n+m^*$, $m^* > 0$. We first review Case (i) in which no unbiased estimator exists for $E[C(n)]$ based on the reference sample itself. Turing's simple and efficient estimator[35,36] is the complement of the proportion of singletons in the reference sample. Chao and Jost[16] refined Turing's estimator by using the additional information of doubletons to obtain a less-biased estimator as follows:

$$\hat{C} \equiv \hat{C}\,(n) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]. \qquad (14b)$$

For a rarefied sample of size $m < n$ in Case (ii), an analytical unbiased estimator $\hat{C}(m)$ of $E[C(m)]$ does exist, and the estimator was first derived by Chao and Jost[16]:

$$\hat{C}(m) = 1 - \sum_{X_i>0} \frac{X_i}{n} \frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}, \quad m < n. \qquad (14c)$$

They also derived for Case (iii) the extrapolated coverage estimator $\hat{C}(n+m^*)$ for the expected coverage of any hypothetical enlarged sample of size $n + m^*$:

$$\hat{C}\,(n+m^*) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]^{m^*+1}, \quad m^* > 0. \qquad (14d)$$

When $m^* = 0$, Eq. (14d) reduces to the sample coverage estimator for the reference sample given in Eq. (16b).

The coverage-based sampling curve includes a rarefaction part (which plots $\hat{S}(m)$ as a function of $\hat{C}(m)$), and an extrapolation part (which plots $\hat{S}(n+m^*)$ as a function of $\hat{C}(n+m^*)$), which join smoothly at the reference sample point ($\hat{C}(n)$, S_{obs}). The confidence intervals based on the bootstrap method[16] also join smoothly. The curve can be extended to the coverage that corresponds to double reference sample size.

The sample-size-based approach plots the estimated species richness as a function of sample size, whereas the corresponding coverage-based approach plots the same richness estimate with respect to sample coverage. Therefore, the two types of sampling curves can be bridged by a sample completeness curve, which shows how the sample coverage estimate varies with sample size and also provides an estimate of the sample size needed to achieve a fixed degree of completeness. The two types of sampling curves along with the associated sample completeness curve are illustrated in Section 4 through an example.

3.2 Rarefaction/Extrapolation for Incidence Data

3.2.1 Sample-size-based Rarefaction and Extrapolation

The rarefaction and extrapolation for incidence data is formulated under the product-Bernoulli model (Equations 9a and 9b) in which the incidence frequency count Y_i based on the incidence data of T sampling units follows a binomial

distribution characterized by T and detection probability π_i for the ith species in any sampling unit. In incidence data, the sample size refers to the number of sampling units. Let $S(t)$ be the number of species in a hypothetical sample of size t randomly selected from the assemblage. If we knew the true species detection probabilities $\pi_1, \pi_2, ..., \pi_S$ of each of the S species in each sampling unit, we could compute the following expected value:

$$E[S(t)] = S - \sum_{i=1}^{S} (1-\pi_i)^t, \quad t = 1,\ 2,... \tag{15a}$$

The plot of $E[S(t)]$ with respect to the number of sampling units t is the expected species accumulation curve for incidence data. The rarefaction (interpolation) part estimates the expected species richness for a smaller number of sampling units $t < T$. On the basis of a reference sample of T sampling units, an unbiased estimator $\hat{S}(t)$ for $E[S(t)]$, $t < T$, is

$$\hat{S}(t) = S_{obs} - \sum_{Y_i>0} \binom{T-Y_i}{t} \Big/ \binom{T}{t}, \ t < T. \tag{15b}$$

This analytic formula was first derived by Shinozaki[57] and rediscovered multiple times[14].

The extrapolation is to estimate the expected number of species $E[S(T+t^*)]$ in a hypothetical sample of $T+t^*$ sampling units ($t^* > 0$) from the assemblage. Chao *et al.*[58] obtained an estimator

$$\hat{S}(T+t^*) = S_{obs} + \hat{Q}_0 \left[1 - \left(1 - \frac{Q_1}{T\hat{Q}_0 + Q_1} \right)^{t^*} \right], \tag{15c}$$

where $\hat{Q}_0$ can be obtained using the Chao2 estimator ($\hat{Q}_0 = \hat{S}_{Chao2} - S_{obs}$) given in Equation (11a). Colwell *et al.*[15] linked rarefaction and extrapolation to form an integrated smooth curve. The corresponding confidence intervals based on a bootstrap method[17] also join smoothly at the reference point (T, S_{obs}). As with abundance data, for a short-range prediction (e.g., t^* is much less than T), the extrapolation formula is independent of the choice of $\hat{Q}_0$ as indicated by the approximation formula $\hat{S}(T+t^*) \approx S_{obs} + (Q_1/T)t^*$. However, the extrapolation can be extended at most to double reference sample size.

3.2.2 Coverage-based Rarefaction and Extrapolation

For incidence data, the sample coverage of a reference sample of T sampling units is defined as

$$C \equiv C(T) = \frac{\sum_{i=1}^{S} \pi_i I(Y_i > 0)}{\sum_{i=1}^{S} \pi_i} = 1 - \frac{\sum_{i=1}^{S} \pi_i I(Y_i = 0)}{\sum_{i=1}^{S} \pi_i}, \quad (16a)$$

which represents the fraction of the total incidence probabilities of the detected species in the reference sample. This type of sample coverage was first defined in Chao *et al*.[52] for capture-recapture data. Chao *et al*.[17] derived an accurate estimator of the sample coverage for the reference sample size:

$$\hat{C}\ (T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right], \quad (16b)$$

where $U = \sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$ denotes the total number of incidences in the reference sample.

In addition to the reference sample, we also need to consider the estimation of the expected sample coverage, $E[C(t)]$, for any hypothetical sample of t sampling units, $t = 1, 2, \ldots$. This expected sample coverage is a function of t as given below:

$$E[C(t)] = 1 - \frac{\sum_{i=1}^{S} \pi_i (1-\pi_i)^t}{\sum_{i=1}^{S} \pi_i}, \ t \geq 1. \quad (16c)$$

For a rarefied sample ($t < T$), an unbiased estimator exists for the denominator and numerator in Eq. (16c), respectively, but their ratio $\hat{C}(t)$, given below, is only a nearly unbiased estimator of $E[C(t)]$:

$$\hat{C}(t) = 1 - \sum_{Y_i > 0} \frac{Y_i}{U} \frac{\binom{T - Y_i}{t}}{\binom{T-1}{t}}, \ t < T. \quad (16d)$$

This equation is analogous to Eq. (14c) for abundance data. An estimator for the expected coverage of an extrapolated sample with $T + t^*$ sampling units is

$$\hat{C}(T + t^*) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]^{t^*+1}. \quad (16e)$$

This equation is analogous to Eq. (14d) for abundance data. When $t^* = 0$, Eq. (16e) reduces to the sample coverage estimator for the reference sample as given in Eq. (16b). As with abundance data, smooth coverage-based interpolation and extrapolation curves with confidence intervals can be constructed for incidence data up to the coverage that corresponds to the double reference sample size.

4 Example

We apply both asymptotic and non-asymptotic analyses to the rain forest tree data described and discussed by Magnago *et al*.[59]. The tree species abundance data were collected during January 2011 to January 2012 from 12 forest fragments in south-eastern Brazil. Sampling data in each fragment includes one edge and one interior transect, where an edge transect is about 5m inside the fragment and parallel to the forest edge, and an interior transect is located more than 300m from the nearest edge. Within each transect, every living tree with a diameter at breast height > 4.8 cm and 1.3m height was collected and recorded. One of the goals was to compare the diversity of the Edge habitat with that of the Interior habitat.

The original data with species functional traits were given in Table S2 of Magnago *et al*.[59]. The species abundance frequency counts for the two habitats are summarized in Table1. In the Edge habitat, the reference sample includes 334 species (113 singletons and 50 doubletons) among 1978 individuals, and in the Interior Habitat, the reference sample includes 371 species (129 singletons and 49 doubletons) among 2171 individuals. Based on Equation (14b), the estimated sample coverage values for the two habitats are nearly equal at a level of around 94% (94.29% for the Edge habitat and 94.06% for the Interior habitat) in spite of the difference in sample sizes. Thus, the raw data implies that the Interior Habitat is more diverse than the Edge habitat for a standardized coverage of 94% or for a fraction of 94% of the individuals in each assemblage. Below we apply two R packages to analyze the data through both asymptotic and non-asymptotic analyses, thereby demonstrating more informative comparisons between these two sites.

4.1 Asymptotic Analysis: Species Richness Estimation

We use the function `ChaoSpecies` in the R package SpadeR (Species-richness Prediction And Diversity Estimation in R) to infer the species richness in each habitat. SpadeR is available from Github and can also be downloaded from Anne Chao's website http://chao.stat.nthu.edu.tw/wordpress/software_download/. The installation and procedures are shown in the following commands. Copying these commands into the R Console, we obtain various species richness estimates, their standard errors, along with 95% confidence intervals for each site as shown below. The output for the Edge habitat is shown first, followed by the output for the Interior habitat.

```
install.packages('devtools')
library(devtools)
install_github('AnneChao/SpadeR')
library(SpadeR)
Edge = rep(c(1:21,23,25,27,28,30,32,36,37,41,45,46,49,89,110),c(113,50,
  39,29,15,11,13,5,6,6,3,4,3,5,2,5,2,2,2,2,1,2,1,1,1,1,1,2,1,1,1,1,1,1
  ))
Interior = rep(c(1:21,23,25,27,28,30,32,34,35,52,123,140),c(129,49,42,
  32,19,17,7,9,7,7,6,3,3,3,4,4,2,2,3,4,6,2,1,2,1,1,1,1,1,1,1))
```

```
Forest = list("Edge" = Edge, "Interior" = Interior)
#Output for Edge data
out1 = ChaoSpecies(Forest$"Edge",datatype = "abundance", k = 10, conf
  = 0.95)
#Output for Interior data
out2 = ChaoSpecies(Forest$"Interior",datatype = "abundance", k = 10,
  conf = 0.95)
#Show the output of Edge data
out1


(1) BASIC DATA INFORMATION:

                                                 Variable Value
Number of observed individuals                          n  1978
Number of observed species                              D   334
Coverage estimate for whole data                        C 0.943
CV for whole data                                      CV 1.796
Cut-off point                                           k    10
Number of observed individuals for rare species       n_rare   832
Number of observed species for rare species           D_rare   287
Estimation of the sample coverage for rare species    C_rare 0.864
Estimation of CV for rare species in ACE             CV_rare 0.703
Estimation of CV1 for rare species in ACE-1         CV1_rare 0.886
Number of observed individuals for abundant species   n_abun  1146
Number of observed species for abundant species       D_abun    47

                   f1 f2 f3 f4 f5 f6 f7 f8 f9 f10
Rare.Species.Group 113 50 39 29 15 11 13  5  6   6


(2) SPECIES RICHNESS ESTIMATORS TABLE:

                         Estimate   s.e. 95%Lower 95%Upper
Homogenous Model          379.106  9.234  364.322  401.097
Homogeneous (MLE)         334.912  0.964  334.167  338.980
Chao1 (Chao, 1984)        461.625 32.093  412.553  541.353
Chao1-bc                  458.016 31.128  410.397  535.314
iChao1 (Chiu et al. 2014) 488.313 21.764  451.204  537.173
ACE (Chao & Lee, 1992)    443.684 21.695  408.712  495.024
ACE-1 (Chao & Lee, 1992)  481.667 33.005  429.796  561.625
1st order jackknife       446.943 15.028  421.116  480.426
2nd order jackknife       509.904 26.021  465.839  568.698


#Show the output of Interior data
out2

(1) BASIC DATA INFORMATION:

                                                 Variable Value
Number of observed individuals                          n  2171
Number of observed species                              D   371
Coverage estimate for whole data                        C 0.941
CV for whole data                                      CV 1.979
Cut-off point                                           k    10
Number of observed individuals for rare species       n_rare   932
Number of observed species for rare species           D_rare   318
Estimation of the sample coverage for rare species    C_rare 0.862
Estimation of CV for rare species in ACE             CV_rare 0.716
Estimation of CV1 for rare species in ACE-1         CV1_rare 0.909
Number of observed individuals for abundant species   n_abun  1239
```

```
Number of observed species for abundant species       D_abun    53

                    f1 f2 f3 f4 f5 f6 f7 f8 f9 f10
Rare.Species.Group 129 49 42 32 19 17  7  9  7   7


(2) SPECIES RICHNESS ESTIMATORS TABLE:

                          Estimate   s.e. 95%Lower 95%Upper
Homogenous Model           422.086  9.857  406.120  445.310
Homogeneous (MLE)          372.088  1.054  371.221  376.358
Chao1 (Chao, 1984)         540.728 40.631  477.860  640.582
Chao1-bc                   536.044 39.372  475.074  632.732
iChao1 (Chiu et al. 2014)  572.505 29.327  522.722  638.622
ACE (Chao & Lee, 1992)     498.834 23.842  459.971  554.673
ACE-1 (Chao & Lee, 1992)   545.927 37.058  487.012  634.759
1st order jackknife        499.941 16.057  472.111  535.430
2nd order jackknife        579.889 27.804  532.108  641.842
```

Nearly all of the output is self-explanatory. We only note that the "Homogeneous Model" estimate is obtained from Eq. (7a), and the "Homogeneous (MLE)" is based on Eq. (3). These two estimators, derived under the assumption that all species detection probabilities are the same, usually severely underestimate the true species richness if heterogeneity exists. The CV estimates for the Edge and Interior habitats are, respectively, 1.796 and 1.979 (see the output above), indicating the presence of heterogeneity. Consequently, the two estimates yield substantially low estimates in each habitat.

The Chao1, iChao1, and ACE all give consistent estimates between 440 and 490 for the Edge habitat, whereas their corresponding estimates for the Interior habitat are between 500 and 570. We do not include the ACE-1 in our comparison because all the species richness estimates are less than 1000 and the estimated CV values for both sites are not extremely high. Each of the estimators reveals that the species richness of the Interior habitat is higher than that in the Edge habitat, although the 95% confidence intervals overlap. The first order and second order jackknife estimators also show the same ordering, but the second order jackknife estimate in each habitat is higher than any of the Chao1, iChao1 and ACE estimates.

We present the Chao1 estimates for illustration. The species richness estimates in the Edge and Interior habitats are, respectively, 462 with a 95% confidence interval of (413, 541) in the Edge habitat, and 541 with a 95% confidence interval of (478, 641) in the Interior habitat. The confidence intervals for the two habitats overlap, implying that significant difference is not guaranteed for comparing species richness for complete assemblages. However, data do support significance difference in species richness if only a fraction of the assemblage is compared, as shown in the coverage-based rarefaction and extrapolation in the next section.

4.2 Non-asymptotic Analysis: Rarefaction/Extrapolation

The sample-size- and coverage-based rarefaction and extrapolation sampling curves along with the sample completeness curves can be obtained using the R package iNEXT (iNterpolation and EXTrapolation) which is available on CRAN and also on Anne Chao's website. The following commands return the three sampling curves as shown in Figure 1 to Figure 3, along with some related statistics (omitted here). The omitted output includes basic data information and species richness estimates for some rarefied and extrapolated samples.

```
install.packages ("iNEXT")
library(iNEXT)
library (ggplot2)
out <- iNEXT(Forest, q=0, datatype ="abundance", endpoint=4000)
#plot sample-size-based curve (as shown in Fig. 1)
ggiNEXT(out, type=1)
#plot sample completeness curve (as shown in Fig. 2)
ggiNEXT(out, type=2)
#plot coverage-based curve (as shown in Fig. 3)
ggiNEXT(out, type=3)
#to show the detailed output for related statistics
out
```

In the sample-size-based rarefaction and extrapolation sampling curve (Fig. 1), we compare two equally-large samples. For each site, the extrapolation is extended to a maximum size of 4000 (by specifying endpoint=4000 in the iNEXT function as shown in the above commands). The maximum size of 4000 is approximately double that of each reference sample size. Extrapolation beyond the double reference sample size could theoretically be computed and used for ranking species richnesses, but the estimates may be subject to some prediction biases and should be used with caution in estimating species richness ratios or other measures. Figure 1 clearly reveals that curve of the Interior habitat lies above that of the Edge habitat. However, the confidence intervals of the two sites overlap, implying that comparing two equally-large samples is inconclusive regarding the test of significant difference in species richness between the two habitats.

The sample completeness curve in Figure 2 shows how the sample coverage varies with sample size. Although the curve of the Edge habitat lies above that of the Interior habitat, there is little difference between the two curves for any sample size. For the Edge habitat, when the sample size is extended from 1978 to 4000, the sample coverage is extended from 94.29% to 97.69% (as shown in Figure 2 or the unreported iNEXT output). For the Interior habitat, when the sample size is extended from 2171 to 4000 the coverage is extended from 94.06% to 96.89% (as shown in Fig. 2 or the unreported iNEXT output), with a very small increment.

In the coverage-based rarefaction and extrapolation sampling curve (Figure 3), we compare two equally-complete samples (or equal fractions of population individuals). The extrapolation is extended to 97.69% for the Edge habitat and to 96.89% for the Interior habitat, as explained in the preceding paragraph. Except for the very low initial coverage values, the Interior habitat is significantly more diverse than the Edge habitat as evidenced by the non-overlapping confidence intervals for any fixed coverage up to about 97% in Figure 3. This implies that if we compare species richness for equal population fractions up to 97%, the data do provide sufficient information to conclude that the Interior habitat is significantly more diverse. Note that significant difference cannot be guaranteed based on each of the three asymptotic species richness estimators (Chao1, iChao1 and ACE) due to the overlapping confidence intervals. Thus, if we compare species richness for the two complete assemblages (i.e., data are extrapolated to a coverage of unity), the data do not provide sufficient evidence to conclude significant difference in species richness between the two sites. In other words, data do not have sufficient information to infer the rarest 3% of each assemblage. Unlike the sample-sized-based standardization in which size is determined by samplers, here the coverage-based standardization compares equal population fractions of each assemblage. The population fraction is an assemblage-level characteristic that can be reliably estimated from data.

As demonstrated in the above-described example, the two R packages (SpadeR and iNEXT) supply useful information for both asymptotic and non-asymptotic analyses. These methods efficiently use all available data to make robust and meaningful comparisons of species richness between assemblages for a wide range of sample sizes/completeness. These methods have also been generalized to diversity measures that incorporate species abundances[17] and those that take into account the evolutionary history among species[60].

Acknowledgements

This work is supported by the Taiwan Ministry of Science and Technology under Contract 103-2628-M007-007 (for AC) and 104-2118-M-007-006-MY3 (for CHC).

6 Related Articles

See also **Diversity Indices**; **Zipf's Law**; **Capture--recapture sampling designs**; **Capture--recapture methodology**; **Wildlife Ecology, **; **Spatial analysis in ecology**; **Ecological statistics**; **Species overlap**; **Species diversity**; **Species distribution, monitoring changes in**.

References

[1] Colwell, R.K. and Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. Royal Soc., London, Series B*, **345**, 101–118.
[2] Magurran, A.E. (2004) *Measuring Biological Diversity.* Blackwell, Oxford.
[3] Chao, A. (2005) Species estimation and applications, in *Encyclopedia of Statistical Sciences* (eds N. Balakrishnan, C.B. Read, and B. Vidakovic), Wiley, New York, pp. 7907–7916.
[4] Magurran, A. E. and McGill, B. J. (2011) *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, Oxford.
[5] Chao, A. and Chiu, C.-H. (2012) Estimation of species richness and shared species richness, in *Methods and Applications of Statistics in the Atmospheric and Earth Sciences* (ed N. Balakrishnan), Wiley, New York, pp. 76–111.
[6] Gotelli, N. J. and Chao, A. (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data, in *The Encyclopedia of Biodiversity* 2nd Edition, (ed S.A. Levin), Elsevier, New York, pp. 195–211.
[7] Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. *J. Am. Stat. Assoc.*, **88**, 364–373.
[8] 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.

[9] Sanders, H.L. (1968) Marine benthic diversity: a comparative study. *Am. Nat*., **102**, 243–282.
[10] Hurlbert, S.H. (1971) The Nonconcept of species diversity: A critique and alternative parameters. *Ecology*, **52**, 577–586.
[11] Simberloff, D. (1979) Rarefaction as a distribution-free method of expressing and estimating diversity, in *Ecological Diversity in Theory and Practice* (eds *J.F.* Grassle, G.P. Patil, W.K. Smith, and C. Taillie), International Cooperative Publishing House, Fairland, MD, pp. 159–176.
[12] 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.
[13] Gotelli, N. J., and Colwell, R. K. (2011) Estimating species richness, in *Biological Diversity: Frontiers in Measurement and Assessment* (eds A. Magurran, and B. McGill), Oxford University Press, Oxford, pp. 39–54.
[14] 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.
[15] Colwell, R.K., Chao, A., Gotelli, N.J., *et al.* (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblage. *J. Plant Ecol.*, **5**, 3–21.
[16] Chao, A. and Jost. L. (2012) Coverage-based rarefaction: standardizing samples by completeness rather than by sample size. *Ecology*, **93**, 2533–2547.
[17] Chao, A., Gotelli, N.J., Hsieh, T.C., *et al*. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monog.,* **84**, 45–67.
[18] Flather, C.H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *J. Biogeogr.*, **23**, 155–168.
[19] Preston, F.W. (1948) The commonness and rarity of species. *Ecology*, **29**, 254–283.
[20] Coleman, B.D. (1981) On random placement and species-area relations. *Math. Biosci.*, **54**, 191–215.
[21] Sanathanan, L. (1977) Estimating the size of a truncated sample. *J. Am. Stat. Assoc.*, **72**, 669–672.
[22] Pielou, E. (1977) *Mathematical Ecology*. New York: Wiley.
[23] Bulmer, M.G. (1974) On fitting the Poisson lognormal distribution to species abundance data. *Biometrics*, **30**, 101–110.
[24] 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.
[25] 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.
[26] 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.
[27] Bunge, J., Fitzpatrick. M. and Handley, J. (1995) Comparison of three estimators of the number of species. *J. Appl. Stat.*, **22**, 45–59.

[28] Chao, A., Chiu, C.-H. and Jost, L. (2014) Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annu. Rev. Ecol. Evol. Syst.*, **45**, 297–324.
[29] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scan. J. Statist.*, **11**, 265–270.
[30] Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.
[31] 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/.
[32] Chao, A. and Chiu, C.-H. (2015) Nonparametric estimation and comparison of species richness. In: eLS. John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0026329.
[33] Chiu, C.-H., Wang, Y.T., Walther, B.A., *et al.* (2014) An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. *Biometrics*, **70**, 671–682.
[34] Chao, A. and Lin, C.-W. (2012) Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics,* **68**, 912–921.
[35] Good, I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.
[36] 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.
[37] 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.
[38] Esty, W.W. (1986) The efficiency of Good's nonparametric coverage estimator. *Ann. Stat.,* **14**, 1257–1260.
[39] Chao, A. and Lee, S.-M. (1992) Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.*, **87**, 210–217.
[40] Darroch, J.N. and Ratcliff, D. (1980) A note on capture-recapture estimation. *Biometrics*, **36**, 149–153.
[41] Cormack, R.M. (1989) Log-linear models for capture-recapture. *Biometrics,* **45**, 395–413.
[42] 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.
[43] Burnham, K.P., and Overton, W.S. (1979) Robust estimation of population size when capture probabilities vary among animals. *Ecology*, **60**, 927–936.
[44] Seber, G.A.F. (1982) *The Estimation of Animal Abundance* (2nd Edition), Griffin, London.
[45] Schwarz, C.J. and Seber, G.A.F. (1999) A review of estimating animal abundance III. *Stat. Sci.*, **14**, 427–456.
[46] Amstrup, S.C., McDonald, T.L. and Manly, B.F.J. (2005) *Handbook of Capture-Recapture Analysis*. Princeton University Press, Princeton, USA.
[47] Chao, A. (2001) An overview of closed capture-recapture models. *J. Agric. Bio. Environ. Stat.*, **6**, 158–175.

[48] Chao, A. and Huggins, R. M. (2005a) Modern closed population capture-recapture models, in *Handbook of Capture-Recapture Analysis* (eds B. Manly, T. McDonald, and S. Amstrup), Princeton University Press, Princeton, pp. 22–35.
[49] Chao, A. and Huggins, R. M. (2005b) Classical closed population models, in *Handbook of Capture-Recapture Analysis* (eds B. Manly, T. McDonald, and S. Amstrup), Princeton University Press, Princeton, pp. 58–87.
[50] 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.
[51] Huggins, R.M. (1989) On the statistical analysis of capture experiments. *Biometrika*, **76**, 133–140.
[52] Chao, A., Lee, S.M. and Jeng, S.L. (1992) Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, **48**, 201–216.
[53] Lee, S-M. and Chao, A. (1994) Estimating population size for closed capture-recapture models via sample coverage. *Biometrics*, **50**, 88–97.
[54] Gotelli, N.J. and Chao, A. (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data, in *Encyclopedia of Biodiversity* (ed S.A. Levin*)*, 2nd edn, vol. 5, Academic Press, Waltham, MA, pp. 195–211.
[55] Smith, W. and Grassle, J.F. (1977) Sampling properties of a family of diversity measures. *Biometrics*, **33**, 283–292.
[56] Shen, T.-J., Chao, A. and Lin, J.-F. (2003) Predicting the number of new species in further taxonomic sampling. *Ecology,* **84**, 798–804.
[57] Shinozaki, K. (1963) Notes on the species-area curve, 10th Annual Meeting of the Ecological Society of Japan (Abstract) (p. 5).
[58] 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.
[59] Magnago, L.F.S., Edwards, D.P., Edwards, F.A., *et al.* (2014) Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. *J. Ecol.*, **102**, 475–485.
[60] Chao, A., Chiu, C.-H., Hsieh, T.C., *et al.* (2015) Rarefaction and extrapolation of phylogenetic diversity. *Methods Ecol. Evol.*, **6**, 380–388.

Table 1: The tree species frequency counts for the data of two habitats (Edge and Interior) in south-eastern Brazil[59], where f_i denotes the number of species represented by exactly i individuals in the sample.

Habitat	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}
Edge	113	50	39	29	15	11	13	5	6	6	3	4	3	5
Interior	129	49	42	32	19	17	7	9	7	7	6	3	3	3

Habitat	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}	f_{21}	f_{23}	f_{25}	f_{27}	f_{28}	f_{30}	f_{32}
Edge	2	5	2	2	2	2	1	2	1	1	1	1	1
Interior	4	4	2	2	3	4	6	2	1	2	1	1	1

Habitat	f_{34}	f_{35}	f_{36}	f_{37}	f_{41}	f_{45}	f_{46}	f_{49}	f_{52}	f_{89}	f_{110}	f_{123}	f_{140}
Edge	0	0	2	1	1	1	1	1	0	1	1	0	0
Interior	1	1	0	0	0	0	0	0	1	0	0	1	1

Habitat	n	S_{obs}	Sample coverage	CV estimate
Edge	1978	334	94.29%	1.796
Interior	2171	371	94.06%	1.979

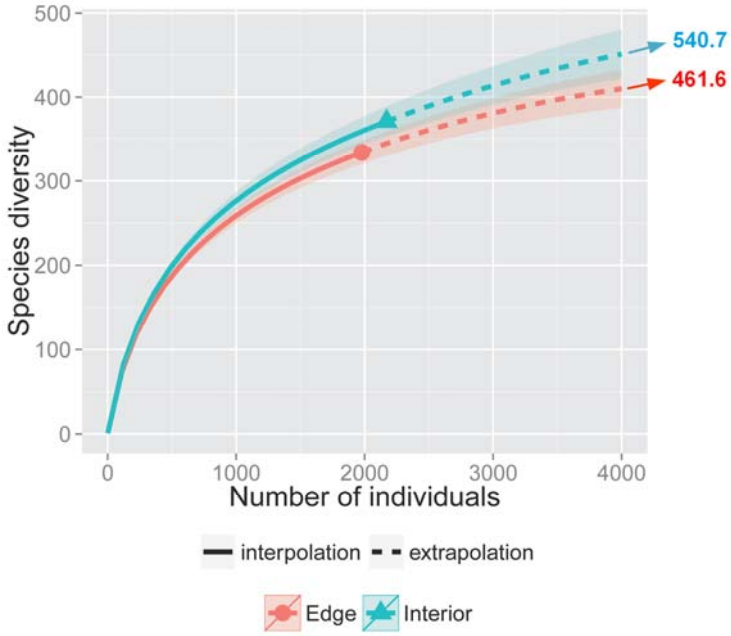

Fig. 1. Sample-size-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing tree richness for data of two habitats (Edge and Interior) in south-eastern Brazil[59]. Observed samples are denoted by the solid dot and triangle. The extrapolation extends up to a maximum sample size of 4000. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.

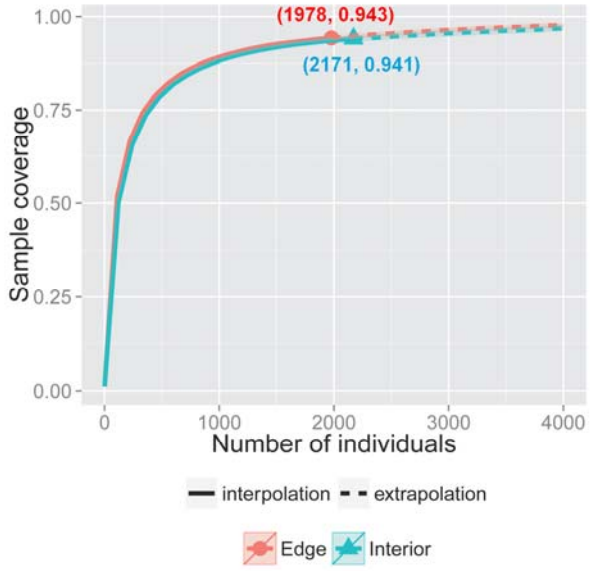

Fig. 2. Sample completeness curve which depicts how sample completeness (measured by sample coverage) increases with sample size for tree data of two habitats (Edge and Interior) in south-eastern Brazil[59]. For each habitat, the plot of sample coverage for rarefied samples (solid lines) and extrapolated samples (dashed lines) with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) is extrapolated up to a maximum sample size of 4000. The observed samples are denoted by the solid dot and triangle. For each reference sample point, the numbers in parentheses show the x- and y-axis coordinate.

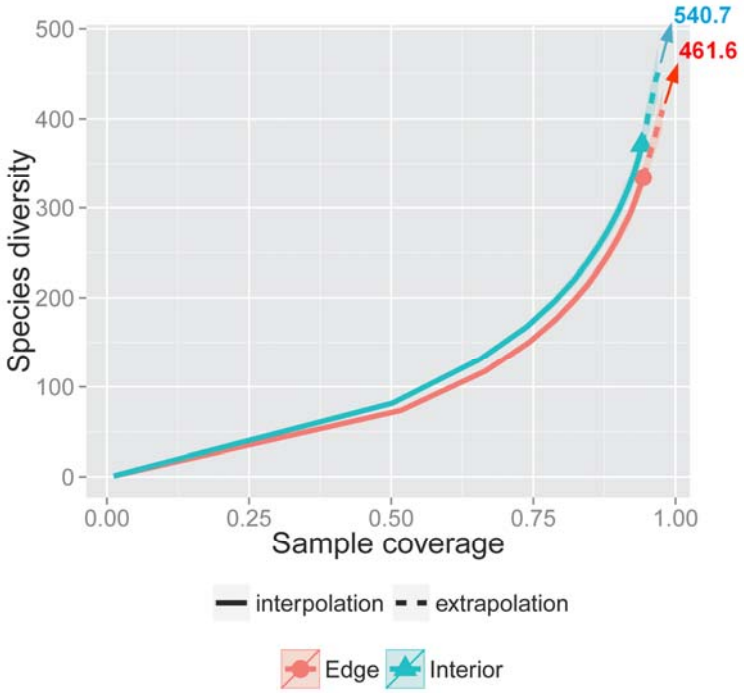

Fig. 3. Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing tree richness for data of two habitats (Edge and Interior) in south-eastern Brazil[59]. Observed samples are denoted by the solid dot and triangle. The extrapolation extends up to the coverage value of the corresponding maximum sample size of 4000 in Fig. 2 (97.69% in the Edge habitat, and 96.89% in the Interior habitat). The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.