Nonparametric Estimation and Comparison of Species Richness

Anne Chao, *Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan*

Chun-Huo Chiu, *Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan*

Advanced article

Article Contents

Online posting date: 16th May 2016

Species richness (the number of species) in an assemblage is a key metric in many research fields of ecology. Simple counts of species in samples typically underestimate the true species richness and strongly depend on sampling effort and sample completeness. Based on possibly unequal-sampling effort and incomplete samples that miss many species, there are two approaches to infer species richness and make fair comparisons among multiple assemblages,: (1) An asymptotic approach via species richness estimation. This approach aims to compare species richness estimates across assemblages. We focus on the nonparametric estimators that are universally valid for all species abundance distributions. (2) A non-asymptotic approach via the sample-size- and coverage-based rarefaction and extrapolation on the basis of standardised sample size or sample completeness (as measured by sample coverage). This approach aims to compare species richness estimates for equally large or equally complete samples. Two R packages (SpadeR and iNEXT) are applied to beetle data for illustration.

Introduction

The measurement and assessment of biodiversity is a central objective of many studies in ecological research. The simplest and still most frequently used measure of biodiversity is species richness (the number of species) of an assemblage. Species richness possesses intuitive mathematical properties, and features prominently in foundational models of community ecology; it is a key metric in conservation biology and historical biogeography. Comparisons of species richness among multiple assemblages help researchers and other professionals to understand the causes and processes of biodiversity, assess the effect of human disturbances on biodiversity and make environmental policy decisions. **See also**: **Species Richness: Small Scale**; **Conservation Biology and Biodiversity**; **Elevational Gradients in Species Richness**

In practice, nearly all biodiversity studies and analyses are based on sampling data taken from the focal assemblages. However, due to sampling limitation, it is virtually impossible to detect all species with limited sampling efforts especially in highly diverse assemblages with many rare species. There are undetected species in almost every taxonomic survey or species inventory. Consequently, the simple count of species (empirical or the observed richness) in a sample underestimates the true species richness (observed plus undetected), with the magnitude of the negative bias possibly substantial. Also, empirical richness strongly depends on sampling effort and thus also depends on sample completeness. **See also**: **Community Ecology: An Introduction**

Generally, there are two statistical approaches to infer species richness and make fair comparisons among assemblages based on possibly unequal-sampling effort and incomplete samples that miss many species: (1) an asymptotic approach based on species richness estimators (Gotelli and Colwell, 2011; Chao and Chiu, 2012), and (2) a non-asymptotic approach based on standardisation of sampling effort or sample completeness (Colwell *et al*., 2012; Chao and Jost, 2012). We describe the background of each approach as follows.

An asymptotic approach based on species richness estimators

This approach aims to estimate the asymptote of a species accumulation curve. The estimated asymptote is then used as a species richness estimate, which can be compared across assemblages. Species richness estimation based on sampling data has a long history. In general contexts, 'species' can be defined in a broad sense: they may be biological species, individuals of a target

eLS subject area: Ecology

How to cite:
Chao, Anne and Chiu, Chun-Huo (May 2016) Nonparametric Estimation and Comparison of Species Richness. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0026329

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; see Bunge and Fitzpatrick (1993), Colwell and Coddington (1994), Magurran (2004), Chao (2005) and Chao and Chiu (2012) for reviews.

The traditional curve-fitting approach uses parametric curves to fit a species-accumulation or species-area curve to predict its asymptote. Among the proposed asymptotic functions are the negative exponential function, the Weibull function, the logistic function, and the Michaelis–Menten function (Colwell and Coddington, 1994). 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 distribution or functional form to the observed species frequencies to obtain an estimate of species richness. The earliest such approach was proposed by Preston (1948), who fitted a log-normal curve to the (properly grouped) observed frequencies in order to estimate the portion of the assemblage behind 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; see Magurran, 2004) can also be applied. Although this approach uses information on the frequencies of common and rare species, it simply fits a curve to the observed frequency data.

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.

The pioneering work by R. A. Fisher (Fisher *et al*., 1943) led to the founding of the sampling-theory-based approach. Since then, an enormous number of models and methods based on statistical sampling theory have been proposed in the literature to estimate species richness. Generally, there are two frameworks: parametric and nonparametric (see Magurran (2004) and Magurran and McGill (2011) for a review). In the parametric framework, it is assumed that species abundance follows a statistical model with one or two parameters. For example, Fisher *et al*. (1943) adopted a parametric approach and assumed that individuals of any species arrive in the sample according to a Poisson process with a mean occurrence rate that follows a gamma distribution with two parameters. This model, under the extreme case that the degree of heterogeneity among species abundances tends to infinity, leads to the log-series for species-rank distribution (one of the parameters for the log-series distribution is the well-known 'Fisher's alpha'). This approach does not yield a species richness estimate (Pielou, 1977, p. 274), although Fisher's alpha has been used by some researchers to characterise species diversity of an assemblage. Other parametric models assume that the Poisson mean rate follows the log-normal (Bulmer, 1974), inverse-Gaussian (Ord and Whitmore, 1986), or generalised inverse-Gaussian (Sichel, 1997) distribution. Extensive numerical procedures are typically required to find the species richness estimates under these models.

The chief weakness of the parametric approach is that simulations show that they work well only when the correct form of the species abundance distribution is already known (O'Hara, 2005; Chiu *et al*., 2014), but this is never the case for empirical data. Furthermore, as with the curve-fitting method, it may be difficult to select an adequate parametric model. The parametric approach also does not permit meaningful comparisons of assemblages with different distribution functions (e.g. a log-normal assemblage cannot be compared to an assemblage whose species-rank distribution follows a geometric series). A practical problem is that in some cases the iterative steps fail to converge properly and thus species richness estimates may not be obtainable.

The nonparametric approach, which makes no assumptions about the mathematical form of the underlying species abundance distributions, avoids the above-mentioned drawbacks and is more robust in applications. In this article, we focus on reviewing some analytic nonparametric estimators that are universally valid for all species abundance distributions and allow for comparison among multiple assemblages.

A non-asymptotic approach based on standardisation

The objective of this approach is to control the dependence of the empirical species counts on sampling effort and sample completeness. The earliest development of standardisation of sample size for abundance data by rarefaction was proposed by Sanders (1968), but Chiarucci *et al*. (2008) can be referred for a historical review. Subsequent developments include Hurlbert (1971), Simberloff (1979), Heck *et al*. (1975) and Coleman *et al*. (1982); see Gotelli and Colwell (2001, 2011) for details. Ecologists typically use rarefaction to down-sample the larger samples until they are the same size as the smallest sample. Ecologists then compare the richnesses of these equally large samples, but this implies that some data in larger samples are thrown away. To avoid discarding data, Colwell *et al*. (2012) proposed using a sample-size-based rarefaction (interpolation) and extrapolation (prediction) sampling curve for species richness that can be rarefied to smaller sample sizes or extrapolated to larger sample sizes.

Chao and Jost (2012) indicated that a sample of a given size may be sufficient to fully characterise a low-diversity assemblage, but insufficient to characterise a rich-assemblage. Thus, when the species counts of two equally large samples are compared, one might be comparing a nearly complete sample to a very incomplete one. This will generally lead to underestimation of the difference in diversity between the sites. The authors showed that, when compared to the traditional method

of standardisation to equal sample sizes, rarefaction and extrapolation to a given degree of sample completeness (as measured by sample coverage; see below) was superior both in terms of judging the magnitude of the differences in richness among assemblages and ranking assemblages efficiently. In order to implement this method, the authors developed a coverage-based rarefaction and extrapolation methodology for species richness. The sample-size- and coverage-based integration of rarefaction and extrapolation of species richness represent a unified sampling framework for quantifying and comparing species richness across multiple assemblages based on finite samples.

In this article, we first review some nonparametric species richness estimators based on species abundance or incidence data (to be described further), and use real data to demonstrate the application of the R package SpadeR (Species-richness Prediction And Diversity Estimation in R) to obtain estimates. We also review the sample-size- and coverage-based rarefaction and extrapolation of species richness, and illustrate the use of the R package iNEXT (iNterpolation/EXTrapolation) to compute and plot the seamless sampling curves. These methods allow researchers to efficiently use all 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.

Two Types of Biodiversity Sampling Data and Models

The notation and terminology used here generally follow that of Colwell *et al*. (2012) and Chao *et al*. (2014). Assume that there are S species in the focal assemblage, where S is the estimating target in species richness estimation. In most biological surveys, data can be generally classified into two types: individual-based abundance data and sampling-unit-based incidence data, as described in the following sections.

Individual-based abundance data and model

In the model formulation, it is assumed that the true but unknown species relative abundances or probabilities of the S species are $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Here p_i can also be interpreted as the probability that any individual is classified to the ith species. For abundance data, the sampling unit is an individual. We assume a *reference sample* of n individuals is selected with replacement, i.e. individuals can be repeatedly observed. If individuals are selected by sampling without replacement, then all methods reviewed in this article are still valid if the sample size is relatively small compared to the population size. In the reference sample, let X_i denote the sample abundance or frequency of the ith species in the reference sample, $i = 1, 2, \ldots, S$, and S_{obs} denote the number of observed species. Only those species with abundance $X \geq 1$ are detected in the sample; and those species with abundance $X = 0$ are undetected in the sample and are, therefore, not included in the data.

Define the *abundance frequency count* f_k, $k = 0, 1, 2 \ldots$, as the number of species each represented by exactly k individuals in the reference sample. Thus f_1 is the number of 'singletons' (those species represented by exactly 1 individual in the reference sample), and f_2 is the number of 'doubletons' (those represented by exactly 2 individuals in the reference sample). In this terminology, f_0 is the number of undetected species, i.e. species that are present in the assemblage of S species, but were not detected in the reference sample of n individuals and S_{obs} species.

Sampling-unit-based incidence data and model

For incidence data, the sampling unit is usually a trap, net, quadrat, plot, or timed survey, and it is these sampling units, not the individual organisms, which are sampled randomly and independently. Because it is not always possible to count individuals within a sampling unit, the estimation can be based on a set of T sampling units and only the incidence (detection or non-detection) of species in each sampling unit is recorded. In a typical study, these sampling units are deployed randomly in space within the area encompassing the assemblage. However, in a temporal study of diversity, the T sampling units would be deployed in one place at different independent points in time (such as an annual breeding bird census at a single site).

For any sampling unit, the model assumes that the ith species has its own unique incidence (or occurrence) 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 incidence probability π_i is analogous to p_i in the abundance data, but $\sum_{i=1}^{S} \pi_i$ may not necessarily be equal to unity.

A reference sample for incidence data consists of a species-by-sampling-unit incidence matrix $\{W_{ij};\ i = 1, 2, \ldots, S, j = 1, 2, \ldots, T\}$ with S rows and T columns; here $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise. Let Y_i be the number of sampling units in which species i is detected, $Y_i = \sum_{j=1}^{T} W_{ij}$; here Y_i is referred to as the sample *species incidence frequency* and is analogous to X_i in the abundance data. Species present in the assemblage but not detected in any sampling unit yield $Y = 0$.

Denote the *incidence frequency counts* by $(Q_0, Q_1, \ldots, Q_T)$, where Q_k is the number of species detected in exactly k sampling units in the data, $k = 0, 1, \ldots, T$. The count Q_k is analogous to f_k in the abundance data. Here, Q_1 represents the number of 'unique' species (those that are detected in only one sampling unit), and Q_2 represents the number of 'duplicate' species (those that are detected in only two sampling units). The unobservable zero frequency count Q_0 denotes the number of species among the S species present in the assemblage that are not detected in any of the T sampling units.

Asymptotic Approach: Nonparametric Species Richness Estimation

Because of the relationship $S_{obs} + f_0 = S$ for abundance data, and a similar relationship $S_{obs} + Q_0 = S$ for incidence data, estimating species richness is equivalent to predicting the number of undetected species (f_0 or Q_0) in the reference sample. An intuitive and basic concept is that abundant or frequent 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 counts of the detected rare species, especially the numbers of singletons and doubletons for abundance data, or the numbers of uniques and duplicates for incidence data.

Chao1- and Chao2-type estimators

When there are many undetectable or 'invisible' species in a highly diverse assemblage, it will be 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 (1984, 1987) 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* (Colwell and Coddington, 1994):

$$\widehat{S}_{\text{Chao1}} = \begin{cases} S_{obs} + \left[(n-1)/n\right][f_1^2/(2f_2)], & \text{if } f_2 > 0 \\ S_{obs} + \left[(n-1)/n\right] f_1(f_1-1)/2, & \text{if } f_2 = 0 \end{cases} \approx \begin{cases} S_{obs} + f_1^2/(2f_2), & \text{if } f_2 > 0 \\ S_{obs} + f_1(f_1-1)/2, & \text{if } f_2 = 0 \end{cases} \quad (1)$$

A greater lower bound using the additional information of tripletons and quadrupletons to estimate undetected species richness was recently derived by Chiu *et al*. (2014); the corresponding lower bound of species richness is referred to as *iChao1 estimator* (here the sub-index *i* stands for 'improved'):

$$\widehat{S}_{\text{iChao 1}} = \widehat{S}_{\text{Chao 1}} + \frac{(n-3)}{n}\frac{f_3}{4f_4} \times \max\left(f_1 - \frac{(n-3)}{(n-1)}\frac{f_2 f_3}{2f_4},\ 0\right) \approx \widehat{S}_{\text{Chao 1}} + \frac{f_3}{4f_4} \times \max\left(f_1 - \frac{f_2 f_3}{2f_4},\ 0\right) \quad (2)$$

Although both the Chao1 and iChao1 estimators were derived as lower bounds of species richness, they generally work satisfactorily as point estimators when (1) the abundances of *rare* species are nearly homogeneous, or (2) the size *n* is sufficiently large relative to species richness so that singletons and undetected species have approximately the same mean abundances (Chiu and Chao, 2016).

For incidence data, the corresponding estimator of species richness is called the *Chao2 estimator* given in the following formula (Chao, 1987):

$$\widehat{S}_{\text{Chao2}} = \begin{cases} S_{obs} + \left[(T-1)/T\right] Q_1^2/(2Q_2), & \text{if } Q_2 > 0 \\ S_{obs} + \left[(T-1)/T\right] Q_1(Q_1-1)/2, & \text{if } Q_2 = 0 \end{cases} \quad (3)$$

Unlike in the Chao1 estimator, here the factor $(T-1)/T$ cannot be neglected because T may not be sufficiently large for incidence data. Chiu *et al*. (2014) derived the corresponding *iChao 2 estimator*:

$$\widehat{S}_{\text{iChao2}} = \widehat{S}_{\text{Chao2}} + \frac{(T-3)}{4T}\frac{Q_3}{Q_4} \times \max\left(Q_1 - \frac{(T-3)}{2(T-1)}\frac{Q_2 Q_3}{Q_4},\ 0\right) \quad (4)$$

The variance formulas for the estimators in eqns 1–4 can be evaluated by the standard statistical approximation method. The above estimators were derived under sampling with replacement, i.e. individuals or sampling units can be repeatedly observed. Chao and Lin (2012) generalised the above estimators to data based on sampling without replacement, but the sampling fraction (the ratio of sample size to population size) must be known.

Coverage-based estimators (ACE- and ICE-type estimators)

The *ACE* (abundance-based coverage estimator) and *ICE* (incidence-based coverage estimator) of species richness are 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 (Good, 1953, 2000; Good and Toulmin, 1956). Briefly, the coverage of a sample represents the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample. It 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.

The ACE model assumes that the species relative abundances ($p_1, p_2, \ldots, p_S$) are fully characterised by their mean $\overline{p} = 1/S$ and CV (coefficient of variation), where the squared CV, γ^2, is defined as $\gamma^2 = \left[S^{-1}\sum_{i=1}^{S}\left(p_i - \overline{p}\right)^2\right]/\overline{p}^2$. The CV parameter is used to characterise the degree of heterogeneity among species abundances. The larger the CV, the greater 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 $> \kappa$) groups. The cut-off $\kappa = 10$ works well for many empirical data sets. For highly heterogeneous communities such as bacterial or microbial sequencing data, an alternative choice is $\kappa = \max(10, n/S_{obs})$ (Chao and Chiu, 2012).

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 $S_{\text{rare}} = \sum_{i=1}^{\kappa} f_i$. Because detected rare species contain nearly all the information about the undetected species, the ACE approach estimates the number of undetected species by using information from the rare species group. Let $n_{\text{rare}} = \sum_{i=1}^{\kappa} i f_i$ be the sample size for the rare species group. Turing's coverage estimate for this group is $\hat{C}_{\text{rare}} = 1 - f_1/n_{\text{rare}}$, which measures the sample completeness of the subsample restricted to the rare species. In the special case of homogeneous abundances for rare species (CV = 0), the coverage-based estimator is (Darroch and Ratcliff, 1980):

$$\widehat{S}_0 = S_{\text{abun}} + \frac{S_{\text{rare}}}{\widehat{C}_{\text{rare}}} \tag{5a}$$

The basic idea of the ACE (Chao and Lee, 1992; Chao, 2005) is to adjust the estimator in eqn 5a by accounting for heterogeneity. The ACE is expressed as:

$$\widehat{S}_{\text{ACE}} = S_{\text{abun}} + \frac{S_{\text{rare}}}{\widehat{C}_{\text{rare}}} + \frac{f_1}{\widehat{C}_{\text{rare}}} \widehat{\gamma}^2_{\text{rare}} \tag{5b}$$

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

$$\widehat{\gamma}^2_{\text{rare}} = \max\left\{ \frac{S_{\text{rare}}}{\widehat{C}_{\text{rare}}} \frac{\sum_{i=1}^{\kappa} i\,(i-1) f_i}{\left(\sum_{i=1}^{\kappa} i f_i\right)\left(\sum_{i=1}^{\kappa} i f_i - 1\right)} - 1, \quad 0 \right\} \tag{5c}$$

For species-rich and highly heterogeneous assemblages (say, species richness > 1000 and estimated CV for whole data > 2), the estimator $\widehat{\gamma}_{\text{rare}}$ in eqn 5c and the resulting ACE generally underestimates. In such cases, a modified estimator, ACE-1, was derived in Chao and Lee (1992). An approximate variance for the ACE and ACE-1 can be obtained using standard statistical approximation theory.

For incidence data, there is a corresponding estimator called ICE. 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 $\kappa = 10$ is recommended. Denote the number of species in the frequent group by $S_{\text{freq}} = \sum_{i>\kappa} Q_i$ and the number of species in the infrequent group by $S_{\text{infreq}} = \sum_{i=1}^{\kappa} Q_i$. The estimated sample coverage for the infrequent group is $\widehat{C}_{\text{infreq}} = 1 - Q_1 / \sum_{i=1}^{\kappa} i Q_i$. Let the number of sampling units that include at least one infrequent species be T_{infreq}. Then the ICE is expressed as

$$\widehat{S}_{\text{ICE}} = S_{\text{freq}} + \frac{S_{\text{infreq}}}{\widehat{C}_{\text{infreq}}} + \frac{Q_1}{\widehat{C}_{\text{infreq}}} \widehat{\gamma}^2_{\text{infreq}} \tag{6a}$$

where $\widehat{\gamma}^2_{\text{infreq}}$ is the estimate of the squared CV of the species incidence probabilities in the infrequent species group,

$$\widehat{\gamma}^2_{\text{infreq}} = \max\left\{ \frac{S_{\text{infreq}}}{\widehat{C}_{\text{infreq}}} \frac{T_{\text{infreq}}}{\left(T_{\text{infreq}} - 1\right)} \frac{\sum_{i=1}^{\kappa} i(i-1) Q_i}{\left(\sum_{i=1}^{\kappa} i Q_i\right)\left(\sum_{i=1}^{\kappa} i Q_i - 1\right)} - 1, \quad 0 \right\} \tag{6b}$$

A similar ICE-1 estimator for species-rich and highly heterogeneous assemblages can also be obtained (Gotelli and Chao, 2013).

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 (1989) implied that 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

$$\widehat{S}_{jk1} = S_{\text{obs}} + \frac{n-1}{n} f_1 \approx S_{\text{obs}} + f_1 \tag{7a}$$

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:

$$\widehat{S}_{jk2} = S_{\text{obs}} + \frac{2n-3}{n} f_1 - \frac{(n-2)^2}{n(n-1)} f_2 \approx S_{\text{obs}} + 2f_1 - f_2 \tag{7b}$$

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 (Burnham and Overton, 1978, 1979). For incidence data, the first- and second-order jackknife estimators are obtained by replacing (f_1, f_2) with (Q_1, Q_2), and replacing n with T in the two formulas 7a and 7b.

Extensive simulations conducted by Chiu *et al*. (2014) 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 (Palmer, 1991; Chiarucci *et al*., 2003; Walther and Moore, 2005; Xu *et al*., 2012) found the jackknife estimators to exhibit a relatively good performance. However, as this narrow range of good performance changes with each model, the theoretical behaviour is not predictable. Outside this range, the two jackknife estimators may have appreciable biases. The

jackknife estimators also exhibit counter-intuitive patterns: their bias, accuracy and coverage probability regularly do not improve as sample size increases whereas the other estimators presented above always improve.

Non-asymptotic Analysis: Rarefaction and Extrapolation

As described in the Introduction, the sample-size- and coverage-based integration of rarefaction and extrapolation represent a unified sampling framework from which to make fair and meaningful comparisons of species richness among multiple assemblages. In the following text, only the estimation for abundance data is reviewed; all the corresponding analyses for incidence data are generally parallel and are shown in **Table 1**.

Sample-size-based rarefaction and extrapolation

All samples are standardised by estimating the expected species richness for a common sample size, which can be smaller than an observed sample (traditional rarefaction) or larger than an observed sample (extrapolation). Based on a reference sample of size n, an unbiased estimator for the expected species richness in a rarefied sample of size m, $S(m)$, $m < n$, is (Hurlbert, 1971; Smith and Grassle, 1977)

$$\widehat{S}(m) = S_{\text{obs}} - \sum_{X_i>0} \frac{\binom{n-X_i}{m}}{\binom{n}{m}} \tag{8}$$

Table 1 The theoretical formulas and analytic estimators for rarefaction and extrapolation of species richness (upper half) and expected sample coverage (lower half) based on abundance data or incidence data given a reference sample with the observed species richness S_{obs} and estimated coverage $\hat{C}(n)$ for abundance data and $\hat{C}(T)$ for incidence data

Formula/estimator	Abundance data	Incidence data
Rarefaction and extrapolation of species richness		
Theoretical formula: species accumulation curve as a function of sample size m or t	$S(m) = S - \sum_{i=1}^{S} (1-p_i)^m$	$S(t) = S - \sum_{i=1}^{S} (1-\pi_i)^t$
Rarefaction estimator ($m<n$, $t<T$)	$\widehat{S}(m) = S_{\text{obs}} - \sum_{X_i>0} \frac{\binom{n-X_i}{m}}{\binom{n}{m}}$	$\widehat{S}(t) = S_{\text{obs}} - \sum_{Y_i>0} \frac{\binom{T-Y_i}{t}}{\binom{T}{t}}$
Reference sample of size n or T	$\widehat{S}(n) = S_{\text{obs}}$	$\widehat{S}(T) = S_{\text{obs}}$
Extrapolation estimator	$\widehat{S}(n+m^*) = S_{\text{obs}} + \widehat{f}_0 \left[1 - \left(1 - \frac{f_1}{n\widehat{f}_0 + f_1}\right)^{m^*}\right]$	$\widehat{S}(T+t^*) = S_{\text{obs}} + \widehat{Q}_0 \left[1 - \left(1 - \frac{Q_1}{T\widehat{Q}_0 + Q_1}\right)^{t^*}\right]$
Rarefaction and extrapolation of expected sample coverage		
Theoretical formula: coverage accumulation curve as a function of sample size m or t	$C(m) = 1 - \sum_{i=1}^{S} p_i(1-p_i)^m$	$C(t) = 1 - \frac{\sum_{i=1}^{S} \pi_i(1-\pi_i)^t}{\sum_{i=1}^{S} \pi_i}$
Rarefaction estimator ($m<n$, $t<T$)	$\widehat{C}(m) = 1 - \sum_{X_i>0} \frac{X_i}{n} \frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}$	$\widehat{C}(t) = 1 - \sum_{Y_i>0} \frac{Y_i}{U} \frac{\binom{T-Y_i}{t}}{\binom{T-1}{t}}$
Reference sample of size n or T	$\widehat{C}(n) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]$	$\widehat{C}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]$
Extrapolation estimator	$\widehat{C}(n+m^*) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]^{m^*+1}$	$\widehat{C}(T+t^*) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]^{t^*+1}$

See Colwell *et al*. (2012) and Chao and Jost (2012) for derivation details.

Notes: $U = \sum_{Y_i>0} Y_i = \sum_{j=1}^{T} jQ_j$ denotes the total number of incidences in T sampling units; $\widehat{f}_0$ and $\widehat{Q}_0$ denote the estimated number of undetected species based on the Chao1 estimator in eqn 1 and the Chao2 estimator in eqn 3, respectively.

This is traditional rarefaction form. Colwell *et al*. (2012) and Chao and Jost (2012) followed the approach of Shen *et al*. (2003) and derived the following species richness estimator for the expected number of species in an extrapolated sample of size $n+m^*$:

$$\hat{S}(n+m^*) = S_{\text{obs}} + \hat{f}_0\left[1-\left(1-\frac{f_1}{n\hat{f}_0+f_1}\right)^{m^*}\right], \quad m^*>0 \quad (9)$$

where $\hat{f}_0$ is the estimated zero-frequency based on the Chao1 estimator (eqn 1), and f_1 denotes the number of singletons. For a short-range prediction (e.g. m^* is much less than n), the prediction formula is approximately $\hat{S}(n+m^*) \approx S_{\text{obs}} + (f_1/n)m^*$, which is independent of the choice of $\hat{f}_0$. This implies that the extrapolation formula in eqn 9 is very robust and reliable even though the species richness estimator is subject to bias. Previous experiences by Colwell *et al*. (2012) suggested that the prediction size can be extrapolated at most to double the observed sample size.

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 $n+m^*$), joining smoothly at the reference point (n, S_{obs}). The confidence intervals based on the bootstrap method developed by Colwell *et al*. (2012) also join smoothly.

Coverage-based rarefaction and extrapolation

Here the concept of 'coverage' (as an objective measure of sample completeness) is discussed in the subsection *Coverage-based estimators*. Chao and Jost (2012) proposed standardising 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). Turing (Good, 1953, 2000) showed that the coverage for a reference sample of size n can be accurately estimated by the observed sample itself. Their estimator is surprisingly elegant and simple: it is just the complement of the proportion of singletons. Turing's sample coverage estimator is very efficient (Esty, 1986) and has found wide applications in various research fields. For example, Turing's coverage estimate for the rare species group, $\hat{C}_{\text{rare}} = 1-f_1/n_{\text{rare}}$, was applied in the ACE approach. Chao and Jost (2012) further refined Turing's estimator by using the additional information of doubletons; see **Table 1** for the refined formula.

Like species richness, the expected sample coverage for a hypothetical sample of size m is also a function of m. Chao and Jost (2012) derived an interpolated coverage estimator $\hat{C}(m)$ for any rarefied sample of size $m<n$ and an extrapolated coverage estimator $\hat{C}(n+m^*)$ for any enlarged sample of size $n+m^*$. All formulas are shown in **Table 1**.

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 $\hat{C}(n+m^*)$), joining smoothly at the reference sample point $(\hat{C}(n), S_{\text{obs}})$. The confidence intervals based on the bootstrap method (Chao and Jost, 2012) also join smoothly. The curve can be extended to the coverage of the maximum size used in the sample-size-based sampling curve.

The sample-size-based approach plots the estimated diversity as a function of sample size, whereas the corresponding coverage-based approach plots the same diversity 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 varies with sample size as well as 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 by an example in the following section.

Example: Comparing Beetle Species Richness of Two Sites

Janzen (1973a,b) presented many data sets of tropical foliage insects from sweep samples in Costa Rica. We selected two data sets to compare beetle species richness between an old-growth forest site and a second-growth site. The beetle abundance frequency counts (reference samples) are tabulated in **Table 2**. There were 112 species among 237 individuals in the old-growth site, and there were 140 species among 976 individuals in the second-growth site. In the raw data, fewer species (112 vs. 140) were found in the old-growth site than in the second-growth site mainly because the old-growth site has a much smaller sample size (237 vs. 976) and lower sample coverage (64.6% vs. 92.8%) compared to the second-growth site. Here, the sample coverage estimates are obtained by the formulas given in **Table 1**.

Traditional sample-size-based rarefaction analysis would rarefy the second-growth sample size down to 237 and conclude that for a standardised size of 237, the old-growth site has more species (112 vs. 70). Below we apply two R packages to analyse the data and demonstrate a more informative comparison between these two sites.

Asymptotic approach: 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 at each site. SpadeR is available from Github and can also be downloaded from Anne Chao's website http://chao.stat.nthu.edu.tw/wordpress/software_download/ and soon will be available in CRAN. 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: (Summary data information and some estimators not presented in this article are omitted.)

```
> install.packages("devtools")
> library(devtools)
> install_github("AnneChao/spadeR")
> library(SpadeR)
```

Table 2 The species abundance frequency counts for beetles from two sites on the Osa Peninsula in southwestern Costa Rica (Janzen, 1973a,b), where f_i denotes the number of species represented by i individuals in the sample

(a) Osa old-growth site: $S_{obs} = 112$, $n = 237$, sample coverage estimate = 64.6%

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

(b) Osa second-growth site: $S_{obs} = 140$, $n = 976$, sample coverage estimate = 92.8%

i	1	2	3	4	5	6	7	8	9	10	11	12	14	17	19
f_i	70	17	4	5	5	5	5	3	1	2	3	2	2	1	2
i	20	21	24	26	40	57	60	64	71	77					
f_i	3	1	1	1	1	2	1	1	1	1					

```
> SecondGrowth = rep(c(1:12,14,17,19,20,21,24,
  26,40,57,60,64,71,77),c(70,17,4,5,5,5,3,1,
  2,3,2,2,1,2,3,1,1,1,1,2,1,1,1,1))
> OldGrowth=rep(c(1,2,3,4,5,6,7,8,14,42),
  c(84,10,4,3,5,1,2,1,1,1))
> Forest=list("Second-growth" = SecondGrowth,
  "Old-growth" = OldGrowth)
> out1=ChaoSpecies(Forest$"Second-growth",
  datatype = "abundance", k=10,conf=0.95)
  #Output for the second-growth site
> out2 = ChaoSpecies(Forest$"Old-growth",
  datatype = "abundance", k=10,conf=0.95)
  #Output for the old-growth site
> out1 #Show the output of the
  second-growth site


$Species.Table

              Estimate  s.e.  95%Lower  95%Upper
Chao1         283.970  50.474  213.868   420.600
 (Chao, 1984)
iChao1        296.610  41.835  233.615   401.995
(Chiu et al.
 2014)
ACE (Chao &   273.535  36.345  219.077   365.498
 Lee, 1992)
ACE-1(Chao &  386.977  83.476  269.630   610.555
 Lee, 1992)
1st order     209.928  11.823  190.321   237.175
 jackknife
2nd order     262.837  20.466  228.813   309.895
 jackknife


> out2  #Show the output of the old-growth site
$Species.Table


              Estimate  s.e.    95%Lower 95%Upper
Chao1          463.311  136.273  280.660  843.766
 (Chao, 1984)
iChao1         489.089  125.793  311.502  824.755
(Chiu et al.
 2014)
ACE (Chao &    396.138   82.504  274.687  608.256
 Lee, 1992)
ACE-1 (Chao &  714.185  223.088  410.176 1328.148
 Lee, 1992)
1st order      195.646   12.920  173.906  225.019
 jackknife
2nd order      269.063   22.329  231.033  319.243
 jackknife
```

From the above output, the Chao1, iChao1, and ACE all give consistent estimates between 270 and 300 for the second-growth site, whereas their estimates for the old-growth site are between 400 and 500. 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 (in output not shown here) for both sites are not extremely high. Each of the estimators reveals that the species richness of the old-growth site is higher than that in the second-growth site, although the 95% confidence intervals overlap. In contrast, the first order jackknife estimator shows the reverse ordering, while the second order jackknife estimator implies that the richnesses in the two sites differ little. The two jackknife estimates are much lower than any of the Chao1, iChao1 and ACE estimates. Based on results of our previous simulations for heterogeneous models, the two jackknife estimators might underestimate for relatively low sample sizes such as those in these data sets.

Non-asymptotic approach: rarefaction and 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 **Figures 1**, **2** and **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=1200)
> ggiNEXT(out, type=1) #plot sample-size-based
```

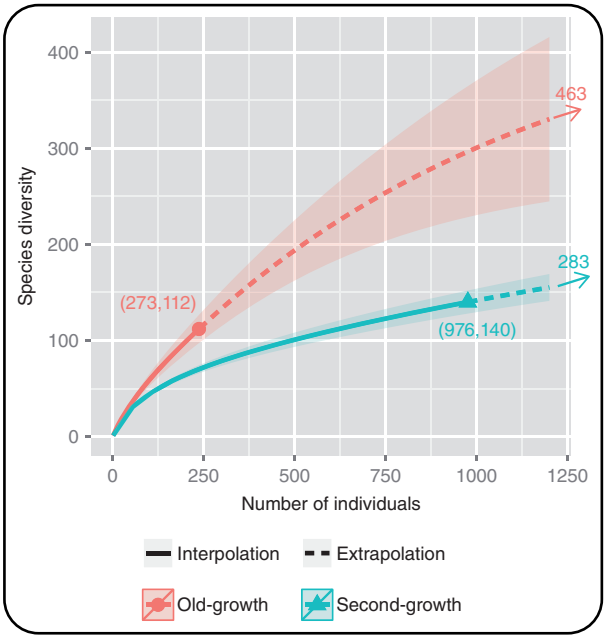

Figure 1 Sample-size-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing beetle species richness between an old-growth site and a second-growth site (Janzen, 1973a,b). Observed samples are denoted by the solid dot and triangle. The extrapolation extends up to a maximum sample size of 1200. For each reference sample point, the numbers in parentheses show the *x*- and *y*-axis coordinate. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.

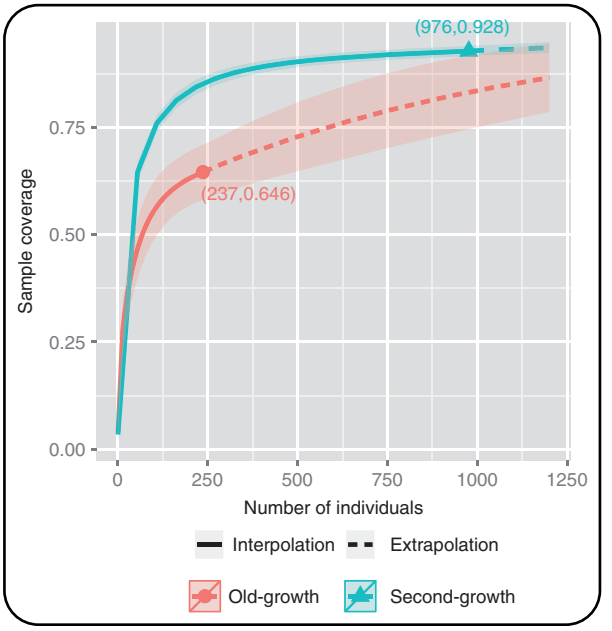

Figure 2 Sample completeness curve which depicts how sample completeness (measured by sample coverage) increases with sample size for beetle species data in an old-growth site and a second-growth site (Janzen, 1973a,b). For each site, the plot of sample coverage for rarefied samples (solid lines) and 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 1200. 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.

```
  curve (as shown in Fig. 1)
> ggiNEXT(out, type=2) #plot sample
  completeness curve (as shown in Fig. 2)
> ggiNEXT(out, type=3) #plot coverage-based
  curve (as shown in Fig. 3)
> out #to show the detailed output for
  related statistics
```

In the sample-size-based rarefaction and extrapolation sampling curve (**Figure 1**), we compare two equally large samples. For each site, the extrapolation is extended to a maximum size of 1200 (by specifying `endpoint = 1200` in the iNEXT function). For the old-growth site, the extrapolation exceeds the suggested maximum size (double reference sample size). Extrapolation beyond the double reference sample size theoretically could 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 the old-growth site is more diverse; the confidence intervals of the two sites do not overlap for any sample size considered in the figure (except for the initial small sizes). This implies that beetle species richness is significantly greater in the old-growth site than that in the second-growth site for sample size up to 1200 individuals.

The sample completeness curve in **Figure 2** shows how the sample coverage varies with sample size. The curve of the second-growth site lies above that of the old-growth site. For the old-growth site, when the sample size is extended from 237 to 1200, the sample coverage is extended from 65% to 86.6% (as shown in **Figure 2** or the unreported iNEXT output). For the second-growth site, when the sample size is extended from 976 to 1200 the coverage is extended from 92.8% to 93.6% (as shown in **Figure 2** or the unreported iNEXT output), an increment of only 0.8%.

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 86.6% for the old-growth site and to 93.6% for the second-growth site, as explained in the preceding paragraph. Except for very low coverage values, the beetle species richness in the old-growth site is significantly higher than that of the second-growth site as evidenced by the non-overlapping confidence intervals for any fixed coverage value up to 86.6%. Unlike the sized-based standardisation in which size is determined by samplers, here the coverage-based standardisation compares equal population fractions of each assemblage. The population fraction is an assemblage-level characteristic that can be reliably estimated from data. Note that significant differences cannot be guaranteed based on each of the three species richness estimators (Chao1, iChao1 and ACE) due to the overlapped confidence intervals. That means, if we compare species richness for the two entire assemblages (i.e. data are extrapolated to coverage

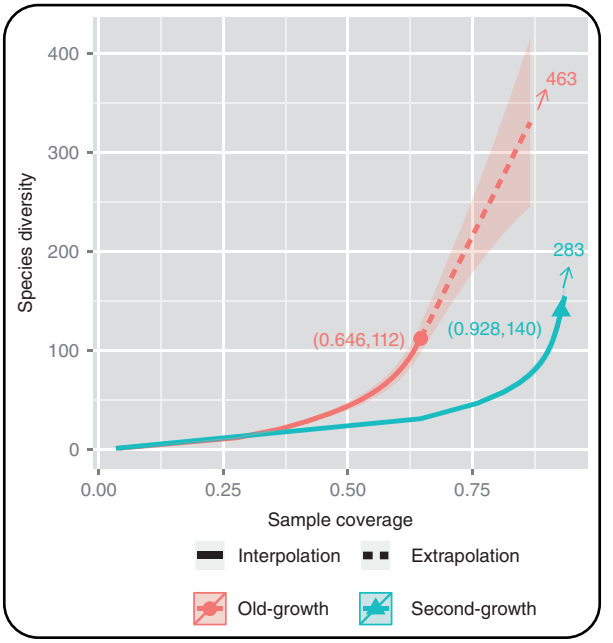

Figure 3 Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing beetle species richness between an old-growth site and a second-growth site (Janzen, 1973a,b). Observed samples are denoted by the solid dot and triangle. The extrapolation extends up to the coverage value of the corresponding maximum sample size of 1200 in **Figure 2** (86.6% in the old-growth site, and 93.6% in the second-growth site). For each reference sample point, the numbers in parentheses show the *x*- and *y*-axis coordinate. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.

of unity), then data do not provide sufficient evidence to conclude significant difference in species richness between the two sites. On the other hand, if we only compare species richness for equal population fractions up to 86.6%, then data do provide sufficient information to conclude significance.

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 more robust and meaningful comparisons of species richness between assemblages for a wide range of sample sizes/completeness. These methods have also been generalised to diversity measures that incorporate species abundances (Chao *et al*., 2014) and those that take into account the evolutionary history among species (Chao *et al*., 2015).

References

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

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

Burnham KP and Overton WS (1978) Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**: 625–633.

Burnham KP and Overton WS (1979) Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60** (5): 927–936.

Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11** (4): 265–270.

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

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

Chao A (2005) Species estimation and applications. In: Balakrishnan N, Read CB and Vidakovic B (eds) *Encyclopedia of Statistical Sciences*, pp. 7907–7916. New York: John Wiley & Sons, Inc.

Chao A and Chiu C-H (2012) Estimation of species richness and shared species richness. In: Balakrishnan N (ed.) *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, pp. 76–111. New York: John Wiley & Sons, Inc.

Chao A and Jost L (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* **93** (12): 2533–2547.

Chao A and Lin C-W (2012) Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics* **68** (3): 912–921.

Chao A, Gotelli NJ, Hsieh TC, *et al.* (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84** (1): 45–67.

Chao A, Chiu C-H, Hsieh TC, *et al.* (2015) Rarefaction and extrapolation of phylogenetic diversity. *Methods in Ecology and Evolution* **6**: 380–388.

Chiarucci A, Enright NJ, Perry GLW, *et al.* (2003) Performance of nonparametric species richness estimators in a high diversity plant community. *Diversity and Distributions* **9** (4): 283–295.

Chiarucci A, Bacaro G, Rocchini D, *et al.* (2008) Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Community Ecology* **9** (1): 121–123.

Chiu C-H and Chao A (2016) Estimating and comparing microbial diversity when singletons are subject to sequencing errors. *PeerJ*. **4**: e1634.

Chiu C-H, Wang YT, Walther BA, *et al.* (2014) An improved nonparametric lower bound of species richness via a modified good–Turing frequency formula. *Biometrics* **70** (3): 671–682.

Coleman BD, Mares MA, Willig MR, *et al.* (1982) Randomness, area, and species richness. *Ecology* **63** (4): 1121–1133.

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

Colwell RK, Chao A, Gotelli NJ, *et al.* (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* **5** (1): 3–21.

Cormack RM (1989) Log-linear models for capture-recapture. *Biometrics* **45** (2): 395–413.

Darroch JN and Ratcliff D (1980) A note on capture-recapture estimation. *Biometrics* **36**: 149–153.

Esty WW (1986) The efficiency of Good's nonparametric coverage estimator. *The Annals of Statistics* **14**: 1257–1260.

Fisher RA, Corbet AS and Williams CB (1943) The relation between the number of species and the number of individuals in a random

sample of an animal population. *Journal of Animal Ecology* **12**: 42–58.

Good IJ (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* **40**: 237–264.

Good IJ and Toulmin G (1956) The number of new species and the increase of population coverage when a sample is increased. *Biometrika* **43**: 45–63.

Good IJ (2000) Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. *Journal of Statistical Computation and Simulation* **66** (2): 101–111.

Gotelli NJ and Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* **4** (4): 379–391.

Gotelli NJ and Colwell RK (2011) Estimating species richness. In: Magurran AE and McGill BJ (eds) Biological Diversity*:* Frontiers in Measurement and Assessment, pp. 39–54. Oxford: Oxford University Press.

Gotelli NJ and Chao A (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin SA (ed.) *Encyclopedia of Biodiversity*, 2nd edn, vol. **5**, pp. 195–211. Waltham, MA: Academic Press.

Heck KL Jr, van Belle G and Simberloff D (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* **56**: 1459–1461.

Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52** (4): 577–586.

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

Janzen DH (1973b) Sweep samples of tropical foliage insects: effects of seasons, vegetation types, elevation, time of day, and insularity. *Ecology* **54**: 687–708.

Magurran AE (2004) *Measuring Biological Diversity*. Oxford: Blackwell.

Magurran AE and McGill BJ (2011) *Biological diversity: Frontiers in Measurement and Assessment*. Oxford: Oxford University Press.

O'Hara RB (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology* **74** (2): 375–386.

Ord JK and Whitmore GA (1986) The Poisson-inverse Gaussian distribution as a model for species abundance. *Communications in Statistics A – Theory and Methods* **15** (3): 853–871.

Palmer MW (1991) Estimating species richness: the second-order jackknife reconsidered. *Ecology* **72** (4): 1512–1513.

Pielou E (1977) *Mathematical Ecology*. New York: John Wiley & Sons, Inc.

Preston FW (1948) The commonness and rarity of species. *Ecology* **29** (3): 254–283.

Sanders HL (1968) Marine benthic diversity: a comparative study. *American Naturalist* **102**: 243–282.

Shen T-J, Chao A and Lin J-F (2003) Predicting the number of new species in further taxonomic sampling. *Ecology* **84** (3): 798–804.

Sichel HS (1997) Modeling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. *South African Statistical Journal* **31** (1): 13–37.

Simberloff D (1979) Rarefaction as a distribution-free method of expressing and estimating diversity. In: Grassle JF, Patil GP, Smith WK and Taillie C (eds) *Ecological Diversity in Theory and Practice*, pp. 159–176. Fairland, MD: International Cooperative Publishing House.

Smith W and Grassle JF (1977) Sampling properties of a family of diversity measures. *Biometrics* **33** (2): 283–292.

Walther BA and Moore JL (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **28** (6): 815–829.

Xu H, Liu S, Li Y, *et al.* (2012) Assessing non-parametric and area-based methods for estimating regional species richness. *Journal of Vegetation Science* **23** (6): 1006–1012.

Further Reading

Chao A, Colwell RK, Lin CW, *et al.* (2009) Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **90** (4): 1125–1133.

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. *Annual Reviews of Ecology, Evolution, and Systematics* **45**: 297–324.

Chao A and Jost L (2015) Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution* **6** (8): 873–882.

Hughes JB, Hellmann JJ, Ricketts TH, *et al.* (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* **67** (10): 4399–4406.

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

Legendre P and Legendre L (2012) *Numerical Ecology*, 3rd edn. Amsterdam: Elsevier.