Journal of
Plant Ecology

VOLUME 5, NUMBER 1,
PAGES 3–21

MARCH 2012

doi: 10.1093/jpe/rtr044

available online at
www.jpe.oxfordjournals.org

Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages

Robert K. Colwell[1,], Anne Chao[2], Nicholas J. Gotelli[3], Shang-Yi Lin[2], Chang Xuan Mao[4], Robin L. Chazdon[1] and John T. Longino[5]*

[1] *Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA*
[2] *Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan*
[3] *Department of Biology, University of Vermont, Burlington, VT 05405, USA*
[4] *School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China*
[5] *Department of Biology, University of Utah, Salt Lake City, UT 84112, USA*
*Correspondence address. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA. E-mail: colwell@uconn.edu

Abstract

Aims

In ecology and conservation biology, the number of species counted in a biodiversity study is a key metric but is usually a biased underestimate of total species richness because many rare species are not detected. Moreover, comparing species richness among sites or samples is a statistical challenge because the observed number of species is sensitive to the number of individuals counted or the area sampled. For individual-based data, we treat a single, empirical sample of species abundances from an investigator-defined species assemblage or community as a reference point for two estimation objectives under two sampling models: estimating the expected number of species (and its unconditional variance) in a random sample of (i) a smaller number of individuals (multinomial model) or a smaller area sampled (Poisson model) and (ii) a larger number of individuals or a larger area sampled. For sample-based incidence (presence–absence) data, under a Bernoulli product model, we treat a single set of species incidence frequencies as the reference point to estimate richness for smaller and larger numbers of sampling units.

Methods

The first objective is a problem in interpolation that we address with classical rarefaction (multinomial model) and Coleman rarefaction (Poisson model) for individual-based data and with sample-based rarefaction (Bernoulli product model) for incidence frequencies. The second is a problem in extrapolation that we address with sampling-theoretic predictors for the number of species in a larger sample (multinomial model), a larger area (Poisson model) or a larger number of sampling units (Bernoulli product model), based on an estimate of asymptotic species richness. Although published methods exist for many of these objectives, we bring them together here with some new estimators under a unified statistical and notational framework. This novel integration of mathematically distinct approaches allowed us to link interpolated (rarefaction) curves and extrapolated curves to plot a unified species accumulation curve for empirical examples. We provide new, unconditional variance estimators for classical, individual-based rarefaction and for Coleman rarefaction, long missing from the toolkit of biodiversity measurement. We illustrate these methods with datasets for tropical beetles, tropical trees and tropical ants.

Important Findings

Surprisingly, for all datasets we examined, the interpolation (rarefaction) curve and the extrapolation curve meet smoothly at the reference sample, yielding a single curve. Moreover, curves representing 95% confidence intervals for interpolated and extrapolated richness estimates also meet smoothly, allowing rigorous statistical comparison of samples not only for rarefaction but also for extrapolated richness values. The confidence intervals widen as the extrapolation moves further beyond the reference sample, but the method gives reasonable results for extrapolations up to about double or triple the original abundance or area of the reference sample. We found that the multinomial and Poisson models produced indistinguishable results, in units of estimated species, for all estimators and datasets. For sample-based abundance data, which allows the comparison of all three models, the Bernoulli product model generally yields lower richness estimates for rarefied data than either the multinomial or the Poisson models because of the ubiquity of non-random spatial distributions in nature.

Keywords: Bernoulli product model • Coleman curve • multinomial model • Poisson model • random placement • species–area relation

Received: 20 July 2011 Revised: 15 October 2011 Accepted: 17 October 2011

INTRODUCTION

Exhaustive biodiversity surveys are nearly always impractical or impossible (Lawton et al. 1998), and the difficulties inherent in estimating and comparing species richness from sampling data are well known to ecologists and conservation biologists. Because species richness increases non-linearly with the number of individuals encountered, the number of samples collected or the area sampled, observed richness is inevitably a downward biased estimate of true richness. 'Adjustment' for differences in sampling effort by calculating simple ratios of species per individual or species per unit of sampling effort seriously distorts richness values and should never be relied upon (Chazdon et al. 1999). For assessing and comparing species accumulation curves or rarefaction curves, methods that are based on an explicit statistical sampling model provide a straightforward resolution for many applications (Gotelli and Colwell 2011).

In many biodiversity studies, the basic units are individuals, ideally sampled randomly and independently, counted and identified to species. We refer to such data as 'individual based'. In many other studies, the sampling unit is not an individual, but a trap, net, quadrat, plot or a fixed period of survey time. It is these sampling units, and not the individual organisms, that are sampled randomly and independently. If the number of individuals for each species appearing within each sampling unit can be measured or approximated, we refer to the resulting data as 'sample-based abundance data'. For many organisms, especially microorganisms, invertebrates or plants, only the incidence (presence or absence) of each species in each sampling unit can be accurately recorded. We refer to such a dataset as 'sample-based incidence data' (Gotelli and Colwell 2001).

For individual-based data, we treat a single, empirical sample of species abundances, which we refer to as the 'abundance reference sample' from an investigator-defined species assemblage or community as a reference point for two estimation problems: (i) estimating the expected number of species (and its variance) in a random sample of a smaller number of individuals or a smaller area sampled and (ii) estimating the number of species (and its variance) that might be expected in a larger number of individuals or a larger area sampled. The first is an 'interpolation' problem that is addressed with classical rarefaction and Coleman rarefaction. The second is an 'extrapolation' problem that we address with sampling-theoretic predictors for the number of species in a larger sample or larger area based on an estimated asymptotic species richness.

For sample-based incidence data, the statistical equivalent of the abundance reference sample is the 'incidence reference sample', the set of incidence frequencies among the sampling units, one frequency for each observed species over all sampling units. For interpolation and extrapolation, we treat these incidence frequencies in nearly the same way that we treat the list of species abundances in a single abundance reference sample (with appropriate statistical modifications), to estimate richness for smaller and larger numbers of sampling units. For sample-based abundance data, the abundances are either first converted to incidences (presence or absence) before applying incidence-based methods or else abundances are summed across sampling units and individual-based (abundance) methods are applied to the sums.

Both interpolation and extrapolation from an empirical reference sample can be viewed as estimating the form of the underlying species accumulation curve. This curve is a plot of species richness as a function of the number of individuals or sampling units, including both smaller and larger numbers of individuals or sampling units than in the reference sample. We model the species accumulation curve as asymptotic to an estimate of the species richness of the larger community or assemblage represented by the empirical reference sample (Fig. 1).

In this paper, for the interpolation (rarefaction) problem for individual-based (abundance) data, we present a unified statistical framework for two distinct approaches: (i) a multinomial model for classical rarefaction (Heck et al. 1975; Hurlbert 1971; Sanders 1968; Simberloff 1979; Smith and Grassle 1977) and (ii) a continuous Poisson model for Coleman's 'random-placement' rarefaction method (Coleman 1981; Coleman et al. 1982). For sample-based (incidence) data, we present a 'Bernoulli product model' for sample-based rarefaction (Shinozaki 1963, Ugland et al. 2003 and Colwell et al. 2004, with an instructive historical perspective by Chiarucci et al. 2008).

For the extrapolation problem, we present—in the same unifying statistical framework as for interpolation—non-parametric methods for projecting rarefaction curves beyond the size of the reference sample under all three models. For the multinomial model, first explored for extrapolation by Good and Toulmin (1956), we rely on published work by Solow and Polasky (1999), Shen et al. (2003) and Chao et al. (2009). For the Poisson model, we follow Chao and Shen (2004), in which the pioneering work by Good and Toulmin (1956) was discussed. Alternative approaches to the extrapolation of individual-based rarefaction curves include the little-used 'abundification' method of Hayek and Buzas (1997) and the mixture model of Mao (2007). Extrapolating sample-based rarefaction curves

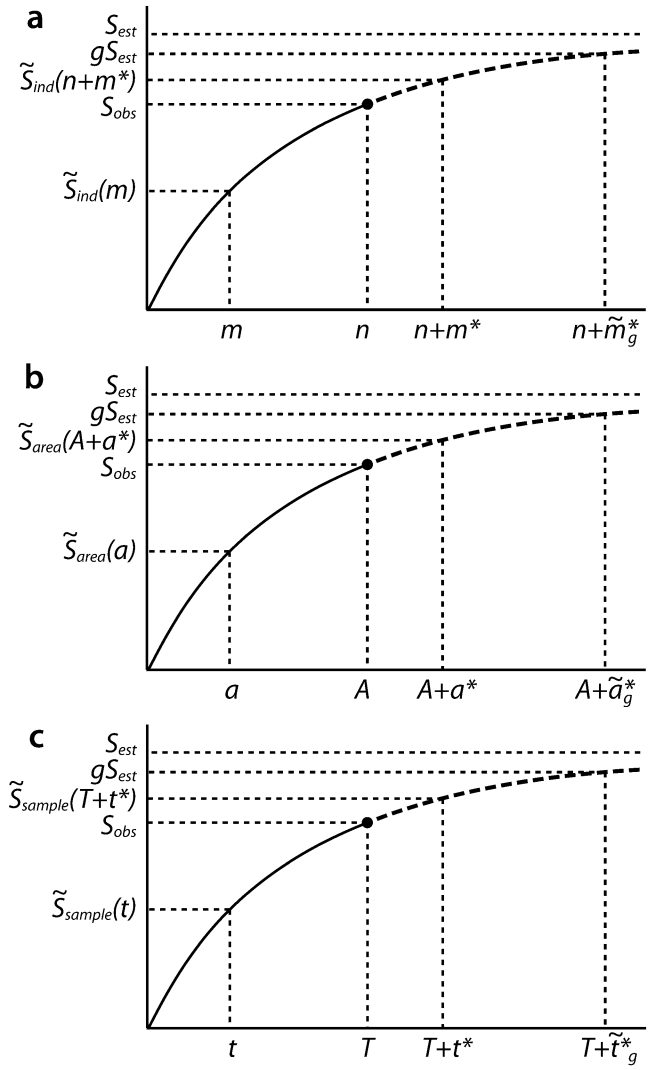

Figure 1: concepts and notation for interpolation (solid curves) and extrapolation (dashed curves) from an abundance reference sample (individual-based models) or an incidence reference sample (sample-based models), indicated by filled black circles, under three statistical models: (**a**) the multinomial model, (**b**) the Poisson model and (**c**) the Bernoulli product model. S_{est} is the estimated asymptotic number of species in the assemblage. The reference sample of n individuals (multinomial), the individuals found in area A (Poisson) or T sampling units (Bernoulli product model) reveals S_{obs} species. Interpolation (rarefaction) shows the estimated number of species $\tilde{S}_{\text{ind}}(m)$ found among m individuals, $m < n$ (multinomial, Equation 4), the estimated number of species $\tilde{S}_{\text{area}}(a)$ found area a, $a < A$ (Poisson, Equation 6), or the estimated number of species $\tilde{S}_{\text{sample}}(t)$ found in t sampling units, $t < T$ (Bernoulli product, Equation 17). Extrapolation shows the estimated number of species $\tilde{S}_{\text{ind}}(n+m^*)$ found among an augmented sample of $n+m^*$ (multinomial, Equation 9) individuals, the estimated number of species $\tilde{S}_{\text{area}}(A+a^*)$ found in a larger area $A+a^*$ (Poisson, Equation 12) or the estimated number of species $\tilde{S}_{\text{sample}}(T+t^*)$ found in $T+t^*$ sampling units (Bernoulli product, Equation 18), For extrapolation, $\tilde{m}^*_g$ estimates the number of additional individuals (multinomial, Equation 11), $\tilde{a}^*_g$ the additional area (Poisson, Equation 14) and $\tilde{t}^*_g$ the additional number of sampling units (Bernoulli product, Equation 20), required to reach proportion g of the asymptotic richness S_{est}.

beyond the incidence reference sample has been investigated by Colwell et al. (2004), Mao et al. (2005), Mao and Colwell (2005) and Mao (2007), but those methods, although theoretically useful and flexible, are based on rather complicated mixture models. Here, we take a simpler approach to extrapolation for the Bernoulli product model (e.g. Burnham and Overton 1978), in hopes that it will be more widely applied.

The principal use of rarefaction curves has long been the comparison of species richness among empirical samples that differ in the total number of individuals (e.g. Lee et al. 2007; Sanders 1968, among many others) or among sample-based datasets that differ in the total number of sampling units (e.g. Longino and Colwell 2011; Norden et al. 2009, among many others). Rigorous comparison of rarefaction curves at a common number of individuals or a common number of sampling units requires computation of confidence intervals for these curves. However, existing variance estimators for individual-based (classical) rarefaction (Heck et al. 1975) and for Coleman rarefaction (Coleman et al. 1982) are not appropriate for this purpose because they are conditional on the reference sample.

For sample-based rarefaction, Colwell et al. (2004) derived an unconditional variance estimator, which we use as a model to develop simple, approximate expressions for the unconditional variance for both classical rarefaction and Coleman's random-placement rarefaction, long missing from the toolkit of biodiversity measurement and estimation for individual-based data (Gotelli and Colwell 2011). These unconditional variance expressions assume that the reference sample represents a random draw from a larger (but unmeasured) community or species assemblage, so that confidence intervals for rarefaction curves remain 'open' at the full-sample end of the curve. In contrast, traditional variance estimators for rarefaction (e.g. Heck et al. 1975; Ugland et al. 2003) are conditional on the sample data, so that the confidence interval closes to zero at the full-sample end of the curve, making valid comparisons of curves and their confidence intervals inappropriate for inference about larger communities or species assemblages. For all three models, we also provide unconditional variance estimators for extrapolation, modeled on the estimators of Shen et al. (2003) and Chao and Shen (2004).

For individual-based methods, we illustrate interpolation, extrapolation and comparison between reference samples from different assemblages using datasets from old-growth and nearby second-growth forests in two regions of Costa Rica. One dataset, from southwestern Costa Rica, is for beetles (Janzen 1973a, 1973b) and the other, from northeastern Costa Rica, is for trees (Norden et al. 2009). We illustrate sample-based methods with biogeographical data for Costa Rican ants sampled at five elevations along an elevational transect (Longino and Colwell 2011). We use the unconditional variance formulas to construct 95% confidence intervals for both interpolated and extrapolated values. For extrapolation, we also show, for all three models, how to estimate the sample size required to reach a specified proportion of the estimated

asymptotic species richness, following the approach of Chao and Shen (2004) and Chao et al. (2009).

THE MODELS

Individual-based (abundance) data

Consider a species assemblage consisting of N individuals, each belonging one of S different species. Species i has N_i individuals, representing proportion $p_i = N_i/N$ of the total, $\sum_{i=1}^{S} N_i = N$. A single, representative sample of n individuals, the reference sample, is drawn at random from the assemblage, from an area A units in size. Each individual in the reference sample is identified to species (or to some other consistently applied taxonomic rank, DNA sequence similarity or functional group assignment). The total number of species observed in the sample is S_{obs}, with the ith species represented by X_i individuals, $\sum_{i=1}^{S} X_i = n$ (only species with $X_i > 0$ contribute to S_{obs} in the reference sample). We define the 'abundance frequency count' f_k as the number of species each represented by exactly $X_i = k$ individuals in the reference sample, $0 \leqslant k \leqslant n$. Formally, $f_k = \sum_{i=1}^{S} I(X_i = k)$, where $I(\cdot)$ is an indicator function that equals 1 when true and 0 otherwise, so that $\sum_{k=1}^{n} k f_k = n$, $S_{\text{obs}} = \sum_{k=1}^{n} f_k$. The number of species present in the assemblage but not detected in the reference sample is thus represented as f_0.

For most assemblages, no sampling method is completely unbiased in its ability to detect individuals of all species (e.g. Longino and Colwell 1997). For this reason, a 'representative' sample is necessarily defined as one that is random within the capabilities of the sampling method in relation to the taxon sampled. We use the term 'assemblage' to refer to the set of all individuals that would be detected with this sampling method in a very large sample. In other words, we assume in this paper that the assemblage is the effectively infinite sampling universe from which the reference sample has been collected.

We consider two alternative sampling models for individual-based (abundance) data. In the 'multinomial model' for classical, individual-based rarefaction (Hurlbert 1971), the reference sample is of fixed size n, within which discrete and countable organisms are assumed to be distributed among species multinomially. The assemblage has S species, in relative abundances (proportions) $p_1, p_2, \ldots p_S$, so that the probability distribution is

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

The multinomial model assumes that the sampling procedure itself does not substantially alter relative abundances of species $(p_1, p_2, \ldots, p_S)$. We assume that, in most biological applications, the biological populations in the assemblage being sampled are sufficiently large that this assumption is met. If this assumption is not met, the hypergeometric model, which describes sampling without replacement, is technically more appropriate (Heck et al. 1975), but in practice the two probability distributions differ little if sample size (n) is small relative to assemblage size (N).

In the 'continuous Poisson model' or Coleman rarefaction (Coleman 1981), the reference sample is defined not by n, the number of individuals sampled, but instead by a specified area A (or a specified period of time), within which the ith species occurs at a species-specific mean rate $A\lambda_i$, so that the probability distribution is

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

Based solely on information in the reference sample of n individuals or the individuals from area A, counted and identified to species, we have these six complementary objectives for abundance-based data (Fig. 1a and b): (i) to obtain an estimator $\tilde{S}_{\text{ind}}(m)$ for the expected number of species in a random sample of m individuals from the assemblage ($m < n$) or (ii) an estimator $\tilde{S}_{\text{area}}(a)$ for the expected number of species in a random area of size a within the reference area of size A ($a < A$); (iii) to obtain an estimator $\tilde{S}_{\text{ind}}(n + m^*)$ for the expected number of species in an augmented sample of $n + m^*$ individuals from the assemblage ($m^* > 0$), given S_{obs}, or (iv) an estimator $\tilde{S}_{\text{area}}(A + a^*)$ for the expected number of species in an augmented area $A + a^*$ ($a^* > 0$), given S_{obs}; and (v) to find an predictor $\tilde{m}_g^*$ for the number of additional individuals or (vi) the additional area $\tilde{a}_g^*$ required to detect proportion g of the estimated assemblage richness S_{est}.

Sample-based incidence data

Consider a species assemblage consisting of S different species, each of which may or may not be found in each of T independent sampling units (quadrats, plots, traps, microbial culture plates, etc.) The underlying data consist of a species-by-sampling-unit incidence matrix, in which $W_{ij} = 1$, if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise. 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 = 1,2, …, S. The frequencies Y_i represent the incidence reference sample to be rarefied or extrapolated. The total number of species observed in the reference sample is S_{obs} (only species with $Y_i > 0$ contribute to S_{obs}). We define the 'incidence frequency count' Q_k as the number of species each represented exactly $Y_i = k$ times in the incidence matrix sample, $0 \leqslant k \leqslant T$. Formally, $Q_k = \sum_{i=1}^{S} I(Y_i = k)$, so that $\sum_{k=1}^{T} k Q_k = \sum_{i=1}^{S} Y_i$, $S_{\text{obs}} = \sum_{k=1}^{T} Q_k$. Thus, Q_1 represents the number of 'unique' species (those that are detected in only one sample) and Q_2 represents the number of 'duplicate' species (those that are detected in only two samples), in the terminology of Colwell and Coddington (1994), while Q_0 denotes the number of species among the S species in the assemblage that were not detected in any of the T sampling units.

For sample-based incidence data, we consider a Bernoulli product model for an incidence reference sample arising from incidence frequencies in a fixed number T of replicate sampling units. Assume that the probability of detecting species i in any

one sample is θ_i, for $i = 1, 2, \ldots, S$. Here, $\sum_{i=1}^{S} \theta_i$ may be greater than 1. (For example, the detection probability of the first species might be 0.6 and for the second species 0.8.) We assume that each W_{ij} is a Bernoulli random variable (since $W_{ij} = 0$ or $W_{ij} = 1$), with probability θ_i that $W_{ij} = 1$. Thus, the probability distribution for the incidence matrix is

$$P\left(W_{ij} = w_{ij};\ \ i = 1, 2, \ldots, S;\ \ j = 1, 2, \ldots, T\right) = \prod_{j=1}^{T} \prod_{i=1}^{S} \theta_i^{w_{ij}} (1 - \theta_i)^{1 - w_{ij}} = \prod_{i=1}^{S} \theta_i^{y_i} (1 - \theta_i)^{T - y_i}. \quad (3)$$

This model has been widely used in the context of capture–recapture models (e.g. Burnham and Overton 1978). The row sums $(Y_1, Y_2, \ldots, Y_S)$ are the sufficient statistics, and our analysis is based on the incidence frequency counts Q_k defined from $(Y_1, Y_2, \ldots, Y_S)$.

Based solely on information in the incidence reference sample of T sampling units, we have these three complementary objectives for sample-based incidence data (Fig. 1c): (i) to obtain an estimator $\tilde{S}_{\text{sample}}(t)$ for the expected number of species in a random set of t sampling units from the T sampling units defining the reference sample ($t < T$), (ii) to obtain an estimator $\tilde{S}_{\text{sample}}(T + t^*)$ for the expected number of species in an augmented set of $T + t^*$ sampling units ($t^* > 0$) from the assemblage, given S_{obs}, and (iii) to find a predictor $\tilde{t}_g^*$ for the number of additional sampling units required to detect proportion g of the estimated assemblage richness S_{est}.

INDIVIDUAL-BASED INTERPOLATION (RAREFACTION)

The multinomial model (classic rarefaction)

For the multinomial model (classical rarefaction), we need to estimate the expected number of species $S_{\text{ind}}(m)$ in a random set of m individuals from the reference sample ($m < n$) (Fig. 1a). If we knew the true occurrence probabilities $(p_1, p_2, \ldots, p_S)$ of each of the S species in the assemblage, we could compute

$$S_{\text{ind}}(m) = \sum_{i=1}^{S} [1 - (1 - p_i)^m] = S - \sum_{i=1}^{S} (1 - p_i)^m.$$

Instead, we have only the reference sample to work from, with observed species abundances X_i. Smith and Grassle (1977) proved that the minimum variance unbiased estimator (MVUE) for $S_{\text{ind}}(m)$ is

$$\tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{X_i > 0} \left[\binom{n - X_i}{m} \Big/ \binom{n}{m} \right].$$

They showed that this expression is also the MVUE for the hypergeometric rarefaction model, which assumes sampling without replacement. Because the MVUE is the same for the hypergeometric and the multinomial models, we can relax our assumption about sampling effects on assemblage abundances. In terms of frequency counts f_k, the estimator becomes

$$\tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{k=1}^{n} \left[\binom{n - k}{m} \Big/ \binom{n}{m} \right] f_k. \quad (4)$$

If we define

$$\alpha_{km} = \binom{n - k}{m} \Big/ \binom{n}{m} = \frac{(n - k)!(n - m)!}{n!(n - k - m)!} \quad \text{for } k \leqslant n - m,$$

$$\alpha_{km} = 0 \text{ otherwise},$$

then

$$\tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{k=1}^{n} \alpha_{km} f_k.$$

Assume that the occurrence probabilities $(p_1, p_2, \ldots, p_S)$ can be treated as a random vector from a multivariate distribution with identical marginal distributions, implying that the abundance frequency counts follow approximately a multinomial distribution. If we can estimate the full richness S of the assemblage with an estimator S_{est}, then an approximate unconditional variance $\sigma^2_{\text{ind}}(m)$ of rarefied richness $\tilde{S}_{\text{ind}}(m)$ is given by

$$\sigma^2_{\text{ind}}(m) = \sum_{k=1}^{n} (1 - \alpha_{km})^2 f_k - \tilde{S}_{\text{ind}}(m)^2 / S_{\text{est}}. \quad (5)$$

This variance is based on an approach similar to that used by Burnham and Overton (1978) for a jackknife estimator of population size in the context of capture–recapture models. Smith and Grassle (1977) provide an unconditional variance formula of $\tilde{S}_{\text{ind}}(m)$, but their expression for the variance is difficult to compute. We postpone specification of S_{est} for a later section.

The Poisson model (Coleman rarefaction)

For the Poisson model (Coleman rarefaction), we need to estimate the expected number of species $S_{\text{area}}(m)$ in a random area of size a within the reference area of size A ($a < A$) (Fig. 1b). If we knew the true Poisson occurrence rate $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ of each of the S species in the assemblage, we could compute

$$S_{\text{area}}(a) = \sum_{i=1}^{S} [1 - \exp(-a\lambda_i)] = S - \sum_{i=1}^{S} [\exp(-a\lambda_i)].$$

Instead, based on species abundances X_i in the reference sample, Coleman (1981) showed that

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

This estimator is the MVUE for $S_{\text{area}}(a)$ (Lehmann and Casella 1998, 108–9). In terms of frequency counts f_k, the estimator becomes

$$\tilde{S}_{\text{area}}(a) = \sum_{k=1}^{n} \left[1 - \left(1 - \frac{a}{A}\right)^k\right] f_k. \quad (6)$$

If we can estimate the full richness S of the assemblage by an estimator S_{est}, then an expression for the unconditional variance $\sigma^2_{\text{area}}(a)$ of rarefied richness $\tilde{S}_{\text{area}}(a)$ is given by

$$\sigma^2_{\text{area}}(a) = \sum_{k=1}^{n}\left[1 - \left(1 - \frac{a}{A}\right)^k\right]^2 f_k - \tilde{S}_{\text{area}}(a)^2/S_{\text{est}}. \quad (7)$$

Coleman et al. (1982) provide an estimator for the variance of $\tilde{S}_{\text{area}}(a)$ conditional on the reference sample. We postpone specification of S_{est} for a later section.

Comparing the multinomial and Poisson models for interpolation

How different are the rarefaction estimates of species richness estimators under the multinomial and the Poisson models? From Equations (4) and (6), the estimates from the two models can be compared by computing

$$\Delta\tilde{S}(a,m) = \tilde{S}_{\text{ind}}(m) - \tilde{S}_{\text{area}}(a) = \left[S_{\text{obs}} - \sum_{k=1}^{n}\alpha_{km}f_k\right] - \left[S_{\text{obs}} - \sum_{k=1}^{n}\left(1 - \frac{a}{A}\right)^k f_k\right].$$

$$\Delta\tilde{S}(a,m) = \sum_{k=1}^{n}\left(1 - \frac{a}{A}\right)^k f_k - \sum_{k=1}^{n}\alpha_{km}f_k. \quad (8)$$

If we assume that individuals are randomly and independently distributed in space, then $a/A \approx m/n$ and

$$\Delta\tilde{S}(m) = \sum_{k=1}^{n}\left[(1 - m/n)^k - \alpha_{km}\right]f_k.$$

Colwell and Coddington (1994) and Brewer and Williamson (1994) showed that, for most datasets, $\Delta\tilde{S}$ is quite small because both $(1 - m/n)^k$ and α_{km} approach zero as subsample size m approaches the reference sample size n, and the frequency count f_k also becomes small at larger k. Thus, $\Delta\tilde{S}$ is very small except for small values of m. In a later section, we compare the two methods using an example from tropical beetles (Janzen 1973a, 1973b). If individuals are not randomly distributed but are aggregated intraspecifically, both methods will overestimate the number of species for a smaller number of individuals m or a smaller area a (Chazdon et al. 1998; Colwell and Coddington 1994; Colwell et al. 2004; Gotelli and Colwell 2001; Kobayashi 1982).

INDIVIDUAL-BASED EXTRAPOLATION

The multinomial model

For the multinomial model, the extrapolation problem is to estimate the expected number of species $S_{\text{ind}}(n+m*)$ in an augmented sample of $n + m^*$ individuals from the assemblage ($m^* > 0$) (Fig. 1a). If we knew the true occurrence probabilities $p_1, p_2, \ldots, p_S$ of each of the S species in the assemblage, given S_{obs}, we could compute

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

Instead, we have only the reference sample to work from, with observed species abundances X_i and their frequency counts f_i. Based on work by Solow and Polasky (1999), Shen et al. (2003) proposed an estimator for $S_{\text{ind}}(n+m^*)$,

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

where $\hat{f}_0$ is an estimator for f_0, number of species present in the assemblage, but not observed in the reference sample.

Any estimator of f_0 is a function of the frequencies $(f_1, f_2, \ldots, f_n)$. Thus, $\tilde{S}_{\text{ind}}(n+m^*)$ can be expressed as a function of $(f_1, f_2, \ldots, f_n)$ and $\partial\tilde{S}/\partial f_i$, the partial derivative of $\tilde{S}_{\text{ind}}(n+m^*)$ with respect to the variable f_i. Based on this expression, a standard asymptotic statistical method gives a variance estimator for $\tilde{S}_{\text{ind}}(n+m^*)$,

$$\hat{\text{var}}(\tilde{S}_{\text{ind}}(n+m^*)) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial\tilde{S}}{\partial f_i}\frac{\partial\tilde{S}}{\partial f_j}\hat{\text{cov}}\left(f_i, f_j\right), \quad (10)$$

where $\hat{\text{cov}}(f_i, f_j) = f_i[1 - f_i/(S_{\text{obs}} + \hat{f}_0)]$ for $i = j$ and $\hat{\text{cov}}(f_i, f_j) = -f_i f_j/(S_{\text{obs}} + \hat{f}_0)$ for $i \neq j$. (For simplicity, we write $\tilde{S}$ for $\tilde{S}_{\text{ind}}(n + m^*)$ in the right-hand side of Equation 10.) See Shen et al. (2003, their Equation 11) for details. We postpone specification of $\hat{f}_0$ for a later section.

Based on the estimator in Equation (9) for $\tilde{S}_{\text{ind}}(n+m^*)$, Chao et al. (2009) showed that we can estimate the number of additional individuals $\tilde{m}^*_g$ required, beyond the reference sample, to detect proportion g of the estimated assemblage richness S_{est} as

$$\tilde{m}^*_g = \frac{nf_1}{2f_2}\log\left[\frac{\hat{f}_0}{(1 - g)S_{\text{est}}}\right], \quad S_{\text{obs}}/S_{\text{est}} < g < 1. \quad (11)$$

The Poisson model

For the Poisson model, the objective is to estimate the expected number of species $\tilde{S}_{\text{area}}(A + a^*)$ in an augmented area $A + a^*$ ($a^* > 0$) (Fig. 1b). If we knew the true Poisson occurrence rates $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ of each of the S species in the assemblage, we could compute, given S_{obs},

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

Working from species abundances X_i in the reference sample, Chao and Shen (2004) proposed an estimator for $\tilde{S}_{\text{area}}(A + a^*)$,

$$\tilde{S}_{\text{area}}(A + a^*) = S_{\text{obs}} + \hat{f}_0\left[1 - \exp\left(-\frac{a^*}{A}\frac{f_1}{\hat{f}_0}\right)\right]. \quad (12)$$

We postpone specification of $\hat{f}_0$, which estimates the species present in the assemblage but not observed in the reference sample, for a later section.

Chao and Shen (2004, their Equation 2.13) also proposed a variance estimator for $\tilde{S}_{\text{area}}(A + a^*)$ (we write $\tilde{S}$ for $\tilde{S}_{\text{area}}(A + a^*)$ in the right-hand side of the following formula),

$$\hat{\text{var}}(\tilde{S}_{\text{area}}(A+a^*)) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial\tilde{S}}{\partial f_i}\frac{\partial\tilde{S}}{\partial f_j}\hat{\text{cov}}\left(f_i,f_j\right), \qquad (13)$$

where $\hat{\text{cov}}(f_i,f_j)$ is defined as for Equation (10) above.

Given this estimator for $S_{\text{area}}(A + a^*)$, it follows from Chao and Shen (2004) that an estimator $\tilde{a}_g^*$ for the additional area required to detect proportion g of the estimated assemblage richness S_{est} is

$$\tilde{a}_g^* = \frac{Af_1}{2f_2}\log\left[\frac{\hat{f}_0}{(1-g)S_{\text{est}}}\right], \quad S_{\text{obs}}/S_{\text{est}} < g < 1. \qquad (14)$$

Estimating the number of species present in the assemblage but not observed in the reference sample for individual-based data

Several estimators in the previous two sections require either an estimate of f_0, the number of species present in the assemblage but not observed in the reference sample or an individual-based estimate for the full richness of the assemblage, S_{est}. Many estimators of the form $S_{\text{est}} = S_{\text{obs}} + \hat{f}_0$ are available (Chao 2005). The simplest (Chao 1984), widely known as Chao1 (Gotelli and Colwell 2011), is $S_{\text{est Chao1}} = S_{\text{obs}} + \hat{f}_{0\text{ Chao1}}$, where

$$\hat{f}_{0\text{ Chao1}} = f_1^2/(2f_2), \quad \text{for } f_2 > 0, \qquad (15a)$$

or

$$\hat{f}_{0\text{ Chao1}} = f_1(f_1-1)/[2(f_2+1)] \quad \text{for } f_2 = 0. \qquad (15b)$$

For the individual-based empirical examples in this paper, we have used the Chao1 estimator, above, which Chao (1984) proved is a minimum estimator of asymptotic species richness.

For assemblages with many rare species, the abundance-based coverage estimator (ACE) (Chao and Lee 1992; Chao et al. 2000; Chazdon et al. 1998) is often a more appropriate estimator of asymptotic richness (Chao and Shen 2004), $S_{\text{est ACE}} = S_{\text{obs}} + \hat{f}_{0\text{ ACE}}$. ACE takes into account the frequency counts for rare species $f_1, f_2, \ldots, f_k, \ldots, f_R$, where R is a cutoff frequency between rare and common species in the reference sample. Thus, $S_{\text{rare}} = \sum_{i=1}^{S} I(0 < X_i \leq R)$ with summed abundance $X_{\text{rare}} = \sum_{i=1}^{S} X_i I(X_i \leq R)$. These counts and an estimate of sample coverage, $\hat{C}_{\text{ACE}} = 1 - f_1/X_{\text{rare}}$, are used to compute a squared coefficient of variation, $\hat{\gamma}^2_{\text{ACE}}$, and the estimator $\hat{f}_{0\text{ ACE}}$,

$$\hat{\gamma}^2_{\text{ACE}} = \max\left\{\frac{S_{\text{rare}}}{\hat{C}_{\text{ACE}}}\frac{\sum_{k=1}^{R}k(k-1)f_k}{\left(\sum_{k=1}^{R}kf_k\right)\left(\sum_{k=1}^{R}kf_k - 1\right)} - 1, 0\right\}.$$

$$\hat{f}_{0\text{ ACE}} = \frac{S_{\text{rare}}}{\hat{C}_{\text{ACE}}} + \frac{f_1}{\hat{C}_{\text{ACE}}}\hat{\gamma}^2_{\text{ACE}} - S_{\text{rare}}. \qquad (16)$$

The expression for $\hat{\gamma}^2_{\text{ACE}}$, above, is for the multinomial model. For the Poisson model, the summation in the denominator should be replaced by $(\sum_{k=1}^{R} kf_k)^2$. Chao and Shen (2004) recommended $R = 10$ as rule of thumb, with exploration of other values suggested for samples with large coefficients of variation.

SAMPLE-BASED INTERPOLATION (RAREFACTION)

The Bernoulli product model

For the Bernoulli product model (sample-based rarefaction), we need to estimate the expected number of species $S_{\text{sample}}(t)$ in a random set of t sampling units from among the T sampling units defining the incidence reference sample ($t < T$) (Fig. 1c). If we knew the true detection probabilities $\theta_1, \theta_2 \ldots, \theta_S$ of each of the S species in the assemblage, we could compute

$$S_{\text{sample}}(t) = \sum_{i=1}^{S}\left[1 - (1-\theta_i)^t\right] = S - \sum_{i=1}^{S}(1-\theta_i)^t.$$

Instead, we have only the incidence reference sample to work from, with observed species incidence frequencies Y_i. The MVUE for $S_{\text{sample}}(t)$ is

$$\tilde{S}_{\text{sample}}(t) = S_{\text{obs}} - \sum_{Y_i>0}\left[\binom{T-Y_i}{t}\Big/\binom{T}{t}\right]. \qquad (17)$$

This analytic formula was first derived by Shinozaki (1963) and rediscovered multiple times (Chiarucci et al. 2008). Colwell et al. (2004, their Equation 5) provide a mathematically equivalent equation in terms of the incidence frequency counts Q_k similar to our Equation (4). This estimator has long been called 'Mao Tau' in the widely used software application 'EstimateS' (Colwell 2011). Colwell et al. (2004, their Equation 6) developed an estimator for the unconditional variance in terms of the frequency counts Q_k, similar to our Equation (5), that requires an incidence-based estimator S_{est} for assembly richness S. We postpone specification of S_{est} for a later section.

SAMPLE-BASED EXTRAPOLATION

The Bernoulli product model

For the Bernoulli product model, the extrapolation problem is to estimate the expected number of species $S_{\text{sample}}(T + t^*)$ in an augmented set of $T + t^*$ sampling units ($t^* > 0$) from the assemblage (Fig. 1c). If we knew the true detection probabilities $\theta_1, \theta_2 \ldots, \theta_S$ of each of the S species in the assemblage, given S_{obs}, we could compute

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

Based on a derivation by Chao et al. (2009), we have the estimator

$$\begin{aligned}\tilde{S}_{\text{sample}}\left(T+t^*\right) &= S_{\text{obs}} + \hat{Q}_0\left[1 - \left(1 - \frac{Q_1}{Q_1 + T\hat{Q}_0}\right)^{t^*}\right] \\ &\approx S_{\text{obs}} + \hat{Q}_0\left[1 - \exp\left(\frac{-t^*Q_1}{Q_1 + T\hat{Q}_0}\right)\right].\end{aligned} \qquad (18)$$

Expressing $\tilde{S}_{\text{sample}}(T + t^*)$ as a function of $(Q_1, Q_2, \ldots, Q_T)$, and using an asymptotic method, we obtain an approximate variance formula

$$\hat{\text{var}}(\tilde{S}_{\text{sample}}(T+t^*)) = \sum_{i=1}^{T}\sum_{j=1}^{T}\frac{\partial\tilde{S}}{\partial Q_i}\frac{\partial\tilde{S}}{\partial Q_j}\,\hat{\text{cov}}\left(Q_i, Q_j\right), \qquad (19)$$

where $\hat{\text{cov}}(Q_i, Q_j) = Q_i[1 - Q_i/(S_{\text{obs}} + \hat{Q}_0)]$ for $i = j$ and $\hat{\text{cov}}(Q_i, Q_j) = -Q_iQ_j/(S_{\text{obs}} + \hat{Q}_0)$ for $i \neq j$. (For simplicity, we write $\tilde{S}$ for $\tilde{S}_{\text{sample}}(T + t^*)$ in the above variance formula.)

Equations (18) and (19), above, both require an estimate of Q_0, the number of species present in the assemblage but not detected in any sampling units. We postpone specification of an estimator for Q_0 for the next section.

Based on the estimator in Equation (18) for $\tilde{S}_{\text{sample}}(T + t^*)$, the number of additional sampling units $\tilde{t}_g^*$ required to detect proportion g of the estimated assemblage richness S_{est} is

$$\tilde{t}_g* \approx \frac{\log\left[1 - \dfrac{T}{(T-1)}\dfrac{2Q_2}{Q_1^2}(gS_{\text{est}} - S_{\text{obs}})\right]}{\log\left[1 - \dfrac{2Q_2}{(T-1)Q_1 + 2Q_2}\right]}, \quad S_{\text{obs}}/S_{\text{est}} < g < 1. \qquad (20)$$

Estimating the number of species present in the assemblage but not observed in the reference sample for sample-based incidence data

The extrapolation estimators for the Bernoulli product model require either an estimate of Q_0, the number of species present in the assemblage but not observed in the sampling units comprising the incidence reference sample, or a sample-based estimate for the full richness of the assemblage, S_{est}. Many estimators of the form $S_{\text{est}} = S_{\text{obs}} + \hat{Q}_0$ are available (Chao 2005). The simplest (Chao 1987), widely known as Chao2 (Gotelli and Colwell 2011), is $\hat{S}_{\text{est Chao2}} = S_{\text{obs}} + \hat{Q}_{0\text{ Chao2}}$, where

$$\hat{Q}_{0\text{ Chao2}} = \left[(T-1)/T\right]\left[Q_1^2/(2Q_2)\right] \quad \text{for } Q_2 > 0 \qquad (21)$$

or

$$\hat{Q}_{0\text{ Chao2}} = [(T-1)/T][Q_1(Q_1 - 1)/[2(Q_2 + 1)] \text{ for } Q_2 = 0. \qquad (22)$$

For the sample-based incidence example in this paper, we have used the Chao2 estimator, above, which Chao (1987) showed is a minimum estimator of asymptotic species richness.

For assemblages with many rare species, the incidence-based coverage estimator (ICE; Chazdon et al. 1998; Lee and Chao 1994) is often a more appropriate estimator of asymptotic species richness (Chao and Shen 2004),

$$\hat{S}_{\text{est ICE}} = S_{\text{obs}} + \hat{Q}_{0\text{ ICE}}.$$

ICE takes into account the frequency counts for rare species $(Q_1, Q_2, \ldots, Q_k, \ldots, Q_R)$, where R is a cutoff frequency between infrequent and frequent species in the reference sample. Thus, the number of species that occur in fewer than R sampling units is $S_{\text{infreq}} = \sum_{i=1}^{S} I(0 < Y_i \leqslant R)$ with summed incidence frequencies $Y_{\text{infreq}} = \sum_{i=1}^{S} Y_i I(Y_i \leqslant R)$. These counts, the number of sampling units that include at least one infrequent species (T_{infreq}), and an estimate of sample coverage, $\hat{C}_{\text{ICE}} = 1 - Q_1/Y_{\text{infreq}}$, are used to compute a squared coefficient of variation, γ^2_{ICE}, and the estimator $Q_{0\text{ ICE}}$:

$$\hat{\gamma}^2_{\text{ICE}} = \max\left\{\frac{S_{\text{infreq}}}{\hat{C}_{\text{ICE}}}\frac{T_{\text{infreq}}}{\left(T_{\text{infreq}} - 1\right)}\frac{\sum_{k=1}^{R}k(k-1)Q_k}{\left(\sum_{k=1}^{R}kQ_k\right)^2} - 1,\ 0\right\}$$

$$\hat{Q}_{0\text{ ICE}} = \frac{S_{\text{infreq}}}{\hat{C}_{\text{ICE}}} + \frac{Q_1}{\hat{C}_{\text{ICE}}}\hat{\gamma}^2_{\text{ICE}} - S_{\text{infreq}}. \qquad (23)$$

We recommend $R = 10$ as rule of thumb, with exploration of other values suggested for samples with large coefficients of variation.

EXAMPLES

Tropical beetles: individual-based rarefaction and extrapolation (multinomial model)

Janzen (1973a, 1973b) tabulated many data sets on tropical foliage insects from sweep samples in southwestern Costa Rica. We selected two beetle data sets ('Osa primary' and 'Osa secondary') to compare beetle species richness between old-growth forest and second-growth vegetation on the Osa Peninsula. The species frequency counts appear in Table 1.

Janzen's study recorded 976 individuals representing 140 species in the Osa second-growth site and 237 individuals of 112 species in the Osa old-growth site. From the unstandardized raw data (the reference samples), one might conclude that the second-growth site has more beetle species than the old-growth site (140 vs. 112; Fig. 2c, solid points). However, the sample sizes (number of individual beetles) for the two samples are quite different (976 vs. 237 individuals, Fig. 2a and b). When the sample size in the second-growth site is rarefied down to 237 individuals to match the size of the old-growth

Table 1: beetle species abundance frequency counts from two sites on the Osa Peninsula in southwestern Costa Rica (Janzen 1973a, 1973b)

(a) Osa second growth: $S_{\text{obs}} = 140$, $n = 976$

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

(b) Osa old growth: $S_{\text{obs}} = 112$, $n = 237$

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

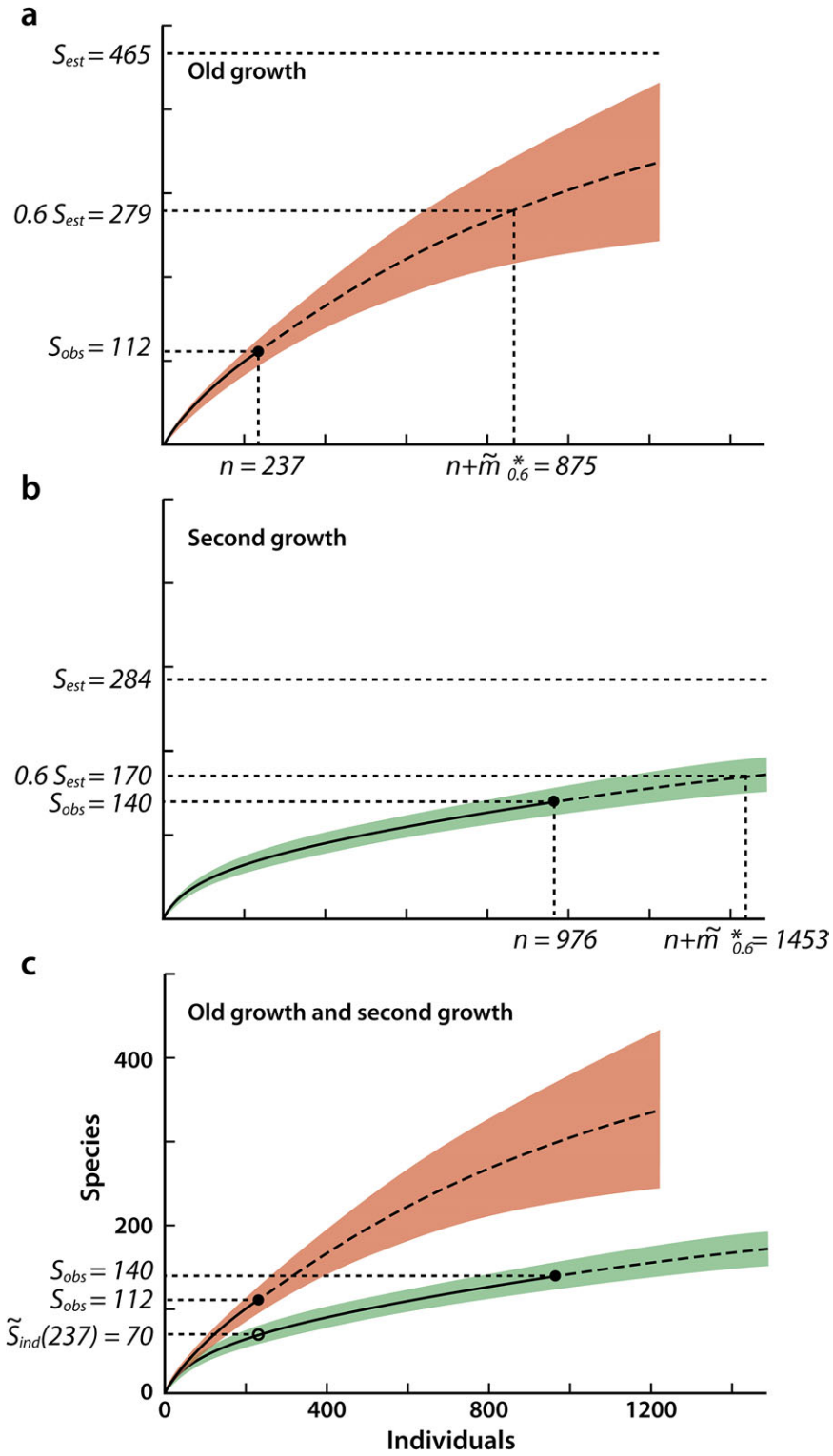

Figure 2: individual-based interpolation (rarefaction) and extrapolation from two reference samples (filled black circles) of beetles from southwestern Costa Rica (Janzen 1973a, 1973b), illustrating the computation of estimators from Fig. 1a for the multinomial model, with 95% unconditional confidence intervals. (**a**) Osa old-growth forest sample. (**b**) Osa second-growth forest sample (**c**) Comparison of the curves from the samples in (**a**) and (**b**). Based on observed richness, S_{obs}, the Osa second-growth assemblage (with 140 species in the reference sample) is richer in species than the Osa second-growth assemblage (with 112 species in the reference sample), but after rarefying the second-growth sample to 237 individuals to match the size of the old-growth sample (open black circle), the second-growth sample has only 70 species. Clearly the old-growth assemblage is richer, based on these samples.

sample (Fig. 2c, open point), using the multinomial model (Equation 4), the ordering of the two sites is reversed. The interpolated species richness for 237 individuals in the second-growth site is only 70, considerably less than primary site, with 112 species. Moreover, the 95% confidence intervals do not overlap (Fig. 2c).

Individual-based rarefaction of abundance data, like the interpolation analysis above, has been carried out in this way for decades. Here, we apply individual-based rarefaction and extrapolation to the same reference sample for the first time. Applying the multinomial model (Equation 9) to the Janzen dataset to increase the sample size (number of individuals) in each site yields the extrapolated curves (broken line curves) for each site is shown in Fig. 2. Even though the mathematical derivations for interpolation and extrapolation are fundamentally different, the interpolation and extrapolation curves join smoothly at the single data point of the reference sample.

In Table 2a, using the multinomial model (classical rarefaction), we show for the Osa old-growth data (S_{obs} = 112, n = 237 in the reference sample): (i) values for the interpolated estimate $\tilde{S}_{ind}(m)$, for values of m from 1 up to the reference sample size of 237 individuals (Equation 4), along with the unconditional standard error (SE, Equation 5) values that are used to construct the 95% confidence intervals shown in Fig. 2a and c; (ii) the extrapolated estimate $\tilde{S}_{ind}(n+m^*)$ (Equation 9), where m^* ranges from 0 to 1 000 individuals, along with the unconditional SE (Equation 10); and (iii) the number of additional individuals $\tilde{m}^*_g$ required to detect proportion g of the estimated assemblage richness (Equation 11), for g = 0.3 to 0.9, in increments of 0.1. In Fig. 2a, we plot the multinomial rarefaction curve and extrapolation curve up to a sample size of 1 200 individuals and show the predicted number of individuals need to reach for g = 0.6. The corresponding values and curves for the Osa second-growth data (S_{obs} = 140, n = 976 in the reference sample) are shown in Table 2b and Fig. 2b.

Table 2: individual-based interpolation, extrapolation and prediction of additional individuals required to reach gS_{est}, under the multinomial model, for beetle samples from two sites on the Osa Peninsula in southwestern Costa Rica (Janzen 1973a, 1973b)

Rarefaction			Extrapolation			Individuals prediction	
m	$\tilde{S}_{ind}(m)$	SE	m^*	$\tilde{S}_{ind}(n+m^*)$	SE	g	$\tilde{m}^*_g$
(a) Osa old-growth site, S_{obs} = 112, n = 237. The extrapolation is extended to more than five times of the reference sample size, in order to compare with the Osa second-growth curve (b); see Table 2(b).							
1	1.00	0.00	0	112	9.22	0.3	80.60
20	15.89	1.95	100	145.74	12.20	0.4	234.04
40	28.44	3.00	200	176.25	15.38	0.5	415.52
60	39.44	3.85	400	228.80	22.58	0.6	637.64
80	49.40	4.57	600	271.77	30.84	0.7	924.00
100	58.62	5.22	800	306.93	39.79	0.8	1327.60
120	67.29	5.83	1 000	335.68	48.96	0.9	2017.56
140	75.54	6.42					
160	83.48	7.00					
180	91.16	7.57					
200	98.63	8.15					
220	105.92	8.73					
237	112.00	9.22					
(b) Osa second-growth site, S_{obs} = 140, n = 976. The extrapolation is extended to double the reference sample size.							
1	1.00	0.00	0	140.00	8.43	0.5	28.91
100	44.30	4.36	100	147.00	8.87	0.6	477.30
200	64.43	5.31	200	153.66	9.34	0.7	1055.37
300	78.83	5.85	400	166.02	10.34	0.8	1870.12
400	90.58	6.25	600	177.21	11.46	0.9	3262.94
500	100.83	6.60	800	187.34	12.68		
600	110.11	6.95	1 000	196.51	13.99		
700	118.72	7.32					
800	126.80	7.70					
900	134.45	8.11					
976	140.00	8.43					

For both samples, the unconditional variance, and thus the 95% confidence interval, increased with sample size. For extrapolation, the SE values are relatively small up to a doubling of the reference sample, signifying quite accurate extrapolation in this range. For the Osa old-growth site (Table 2a; Fig. 2a), the extrapolation is extended to five times of the original sample size in order to compare with the Osa second-growth curve. This long-range extrapolation (>3× the original sample size) inevitably yields very wide confidence intervals. For the Osa second-growth site (Table 2b; Fig. 2b), the extrapolation is extended only to double the reference sample size (not fully shown in Fig. 2b) yielding a quite accurate extrapolated estimate with a narrow confidence interval.

Based on Fig. 2, even though the Osa old-growth site extrapolation for large sample sizes exhibits high variance, the old-growth and second-growth confidence intervals do not overlap for any sample size considered. This implies that beetle species richness for any sample size is significantly greater in the old-growth site than that in the second-growth site for sample size up to at least 1 200 individuals.

Tropical beetles: individual-based rarefaction and extrapolation (Poisson model)

In addition to applying estimators based on the multinomial model, we also analysed the Janzen beetle dataset with estimators based on the Poisson model, including Coleman area-based rarefaction (Equations 6 and 7), area-based extrapolation (Equations 12 and 13), and estimation of the additional area required to detect proportion g of the estimated assemblage richness S_{est} (Equation 14). The results for the Osa old-growth beetle sample appear in Table 3a and the results for the Osa second-growth beetle sample in Table 3b. Comparison of the results for the Poisson model estimators (Table 3) with the corresponding results for the multinomial model estimators (Table 2) reveals a remarkable similarity that makes sense mathematically because the distribution for the Poisson model (Equation 2), conditional on the total number of individuals, is just the

Table 3: individual-based interpolation, extrapolation and prediction of additional area required to reach gS_{est}, under the Poisson model, for beetle samples from two sites on the Osa Peninsula in southwestern Costa Rica (Janzen 1973a, 1973b)

Rarefaction			Extrapolation			Area prediction	
a	$\tilde{S}_{area}(a)$	SE	a^*	$\tilde{S}_{area}(A+a^*)$	SE	g	$\tilde{a}^*_g$
(a) Osa old-growth site, S_{obs} = 112, A = 237. The extrapolation is extended to more than five times of the reference sample size, in order to compare with the Osa second-growth curve (b).							
1	0.98	0.00	0	112.00	9.22	0.3	80.60
20	15.83	1.93	100	145.72	12.20	0.4	234.04
40	28.37	2.99	200	176.22	15.38	0.5	415.52
60	39.38	3.84	400	228.75	22.58	0.6	637.64
80	49.35	4.56	600	271.72	30.84	0.7	924.00
100	58.58	5.21	800	306.86	39.78	0.8	1327.60
120	67.26	5.82	1 000	335.61	48.96	0.9	2017.56
140	75.52	6.41					
160	83.46	6.99					
180	91.15	7.57					
200	98.62	8.15					
220	105.92	8.73					
237	112.00	9.22					
(b) Osa second-growth site, S_{obs} = 140, A = 976. The extrapolation is extended to double the reference sample size.							
1	0.98	0.17	0	140.00	8.43	0.5	28.91
100	44.23	4.35	100	147.00	8.87	0.6	477.30
200	64.38	5.31	200	153.65	9.34	0.7	1055.37
300	78.81	5.85	400	166.01	10.34	0.8	1870.12
400	90.57	6.24	600	177.20	11.46	0.9	3262.94
500	100.82	6.60	800	187.33	12.68		
600	110.10	6.95	1 000	196.50	13.99		
700	118.71	7.32					
800	126.79	7.70					
900	134.44	8.11					
976	140.00	8.43					

multinomial model (Equation 1). Moreover, the similarity applies not only to rarefaction (as previously noted by Brewer and Williamson 1994) but also to extrapolation. Figure 3 shows just how close the results based on the two models are for this example. For interpolation and extrapolation, the difference is always less than one-tenth of one individual (assuming for the Poisson model that individuals are randomly and independently distributed in space, so that $a/A \approx m/n$). This means that rounding to the nearest individual consistently yields precisely the same values under both models. For this reason, we do not plot the results from the Poisson model because the figure would be identical to Fig. 2, with the Poisson variables in Fig. 1b substituted for the multinomial variables in Fig. 1a. In addition, estimates of the additional area required to detect proportion g of the estimated assemblage richness under the Poisson model (Table 3, from Equation 14) are identical to the estimates of the additional number of individuals required to reach proportion g of the estimated assemblage richness under the multinomial model (Table 2, from Equation 11).

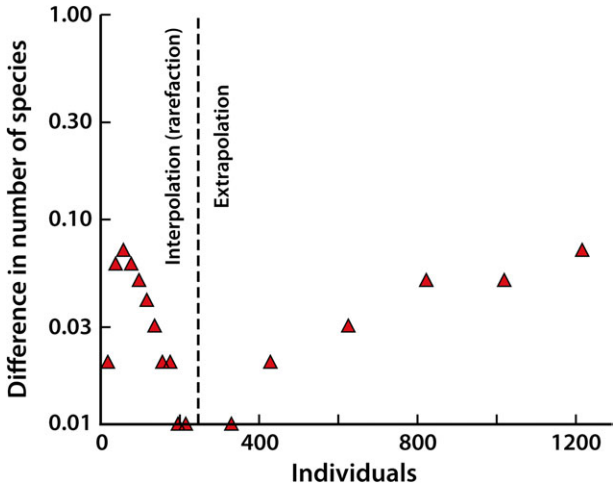

Figure 3: richness estimated by the multinomial model versus the Poisson model for the Osa old-growth beetle sample (Janzen 1973a, 1973b). The numbers on the ordinate show the magnitude of the multinomial estimate minus the Poisson estimate, in ordinary arithmetic units, scaled logarithmically only to spread out the values vertically so they can be seen. Although the multinomial estimate is consistently higher, the difference never exceeds one tenth of one species, so the results rounded to the nearest species are identical.

Tropical trees: individual-based rarefaction and extrapolation (multinomial model)

Norden et al. (2009) compared species composition of trees, saplings and seedlings in six 1-ha forest plots spanning three successional stages in lowland forests of northeastern Costa Rica. We selected data for tree stems ⩾5 cm diameter at breast height in three samples from this dataset, all located within La Selva Biological Station. One of the samples represents an old-growth plot (Lindero El Peje [LEP] old growth, S_{obs} = 152, n = 943) and two were from second-growth forest plots, one of them 29 years old (LEP second growth, S_{obs} = 104, n = 1 263) and the other 21 years old in 2006 (Lindero Sur, S_{obs} = 76, n = 1 020), following pasture abandonment. The species frequency counts for the three plots appear in Table 4.

The results for interpolation and extrapolation from these three reference samples, under the multinomial model, appear in Table 5 and Fig. 4a. For each of the three samples, Table 5 shows: (i) species richness values for the interpolated estimate $\tilde{S}_{ind}(m)$, under the multinomial model (classical rarefaction, Equation 4), for values of m from 1 up to the reference sample size for each sample (n = 943, 1 263 or 1 020 individuals), along with the unconditional SE (Equation 5) values that are used to construct the 95% confidence intervals shown in Fig. 4a, and (ii) the extrapolated estimate $\tilde{S}_{ind}(n+m^*)$, where m^* ranges from 0 to 1 500, 1 200 or 1 400 individuals (for the three samples), so that all samples are extrapolated to roughly 2 400 individuals, along with the unconditional SE (Equation 10).

In Fig. 4a, we plot the multinomial rarefaction curves and extrapolation curves up to a sample size of 2 400 individuals. Clearly the number of species at any plotted sample size (beyond very small samples) is significantly greater for LEP old growth than in either of the two samples from second-growth forest. The number of species in the plot of intermediate age, LEP second growth, significantly exceeds the number of species in the youngest plot, Lindero Sur, for sample sizes

Table 4: Species abundance frequency counts for tree samples from three forest sites in northeastern Costa Rica (Norden et al. 2009)

(a) LEP old growth, S_{obs} = 152, n = 943

i	1	2	3	4	5	6	7	8	9	10	11	13	15	16	18	19	20	25	38	39	40	46	52	55
f_i	46	30	16	12	6	5	3	4	5	4	1	3	1	1	1	1	4	3	1	1	1	1	1	1

(b) LEP older (29 years) second growth, S_{obs} = 104, n = 1 263

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	20	22	39	45	57	72	88	132	133	178
f_i	33	15	13	4	5	3	3	1	2	1	4	2	2	1	2	1	1	1	1	1	1	1	1	2	1	1	1

(c) Lindero Sur younger (21 years) second growth, S_{obs} = 76, n = 1 020

i	1	2	3	4	5	7	8	10	11	12	13	15	31	33	34	35	66	72	78	127	131	174
f_i	29	13	5	2	3	4	1	2	2	1	2	2	1	1	1	1	1	1	1	1	1	1

Table 5: Individual-based interpolation and extrapolation, under the multinomial model, for tree samples from three forest sites in northeastern Costa Rica (Norden et al. 2009)

Rarefaction			Extrapolation		
m	$\tilde{S}_{\text{ind}}(m)$	SE	m^*	$\tilde{S}_{\text{ind}}(n+m^*)$	SE
(a) LEP old growth, S_{obs} = 152, n = 943					
1	1.00	0.00	0	152.00	5.35
20	16.72	1.82	100	156.56	5.55
40	28.88	2.72	200	160.53	5.79
60	38.51	3.24	300	163.98	6.08
80	46.57	3.59	400	166.99	6.42
100	53.54	3.84	500	169.61	6.79
200	79.28	4.46	600	171.90	7.18
300	96.99	4.69	700	173.88	7.59
400	110.49	4.81	800	175.61	8.00
500	121.32	4.88	900	177.12	8.40
600	130.28	4.96	1 000	178.43	8.80
700	137.83	5.04	1 100	179.57	9.18
800	144.28	5.15	1 200	180.57	9.55
900	149.84	5.28	1 300	181.43	9.89
943	152.00	5.35	1 400	182.19	10.22
			1 500	182.84	10.52
(b) LEP older (29 years) second growth, S_{obs} = 104, n = 1 263					
1	1.00	0.00	0	104.00	5.19
20	12.96	2.17	100	106.52	5.33
40	20.14	2.72	200	108.87	5.49
60	25.62	3.02	300	111.05	5.66
80	30.25	3.24	400	113.08	5.86
100	34.30	3.43	500	114.97	6.08
200	49.47	3.99	600	116.73	6.31
300	59.92	4.25	700	118.37	6.56
400	67.94	4.40	800	119.89	6.83
500	74.50	4.50	900	121.31	7.11
600	80.06	4.58	1 000	122.63	7.39
700	84.88	4.65	1 100	123.86	7.69
800	89.14	4.73	1 200	125.00	7.99
900	92.93	4.81			
1 000	96.35	4.90			
1 100	99.46	5.00			
1 200	102.32	5.11			
1 263	104.00	5.19			
(c) Lindero Sur younger (21 years) second growth, S_{obs} = 76, n = 1 020					
1	1.00	0.00	0	76.00	4.76
20	11.51	2.19	100	78.72	4.95
40	17.08	2.68	200	81.22	5.16
60	21.16	2.94	300	83.50	5.40
80	24.52	3.12	400	85.59	5.66
100	27.41	3.26	500	87.51	5.95
200	38.15	3.65	600	89.26	6.25
300	45.72	3.83	700	90.87	6.58
400	51.77	3.95	800	92.34	6.91
500	56.90	4.05	900	93.69	7.26
600	61.41	4.16	1 000	94.92	7.61
700	65.44	4.28	1 100	96.05	7.97
800	69.09	4.42	1 200	97.09	8.33
900	72.40	4.56	1 300	98.03	8.68
1 000	75.43	4.73	1 400	98.90	9.03
1 020	76.00	4.76			

between 500 and 1 600 individuals, based conservatively on non-overlapping confidence intervals. Due to the prevalence of rare species in old-growth tropical forests and widespread dispersal limitation of large-seeded animal-dispersed species, tree species richness is slow to recover during secondary succession and may require many decades to reach old-growth levels, even under conditions favorable to regeneration.

Tropical ants: sample-based rarefaction and extrapolation for incidence data (Bernoulli product model)

Longino and Colwell (2011) sampled ants at several elevations on the Barva Transect, a 30-km continuous gradient of wet forest on Costa Rica's Atlantic slope. For this example, we use results from five sites, at 50-, 500-, 1 070-, 1 500- and 2 000-m elevation, to illustrate sample-based rarefaction and extrapolation. The sampling unit consisted of all worker ants extracted from a 1-m^2 forest floor plot, applying a method called 'mini-Winkler extraction'. Because ants are colonial and the colony is the unit of reproduction, scoring each sampling unit for presence or absence of each species makes more sense than using abundance data (Gotelli et al. 2011). A sample-by-species incidence matrix was therefore produced for each of the five sites. The incidence frequency counts for the five sites appear in Table 6.

The results for sample-based interpolation and extrapolation from these five sites (at five elevations), under the Bernoulli product model, appear in Table 7 and Fig. 4b. For each of the five samples, Table 7 shows: (i) values for the interpolated estimate $\tilde{S}_{\text{sample}}(t)$, under the Bernoulli product model (Equation 17), for values of t from 1 up to the reference sample size T for each elevation (T = 599, 230, 150, 200, 200 sampling units), along with the unconditional SE values (Colwell et al. 2004, their Equation 6) that are used to construct the 95% confidence intervals shown in Fig. 4b; and (ii) the

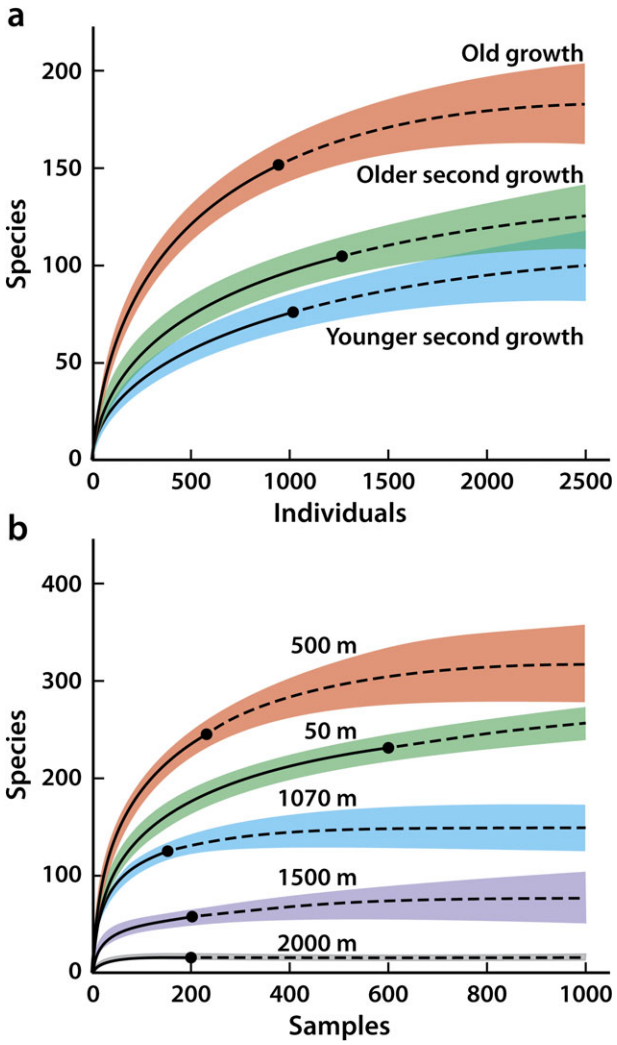

Figure 4: **(a)** individual-based interpolation (rarefaction) and extrapolation from three reference samples (filled black circles) from 1-ha tree plots in northeastern Costa Rica (Norden et al. 2009) under the multinomial model, with 95% unconditional confidence intervals. Species richness in the old-growth plot (LEP old growth, shown in red) consistently exceeds the richness in second-growth plot, LEP second growth (29 years old, shown in green) and Lindero Sur second growth (21 years old, shown in blue). Richness in LEP (green) significantly exceeds richness in Lindero Sur (blue) for sample sizes between 500 and 1 600 individuals, based conservatively on non-overlapping confidence intervals. **(b)** Sample-based interpolation (rarefaction) and extrapolation for reference samples (filled black circles) for ground-dwelling ants from five elevations on the Barva Transect in northeastern Costa Rica (Longino and Colwell 2011) under the Bernoulli product model, with 95% unconditional confidence intervals. Because each sampling unit is a 1-m^2 plot, what Fig. 4b plots on the 'species' axis are actually estimates of species density, the number of species in multiples of a 1-m^2 area. (See the Discussion for information on approximating species richness from species density.) Maximum species density is found at the 500-m elevation site, consistently exceeding the species density at both higher and lower elevations. Species density drops significantly with each increase in elevation above 500 m, based conservatively on non-overlapping confidence intervals.

extrapolated estimate $\tilde{S}_{\text{sample}}(T+t^*)$, where t^* ranges from 401 to 800 sampling units, to extrapolate all elevations to 1 000 sampling units (Equation 18), along with the unconditional SE (Equation 19).

DISCUSSION

In this paper, we developed a unified theoretical and notational framework for modeling and analyzing the effects on observed species richness of the number of individuals sampled or the number of sampling units examined in the context of a single, quantitative, multispecies sample (an abundance reference sample) or a single set of incidence frequencies for species among sampling units (an incidence reference sample). We compared three statistically distinct models, one based on the multinomial distribution, for counts of individuals (Fig. 1a), the second based on the Poisson distribution, for proportional areas (Fig. 1b), and the third based on a Bernoulli product distribution, for incidence frequencies among sampling units (Fig. 1c).

For interpolation to samples smaller than the reference sample, these correspond to classical rarefaction (Hurlbert 1971), Coleman rarefaction (Coleman 1981) and sample-based rarefaction (Colwell et al. 2004). For the first time, we have linked these well-known interpolation approaches with recent sampling-theoretic extrapolation approaches, under both the multinomial model (Shen et al. 2003) and the Poisson model (Chao and Shen 2004), as well as to methods for predicting the number of additional individuals (multinomial model, Chao et al. 2009) or the amount of additional area (Poisson model, Chao and Shen 2004) needed to reach a specified proportion of estimated asymptotic richness. For the Bernoulli product model, we have developed new estimators, using a similar approach, for sample-based extrapolation (Fig. 1c). The fundamental statistics for all these estimators are the abundance frequency counts f_k—the number of species each represented by exactly $X_i = k$ individuals in a reference sample (e.g. Tables 1 and 4)—for individual-based models, or the incidence frequency counts Q_k—the number of species that occurred in exactly $Y_i = k$ sampling units (e.g. Table 6)—for sample-based models.

This novel integration of mathematically distinct approaches allowed us to link interpolated (rarefaction) curves and extrapolated curves to plot a unified species accumulation curve for empirical examples (Figs 2 and 4). Perhaps the most surprising (and satisfying) result is how smoothly the interpolated and extrapolated moieties of the curve come together at the reference sample, in all examples we have investigated. The remarkable degree of concordance between multinomial and Poisson estimators (e.g. Fig. 3), not only for interpolation (as anticipated by Brewer and Williamson [1994] and Colwell and Coddington [1994]) but also for extrapolation (as first shown here), was a second surprise, although the two models are closely related, as discussed earlier. We see little reason, for individual-based data, to recommend computing estimators based on one model over the other (although Coleman curves are computationally

Table 6: species incidence frequency counts for ant samples from five elevations in northeastern Costa Rica (Longino and Colwell 2011)

(a) Elevation 50 m, S_{obs} = 227, T = 599

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Q_i	49	23	18	14	9	10	4	8	6	2	1	2	2	5	2	4	3	2	2	3	1	1	2

i	25	27	29	30	31	33	39	40	43	46	47	48	51	52	56	58	61	65	69	72	77	79	82
Q_i	1	2	1	2	1	1	1	1	1	2	1	1	1	2	2	2	2	1	1	1	1	1	1

i	83	84	86	91	95	97	98	106	113	124	126	127	128	129	182	183	186	195	222	236	263	330	
Q_i	1	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	

(b) Elevation 500 m, S_{obs} = 241, T = 230

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	23	24	25	26	27	30
Q_i	71	34	12	14	9	11	8	4	7	5	2	3	4	2	1	2	4	1	1	1	2	1	1	3	1	1	1

i	31	32	33	34	36	37	38	39	41	42	43	44	45	46	47	49	52	53	54	56	60	65	73	78	123	131	133
Q_i	2	2	1	1	1	1	4	2	1	1	1	1	1	1	1	1	2	1	1	1	2	1	1	1	1	1	1

(c) Elevation 1 070 m, S_{obs} = 122, T = 150

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	21	22	23
Q_i	28	16	13	3	1	3	6	1	1	1	4	3	4	1	1	4	1	2	1	1	1

i	24	25	26	30	31	32	34	36	38	39	43	45	46	54	60	68	74	80	96	99	
Q_i	1	3	1	1	3	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	

(d) Elevation 1 500 m, S_{obs} = 56, T = 200

i	1	2	3	4	5	6	9	11	17	18	19	23	24	25	29	30	32	33	43	50	53	73	74	76	79	113	144
Q_i	13	4	2	2	4	2	4	2	2	1	1	2	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1

(e) Elevation = 2 000 m, S_{obs} = 14, T = 200

i	1	2	3	4	8	13	15	19	23	34	59	80
Q_i	1	2	1	1	2	1	1	1	1	1	1	1

less demanding than classical rarefaction), and no reason whatsoever to compute both.

The ability to link rarefaction curves with their corresponding extrapolated richness curves, complete with unconditional confidence intervals, helps to solve one of most frustrating limitations of traditional rarefaction: 'throwing away' much of the information content of larger samples, in order to standardize comparisons with the smallest sample in a group of samples being compared. The ant dataset (Fig. 4b) (Longino and Colwell 2011), which spans an elevation gradient from lowland rainforest at 50-m elevation to montane cloud forest at 2 000 m, is an excellent example. Typical of tropical mountains, ants are scarce and represent few species above ~1 500 m on this transect. Datasets range from 200 sampling units (with only 270 incidences) at the 2 000 m site, up to 599 sampling units (with 5 346 incidences) at the 50-m site. (Each incidence is the occurrence of one species in one sampling unit.)

Because each sampling unit is a 1-m^2 plot, in the ant study, what Fig. 4b plots on the 'species' axis are actually estimates of ant species density (species per area) at multiples of 1-m^2 spatial scale. To convert the plot to approximations of species richness for the local assembly at each elevation, the curves could be rescaled from 'samples' to 'incidences' for each elevation separately and replotted together on a new graph with 'incidences' as the X-axis (Longino and Colwell 2011). Rescaling to incidences can also be useful for any organisms that, like ants, live colonially or that cannot be counted individually (e.g. multiple stems of stem-sprouting plants or cover-based vegetation data).

The same approach to approximating species richness is recommended, but with re-scaling to individuals instead of incidences, for rarefaction of sample-based abundance datasets. For these datasets, abundances can first be converted to incidences (presence or absence) before applying incidence-based rarefaction. Then, differences in density (the number of individuals per sampling unit) among datasets can be accounted for by rescaling the X-axis of sample-based rarefaction and extrapolation curves to individuals (Chazdon et al. 1998; Gotelli and Colwell 2001, 2011; Norden et al. 2009). With rescaling to individuals, however, strong among-sample differences in dominance can produce misleading results.

Analytical methods (classical rarefaction and Coleman rarefaction) have existed for decades for estimating the number of species in a subset of samples from an individual-based dataset. Confidence intervals for those estimates have always been based on conditional variances because unconditional variances for individual-based classical rarefaction and Coleman curves have until now remained elusive. Suppose we wish to compare two reference samples differing in number of individuals, with sample Y larger than sample X. The two samples may be drawn from either the same assemblage or from two different assemblages. The conditional variance of the larger sample Y is appropriate for answering the question: 'Is the number of species recorded in the smaller sample, X,

Table 7: sample-based interpolation, extrapolation and prediction of number of additional sampling units required to reach gS_{est}, under the multinomial product model, for ant samples from five elevations in northeastern Costa Rica (Longino and Colwell 2011)

Rarefaction			Extrapolation			Sampling units prediction	
t	$\tilde{S}_{sample}(t)$	SE	t^*	$\tilde{S}_{sample}(T+t^{*})$	SE	g	$\tilde{t}_g^*$
(a) Elevation 50 m, S_{obs} = 227, T = 599							
1	9.98	1.27	0	227.00	6.51	0.82	23.29
50	109.64	6.17	100	234.57	6.81	0.86	183.50
100	140.09	6.39	200	241.03	7.24	0.90	398.00
150	159.30	6.41	300	246.56	7.79	0.94	723.65
200	173.30	6.37	400	251.29	8.43	0.98	1424.02
250	184.27	6.33	401	251.33	8.44		
300	193.23	6.30					
350	200.79	6.28					
400	207.32	6.27					
450	213.06	6.29					
500	218.19	6.34					
550	222.83	6.41					
599	227.00	6.51					
(b) Elevation 500 m, S_{obs} = 241, T = 230							
1	12.80	1.42	0	241.00	7.52	0.82	63.33
20	98.96	6.03	100	266.19	8.93	0.86	123.55
40	132.85	6.57	200	282.78	11.08	0.90	204.17
60	155.08	6.76	300	293.71	13.45	0.94	326.56
80	171.85	6.84	400	300.91	15.64	0.98	589.79
100	185.42	6.88	500	305.65	17.49		
120	196.90	6.92	600	308.78	18.97		
140	206.90	6.97	700	310.84	20.11		
160	215.80	7.04	770	311.84	20.75		
180	223.83	7.14					
200	231.15	7.27					
230	241.00	7.52					
(c) Elevation 1 070 m, S_{obs} = 122, T = 150							
1	11.53	1.51	0	122.00	4.50	0.84	5.06
20	68.06	4.59	100	135.00	5.95	0.86	22.53
40	85.18	4.57	200	141.06	8.00	0.88	42.71
60	95.91	4.44	300	143.88	9.63	0.90	66.57
80	103.97	4.36	400	145.19	10.69	0.92	95.77
100	110.41	4.34	500	145.80	11.32	0.94	133.42
120	115.68	4.37	600	146.09	11.67	0.96	186.49
140	120.07	4.44	700	146.22	11.87	0.98	277.20
150	122.00	4.50	800	146.28	11.97		
			850	146.30	12.00		
(d) Elevation 1 500 m, S_{obs} = 56, T = 200							
1	5.85	1.17	0	56.00	3.91	0.74	15.70
20	31.68	3.49	100	61.58	4.61	0.78	69.80
40	38.64	3.65	200	65.68	5.74	0.82	134.79
60	42.71	3.68	300	68.70	7.10	0.86	216.19
80	45.65	3.69	400	70.91	8.50	0.90	325.16
100	47.98	3.69	500	72.53	9.82	0.94	490.60
120	49.94	3.70	600	73.72	11.00	0.98	846.42
140	51.67	3.73	700	74.60	12.01		
160	53.22	3.77	800	75.24	12.87		
180	54.66	3.83					
200	56.00	3.91					
(e) Elevation 2 000 m, S_{obs} = 14, T = 200							
1	1.36	0.43	0	14.00	0.49	0.99	28.00
20	8.60	1.25	100	14.21	0.63		
40	10.59	1.12	200	14.24	0.70		
60	11.62	0.95	300	14.25	0.72		
80	12.31	0.82	400	14.25	0.73		
100	12.81	0.70	500	14.25	0.73		
120	13.19	0.62	600	14.25	0.73		
140	13.49	0.56	700	14.25	0.73		
160	13.71	0.52	800	14.25	0.73		
180	13.88	0.50					
200	14.00	0.49					

The extrapolation is extended to 1 000 samples for each elevation.

statistically different from the richness of a random sample of the same size drawn from the larger reference sample, Y ? ' (The conditional variance of sample X is zero for the full sample.) In contrast, ecologists would usually prefer to answer the question, 'Are the numbers of species recorded in samples X and Y statistically different from the richness of random samples, matching the smaller sample X in number of individuals, from the assemblage or assemblages they represent?' (Simberloff 1979). The latter question requires an estimate of the unconditional variance for both samples. We present, for the first time, simple and explicit variance estimators for both fixed size (multinomial, Equation 5) and random-size (Poisson, Equation 7) individual-based rarefaction models, and we extend the potential for statistical comparison beyond the size of reference samples by extrapolation.

Even when based on unconditional variances, the use of confidence intervals to infer statistical significance (or lack of it) between samples is not straightforward. In general, lack of overlap between 95% confidence intervals (mean plus or minus 1.96 SE) does indeed guarantee significant difference in means at $P \leqslant 0.05$, but this condition is overly conservative: samples from normal distributions at the $P = 0.05$ threshold have substantially overlapping 95% confidence intervals. Payton et al. (2004) show that, for samples from two normal

distributions with approximately equal variances, overlap or non-overlap of 84% confidence intervals (mean plus or minus 1.41 SE) provide a more appropriate rule of thumb for inferring a difference of mean at $P = 0.05$, and this approach has been suggested by two of us for comparing unconditional confidence intervals around rarefaction curves (Gotelli and Colwell 2011). Unfortunately, the statisticians among us (A.C., C.X.M. and S.-Y.L.) doubt that this approach is likely to be accurate for the confidence intervals around rarefaction (or extrapolation) curves, so the matter of a simple method must be left for further study. Meanwhile, non-overlap of 95% confidence intervals constructed from our unconditional variance estimators can be used as a simple but conservative criterion of statistical difference. Mao and Li (2009) developed a mathematically complicated method for comparing entire rarefaction curves, but it has so far been little used.

All our examples (Tables 2, 3, 5 and 7; Figs 2 and 4) reveal that the unconditional variance increases sharply with sample size for extrapolated curves, and thus, the confidence interval expands accordingly. As with any extrapolation, the estimate becomes more uncertain the further it is extended away from the reference sample. As a consequence, confidence intervals that do not overlap at moderate sample sizes may do so at larger sample sizes, even if the extrapolated curves are not converging. An example of this phenomenon can be seen in the lower two curves of Fig. 4a. We would suggest that extrapolation is reliable, at most, only up to a tripling of the reference sample size, or more conservatively, a doubling of sample size. We have carried out simulations to investigate the performance of the unconditional variance estimators (Equations5, 7, 10 and 13). The proposed unconditional variances perform satisfactorily when sample size is relatively large because they were derived by an asymptotic approach (i.e. assuming the sample size is large). When sample size is not sufficiently large, the unconditional variances tend to overestimate and, thus, produce a conservative confidence interval. For small samples, we suggest estimating variance by non-parametric bootstrapping.

Under all three of the models we discuss, all our estimators for extrapolated richness, as well as all our unconditional variance estimators, require an estimate of asymptotic species richness for the assemblage sampled. For this reason, the accuracy of our extrapolation and variance estimators is of course dependent upon the accuracy of the asymptotic richness estimates they rely upon. However, if and when better estimators of assemblage richness become available, they can simply be plugged into our equations wherever S_{est}, $\hat{f}_{0,}$ or $\hat{Q}_0$ appear in our equations.

Under the Poisson model, individual-based rarefaction curves and species accumulation curves, because they rely on area, assume that individuals are randomly distributed in space, within and between species. The multinomial model can be viewed as having the same assumption, or alternatively, may be viewed as assuming that species need not be randomly distributed, but that individuals have been recorded randomly without regard to their position in space. Neither assumption is realistic for a practical study of any natural assemblage, which routinely exhibit spatial aggregation within species, as well as spatial patterning in association and dissociation between species. All such violations of the assumptions of spatial randomness lead to an overestimation of richness for a given number of individuals or a given amount of accumulated space, compared with what richness would be for actual smaller or larger samples (Chazdon et al. 1998; Colwell and Coddington 1994; Kobayashi 1982).

Sample-based approaches (e.g. estimators based on the Bernoulli product model), using replicated incidence data (or sample-based abundance data converted to incidence), perform better in this regard as they retain some aspects of the spatial (or temporal) structure of assemblages (Colwell et al. 2004; Gotelli and Colwell 2001; Smith et al. 1985), although sampling designs are nonetheless critical to avoiding bias from spatial structure (Collins and Simberloff 2009; Chiarucci et al. 2009). It may at first appear paradoxical that a simple list of incidence frequencies (e.g. Table 6) retains any information on the spatial structure of the biological populations sampled. But consider two equally abundant species in the same assemblage, one with a very patchy spatial distribution and the other with all individuals distributed independently and at random. With individual-based rarefaction, the two species will be indistinguishable. In a sample-based study of the same assemblage, however, the aggregated species will generally have a lower incidence frequency (since many individuals will end up some samples and none in others) than the randomly distributed species. While accounting for within-species aggregation, however, sample-based rarefaction is blind to interspecific association or dissociation (Colwell et al. 2004, their Table 2).

When sample-based (replicate) data are not available, the individual-based methods we present here can be applied, with the understanding that spatial structure is ignored. To model species aggregation explicitly, the current models could be extended to a negative binomial model (a generalized form of our Poisson model; Kobayashi 1982, 1983) and to a multivariate negative binomial model (a generalized form of our multinomial) model. Extra parameters that describe spatial aggregation would need to be introduced in the generalized model, and thus, statistical inference would become more complicated.

We plan to implement the rarefaction and extrapolation estimators discussed in this paper in the freeware applications EstimateS (Colwell 2011) and in iNEXT (http://chao.stat.nthu.edu.tw/softwareCE.html).

FUNDING

US National Science Foundation (DEB 0639979 and DBI 0851245 to R.K.C.; DEB-0541936 to N.J.G.; DEB-0424767 and DEB-0639393 to R.L.C.; DEB-0640015 to J.T.L.); the US Department of Energy (022821 to N.J.G.); the Taiwan

National Science Council (97-2118-M007-MY3 to A.C.); and the University of Connecticut Research Foundation (to R.L.C.).

ACKNOWLEDGEMENTS

We are grateful to Fangliang He and Sun Yat-sen University for the invitation to contribute this paper to a special issue of JPE and to an anonymous reviewer for helpful comments.
Conflict of interest statement. None declared.

REFERENCES

Brewer A, Williamson M (1994) A new relationship for rarefaction. *Biodiversity Conserv* **3**:373–9.

Burnham KP, Overton WS (1978) Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**:625–33.

Chao A (1984) Non-parametric estimation of the number of classes in a population. *Scand J Stat* **11**:265–70.

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**:783–91.

Chao A (2005) Species estimation and applications. In: Kotz S, Balakrishnan N, Read CB, Vidakovic B (eds). *Encyclopedia of Statistical Sciences*, 2nd edn. New York: Wiley, 7907–16.

Chao A, Colwell RK, Lin C-W, *et al.* (2009) Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **90**: 1125–33.

Chao A, Hwang W-H, Chen Y-C, *et al.* (2000) Estimating the number of shared species in two communities. *Stat Sin* **10**:227–46.

Chao A, Lee S-M (1992) Estimating the number of classes via sample coverage. *J Am Stat Assoc* **87**:210–7.

Chao A, Shen TJ (2004) Nonparametric prediction in species sampling. *J Agric Biol Environ Stat* **9**:253–69.

Chazdon RL, Colwell RK, Denslow JS, *et al.* (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In: Dallmeier F, Comiskey JA (eds). *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies*. Paris: Parthenon Publishing, 285–309.

Chazdon RL, Colwell RK, Denslow JS. (1999) Tropical tree richness and resource-based niches. *Science* **285**:1459.

Chiarucci A, Bacaro G, Rocchini D, *et al.* (2008) Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Commun Ecol* **9**:121–3.

Chiarucci A, Bacaro G, Rocchini D, *et al.* (2009) Spatially constrained rarefaction: incorporating the autocorrelated structure of biological communities into sample-based rarefaction. *Commun Ecol* **10**: 209–14.

Coleman BD (1981) On random placement and species-area relations. *Math Biosci* **54**:191–215.

Coleman BD, Mares MA, Willig MR, *et al.* (1982) Randomness, area, and species richness. *Ecology* **63**:1121–33.

Collins MD, Simberloff D (2009) Rarefaction and nonrandom spatial dispersion patterns. *Environ Ecol Stat* **16**:89–103.

Colwell RK (2011) Estimates: Statistical Estimation of Species Richness and Shared Species from Samples. Version 9. User's Guide and application published at http://purl.oclc.org/estimates (13 November 2011, date last accessed).

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B Biol Sci* **345**:101–18.

Colwell RK, Mao CX, Chang J (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* **85**:2717–27.

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

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* **4**:379–91.

Gotelli NJ, Colwell RK (2011) Estimating species richness. In: Magurran AE, McGill BJ (eds). *Frontiers in Measuring Biodiversity*. New York: Oxford University Press, 39–54.

Gotelli NJ, Ellison AM, Dunn RR, *et al.* (2011) Counting ants (Hymenoptera: Formicidae): biodiversity sampling and statistical analysis for myrmecologists. *Myrmecol News* **15**:13–9.

Hayek L-A, Buzas MA (1997) *Surveying Natural Populations*. New York: Columbia University Press.

Heck KL, Jr., van Belle G, Simberloff D (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* **56**:1459–61.

Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**:577–86.

Janzen DH (1973a) Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions. *Ecology* 54:659–86.

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

Kobayashi S (1982) The rarefaction diversity measurement and the spatial distribution of individuals. *Jpn J Ecol* **32**:255–8.

Kobayashi S (1983) Another calculation for the rarefaction diversity measurement for different spatial distributions. *Jpn J Ecol* **33**:101–2.

Lawton JH, Bignell DE, Bolton B, *et al.* (1998) Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature* **391**:72–6.

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

Lee TM, Sodhi NS, Prawiradilaga DM (2007) The importance of protected areas for the forest and endemic avifauna of Sulawesi (Indonesia). *Ecol Appl* **17**:1727–41.

Lehmann EL, Casella G (1998) *Theory of Point Estimation*, 2nd edn. New York: Springer-Verlag.

Longino JT, Colwell RK (1997) Biodiversity assessment using structured inventory: capturing the ant fauna of a lowland tropical rainforest. *Ecol Appl* **7**:1263–77.

Longino JT, Colwell RK (2011) Density compensation, species composition, and richness of ants on a Neotropical elevational gradient. *Ecosphere* **2**:art29.

Mao CX, Colwell RK (2005) Estimation of species richness: mixture models, the role of rare species, and inferential challenges. *Ecology* **86**:1143–53.

Mao CX, Colwell RK, Chang J (2005) Estimating species accumulation curves using mixtures. *Biometrics* **61**:433–41.

Mao CX (2007) Estimating species accumulation curves and diversity indices. *Stat Sin* **17**:761–74.

Mao CX, Li J (2009) Comparing species assemblages via species accumulation curves. *Biometrics* **65**:1063–7.

Norden N, Chazdon RL, Chao A, *et al.* (2009) Resilience of tropical rain forests: tree community reassembly in secondary forests. *Ecol Lett* **12**:385–94.

Payton ME, Greenstone MH, Schenker N (2004) Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance. *J Insect Sci* **3**:34, http://insectscience.org/3.34 (13 November 2011, date last accessed).

Sanders H (1968) Marine benthic diversity: a comparative study. *Am Nat* **102**:243.

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

Simberloff D (1979) Rarefaction as a distribution-free method of expressing and estimating diversity. In: Grassle JF, Patil GP, Smith WK, Taillie C (eds). *Ecological Diversity in Theory and Practice*. Fairland, MD: International Cooperative Publishing House, 159–76.

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

Smith EP, Stewart PM, Cairns J (1985) Similarities between rarefaction methods. *Hydrobiologia* **120**:167–70.

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

Solow A, Polasky S (1999) A quick estimator for taxonomic surveys. *Ecology* **80**:2799–803.

Ugland KI, Gray JS, Ellingsen KE (2003) The species–accumulation curve and estimation of species richness. *J Anim Ecol* **72**:888–97.