Biometrics 68, 912–921
September 2012

DOI: 10.1111/j.1541-0420.2011.01739.x

Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement

Anne Chao[1,*] **and Chih-Wei Lin**[1,2]

[1]Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043
[2]Department of Leisure Services Management, Chaoyang University of Technology, Taichung, Taiwan 41349
[*]*email:* chao@stat.nthu.edu.tw

Summary. A number of species richness estimators have been developed under the model that individuals (or sampling units) are sampled with replacement. However, if sampling is done without replacement so that no sampled unit can be repeatedly observed, then the traditional estimators for sampling with replacement tend to overestimate richness for relatively high-sampling fractions (ratio of sample size to the total number of sampling units) and do not converge to the true species richness when the sampling fraction approaches one. Based on abundance data or replicated incidence data, we propose a nonparametric lower bound for species richness in a single community and also a lower bound for the number of species shared by multiple communities. Our proposed lower bounds are derived under very general sampling models. They are universally valid for all types of species abundance distributions and species detection probabilities. For abundance data, individuals' detectabilities are allowed to be heterogeneous among species. For replicated incidence data, the selected sampling units (e.g., quadrats) need not be fully censused and species can be spatially aggregated. All bounds converge correctly to the true parameters when the sampling fraction approaches one. Real data sets are used for illustration. We also test the proposed bounds by using subsamples generated from large real surveys or censuses, and their performance is compared with that of some previous estimators.

Key words: Abundance data; Biological sampling; Quadrat sampling; Replicated incidence data; Shared species; Species richness.

1. Introduction

One of the most fundamental and also most popular measures of diversity is the number of species present in a community. The estimation of species richness and its applications have been extensively discussed in the literature; see, for example, Bunge and Fitzpatrick (1993), Colwell and Coddington (1994), Magurran (2004), Chao (2005), Royle and Dorazio (2008), and Gotelli and Colwell (2011) for reviews and additional references.

Both theoretical and practical works show that if there are many almost undetectable species in a hyper-diverse community, then it becomes very difficult to obtain an accurate point estimate of species richness. Given this difficulty, the determination of a lower bound for species richness and shared species richness may be of more practical use, especially if the accuracy of this lower bound is much better than the point estimate of species richness.

Two different formulas for such a lower bound for a single community were derived by Chao (1984, 1987, 1989) for abundance-based data and for replicated incidence (i.e., presence–absence) data, respectively. These lower bounds were "nonparametric," which means that we do not assume any particular parametric distribution for the underlying distributions for species abundance or species detection probabilities. These two lower bounds are referred to as the Chao1 estimator (for abundance data) and Chao2 estimator (for replicated incidence data) in the ecological literature (e.g., Colwell and Coddington, 1994). Although these two estimators were originally derived as lower bounds, they have been used as species richness point estimators, applied within various disciplines and featured in several software programs. For example, microbiologists used them to count the "uncountable" micro-organisms (Bohannan and Hughes, 2003).

Compared with species richness in a single community, the estimation of species richness shared by multiple communities has received relatively little attention. When there are multiple communities, the number of shared species among communities can be used to describe community overlap and forms a basis to construct various types of similarity or dissimilarity indices (Colwell and Coddington, 1994; Magurran, 2004). Pan, Chao, and Foissner (2009) developed a unified approach to obtain lower bounds for the shared species richness.

All the proposed lower bounds mentioned above were derived under the assumption of sampling with replacement, in which individuals (or any sampling unit) can be repeatedly observed. In this paper, we consider another type of sampling scheme, sampling without replacement. This sampling scheme is widely used in trap/net surveys when multiple individuals such as, insects, are killed when sampled, so that no sampled individual can be repeatedly observed. It has been also applied to other sampling protocols, e.g., forestry surveys, in which trees are censused by plots or quadrats that are selected without repetition. In this type of sampling scheme, any individual (or any sampling unit such as plot or quadrat) can be surveyed at most once.

Only a few species richness estimators for sampling without replacement have been described in the literature, and each has been derived from different models. Here, we only review those approaches that take into account the sampling fraction (i.e., the ratio of sample size divided by the total number of sampling units in the community) in estimation. Goodman (1949) was the first to propose an unbiased estimator under a restrictive condition, and Shlosser (1981) also proposed an estimator. However, the variance of Goodman's (1949) estimator is very large and Shlosser's (1981) estimator has a large root mean square error (Haas and Stokes, 1998). In our simulations, in some cases Shlosser's estimator has low bias but in other cases it has quite a large bias. Quadrat sampling protocols in which quadrats are selected without replacement and surveyed represent a special application of sampling without replacement. Under quadrat sampling, Mingoti and Meeden (1992) and Shen and He (2008) assume a parametric beta distribution for species detection probability in each quadrat. Iterative computation is required to obtain the resulting estimates. There are two nonparametric estimators based on the generalized jackknife techniques: the first- and second-order jackknife estimators derived by Haas and Stokes (1998) for abundance data, and those derived by Haas, Liu, and Stokes (2006) for replicated incidence data. Our numerical comparisons will be focused on these two nonparametric jackknife estimators.

So far, no universal lower bounds have been proposed for species richness and shared species richness for sampling without replacement. We here develop such estimators for the first time. Specifically, for abundance-based or replicated incidence data, we develop a nonparametric lower bound for species richness in a single community as well as a lower bound for the number of species shared by multiple communities. The proposed bounds are derived under very general sampling models and are universally valid for all types of species abundance distributions. Variance estimators for these lower bounds are also developed.

Under sampling with replacement, we only model species' relative abundances, which are independent of the population size of the community. In contrast, for sampling without replacement, we need to model the species' absolute abundances. For statistical reasons, modeling for the latter sampling scheme is unavoidably more complicated. Therefore, some researchers have been using the statistical models for sampling with replacement even when their sampling protocols were actually based on sampling without replacement. When the sampling fraction is small and the total sampling units is very large, there is little difference in the inferences for the two types of sampling schemes (Section 2). But when the sampling fraction is relatively high, the traditional estimators for sampling with replacement tend to overestimate richness when sampling is actually done without replacement; see Shen and He (2008) and Section 4 for numerical evidence. When the sampling fraction approaches one, implying all sampling units have been observed, we should expect that any estimator closely approaches the observed species richness (i.e., the true parameter). However, in this extreme case, all the traditional estimators for sampling with replacement do not estimate the true species richness reliably when the samples are taken without replacement, as will be shown in Section 5. The sampling fraction plays an important role in our proposed lower bounds. When the sampling fraction approaches one, our bounds correctly yield the true species richness whereas when the sampling fraction approaches zero, our bounds are identical to those based on sampling with replacement.

Section 2 introduces the proposed lower bound for species richness for abundance-based data (Section 2.1) as well as for replicated incidence data (Section 2.2). Section 3 develops the lower bound for the number of species shared by multiple communities for abundance-based data (Section 3.1) as well as for replicated incidence data (Section 3.2). Section 4 uses real data sets for illustration of the proposed lower bounds. Section 5.1 examines the performance of the proposed bounds by using data sets simulated from large real surveys or censuses. The performance of our lower bounds is compared with that of two jackknife estimators. Simulation results are also used to examine the effects of some model assumptions on the proposed bounds in Section 5.2. Section 6 provides some concluding remarks and discussion.

2. Species Richness

2.1 *Sampling by Individuals (Abundance Data)*

Assume that there are S species indexed from 1 to S, with S unknown. Let N_i (true species abundance or absolute abundance) be the "unknown" number of individuals of the ith species in the community, $i = 1, 2, \ldots, S$, $N_i > 0$. The total population size is then $N = \sum_{i=1}^{S} N_i$. We assume the total size N is "known" so that the sampling fraction is known. Haas and Stokes (1998) described some applications in which N is known (see Section 6). (Generally, N denotes the total number of sampling units. When we discuss quadrat sampling in Section 2.2, N becomes the number of quadrats, which is known by design.) See Section 4 for an example and Section 6 for discussion of this assumption.

Assume that a sample of n individuals is taken from the community, with individuals being drawn without replacement. Let X_i (sample species frequency) be the number of individuals of the ith species which are observed in the sample, $i = 1, 2, \ldots, S$. Only those species with $X_i > 0$ are observable in the sample. Let f_k (sample frequency counts), $k = 0, 1, \ldots, n$, be the number of species represented by exactly k individuals in the sample. Here, f_0 denotes the number of undetected species in the sample. Thus, we have $n = \sum_{i=1}^{S} X_i = \sum_{k \geq 1} k f_k$. The sample fraction is defined as $q = n/N$, the ratio of sample size to the population size. Let D denote the number of distinct species observed in the sample, i.e., $D = \sum_{k \geq 1} f_k$.

Generally, the species detection probability or rate is a combination of species abundance and individual detectability, which is determined by many possible factors (such as individual movement patterns, color, size, and vocalizations). Traditional model assumes that all individuals have the same detectability so that the sample species frequencies

$(X_1, X_2, \ldots, X_S)$ follow a generalized hypergeometric distribution:

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

In this special case, the species detection rate for the ith species is simply the true relative abundance $p_i = N_i/N$. Here we consider a more general model which allows that individuals' detectabilities vary across species. We assume that the detectability of any individual within the ith species is $\theta_i > 0$. Under this general model, the species detection rate for the ith species becomes $\psi_i = N_i\theta_i / \sum_{k=1}^{S} N_k\theta_k = p_i\theta_i / \sum_{k=1}^{S} p_k\theta_k$, $i = 1, 2, \ldots, S$. That is, the detection probability for species i is normalized product of species relative abundance (i.e., p_i) and individual detectability (i.e., θ_i). Intuitively, the number of individuals which have equal chance to be observed in the sample for species i is thus approximately $N\psi_i$. But $N\psi_i$ may not be an integer, so we define the integer-valued variable Z_i for the ith species as the unknown number of individuals which have equal chance to be observed in sample. Since $Z_i \geq 1$ (otherwise this species will have no chance to be included in sample and thus should be excluded in the estimating target) and the sampling fraction is n/N, the vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_S)$ can be modeled as a truncated multinomial distribution with cell total N and cell probabilities $(\psi_1^*, \psi_2^*, \ldots, \psi_S^*)$, where $\psi_i^* = \psi_i / P\{z; z_i \geq 1, i = 1, 2, \ldots, S\}$, $\mathbf{z} = (z_1, z_2, \ldots, z_S)$, and $\sum_{i=1}^{S} z_i = N$. For any given value of $\mathbf{z} = (z_1, z_2, \ldots, z_S)$, the sample species frequencies $(X_1, X_2, \ldots, X_S)$ follow a generalized hypergeometric distribution:

$$P(X_i = x_i, i = 1, 2, \ldots, S) = \binom{z_1}{x_1}\binom{z_2}{x_2}\cdots\binom{z_S}{x_S}\Big/\binom{N}{n}, z_i \geq 1, \sum_{i=1}^{S} z_i = N. \tag{1b}$$

If all N_i's are infinitely large and the sampling fraction is relatively small (i.e., $N >> n$), then equation (1b) approaches the following multinomial model

$$P(X_i = x_i,\ i = 1,\ 2., \ldots, S) \to \frac{n!}{x_1! \ldots x_S!}\psi_1^{x_1}\psi_2^{x_2}\ldots\psi_S^{x_S}. \tag{1c}$$

This is a model for sampling with replacement with cell probabilities $(\psi_1, \psi_2, \ldots, \psi_S)$. Equation (1c) shows that if all species are very abundant and only a small portion of the community is sampled, then the inferences for the two types of sampling schemes differ little.

Based on the general model (1b), the marginal distribution for each species frequency is a hypergeometric distribution:

$$P(X_i = x_i) = \binom{z_i}{x_i}\binom{N - z_i}{n - x_i}\Big/\binom{N}{n}. \tag{1d}$$

The expected value of the frequency counts using (1d) is

$$E(f_k) = \sum_{i=1}^{S} P(X_i = k) = \sum_{i=1}^{S}\binom{z_i}{k}\binom{N - z_i}{n - k}\Big/\binom{N}{n}. \tag{2}$$

In particular, we have

$$E(f_0) = \sum_{i=1}^{S}\binom{N - z_i}{n}\Big/\binom{N}{n},$$

$$E(f_1) = \sum_{i=1}^{S}\binom{z_i}{1}\binom{N - z_i}{n - 1}\Big/\binom{N}{n}$$

$$= \sum_{i=1}^{S}\frac{nz_i}{N - z_i - n + 1}\binom{N - z_i}{n}\Big/\binom{N}{n},$$

$$E(f_2) = \sum_{i=1}^{S}\binom{z_i}{2}\binom{N - z_i}{n - 2}\Big/\binom{N}{n}$$

$$= \sum_{i=1}^{S}\frac{n(n-1)z_i(z_i-1)}{2(N-z_i-n+1)(N-z_i-n+2)}\binom{N-z_i}{n}\Big/\binom{N}{n}.$$

The Cauchy–Schwarz inequality leads to

$$\left\{\sum_{i=1}^{S}\frac{nz_i}{N - z_i - n + 1}\binom{N - z_i}{n}\Big/\binom{N}{n}\right\}^2 \leq \left\{\sum_{i=1}^{S}\binom{N - z_i}{n}\Big/\binom{N}{n}\right\} \times \left\{\sum_{i=1}^{S}\left(\frac{nz_i}{N - z_i - n + 1}\right)^2\binom{N - z_i}{n}\Big/\binom{N}{n}\right\},$$

with equality achieved when all z_i's are equal. The left side in the above inequality is $\{E(f_1)\}^2$, and the first sum on the right side is $E(f_0)$. We need a little algebra for the second sum. Rewrite

$$\left(\frac{nz_i}{N - z_i - n + 1}\right)^2 = \frac{n}{n-1}\left(\frac{n(n-1)z_i(z_i - 1)}{(N - z_i - n + 1)^2}\right) + \frac{n^2 z_i}{(N - z_i - n + 1)^2}.$$

Thus the second sum becomes

$$\left\{\sum_{i=1}^{S}\left(\frac{nz_i}{N - z_i - n + 1}\right)^2\binom{N - z_i}{n}\Big/\binom{N}{n}\right\} \approx \frac{2n}{n-1}E(f_2) + \sum_{i=1}^{S}\left[\frac{n}{N - z_i - n + 1}\right]\frac{nz_i}{N - z_i - n + 1}\binom{N - z_i}{n}\Big/\binom{N}{n}.$$

The contribution of those species with large z_i to the last term in the above equation is almost negligible. For those species with z_i being much less than N, we have

$$\frac{n}{N - z_i - n + 1} = \frac{n/N}{(N - z_i - n + 1)/N} \approx \frac{n/N}{1 - (n/N)} = \frac{q}{1 - q}.$$

We then obtain the following approximate inequality

$$\{E(f_1)\}^2 \leq \{E(f_0)\}\left\{\frac{n}{n-1}2E(f_2) + \frac{q}{1 - q}E(f_1)\right\},$$

which is equivalent to

$$E(f_0) \geq \frac{E\left(f_1^2\right)}{\dfrac{n}{n-1}2E(f_2) + \dfrac{q}{1-q}E(f_1)}. \quad (3)$$

Replacing the expected value by the observed frequencies, we thus obtain the following lower bound for the true species richness. We call it $\hat{S}_{\text{wor1}}$, where the subscript "wor" refers to "without replacement."

$$\hat{S}_{\text{wor1}} = D + \frac{f_1^2}{\dfrac{n}{n-1}2f_2 + \dfrac{q}{1-q}f_1} \equiv D + \frac{f_1^2}{2wf_2 + rf_1}, \quad (4)$$

where $w = n/(n-1)$ and $r = q/(1-q)$.

When only a small portion of individuals are taken from the entire universe of N individuals in the community, so that the sample fraction q approaches zero, our lower bound approaches the Chao1 estimator:

$$\hat{S}_{\text{Chao1}} = D + \frac{(n-1)}{n}\frac{f_1^2}{2f_2}. \quad (5)$$

On the other hand, when q approaches 1, $q/(1-q)$ approaches infinity and our lower bound reduces to the number of observed species, which equals the true parameter.

An approximate variance formula for $\hat{S}_{\text{wor1}}$ can be obtained by using an asymptotic approach based on the hypergeometric distribution. The resulting variance estimator is:

$$\hat{\text{var}}(\hat{S}_{\text{wor1}}) = \hat{f}_0 + \frac{\left(2wf_2\hat{f}_0^2 + f_1^2\hat{f}_0\right)^2}{f_1^5} + 4w^2f_2\left(\frac{\hat{f}_0}{f_1}\right)^4, \quad (6)$$

where $\hat{f}_0 = f_1^2/(2wf_2 + rf_1)$ denotes the estimator of the undetected species in the sample. The performance of this variance estimator is investigated in Section 5. When $\hat{S}_{\text{wor1}}$ is used as an estimator of species richness, a confidence interval of S can thus be constructed by a log-transformation so that the lower bound is always greater than the number of observed species (Chao, 1987).

2.2 *Sampling by Quadrats (Replicated Incidence Data)*

In many biodiversity studies, the sampling unit is not an individual, but a trap, net, quadrat, plot, or timed survey. It is these sampling units, and not the individual organisms, that are actually sampled randomly and independently. Counting the exact number of individuals for each species appearing within each sampling unit may often become impossible for micro-organisms, invertebrates or plants. In most cases, only their incidence (presence or absence) can be recorded. In this subsection, we discuss the estimation based on a set of replicated samples in which the incidence of each species is recorded in each sample unit instead of its abundance. Although we use the term "quadrat" in the example given below, the sampling unit "quadrat" may as well refer to a trap, net, team, observer, occasion, transect line or fixed period of time in other sampling protocols. Suppose the study area is divided into T quadrats with roughly equal area (T is known) indexed 1, 2, ..., T.

Let M_i (the true incidence-based species frequency) be the unknown number of occupied quadrats by the ith species, $i = 1, 2, \ldots, S$. We assume that in each quadrat, the "conditional" probability of detecting species i in any selected quadrat (given species i is present) is $0 < \alpha_i \leq 1$. That is, any selected quadrat need not be completely censused. Our model can thus be applied to not only surveys of sessile plants but also surveys of mobile animals. The model assumes that out of these M_i quadrats, species i can only be detected in U_i quadrats (U_i is also unknown and $M_i \geq U_i \geq 1$). We restrict to the case $U_i \geq 1$. (For any species with $U_i = 0$, there is no chance to detect this species in any sample, so it should be excluded in the estimating target.) Here, U_i is a truncated binomial distribution with probability $P(U_i = k) = [M_i!/\{k!(M_i-k)!\}]$ $\alpha_i^k(1-\alpha_i)^{M_i-k}/\{1-(1-\alpha_i)^{M_i}\}$ for $k = 1, 2, \ldots, M_i$. In the other $T - U_i$ quadrats, either species i is absent or it is present but cannot be detected. In our simulation (in Section 5.2), we considered three types of distributions for species detection probability α_i (constant, uniform distribution and beta distributions). Because M_i may vary with species, our model holds even if species are spatially aggregated in the study area. See Sections 5.2 and 6 for discussion.

Assume a sample of t quadrats is randomly selected without replacement. The presence or absence of any species for each of these t quadrats is recorded to form a species-by-quadrat incidence matrix. Let Y_i (sample incidence-based species frequency) be the number of quadrats in which the ith species is observed in the sample, $i = 1, 2, \ldots, S$. Then the sample frequencies $(Y_1, Y_2, \ldots, Y_S)$ given $U_i = u_i$ follow a product-hypergeometric distribution:

$$P(Y_i = y_i, i = 1, 2, \ldots, S) = \prod_{i=1}^{S}\left\{\binom{u_i}{y_i}\binom{T-u_i}{t-y_i}\Big/\binom{T}{t}\right\}, \quad 1 \leq u_i \leq M_i. \quad (7)$$

That is, $(Y_1, Y_2, \ldots, Y_S)$are independent but nonidentically distributed random variables and each follows a hypergeometric distribution.

Denote the sample incidence-based frequency counts by $(Q_1, Q_2, \ldots, Q_t)$, where Q_k is the number of species that are detected in exactly k quadrats in the data, $k = 1, 2, \ldots, t$. Hence, Q_1 represents the number of "unique" species (those that are detected in only one quadrat) and Q_2 represents the number of "duplicate" species (those that are detected in only two quadrats). It follows from the distribution of Y's that

$$E(Q_k) = \sum_{i=1}^{S} P(Y_i = k) = \sum_{i=1}^{S}\binom{u_i}{k}\binom{T-u_i}{t-k}\Big/\binom{T}{t}. \quad (8)$$

The sampling fraction here is defined as $q = t\ /T$. The expectation (8) has a similar form to the one presented in (2). Let D denote the number of distinct species that are observed in at least one quadrat. Parallel derivations to those in Section 2.1 can be made with n being replaced by t, and the counts $(f_1, f_2, \ldots, f_n)$ being replaced by $(Q_1, Q_2, \ldots, Q_t)$, yielding a lower bound (called $\hat{S}_{\text{wor2}}$) of the true species richness.

$$\hat{S}_{\text{wor2}} = D + \frac{Q_1^2}{\dfrac{t}{t-1}2Q_2 + \dfrac{q}{1-q}Q_1}. \quad (9)$$

When $q \to 0$, our lower bound approaches the Chao2 estimator which is the lower bound for sampling with replacement (Chao, 1989):

$$\hat{S}_{\text{Chao2}} = D + \frac{(t-1)}{t}\frac{Q_1^2}{2Q_2}. \tag{10}$$

When $q \to 1$, our lower bound converges to the true species richness. A variance estimator and confidence interval can be similarly obtained as in the individual-based case; see equation (6).

3. Shared Species Richness

3.1 *Sampling by Individuals (Abundance Data)*

Below, we mainly discuss the case of two communities. Assume that there are S_1 species in Community I and there are S_2 species in Community II. Let the number of shared species be S_{12}. Let N_{ij} (true species abundance) be the unknown number of individuals of the ith species in Community j, $i = 1, 2, \ldots, S_j$, $j = 1, 2$. The total population size in each community is $N_j = \sum_{i=1}^{S_j} N_{ij}$. We assume the two total sizes N_1 and N_2 to be known.

Two random samples of individuals (sample I with size n_1 and sample II with size n_2) are independently taken without replacement from Communities I and II, respectively. Assume that D_{12} shared species are observed. The observed frequencies in the two communities are given by $(X_{11}, X_{21}, \ldots, X_{S_1,1})$ and $(X_{12}, X_{22}, \ldots, X_{S_2,2})$, respectively. Without loss of generality, we assume that the first S_{12} species of the two sets are the shared species. In each community, the sample species frequencies then follow a generalized hypergeometric distribution as in (1b). Our model allows that individuals' detectabilities vary across species in each community.

Given any two non-negative integers j and k, let f_{jk} denote the number of "shared" species that are represented by j individuals in sample I and k individuals in sample II. In particular, f_{11} denotes the number of shared species that are singletons in both samples, and f_{00} denotes the number of shared species that are undetected in both samples. Also, f_{j+} is defined as the number of shared species that are represented by j individuals in sample I and present (by at least one individual) in sample II, and an analogous definition is used for f_{+k}. Here, f_{+0} is the number of shared species that are observed in sample I but not observed in sample II, and a similar interpretation holds for f_{0+}. Following a similar approach as in Section 2.1 and assuming the independence of the two samples, we obtain

$$E(f_{jk}) = \sum_{i=1}^{S_{12}} \binom{z_{i1}}{j}\binom{N_1 - z_{i1}}{n_1 - j}\binom{z_{i2}}{k}\binom{N_2 - z_{i2}}{n_2 - k} \Bigg/ \binom{N_1}{n_1}\binom{N_2}{n_2}, \tag{11}$$

where z_{ij} denotes the unknown number of individuals which have equal chance to be observed for the ith species in Community j. Since $S_{12} = D_{12} + E(f_{+0}) + E(f_{0+}) + E(f_{00})$ and only D_{12} is observable, our goal is to derive a lower bound for each of the other three terms. Define the two sampling fractions q_1 and q_2 as $q_j = n_j/N_j$, $j = 1, 2$. For $E(f_{+0})$ and $E(f_{0+})$, similar lower bounds as in (4) can be obtained, and these two lower bounds are shown in the second and third terms in the right hand side of (12). Therefore, we only need to derive a lower bound for $E(f_{00})$. Using the moment formula $E(f_{jk})$ in (11) and the Cauchy–Schwarz inequality, we derive a lower bound in Web Appendix A for $E(f_{00})$. The resulting lower bound for the shared species richness S_{12} is given below. We call this lower bound $\hat{S}_{12,\text{wor1}}$.

$$\hat{S}_{12,\text{wor1}} = D_{12} + \frac{{f_{1+}}^2}{2w_1 f_{2+} + r_1 f_{1+}} + \frac{{f_{+1}}^2}{2w_2 f_{+2} + r_2 f_{+1}} + \frac{{f_{11}}^2}{4w_1w_2f_{22} + 2w_1r_2f_{21} + 2r_1w_2f_{12} + r_1r_2f_{11}}, \tag{12}$$

where $w_j = n_j/(n_j - 1)$, and $r_j = q_j/(1 - q_j)$, $j = 1, 2$. If q_1, $q_2 \to 0$, then

$$\hat{S}_{12,\text{wor1}} \to D_{12} + \frac{{f_{1+}}^2}{2w_1 f_{2+}} + \frac{{f_{+1}}^2}{2w_2 f_{+2}} + \frac{{f_{11}}^2}{4w_1w_2f_{22}}, \tag{13}$$

which is identical to the lower bound for sampling with replacement (Pan et al., 2009). If q_1, $q_2 \to 1$, then $\hat{S}_{12,\text{wor1}}$ correctly approaches the true parameter. The above approach represents a unified framework of constructing a lower bound for shared species richness. The extension to the case of more than two communities is shown in Web Appendix B. Instead of deriving an asymptotic variance, we adopt a bootstrap approach to obtain a variance estimator for $\hat{S}_{12,\text{wor1}}$ (details are given in Web Appendix C).

3.2 *Sampling by Quadrats (Replicated Incidence Data)*

We again use the term "quadrat" as in Section 2.2. Suppose the two study areas are divided into T_1 and T_2 quadrats with roughly equal size, respectively, where both T_i's are known. We randomly select t_1 quadrats from the first area and t_2 quadrats from the second area. As described in Section 2.2, our model does not require that all selected quadrats be totally sampled. The incidence of any species for each sampled quadrat is recorded to form a species-by-quadrat incidence matrix for each community. Let Q_{jk} denote the number of shared species that are detected in j quadrats in Community I and k quadrats in Community II. By applying a method analogous to that used in Section 3.1, it can be shown that the lower bound $\hat{S}_{12,\text{wor2}}$ for the number of shared species based on incidence counts has the same form as in (12) except that the samples sizes n_1 and n_2 should be, respectively, replaced by t_1 and t_2, the sampling fractions replaced by t_1/T_1 and t_2/T_2, and the abundance counts replaced by the incidence-based counts Q_{jk}. Extension to the more general case and the associated inference procedures are parallel to those in Section 3.1.

4. Applications to Real Data

We applied the proposed bounds to infer species richness and shared species richness for some real data sets. Here we only present species richness estimation. Shared species richness estimation is provided in Web Appendix D (data in Web Table 1 and estimates in Web Table 2). We considered a small benthic infaunal data set which was originally used in Heltshe and Forrester (1983) and later discussed by Mingoti and Meeden (1992) and Haas et al. (2006). The data consist of species frequencies (number of individuals) in 10 quadrats taken from a subtidal marsh creek in Rhode Island in 1978.

Table 1

Species frequency in 10 *benthic infaunal quadrats taken from a subtidal marsh creek in Rhode Island* (*Heltshe and Forrester,* 1983). *There were* 361 *individuals of* 14 *species. The last column shows the total frequency for each species observed in the data*

	Quadrat number										
Species list	1	2	3	4	5	6	7	8	9	10	Row total
Streblospio benedicti		13	21	14	5	22	13	4	4	27	123
Nereis succines	2	2	4	4	1	1	1		1	6	22
Polydora ligni		1						1			2
Scoloplos robustus	1		1	2		6			1	2	13
Eteone heteropoda			1	2			1			1	5
Heteromastus filiformis	1	1	2	1		1			1	5	12
Capitella capitata	1										1
Scolecolepides viridis	2										2
Hypaniola grayi		1									1
Branis clavata			1								1
Macoma balthica			3							2	5
Ampelisca abdita			5	1		2				3	11
Neopanope texana								1			1
Tubifocodies sp.	8	36	14	19	3	22	6	8	5	41	162

Table 2

The influence of hypothetical quadrat number (T) *on four species richness estimates based on replicated incidence data in* 10 *quadrats shown in Table* 1. *The hypothetical sizes are assumed so that the sampling fraction is decreased from* 0.5 *to* 0.001

$\hat{S}_{uj1}$, $\hat{S}_{uj2}$*: the first- and second-order generalized jackknife by Haas et al.* (2006)*;*

$\hat{S}_{MM}$*: estimates taken from Table 2 of Mingoti and Meeden* (1992)*;*

$\hat{S}_{\text{wor2}}$*: Proposed in equation* (9)*; s.e. is discussed in Section* 2.2

Hypothetical sampling fraction q	Hypothetical quadrat number T	$\hat{S}_{uj1}$	$\hat{S}_{uj2}$	$\hat{S}_{MM}$	$\hat{S}_{\text{wor2}}$ (s.e.)
0.50	20	14.6	16.7	17.3	16.7 (2.5)
0.33	30	14.9	17.3	19.3	17.6 (3.7)
0.20	50	15.0	17.7	21.7	18.4 (3.8)
0.142	70	15.1	17.8	23.3	18.7 (4.7)
0.10	100	15.2	17.9	25.0	19.0 (5.3)
0.01	1000	15.3	18.2	36.1	19.6 (5.7)
0.001	10000	15.3	18.2	47.2	19.6 (6.7)

Table 3

The influence of hypothetic population size (N) *on three species richness estimates based on the abundance data* (*the last column in Table* 1). *The hypothetical sizes are assumed so that the sampling fraction is decreased from* 0.5 *to* 0.001

$\hat{S}_{uj1}$, $\hat{S}_{uj2}$*: the first- and second-order generalized jackknife by Haas and Stokes* (1998)*;*

$\hat{S}_{\text{wor1}}$*: proposed in equation* (4)*; s.e. formula in equation* (6)

Hypothetical sampling fraction q	Hypothetical population size N	$\hat{S}_{uj1}$	$\hat{S}_{uj2}$	$\hat{S}_{\text{wor1}}$ (s.e.)
0.50	722	14.1	24.0	16.0 (2.2)
0.33	1094	14.1	25.8	16.7 (3.1)
0.20	1805	14.1	27.0	17.2 (3.8)
0.142	2542	14.1	27.4	17.4 (4.2)
0.10	3610	14.1	27.8	17.6 (4.5)
0.01	36100	14.2	28.5	18.0 (5.2)
0.001	361000	14.2	28.6	18.0 (5.3)

The species frequencies in each quadrat are reproduced in Table 1. There were 361 individuals of 14 species.

Heltshe and Forrester (1983) obtained a species richness estimate of 18.5 with an estimated s.e. of 4.05, but sampling fraction was not considered in their method. In these data, we have $Q_1 = 5$ (there were five species that were found in only one quadrat) and $Q_2 = 2$ (there were two species that were found in only two quadrats). Based on (10), the Chao2 estimate is 19.6. To apply the estimator $\hat{S}_{\text{wor2}}$ in equation (9), we need the number of quadrats (T) in the study area, but this information was not provided in the data sources. Following the approach of Mingoti and Meeden (1992), we considered several hypothetical values of T such that the sampling fraction is between 0.001 and 0.5. For each value of T in Table 2, we compare the species richness estimates of four methods: the first- and second-order generalized jackknife by Haas et al. (2006, p. 136), the empirical Bayes estimate by Mingoti and Meeden and the proposed estimator $\hat{S}_{\text{wor2}}$ with an estimated s.e. As indicated by Mingoti and Meeden, the empirical Bayes estimate is increasing with T with a rate of $\log(T)$. The first-order jackknife estimator differs little from the observed species richness. The second-order jackknife and our proposed estimator behave similarly and are slowly increasing with T. When the number of quadrats is increased from 20 to 10,000 so that sampling fraction is decreased from 0.5 to 0.001, our proposed estimate is steadily increased from 16.7 (s.e. 2.5) to 19.6 (s.e. 6.7). Our estimates approach the Chao2 estimate as T becomes large. Mingoti and Meeden (1992) in their empirical Bayes approach assumed that the detection probability of any species in a quadrat follows a beta distribution. In our model formulation, the lower bound $\hat{S}_{\text{wor2}}$ is valid for all types of species detection probability (see Section 5.2 for simulation).

In the above quadrat data analysis, only species presence/absence is considered while the species frequencies in each quadrat are not used. An alternative approach is to pool frequencies over quadrats to obtain species abundant data (as shown in the last column of Table 1) and infer species richness based on the method presented in Section 2.1. For the pooled frequencies, we have $f_1 = 4$ (there were four singletons) and detected and $f_2 = 2$ (there were two doubletons), and thus the Chao1 estimate given in equation (5) is 18.0. We need the true population size N to obtain the lower bound $\hat{S}_{\text{wor1}}$ in equation (4). We again considered several hypothetical values of population size such that the sampling fraction is between 0.001 and 0.5. Table 3 shows how the two jackknife estimators and the proposed estimator $\hat{S}_{\text{wor1}}$ vary with population size. When the hypothetical population size is increased from 722

Table 4

Comparison of four species richness estimates for quadrat incidence data generated from the 50 *ha BCI incidence data* (*Hubbell et al.,* 2005*;* 100 × 100 *m quadrats*)*. The true parameter is* $S = 299$ *and the total number of quadrats is* $T = 50$ (*Simulation trials* = 1000)*. Detection rate of each species in any selected quadrat follows a beta* (4, 1) *distribution with mean of* 0.8 $\hat{S}_{\text{Chao2}}$*: see equation* (10)*.* $\hat{S}_{uj1}$, $\hat{S}_{uj2}$, $\hat{S}_{\text{wor2}}$*: see Table* 2

Sampling fraction q	Sample size	Average of observed species	Estimator	Average estimate	Average bias	Sample s.e.	Sample RMSE
0.1	5	223.3	$\hat{S}_{\text{Chao2}}$	245.3	−53.7[a]	11.1	54.8[b]
			$\hat{S}_{uj1}$	237.8	−61.2	7.9	61.7
			$\hat{S}_{uj2}$	237.9	−61.1	8.1	61.6
			$\hat{S}_{\text{wor2}}$ (s.e.)	244.1	−54.9	10.6(8.1)	55.9
0.3	15	260.8	$\hat{S}_{\text{Chao2}}$	289.4	−9.6[a]	14.3	17.2[b]
			$\hat{S}_{uj1}$	263.8	−35.2	5.3	35.6
			$\hat{S}_{uj2}$	271.2	−27.8	6.6	28.6
			$\hat{S}_{\text{wor2}}$ (s.e.)	281.2	−17.8	9.7(8.3)	20.2
0.5	25	276.5	$\hat{S}_{\text{Chao2}}$	307.1	8.1	14.1	16.3
			$\hat{S}_{uj1}$	277.8	−21.2	4.1	21.6
			$\hat{S}_{uj2}$	285.9	−13.1	5.2	14.1
			$\hat{S}_{\text{wor2}}$ (s.e.)	291.6	−7.4[a]	6.5(5.9)	9.9[b]
0.7	35	286.6	$\hat{S}_{\text{Chao2}}$	317.7	18.7	13.7	23.1
			$\hat{S}_{uj1}$	287.1	−11.9	3.2	12.3
			$\hat{S}_{uj2}$	294.0	−5.0	3.9	6.4
			$\hat{S}_{\text{wor2}}$ (s.e.)	295.6	−3.4[a]	4.1(3.7)	5.3[b]
0.9	45	294.2	$\hat{S}_{\text{Chao2}}$	323.3	24.3	11.4	26.9
			$\hat{S}_{uj1}$	294.4	−4.6	2.1	5.1
			$\hat{S}_{uj2}$	298.1	−0.9[a]	2.3	2.5[b]
			$\hat{S}_{\text{wor2}}$ (s.e.)	297.2	−1.8	2.2(1.8)	2.9

[a] Denotes the smallest absolute bias; [b] Denotes the smallest RMSE.

to 361,000 so that sampling fraction is decreased from 0.5 to 0.001, our proposed estimate is steadily increased from 16.0 (s.e. 2.2) to 18.0 (s.e. 5.3). The first-order jackknife estimates implies that there were no undetected species. The second-order jackknife estimates are relatively much higher than the corresponding estimates obtained from the quadrat analyses. Our estimates based on the two types of data are generally consistent.

5. Testing by Simulations

5.1 *Comparison of Estimators*

The performance of the proposed lower bounds was investigated by examining their behaviors when tested with data sets generated from a number of real biodiversity surveys or censuses. We treated the data from each of several large surveys and censuses as the "true community," so that the number of observed species in each survey is regarded as the known "true species richness." We generated subsamples from it and compared the proposed lower bound with the known species richness of the surveys and censuses. All these cases represent highly heterogeneous communities in which the species abundance-based or incidence-based frequencies vary greatly among species. Due to space limit, here we only present species richness estimation under quadrat sampling. The testing for abundance models and for shared species richness is described respectively in Web Appendix E and Appendix F. A test data set is given in Web Table 3 and simulation results are shown in Web Tables 4−7.

Here we analyzed quadrats of the size 100 × 100 m from the 50 ha (1000 × 500 m) Barro Colorado Island (BCI) plot, Panama, censused in 1985 (Hubbell, Condit, and Foster, 2005). The BCI census set included 238,018 individual trees and shrubs (≥ 1 cm in diameter at breast height) representing 299 species in 50 quadrats. Our simulation was based on the model in Section 2.2. In each selected quadrat, the detection probability of any species follows a beta (4, 1) distribution with mean detection probability 0.8; see Section 5.2 for additional types of detection probabilities. Only species presence–absence data in each quadrat were used for analyses.

We considered nine sampling fractions from 10% to 90% in an increment of 10%, but we only report five cases (10%, 30%, 50%, 70% and 90%) in Table 4. All subsampling was done by selecting quadrats without replacement from a set of 50 quadrats. For example, in the case of $q = 10\%$, we randomly selected without replacement 5 quadrats (50×10% = 5) from the whole set of 50 quadrats. For each fixed sampling fraction, 1000 simulated sets of sample with the same number of quadrats were generated. Then for each simulated data set, we recorded the number of observed species and calculated the following four estimators: the Chao2 lower bound, the two jackknife estimators, and the lower bound $\hat{S}_{\text{wor2}}$ along with its

Table 5

The effects of individual detectability on the proposed estimator $\hat{S}_{\mathrm{wor1}}$ for abundance data generated from the butterfly community given in Web Table 3. The true parameter is $S = 620$. The total size $N = 9031$ individuals. Simulation trials = 1000. Beta (α, β): beta distribution with parameters α and β, mean = $\alpha/(\alpha+\beta)$; U(a, b): uniform distribution between a and b, mean = $(a+b)/2$

Sampling fraction q	Sample size	Mean individual detectability	Individual detectability distribution	Average observed species	Average estimate $\hat{S}_{\mathrm{wor1}}$
0.2	1806	1	constant = 1	414.8	518.0
		0.8	constant = 0.8	396.5	509.2
			Beta (4, 1)	394.0	509.0
			U(0.7, 0.9)	395.6	508.9
		0.6	constant = 0.6	372.0	499.0
			Beta (3, 2)	365.2	496.3
			U(0.4, 0.8)	369.1	496.5
0.5	4515	1	constant = 1	534.1	597.1
		0.8	constant = 0.8	522.1	598.7
			Beta (4, 1)	520.4	603.4
			U(0.7, 0.9)	521.9	599.4
		0.6	constant = 0.6	506.4	583.5
			Beta (3, 2)	501.3	581.7
			U(0.4, 0.8)	504.8	582.6
0.9	8127	1	constant = 1	607.6	619.3
		0.8	constant = 0.8	599.9	613.2
			Beta (4, 1)	598.7	612.2
			U(0.7, 0.9)	599.9	613.3
		0.6	constant = 0.6	589.7	604.9
			Beta (3, 2)	586.5	602.2
			U(0.4, 0.8)	589.0	604.4

approximate s.e. based on an asymptotic formula. The sample standard error and sample root mean squared error (RMSE) based on the 1000 simulation samples are provided in the last two columns.

Based on Table 4 and other simulation results in Web Tables 4−7, we summarize our findings as follows: the traditional approach of using the number of observed species or shared species in the samples (shown in the third column in all tables) as an estimator of the true species richness is clearly not appropriate. It exhibits large negative bias, as would be expected and has been shown in various studies before (e.g., Colwell and Coddington, 1994).

The first-order jackknife estimator $\hat{S}_{uj1}$ is severely negatively biased in the cases of low sampling fractions. The Chao2 estimator is monotonically increasing with sampling fraction, but they overestimate the true species richness when sampling fraction is >30%. Although it has the smallest RMSE for $q = 10\%$ and 30%, it does not converge to the true species richness as the sampling fraction q approaches 1. The second-order jackknife performs reasonably in Table 4 when the study area is divided as 50 quadrats, but it gives inconsistent estimates with severe positive biases when the study area is divided as 1250 quadrats (Web Table 5). Each of the other three estimators generally produces similar results for the two quadrat sizes.

As implied by theory, our proposed lower bounds in all cases are less than the true species richness. In all cases we examined, the bound is stably and monotonically increasing as q is increased, and it converges correctly to the true species richness when q approaches one. The bias, standard error and RMSE of the proposed bounds all show the expected decreasing pattern when the sampling fraction is increased. As the sampling fraction exceeds 30%, the lower bound $\hat{S}_{\mathrm{wor2}}$ generally performs best among the four candidate estimators in terms of both bias and RMSE. Thus it can be used as species-richness estimator if the sampling fraction is over 30% and the relative biases in nearly all cases are within 10%. The magnitude of bias typically depends on the sampling fraction as well as on the average and heterogeneity of the species frequency distributions.

Comparing our estimated standard error with the sample standard error (both are shown in the second last column of Table 4 and Web Tables 4−7), we find that the two values for most cases are very close. This shows that the estimated standard errors using the asymptotic method (for species richness) and the bootstrap method (for shared species richness in Web Tables 6−7) are generally satisfactory even if they are slightly negatively biased.

5.2. *The Effects of Detection Rates on Estimators*

Traditional model for abundance data assumes that individuals' detectabilities are homogeneous among species. This assumption is relaxed in our model and we have theoretically justified in Section 2.1 the validity of our method when individual detectabilities vary from species to species. To numerically investigate the effects of heterogeneous detectability on our bounds, we conducted simulations by generating subsamples from the data of a Malayan butterfly survey with 620 species, 9031 individuals (Fisher, Corbet, and Williams, 1943; data are given in Web Table 3). We considered the following four cases. (1) All individuals have a constant detectability of unity. (2) All individuals have a constant detectability of

Table 6

The effects of species detection rates on the proposed estimator $\hat{S}_{\mathrm{wor2}}$ for replicated incidence data generated from 50 *ha incidence BCI census. The true parameter is $S = 299$ and the total number of quadrats is $T = 50$. Simulation trials* = 1000. *Beta* (α, β) *and* $U(a, b)$: *see Table* 5

Sampling fraction q	Sample size	Mean species detection rate	Species detection rate distribution	Average observed species	Average estimate $\hat{S}_{\mathrm{wor2}}$
0.2	10	1	constant = 1	255.4	276.7
		0.8	constant = 0.8	249.4	271.6
			Beta (4, 1)	247.8	270.0
			U(0.7, 0.9)	249.0	270.8
		0.6	constant = 0.6	240.4	262.7
			Beta (3, 2)	235.3	259.5
			U(0.4, 0.8)	239.4	261.6
0.5	25	1	constant = 1	281.5	294.5
		0.8	constant = 0.8	277.3	292.3
			Beta (4, 1)	276.5	291.6
			U(0.7, 0.9)	277.4	292.3
		0.6	constant = 0.6	271.3	288.6
			Beta (3, 2)	268.7	285.9
			U(0.4, 0.8)	270.8	288.3
0.9	45	1	constant = 1	296.5	298.8
		0.8	constant = 0.8	295.1	298.1
			Beta (4, 1)	294.2	297.2
			U(0.7, 0.9)	295.0	298.0
		0.6	constant = 0.6	291.6	295.2
			Beta (3, 2)	289.2	292.8
			U(0.4, 0.8)	291.0	294.6

0.8 (or 0.6). (3) The detectability of any individual within a species is a parameter randomly chosen from a beta distribution with mean detectability of 0.8 (or 0.6); (4) The detectability of any individual within a species is a parameter randomly chosen from a uniform distribution with mean detectability of 0.8 (or 0.6). We show part of the numerical results in Table 5 (full table in Web Table 8). Other test data sets generally yield consistent results. The results show that when mean detectabiliy is 0.8 or 0.6, the estimator $\hat{S}_{\mathrm{wor1}}$ unavoidably exhibits larger bias than that based on data with detectability of unity because the latter contains more data information. However, when mean detectability is fixed so that the data for the three cases (2), (3), and (4) are comparable, the estimates for the three cases yield very close estimates. In each of (2), (3) and (4) cases, as expected, estimates are monotonically increasing to the true value as sampling fraction is increased to one, though the convergence rate is slower. These results numerically confirm that our method is valid for individual detectabilities that vary across species. Also, our approach is independent of the distribution type of individuals' detectabilities.

For incidence data, our sampling model allows that the conditional probability of detecting any species in a selected quadrat may be less than 1 given its presence. In Table 6 (details are given in Web Table 9), we show simulation results based on generated data from BCI census for the four cases of species detection rates mentioned above. Similar conclusions as in the abundance data are obtained.

6. Concluding Remarks and Discussion

When individuals or sampling units are sampled without replacement from target communities, no universal lower bounds for species richness and shared species richness existed before. Based on abundance and replicated incidence data, we developed in this paper simple and useful species richness lower bounds for single communities, when the population size is known. We also derived similar lower bounds for shared species richness. Simulations showed that if we allow the relative bias to be within 10%, all proposed lower bounds can be used as point estimators when the sampling fraction in each community is higher than 30%. The bounds and estimators discussed in this paper will be featured in the Program SPADE (Species Prediction And Diversity Estimation, Chao and Shen, 2010) following publication of this paper.

Our proposed lower bounds are derived from very general sampling models. For abundance data, it is assumed that species detectation rate is proportional to the product of species abundance and individual detectability of that species. Here individuals' detectabilities may vary from species to species. For incidence data based on quadrat sampling, species detection probability in any selected quadrat given its presence may be less than 1 (i.e., each selected quadrat need not be completely surveyed). The number of occupied quadrats for any species is allowed to vary with species, implying that even species are spatially aggregated or clustered, our method is still valid. Since the bounds tend to the Chao1 and Chao2 estimators when sampling fraction is small, our results also imply that both the Chao1 and Chao2 estimators are valid under similar general models.

A critical assumption in all of our proposed models is that the sampling fraction should be known. Although this assumption is usually satisfied under quadrat sampling or surveys based on areas, sampling fraction information is generally not available for animal abundance surveys because population size is often unknown. In such a case, we suggest that a sequence of estimates for several hypothetical values of sampling fraction should be examined as in the data analysis in Section 4. Haas and Stokes (1998) mentioned some other interesting applications in which population size is known. These include the following estimation issues based on a nonrepeated sample. (1) The estimation of the number of "different" people who have entered a contest. Some people might have entered multiple times. Here the total number of contest entries is known (Sudman, 1976). (2) The estimation of total number of "distinct" units in a combined list which is formed by merging several possibly overlapped lists. Here the total number of units over all lists is known. (3) The estimation of the number of distinct values of an attribute in large database management system. The total number of records in the database is known. Thus, our methods can be applied to the above issues.

Chao et al. (2009) developed a nonparametric method for estimating the minimum amount of sampling effort (additional individuals or quadrats/samples) required to detect any arbitrary proportion (including 100%) of the estimated lower bound in a single community. Their models, however, were based on sampling "with" replacement. The extension of their method to sampling without replacement is a worthwhile research topic.

7. Supplementary Materials

Web Appendices and Web Tables referenced in Sections 3, 4, and 5 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank the Editor (Tom Louis), an Associate Editor and two reviewers for carefully reading an earlier version and providing insightful and thoughtful suggestions and comments, which substantially improved the paper. Especially, the reviewers' comments inspired us to develop very general sampling models. The development has greatly broadened the scope of our methods. We also sincerely thank Lou Jost and Bruno Walther for editing and offering very helpful comments. This work was supported by the Taiwan National Science Council.

References

Bohannan, B. J. M. and Hughes, J. (2003). New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology* **6,** 182–187.

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

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

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

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

Chao, A. (2005). Species estimation and applications. *Encyclopedia of Statistical Sciences*, 2nd Edition, Vol. **12,** S. Kotz, N. Balakrishnan, C. B. Read and B. Vidakovic, (eds), 7907—7916. New York: Wiley.

Chao, A., Colwell, R. K., Lin, C. -W., and Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **90,** 1125–1133.

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

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

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

Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics* **20,** 572–579.

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

Haas, P. J., Liu, Y., and Stokes, L. (2006). An estimator of number of species from quadrat sampling. *Biometrics* **62,** 135–141.

Haas, P. J., and Stokes, S. L. (1998). Estimating the number of classes in a finite population. *Journal of the American Statistical Association*, **93,** 1475–1487.

Helstshe J. F., and Forrester N.E. (1983). Estimating diversity using quadrat sampling. *Biometrics* **39,** 1–11.

Hubbell, S.P., Condit, R., and Foster, R. B. (2005). Barro Colorado Forest Census Plot Data. http://ctfs.si.edu/datasets/bci.

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

Mingoti, S. A., and Meeden, G. (1992). Estimating the total number of distinct species using presence and absence data. *Biometrics* **48,** 863–875.

Pan, H. Y., Chao, A., and Foissner, W. (2009). A non-parametric lower bound for the number of species shared by multiple communities. *Journal of Agricultural, Biological and Environmental Statistics* **50,** 957–970.

Royle, J. A., and Dorazio, R. M. (2008). *Hierarchical Modelling and Inference in Ecology.* Amsterdam: Academic Press.

Shen, T.-J., and He, F. (2008). An incidence-based richness estimator for quadrats sampled without replacement. *Ecology* **89,** 2052–2060.

Shlosser, A. (1981). On estimation of the size of the dictionary of a long text on the basis of a sample. *Engineering Cybernetics* **19,** 97–102.

Sudman, S. (1976). *Applied Sampling.* New York: Academic Press.

Received November 2010. *Revised December* 2011.
Accepted December 2011.