Biometrics 70, 671–682
September 2014

DOI: 10.1111/biom.12200

An Improved Nonparametric Lower Bound of Species Richness via a Modified Good–Turing Frequency Formula

Chun-Huo Chiu,[1] Yi-Ting Wang,[1] Bruno A. Walther,[2] and Anne Chao[1,*]

[1]Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan
[2]Master Program in Global Health and Development, College of Public Health and Nutrition, Taipei Medical University, 250 Wu-Hsing St., Taipei 110, Taiwan
email: chao@stat.nthu.edu.tw

Summary. It is difficult to accurately estimate species richness if there are many almost undetectable species in a hyperdiverse community. Practically, an accurate lower bound for species richness is preferable to an inaccurate point estimator. The traditional nonparametric lower bound developed by Chao (1984, *Scandinavian Journal of Statistics* **11**, 265–270) for individual-based abundance data uses only the information on the rarest species (the numbers of singletons and doubletons) to estimate the number of undetected species in samples. Applying a modified Good–Turing frequency formula, we derive an approximate formula for the first-order bias of this traditional lower bound. The approximate bias is estimated by using additional information (namely, the numbers of tripletons and quadrupletons). This approximate bias can be corrected, and an improved lower bound is thus obtained. The proposed lower bound is nonparametric in the sense that it is universally valid for any species abundance distribution. A similar type of improved lower bound can be derived for incidence data. We test our proposed lower bounds on simulated data sets generated from various species abundance models. Simulation results show that the proposed lower bounds always reduce bias over the traditional lower bounds and improve accuracy (as measured by mean squared error) when the heterogeneity of species abundances is relatively high. We also apply the proposed new lower bounds to real data for illustration and for comparisons with previously developed estimators.

Key words: Abundance data; Good–Turing frequency formula; Incidence data; Species richness.

1. Introduction

Species richness (i.e., the number of species present in a community) is the simplest and most intuitive measure of biodiversity. The estimation of species richness based on incomplete samples has been widely applied not only in macroecological and conservation-related studies, but also in many other disciplines; see Bunge and Fitzpatrick (1993), Colwell and Coddington (1994), Walther and Morand (1998), Walther and Martin (2001), Magurran (2004), Royle and Dorazio (2008), Gotelli and Colwell (2011), and Chao and Chiu (2014) for various applications. In general contexts, "species" can be defined in a broad sense: They may be biological species, bugs in software programs, words in a book, genes or alleles in genetic code, or other discrete entities. In this article, we focus on biological applications, namely the unknown number of species within a community. Although we use biological terminology, our framework applies to all other relevant fields.

In most biological surveys, data can be generally classified into two types: (Individual-based) abundance data and (sampling-unit-based) incidence data. For abundance data, the sampling unit is an "individual" and a sample of individuals is randomly taken from the community. For incidence data, the sampling unit is usually a trap, net, quadrat, plot, or timed survey, and it is these sampling units, not the individual organisms, which are sampled randomly and independently. Because it is not always possible to count individuals within a sampling unit, estimation can be based on a set of sampling units in which only the incidence (detection or nondetection) of each species is recorded.

Due to sampling limitation, there are undetected species in almost every survey. Consequently, the observed species richness in a sample is almost always a negatively biased estimate of the true (or total) species richness (observed species plus undetected species). Since the pioneering work by Fisher, Corbet, and Williams (1943), a wide range of statistical sampling-theory-based methodologies have been proposed for both abundance and incidence data to estimate the true species richness, or, equivalently, the number of undetected species in the samples.

In this article, we focus on a nonparametric approach in the sense that no assumptions are made about the underlying distribution of species abundances. One of the first nonparametric approaches was the jackknife method applied by Burnham and Overton (1978, 1979), which obtained a class of species richness estimators for both abundance and incidence data. Their first- and second-order jackknife estimators have been used in biological research, although Cormack (1989) implied that the jackknife method does not have a theoretical basis for bias reduction of species richness estimation. For comparative purposes, these two estimators are considered in our simulation studies (see Section 5).

Both the parametric and nonparametric approaches have so far suffered from increasing inaccuracies (large bias and/or large variance) as the community becomes more

hyper-diverse, especially if the community contains many almost undetectable species (e.g., Novotny and Basset, 2000). In such cases, the confidence intervals of point estimates of species richness often become very wide. Given this difficulty, the determination of a lower bound for 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.

Based on the intuitive idea that, when using abundance data, rare species yield the most information about undetected species, Chao (1984) used only the frequencies of the rarest species (namely, singletons and doubletons) to construct a nonparametric lower bound estimator of the true species richness. Here "singletons" and "doubletons" refer, respectively, to the number of species that are observed only once and twice in the sample. A similar-type lower bound was later derived by Chao (1987) for incidence data, using the information of those species that are detected in only one or two sampling units. Colwell and Coddington (1994) referred to these two lower bounds as the Chao1 estimator (for abundance data) and Chao2 estimator (for incidence data) in the ecological literature (see later sections for mathematical formulas) as these two lower bounds converge toward the true species richness when the sample size or the number of sampling units is sufficiently large. Since then, they have been used as species richness point estimators and have been applied in various disciplines and featured in several software programs. We will therefore use the terms "lower bound" and "estimator" interchangeably for Chao-type estimators throughout the article. Lanumteang and Böhning (2011) derived an extended Chao1 estimator under the gamma-Poisson model. Their estimator is also included in our comparison (see Section 5).

For abundance data, the Chao1 estimator is nearly unbiased for species richness when the species abundances are homogeneous or the sample size is sufficiently large. However, when the species abundances are highly heterogeneous and the sample sizes are not large enough, the Chao1 estimator becomes negatively biased (Chao and Chiu, 2014) because it was derived as a lower bound, rather than as a point estimator. The Chao2 estimator exhibits similar performance for incidence data. Therefore, in this article, we derive an approximation formula for the bias of the Chao1 and Chao2 estimators based on a generalization of the Good–Turing frequency formula originally developed by Turing and Good (Good, 1953, 2000). We then eliminate the biases of these lower bounds and thus obtain new improved lower bounds along with their variance estimators.

We focus on the new improved lower bound for abundance data; all procedures for incidence data are generally parallel. In Section 2, we briefly review the Chao1 lower bound. Since the derivation of our improved lower bound is based on the Good–Turing frequency formula, we provide necessary background information about the Good–Turing frequency theory and propose a more accurate formula in Section 3. The improved lower bound for abundance data is derived in Section 4.1. The corresponding lower bound for incidence data is summarized in Section 4.2. Section 5 examines the performance of the proposed lower bounds by using abundance data sets simulated from several species abundance models. The performance of the proposed new Chao1 lower bound is also compared with those of the Chao1 lower bound, two jackknife estimators and the estimator proposed by Lanumteang and Böhning (2011). In Section 6, we use real data sets to illustrate the proposed lower bound for abundance data (Section 6.1) and incidence data (Section 6.2). Section 7 provides some concluding remarks and discussion.

2. The Chao1 Lower Bound

Here, we briefly review the traditional lower bound of species richness for abundance data (Chao, 1984). Assume that there are S different species in a community, and the species are indexed by 1, 2,..., S, with S unknown. Suppose n individuals are independently selected with replacement from the community, and their species identities are determined. Let p_i denote the detection probability that the ith species is detected in any randomly observed individual, $i = 1, 2, \ldots, S$, $\sum_{i=1}^{S} p_i = 1$. The detection probability is a combination of species abundance and individual detectability, which in turn is determined by species characteristics such as size, color, vocalizations, and movement patterns. The species sample frequencies $X_1, X_2, \ldots, X_S$ (only those species with $X_i > 0$ are observed in the sample) follow a multinomial distribution with the cell total $n = \sum_{i=1}^{S} X_i$ and the cell probabilities (p_1, $p_2, \ldots, p_S$).

Let f_k (abundance frequency counts), $k = 0, 1, \ldots, n$, be the number of species that were observed or represented exactly k times in the sample. Here, f_0 denotes the number of undetected species in the sample while f_1 denotes the number of "singletons" and f_2 denotes the number of "doubletons." Define $S_{\text{obs}} = \sum_{i=1}^{S} I(X_i > 0) = \sum_{k \geq 1} f_k$ as the number of distinct species observed in the sample, where $I(A)$ is the indicator function, that is, $I(A) = 1$ if the event A occurs, and 0 otherwise.

Chao (1984) proposed a lower bound of species richness via estimating $E(f_0)$ by a nonparametric approach. Under the assumption that (p_1, $p_2, \ldots, p_S$) are fixed unknown parameters, the sample frequency X_i follows a binomial distribution. Thus, we have a general expectation formula:

$$E(f_k) = E\left[\sum_{i=1}^{S} I(X_i = k)\right] = \sum_{i=1}^{S} \binom{n}{k} p_i^k (1-p_i)^{n-k}, k = 0, 1, 2, ..., n. \quad (1a)$$

In Web Appendix A, we show that with slight modifications all the derivations and estimators below are also valid under the binomial-mixture and Poisson-mixture models. For simplicity, our presentation throughout the article is focused on the fixed-parameter formula (1a) under a multinomial model for the species sample frequencies (X_1, $X_2, \ldots, X_S$) as described earlier.

Based on (1a), the Cauchy-Schwarz inequality yields

$$\left[\sum_{i=1}^{S} (1-p_i)^n\right]\left[\sum_{i=1}^{S} p_i^2 (1-p_i)^{n-2}\right] \geq \left[\sum_{i=1}^{S} p_i (1-p_i)^{n-1}\right]^2. \quad (1b)$$

Thus, a theoretical lower bound for $E(f_0)$ is derived as

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

The following lower bound for species richness S is thus provided:

$$S = E(S_{\text{obs}}) + E(f_0) \geq E(S_{\text{obs}}) + \frac{n-1}{n}\frac{[E(f_1)]^2}{2E(f_2)}. \quad (2a)$$

The Chao1 lower bound or estimator of species richness is obtained by replacing the expected values in (2a) with the observed data:

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

When $f_2 = 0$, a bias-corrected estimator when species abundances are homogeneous was suggested (e.g., Chao and Shen 2010, p. 15):

$$\hat{S}^*_{\text{Chao1}} = S_{\text{obs}} + \frac{(n-1)}{n}\frac{f_1(f_1-1)}{2(f_2+1)}. \quad (2c)$$

The equality in (2a) holds when species abundances are homogeneous, that is, $p_1 = p_2 = \cdots = p_S$. Thus, the Chao1 estimator is nearly unbiased for species richness in the homogeneous case.

Applying the standard asymptotic approach (Chao, 1987), the following estimated variance estimators can be obtained:

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

$$\hat{var}(\hat{S}^*_{\text{Chao1}}) = \frac{1}{4}\frac{(n-1)^2}{n^2}f_1(2f_1-1)^2 + \frac{1}{2}f_1(f_1-1) - \frac{1}{4}\frac{f_1^4}{\hat{S}^*_{\text{Chao1}}}. \quad (3b)$$

When sample size n is sufficiently large and $\hat{S}_{\text{Chao1}} > S_{\text{obs}}$, the associated 95% confidence interval of species richness based on the Chao1 estimator is constructed using a log-transformation (Chao, 1987) because the distributions of $\hat{S}_{\text{Chao1}}$ (and $\hat{S}_{\text{Chao1}} - S_{\text{obs}}$) are generally skewed to the right. Treating $\log(\hat{S}_{\text{Chao1}} - S_{\text{obs}})$ as an approximately normal random variable, a 95% interval of S is obtained as

$$\left[S_{obs} + (\hat{S}_{Chao1} - S_{obs})/R, \quad S_{\text{obs}} + (\hat{S}_{\text{Chao1}} - S_{\text{obs}})R\right], \quad (4)$$

where $R = \exp\{1.96[1 + \hat{var}(\hat{S}_{\text{Chao1}})/(\hat{S}_{\text{Chao1}} - S_{\text{obs}})^2]^{1/2}\}$. In this case, the resulting lower confidence limit is always greater than or equal to the observed species richness (see Web Appendix B for a sketch of the derivations of equations (3a, 3b, and 4)). A reviewer suggested that an alternative approach would treat $\log(\hat{S}_{\text{Chao1}})$as an approximately normal random variable. A simulation was carried out to compare our and this alternative approach; see Web Appendix B for details.

When the sample size is not sufficiently large so that the community is under-sampled, the Chao1 estimator may have a large negative bias. In this case, it is more sensible to infer a minimum value of species richness by applying a one-sided 95% confidence interval $[S_{\text{obs}} + (\hat{S}_{\text{Chao1}} - S_{\text{obs}})/R^*, \infty)$, where $R^* = \exp\{1.64[1 + \hat{var}(\hat{S}_{\text{Chao1}})/(\hat{S}_{\text{Chao1}} - S_{\text{obs}})^2]^{1/2}\}$. See Section 6.1 for an example.

3. A More Accurate Modified Good-Turing Frequency Formula

The Good–Turing frequency formula was originally developed during the World War II cryptographic analyses by the founder of modern computer science, Turing and Good. Turing never published this theory, but permitted Good to publish it; see Good (1953), Good and Toulmin (1956), and Good (2000). The Good–Turing frequency theory can be formulated as follows: For those species that appeared r times in a sample of size n within a multinomial sample, how can one estimate the true mean relative abundance of those species? Good and Turing focused on the case of small r, that is, rare species (or rare code elements, in Turing's case). For example, Turing recognized that singletons (those species with frequency one) have a mean relative abundance that is not $1/n$, but instead the frequency formula as summarized below.

Turing and Good discovered a surprisingly simple and remarkably effective, although nonintuitive answer. Given data, define $\alpha_r = \sum_{i=1}^{S} p_i I(X_i = r)/f_r$, $r = 0, 1, \ldots$ as the true mean relative abundance of those species that appeared r times in a sample of size n. The Good–Turing frequency formula then states that α_r, $r = 1, 2, \ldots$ is not estimated by its sample frequency r/n, but rather by

$$\tilde{\alpha}_r = \frac{(r+1)}{n}\frac{f_{r+1}}{f_r} \equiv \frac{r^*}{n}, \quad r = 1, 2\ldots \quad (5a)$$

where $r^* = (r+1)f_{r+1}/f_r$. For $r = 0$, $\tilde{\alpha}_0$ is not obtainable because f_0 is unknown. However, the product of $\tilde{\alpha}_0$ and f_0 (i.e., the sum of abundances of all undetected species in the sample) can be well estimated by the proportion of singletons, f_1/n. The Good–Turing formula has been extensively discussed and used in statistics, computer science, linguistics, and many other disciplines (Chao and Jost, 2012). This formula implies that, for those species that appeared as a singleton in a sample, the mean relative abundance should be close to $\tilde{\alpha}_1 = 2f_2/(nf_1)$. The Good–Turing frequency formula is thus contrary to most people's intuition because the estimator in (5a) depends not only on the sample frequency r of the focal species, but also on the frequency information derived from the other species. Good (1953) used a Bayesian approach to obtain (5a) whereas Robbins (1968) derived it as an empirical Bayes estimator.

Below we provide a more direct derivation of (5a) in order to present our extension. Considering the expected total probabilities of those species with frequency r in the sample, the following approximation for small $r = 0, 1, \ldots$ can be derived

from (1a):

$$E\left[\sum_{i=1}^{S} p_i I(X_i = r)\right] = \sum_{i=1}^{S} \binom{n}{r} p_i^{r+1}(1-p_i)^{n-r} \approx \sum_{i=1}^{S} \binom{n}{r} p_i^{r+1}(1-p_i)^{n-(r+1)} = \frac{r+1}{n-r} E(f_{r+1}) \approx \frac{r+1}{n} E(f_{r+1}). \tag{5b}$$

If all the expected values in (5b) are approximated by their observed data, then the definition of α_r implies $(r+1)f_{r+1}/n \approx \alpha_r f_r$, leading to the Good–Turing frequency formula (5a). Chao and Jost (2012) advocated that one notion of sample completeness can be objectively measured by the sample coverage (Good, 1953; Good and Toulmin, 1956), which is defined as $C = \sum_{i-1}^{S} p_i I(X_i > 0) = 1 - \sum_{i-1}^{s} p_i I(X_i = 0) = 1 - \alpha_0 f_0$. Based on the Good–Turing formula, the sample coverage estimator thus becomes $1 - \tilde{\alpha}_0 f_0 = 1 - f_1/n$ (see Chao and Jost, 2012, for a brief review).

Instead of considering the total probabilities of those species with frequency r in the sample in (5b), we evaluate the total odds as follows:

$$E\left[\sum_{i=1}^{S} \frac{p_i}{1-p_i} I(X_i = r)\right] = \sum_{i=1}^{S} \frac{p_i}{1-p_i} \binom{n}{r} p_i^{r}(1-p_i)^{n-r} = \sum_{i=1}^{S} \binom{n}{r} p_i^{r+1}(1-p_i)^{n-(r+1)} = \frac{r+1}{n-r} E(f_{r+1}). \tag{6}$$

An advantage of our approach is that no approximation needs to be used in the above derivation, and thus a more accurate formula can be obtained. Combining (6) and the approximation

$$E\left[\sum_{i=1}^{S} \frac{p_i}{1-p_i} I(X_i = r)\right] \approx \frac{\alpha_r}{1-\alpha_r} E(f_r),$$

we obtain our proposed estimator of α_r:

$$\hat{\alpha}_r = \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}}, r = 1, 2, \ldots \tag{7a}$$

For $r = 0$, the above formula is in terms of f_0 and thus is not obtainable. However, if we replace f_0 with the estimator $\hat{f}_0$ from the Chao1 estimator given in (2b) and (2c), then a more accurate estimator of the sample coverage is measured by $1 - \hat{\alpha}_0 \hat{f}_0$, which can be expressed by this estimator published in Chao and Jost (2012):

$$\hat{C} = \begin{cases} 1 - \dfrac{f_1}{n}\left[\dfrac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right], \text{if } f_2 > 0; \\ 1 - \dfrac{f_1}{n}\left[\dfrac{(n-1)(f_1-1)}{(n-1)(f_1-1)+2}\right], \text{if } f_2 = 0. \end{cases} \tag{7b}$$

4. New Improved Lower Bounds

4.1. *A New Improved Lower Bound for Abundance Data*

Our approach is to evaluate the magnitude of the bias associated with the Chao1 lower bound. The magnitude of the first-order bias from (2a) is:

$$\left|\text{bias}(\hat{S}_{\text{Chao1}})\right| = E(f_0) - \frac{(n-1)}{n}\frac{[E(f_1)]^2}{2E(f_2)} = \frac{E(f_0)\{2E(f_2)/[n(n-1)]\} - [E(f_1)/n]^2}{2E(f_2)/[n(n-1)]}. \tag{8a}$$

Using (1a) and the definition of α_r in the Good–Turing frequency formula, we can then separately approximate each term in the numerator of (8a) as follows:

$$E(f_0) = \sum_{i=1}^{S} \frac{(1-p_i)}{p_i}\frac{1}{n} E[I(X_i = 1)] \approx \frac{1-\alpha_1}{\alpha_1} \sum_{i=1}^{S} p_i(1-p_i)^{n-1}.$$

$$\frac{2E(f_2)}{n(n-1)} = \sum_{i=1}^{S} \frac{(1-p_i)}{p_i} \binom{n}{3}^{-1} E[I(X_i = 3)] \approx \frac{1-\alpha_3}{\alpha_3} \sum_{i=1}^{S} p_i^3(1-p_i)^{n-3}.$$

$$\frac{E(f_1)}{n} = \sum_{i=1}^{S} \left(\frac{1-p_i}{p_i}\right)^2 \binom{n}{3}^{-1} E[I(X_i = 3)] \approx \left(\frac{1-\alpha_3}{\alpha_3}\right)^2 \sum_{i=1}^{S} p_i^3(1-p_i)^{n-3}.$$

Using the above three approximations, the numerator of (8a) becomes

$$E(f_0)\left\{\frac{2E(f_2)}{n(n-1)}\right\} - \left[\frac{E(f_1)}{n}\right]^2 \approx \left[\frac{1-\alpha_1}{\alpha_1}\frac{1-\alpha_3}{\alpha_3} - \left(\frac{1-\alpha_3}{\alpha_3}\right)^2\right] \times \left[\sum_{i=1}^{S} p_i(1-p_i)^{n-1}\right] \times \left[\sum_{i=1}^{S} p_i^3(1-p_i)^{n-3}\right]. \tag{8b}$$

For the last two terms in the above formula, the Cauchy–Schwarz inequality yields:

$$\left[\sum_{i=1}^{S} p_i(1-p_i)^{n-1}\right]\left[\sum_{i=1}^{S} p_i^3(1-p_i)^{n-3}\right] \geq \left[\sum_{i=1}^{S} p_i^2(1-p_i)^{n-2}\right]^2. \tag{8c}$$

From (8b), (8c), and (1a), the magnitude of the bias of the Chao1 lower bound is approximately equal to

$$\left|\text{bias}(\hat{S}_{\text{Chao1}})\right| \approx \frac{1-\alpha_3}{\alpha_3}\left[\frac{1-\alpha_1}{\alpha_1} - \frac{1-\alpha_3}{\alpha_3}\right]\frac{2E(f_2)}{n(n-1)}. \tag{9}$$

The right hand side of the above formula is positive because species that are observed three times in a sample should have a larger mean abundance than that of singletons (i.e., α_3 is larger than α_1). Applying the Good–Turing estimates in (7a) for α_3 and α_1 in (9), we then obtain a lower bound of species richness:

$$\hat{S}_{\text{Chao1}} + \frac{(n-3)}{4n}\frac{f_3}{f_4} \times \max\left(f_1 - \frac{(n-3)}{2(n-1)}\frac{f_2 f_3}{f_4}, 0\right). \tag{10a}$$

When n is large enough, we can omit the two terms $(n-3)/n$ and $(n-3)/(n-1)$ in (10a) and thus obtain the improved lower bound. From hereupon, we refer to it as the iChao1 (here i refers to "improved") lower bound or estimator:

$$\hat{S}_{\text{iChao1}} = \hat{S}_{\text{Chao1}} + \frac{f_3}{4f_4} \times \max\left(f_1 - \frac{f_2 f_3}{2f_4}, 0\right). \tag{10b}$$

If $f_4 = 0$, we suggest replacing f_4 with f_4+1 in (10b) so that the iChao1 lower bound is always obtainable. The standard asymptotic method can be applied to derive a variance estimator for any species richness estimator $\hat{S}$, which is a function of frequency counts $(f_1, f_2, \ldots, f_n)$ (such as the Chao1 or the iChao1 estimator). Using this approach, we derive a variance estimator:

$$\hat{var}(\hat{S}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial \hat{S}}{\partial f_i}\frac{\partial \hat{S}}{\partial f_j}\hat{cov}(f_i, f_j), \tag{11}$$

where

$$\hat{cov}(f_i, f_j) = \begin{cases} f_i(1 - f_i/\hat{S}), & \text{if } i = j; \\ -f_i f_j/\hat{S}, & \text{if } i \neq j. \end{cases}$$

The performance of this variance estimator is examined in Section 5. The associated confidence intervals of species richness based on the iChao1 estimator can be similarly constructed as the one given in (4), Section 2. The iChao1 estimator uses information of the first four frequency counts (f_1, f_2, f_3, and f_4) to estimate the number of undetected species in the samples; it is thus unavoidable that it has a larger associated variation than the Chao1 estimator, which only uses (f_1, f_2) to estimate the number of undetected species.

From this derivation, we can derive the following justifications and properties for the use of the iChao1 bound for estimating species richness. (i) Like the Chao1 estimator, the iChao1 estimator is an approximate lower bound for any sample size; the iChao1 estimator is a greater lower bound because it is always greater than or equal to the Chao1 estimator. (ii) The iChao1 estimator becomes asymptotically unbiased when the sample size n is sufficiently large. Consequently, the use of a two-sided confidence interval is justified in this case. (iii) In the homogeneous case, we can apply (1a) and prove that the expected value of $f_1 - f_2 f_3/(2f_4)$ is approximately equal to 0. Here, the approximation is satisfactory in the following sense: The ratio of the error over the expected value of the number of undetected species tends to zero. In this case, the iChao1 estimator reduces to the original Chao1 estimator, implying that the iChao1 estimator is also a nearly unbiased estimator of species richness in the homogeneous case.

4.2. *A New Improved Lower Bound for Incidence Data*

Assume that there are T sampling units, and that they are indexed 1, 2,…, T. The detection or nondetection of each species within each sampling unit is recorded to form a species-by-sampling-unit incidence matrix (Z_{ij}) with S rows and T columns. Here, $Z_{ij} = 1$ if species i is detected in sampling unit j, and $Z_{ij} = 0$ otherwise. Let Z_{i+} be the number of sampling units in which species i is detected, $Z_{i+} = \sum_{j=1}^{T} Z_{ij}$ whereby Z_{i+} is analogous to X_i in the abundance data. Species present in the community but not detected in any sampling unit yield $Z_{i+} = 0$. Let S_{obs} be the total number of species observed in the T sampling units, that is, only species with $Z_{i+} > 0$ contribute to S_{obs}.

Our model is based on that Z_{i+}, $i = 1, 2, \ldots, S$ follows a binomial distribution with the total number T and the detection probability θ_i, which is defined as the chance of encountering at least one individual of the ith species in any sample. Denote the sample incidence frequency counts by $(Q_1, Q_2, \ldots, Q_T)$, where Q_k is the number of species that are detected in exactly k sampling units in the data, $k = 1, 2, \ldots, T$. Here, Q_k is analogous to f_k in the abundance data. Hence, Q_1 represents the number of "unique" species (those that are detected in only one sampling unit), and Q_2 represents the number of "duplicate" species (those that are detected in only two sampling units). It follows from the distribution of Z_{i+} that

$$E(Q_k) = \sum_{i=1}^{S} P(Z_{i+} = k) = \sum_{i=1}^{S}\binom{T}{k}\theta_i^k(1-\theta_i)^{T-k}. \tag{12a}$$

Comparing (12a) and (1a), we see that the expectations for frequency counts under abundance data and incidence data are similar. Therefore, all derivations are parallel with n being replaced by T, and the frequencies counts $(f_1, f_2, \ldots, f_n)$ being replaced by $(Q_1, Q_2, \ldots, Q_T)$. For example, the Chao2 lower bound or estimator (Chao, 1987) for incidence data has an analogous form to the Chao1 lower bound:

$$\hat{S}_{\text{Chao2}} = \begin{cases} S_{\text{obs}} + [(T-1)/T]Q_1^2/(2Q_2), & \text{if } Q_2 > 0; \\ S_{\text{obs}} + [(T-1)/T]Q_1(Q_1-1)/2, & \text{if } Q_2 = 0. \end{cases} \tag{12b}$$

The variance formulas are also similar to those given in (3a) and (3b). Using exactly the same derivations as for the abundance data (Section 4.1), we obtain the following improved lower bound (which we refer to as the iChao2 lower bound or estimator from hereupon):

$$\hat{S}_{\text{iChao2}} = \hat{S}_{\text{Chao2}} + \frac{(T-3)}{4T}\frac{Q_3}{Q_4} \times \max\left(Q_1 - \frac{(T-3)}{2(T-1)}\frac{Q_2 Q_3}{Q_4}, 0\right). \quad (13)$$

Unlike for the abundance data, T may be a small number. Therefore, we suggest retaining $(T-3)/T$ and $(T-3)/(T-1)$ in the above formula for the iChao2 lower bound. The variance estimator of the iChao2 estimator and the associated confidence interval can be obtained as those in the abundance data.

5. Simulation Results

To investigate the behavior of the new estimators and to compare them with some previously developed estimators, we performed extensive simulations by generating data sets from various species abundance models. Here, we report the results from six representative models for generating abundance data. In each model, we fixed the number of species at 200. The functional forms for species' relative abundances $(p_1, p_2, \ldots, p_{200})$ or the species abundance distributions are given below, whereby c is a normalizing constant in all cases, such that $\sum_{i=1}^{S} p_i = 1$. When species abundances were simulated from a distribution, we first generated a set of 200 random variables, which we regarded as fixed parameters in the simulation. In each case, we also give the CV of the generated set (which is the ratio of the standard deviation over the mean) of $(p_1, p_2, \ldots, p_{200})$. The CV value quantifies the degree of heterogeneity of the probabilities $(p_1, p_2, \ldots, p_S)$. When all probabilities are equal, CV = 0. A larger value of CV indicates a higher degree of heterogeneity among probabilities.

Model 1 (a homogeneous model) with $p_i = 1/S$ and $S = 200$. This is the model with no heterogeneity among species relative abundances (CV = 0).

Model 2 (negative binomial model) with parameter $k = 4$, $r = 0.04$, and $p_i = ca_i$, where $(a_1, a_2, \ldots, a_{200})$ is a random sample from a negative binomial (4, 0.04) distribution with a density function $f(a) = \{(a-1)!/[(k-1)!(a-k)!]\}(1-r)^{a-k}r^k$, $a \geq k$. This is the classical species abundance distribution first used by Fisher et al. (1943) (CV = 0.49).

Model 3 (broken-stick model, MacArthur, 1957) with $p_i = ca_i$, where $(a_1, a_2, \ldots, a_{200})$ is a random sample from an exponential distribution. Or, equivalently, $(p_1, p_2, \ldots, p_{200})$ follows a Dirichlet distribution with parameter 1 (CV = 0.96).

Model 4 (log-normal model) with parameters $\mu = 0$, $\sigma^2 = 1$, and $p_i = ca_i$, where $(a_1, a_2, \ldots, a_{200})$ is a random sample from a log-normal (0, 1) distribution (CV = 1.82).

Model 5 (Zipf–Mandelbrot model) with $p_i = c/(i - 0.1)$, $i = 1, 2, \ldots, 200$. This is a commonly used model in literature and linguistics (Magurran, 2004) (CV = 3.08).

Model 6 (power decay model) with $p_i = c/i^{1.2}$, $i = 1, \ldots, 200$ (CV = 4.20).

The CV values of the above six models range from 0 to 4.20 and thus cover most practical cases in real applications. For each fixed model, we considered a range of sample sizes ($n = 100$–800 in an increment of 100). For each combination of abundance model and sample size, 1000 simulated data sets were generated from the abundance model. For each generated data set, we then computed the observed species richness (S_{obs}) as well as the following nonparametric estimators and their estimated standard errors (s.e.) based on a standard asymptotic method; see (11):

(i) The first- and second-order jackknife estimators: The first-order jackknife estimator ($S_{\text{obs}} + f_1$) uses the number of singletons to estimate the number of undetected species. The second-order jackknife estimator ($S_{\text{obs}} + 2f_1 - f_2$) uses both singletons and doubletons to estimate the number of undetected species.
(ii) A new estimator proposed by Lanumteang and Böhning (2011), which we refer to as the LB estimator from hereupon.
(iii) The Chao1 estimator; see (2b, 2c).
(iv) The proposed iChao1 estimator; see (10b).

In Tables 1 and 2, we show the results for three sample sizes ($n = 200$, 400, and 800) for Model 2 (negative-binomial model) and Model 6 (power-decay model), respectively. The simulation results for the other four models are given in Web Appendix C.

For each estimator, the estimates and their estimated s.e. were averaged over 1000 simulated data sets to give the "average estimate" and the "average estimated s.e." (columns 3 and 5 in Tables 1 and 2). The sample s.e. and the root sample mean squared error (RMSE, see Walther and Moore, 2005) over the 1000 estimates were obtained to give "sample s.e." and "sample RMSE" (columns 4 and 6 in Tables 1 and 2). The percentage of data sets in which the 95% confidence intervals cover the true value is shown in column 7. The average of the number of observed species is also listed in the tables. In Figure 1, we specifically plot the average estimates of our five estimators as a function of sample size (from 100 to 800) so that we can examine and compare the bias behavior as a function of sample size.

A good species richness estimator should have a small magnitude of bias and a high accuracy (i.e., low RMSE; see Walther and Moore, 2005). Furthermore, the coverage probability of its associated confidence interval should be close to the nominal level 95%. As sample size increases, the estimator should also exhibit the following intuitive pattern: Its bias, accuracy (as measured by RMSE) and coverage probability of the confidence interval should generally improve as sample size increases, and its estimates should thus increasingly approach the true species richness. Using these criteria, we generated the following general results which are given in Figure 1, Tables 1 and 2, Web Tables C1–C4, and other unreported simulation results.

As expected, the traditional approach of using the observed richness in a sample as an estimator of species richness seriously underestimates in all cases (Figure 1 and all tables), especially at small sample sizes, thus reiterating prior findings (e.g., Colwell and Coddington, 1994; Walther and Morand, 1998; Walther and Martin, 2001; Walther and Moore, 2005).

Table 1

Comparison of five species richness estimators based on 1000 simulation trials under a negative binomial (4, 0.04) model, with $S = 200$ and CV $= 0.49$. The five estimators are: The iChao1 estimator as given in (10b), the Chao1 estimator as given in (2b, 2c), Jackknife1, Jackknife2 = the first- and second-order jackknife estimators (Burnham and Overton, 1979), and the LB estimator (Lanumteang and Böhning, 2011); see text for details.

Size n (species seen)	Estimator	Average estimate	Sample sample s.e.	Average estimated s.e.	Sample RMSE	95% CI coverage
200 (117.29)	Jackknife1	182.03	10.35	11.36	20.73	0.76
	Jackknife2	214.40	18.35	19.68	23.31[b]	0.85
	LB	225.03	90.54	70.93	93.85	0.88
	Chao1	184.88	22.32	22.23	26.94	0.90
	iChao1	194.94[a]	28.14	28.44	28.57	0.93[c]
400 (159.30)	Jackknife1	212.10	9.38	10.26	15.30	0.71
	Jackknife2	221.96	16.56	17.77	27.49	0.62
	LB	209.31	35.22	30.86	36.40	0.91
	Chao1	193.05	12.22	12.08	14.04[b]	0.92
	iChao1	198.91[a]	15.19	15.36	15.22	0.94[c]
800 (186.94)	Jackknife1	213.33	6.24	7.24	14.71	0.30
	Jackknife2	208.28	12.41	12.54	14.91	0.72
	LB	206.34	17.75	15.62	18.83	0.89
	Chao1	198.83	6.37	6.08	6.47[b]	0.94[c]
	iChao1	200.19[a]	7.69	7.54	7.69	0.93

[a]Denotes the smallest bias. [b]Denotes the smallest RMSE. [c]Closest to 95% coverage.

Table 2

Comparison of five species richness estimators based on 1000 simulation trials under a power decay model $p_i = c/i^{1.2}$, with $S = 200$ and CV $= 4.20$. See Table 1 for the abbreviations of the estimators.

Size n (species seen)	estimator	Average estimate	Sample s.e.	Average Estimated s.e.	Sample RMSE	95% CI coverage
200 (59.84)	Jackknife1	95.55	9.19	8.45	104.85	0.00
	Jackknife2	121.64	14.60	14.64	79.71	0.01
	LB	501.05	2067.20	683.99	2087.98	0.79
	Chao1	135.06	42.64	37.75	77.68	0.63
	iChao1	147.03[a]	47.88	43.57	71.39[b]	0.80[c]
400 (88.09)	Jackknife1	135.39	10.81	9.66	65.51	0.00
	Jackknife2	165.79	17.09	16.73	38.24[b]	0.60
	LB	289.81	274.68	200.07	288.86	0.83
	Chao1	160.87	30.72	29.46	49.74	0.74
	iChao1	172.79[a]	34.98	34.74	44.31	0.88[c]
800 (123.63)	Jackknife1	175.72	10.76	10.25	26.56	0.49
	Jackknife2	203.23[a]	17.60	17.75	17.89[b]	0.93
	LB	243.25	125.61	94.53	132.79	0.86
	Chao1	181.48	22.40	21.44	29.06	0.87
	iChao1	194.70	26.31	26.09	26.62	0.94[c]

[a]Denotes the smallest bias. [b]Denotes the smallest RMSE. [c]Closest to 95% coverage.

The two jackknife estimators typically underestimate when the sample size is relatively small, but then exceed the true species richness and overestimate at larger sample sizes. For example, in the negative binomial model (Figure 1b and Table 1), the first-order jackknife estimator for $n < 200$ has negative bias, crosses the true parameter line around $n = 300$ whereabouts it appears nearly unbiased; however, for $n > 400$, it becomes appreciably positively biased. The second-order jackknife exhibits a similar behavior with an earlier crossing point at around $n = 130$. Similar patterns exist for the other models. In some cases, their crossing points are not shown because they exceed the maximum sample sizes displayed in our

figures. It is thus clear that, for each model, there is a limited range of sample sizes (near crossing points) where jackknife estimators are close to the true species richness. This is the likely reason why many studies (e.g., Palmer, 1991; Chiarucci et al., 2003; Walther and Moore, 2005; Xu et al., 2012) found a relatively good performance of the jackknife estimators. However, this narrow range of good performance changes with each model and is therefore not predictable. Outside this range,

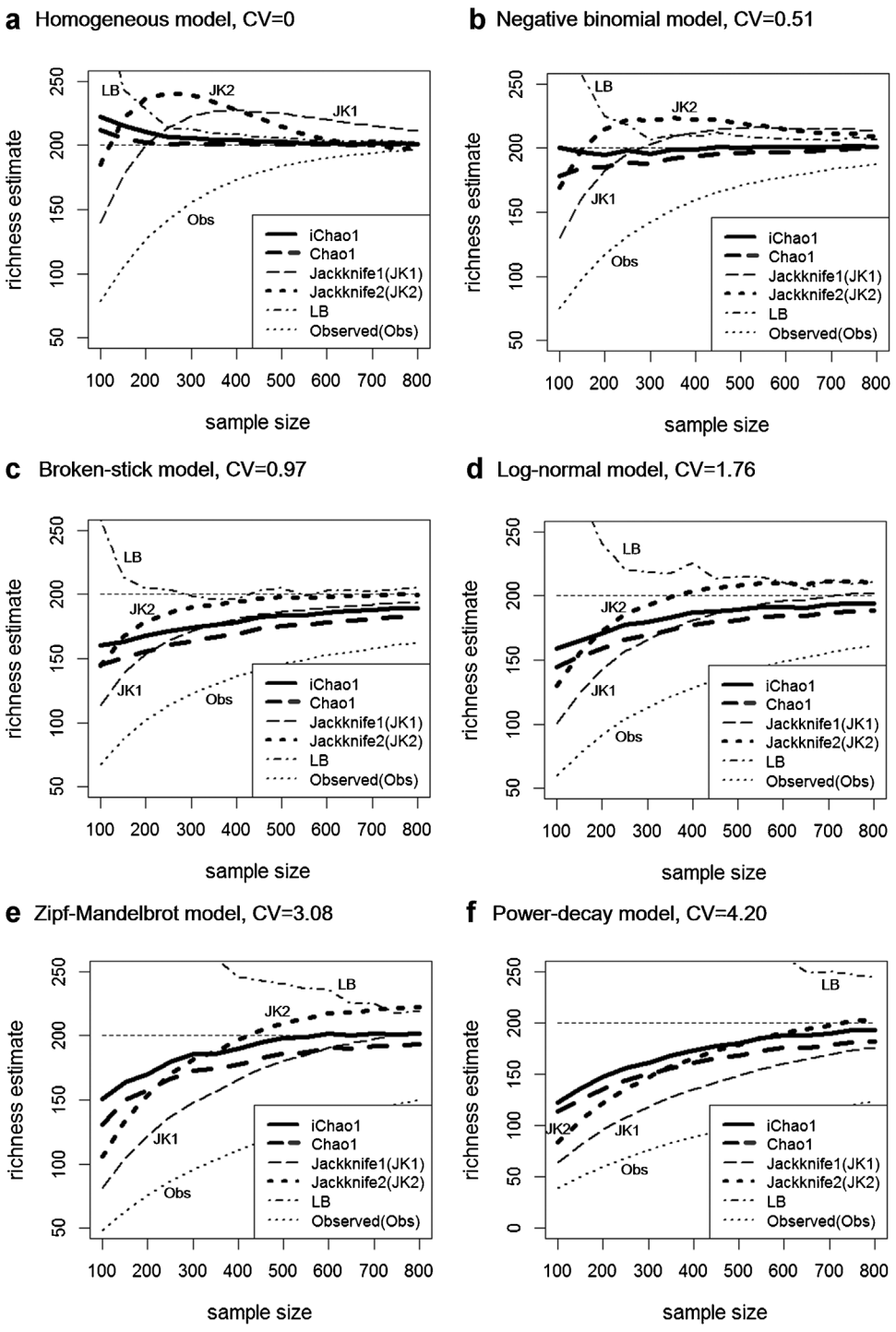

Figure 1. Comparison of the average biases for five species richness estimators and the number of observed species richness in samples when sample size ranges between 100 and 800 under six different models as indicated from panel a to panel f. The biases were obtained by averaging over 1000 data sets generated from six different species abundance models (see text for details). The true parameter is $S = 200$. A good estimator should closely approach the $S = 200$ line (the horizontal dotted line in each panel). See Table 1 for the abbreviations of the estimators.

the two jackknife estimators have appreciable biases (Figure 1 and all tables) and exhibit counter-intuitive patterns: Their bias, accuracy and coverage probability regularly do not improve as sample size increases (e.g., bias and coverage probability in Table 1). Although the jackknife estimators have the smallest RMSE in some cases, their coverage probabilities are significantly lower than the nominal level. Burnham and Overton (1978, 1979) proposed a testing procedure to obtain an interpolated jackknife estimator up to the fifth-order. The general performance of the interpolated jackknife estimator is examined and summarized in Web Appendix C and Web Figure C1.

The LB estimator was derived from a Poisson-gamma model, which includes the broken stick model as a special case. Thus, as expected, the LB estimator performs well for the broken stick model when sample sizes are sufficiently large (Figure 1c and Web Table C2), and it has the smallest bias when applying the broken stick model. However, it is not robust to departures from its base model. With other models, it seriously overestimates, especially when sample sizes are small and the CV is large. In some cases, its variance is extremely large, leading to a large RMSE (Table 2 and Web Tables C3 and C4), which indicates that the LB estimator is only useful when its model assumption is satisfied.

In the homogeneous model (Figure 1a and Web Table C1), the Chao1 and the iChao1 estimators are close to each other (the latter has slightly higher positive bias), and both are nearly unbiased. For heterogeneous models ($CV > 0$), both estimators have negative biases because they are derived as lower bounds. Nevertheless, they follow the intuitive pattern: The magnitude of bias and RMSE decrease as sample size increases. The iChao1 estimator has smaller bias (Figure 1), but larger variance (in all tables) than the Chao1 estimator due to its use of more frequencies to estimate the number of undetected species. In terms of bias, the proposed iChao1 estimator always outperforms the Chao1 estimator (Figure 1). In terms of RMSE and coverage probabilities, the iChao1 estimator is preferable to the Chao1 estimator when CV is relatively large (say, $CV > 1$), as shown in Table 2 and Web Tables C2–C4. When sample sizes are sufficiently large, the coverage probabilities are close to the nominal levels.

The proposed asymptotic variance formulas for both the Chao1 and iChao1 estimators generally work well because the estimated s.e. (column 5 in each table) are close to the corresponding sample s.e. (column 4 in each table) in all cases. Based on the simulations (Tables 1 and 2 and Web Appendix C), the estimated s.e. for the Chao1 estimator slightly underestimates, whereas the estimated s.e. performs better for the iChao1 estimator.

6. Applications

6.1. *Tropical Insect Data*

Janzen (1973a, 1973b) presented data sets of tropical foliage insects from sweep samples in Costa Rica. We selected two beetle data sets to illustrate the performance of our estimators and to compare day-time and night-time beetle species richness determined at the site "Osa-primary-hill, dry season, 1967." The abundance frequency counts of beetles collected during day-time were $(f_1, f_2, f_3, f_4, f_5, f_6, f_{11}) = (59, 9, 3, 2, 2, 2, 1)$, and other frequencies were 0. There were 78 species among 127 individuals. The sample coverage using (7b) is estimated to be 53.7%. The abundance frequency counts during night-time were $(f_1, f_2, f_3, f_5, f_7, f_{10}, f_{14}, f_{16}, f_{18}) = (56, 9, 7, 2, 1, 1, 1, 1, 1)$, and other frequencies were 0. There were 79 species among 170 individuals. The corresponding sample coverage is estimated to be 67.1%. As is evident from these two data sets, most species were observed only once or twice, and there are only a few abundant species. Both communities were therefore clearly under-sampled, as indicated by the low estimated sample coverage, implying there were still many rare species, which were not recorded. Using the CV estimator proposed in Chao and Lee (1992) based on the frequency counts, the estimated CVs for these two data sets are 2.973 and 7.460, respectively. These relatively large CVs indicate that the community was highly heterogeneous in species abundances. Therefore, both the Chao1 and iChao1 estimators may still return severely negatively biased estimates for species richness (as discussed in our simulation studies in Section 5). Nevertheless, we use these data sets to compare various estimates and to demonstrate the use of the proposed iChao1 estimate as a lower bound of species richness. In such under-sampled cases, it is statistically infeasible to obtain accurate point and interval estimates of species richness, especially when the heterogeneity among species abundances is high. At best, we can obtain a more accurate lower bound and lower confidence interval based on the proposed iChao1 lower bound.

Table 3
Comparison of five species richness estimators for the beetle data (Janzen, 1973a, 1973b). The estimated s.e. is shown in parentheses. See Table 1 for the abbreviations of the estimators.

Estimator	Day-time	Night-time
Jackknife1	137.0 (10.9)	135 (10.6)
Jackknife2	187.0 (18.8)	182 (18.3)
LB	464.8 (770.7)	1343 (1436.9)
Chao1	271.4 (82.9)	253.2 (75.7)
iChao1	290.9 (89.1)	296.3 (81.3)

In Table 3, we compare five estimators, namely the Chao1, iChao1, the first-order jackknife, second-order jackknife and the LB estimator. For each estimator, the estimated s.e. computed from (11) is also given. The LB estimate is much higher than the other four estimates and has extremely large s.e. for this data set, suggesting that the true model may deviate greatly from the Poisson-gamma model. The two jackknife estimates are much lower than the Chao1 and iChao1 estimates. Our simulations for highly heterogeneous models and small sample sizes (Table 2, Figure 1e and f, and Web Table C4) show that the two jackknife estimators are generally lower than the Chao1 and iChao1 estimators in highly heterogeneous cases and thus have larger negative biases.

Using the Chao1 and iChao1 estimators, we obtain respective richness estimates of 271.4 (s.e. 82.9) and 290.9 (s.e. 89.1) for the day-time data and of 253.2 (s.e. 75.7) and 296.3 (s.e. 81.3) for the night-time data (Table 3). For each sample, the

Table 4

Comparison of five species richness estimators for the ciliates data (Foissner et al., 2002). The estimated s.e. is shown in parentheses. The four estimators for incidence data are: The iChao2 estimator as given in (13), the Chao2 estimator as given in (12b), Jackknife1, Jackknife2 = the first- and second-order jackknife estimator (Burnham and Overton, 1978) and the LB estimator (Lanumteang and Böhning, 2011).

Estimator	Southern Namibia	Central Namibia	Etosha Pan
Jackknife1	233.33 (12.38)	200.94 (11.23)	352.42 (15.18)
Jackknife2	283.66 (20.58)	238.66 (18.74)	427.08 (25.47)
LB	418.39 (183.66)	281.91 (105.69)	681.10 (249.54)
Chao2	270.26 (34.90)	216.50 (26.07)	402.21 (41.43)
iChao2	290.69 (38.46)	235.55 (33.74)	436.88 (46.83)

iChao1 estimate is higher than the Chao1 estimate, as expected from our theory and simulations when the CV is high. As shown in our simulations, the iChao1 estimate is a greater lower bound than the Chao1 estimate. Therefore, we can conclude for these data based on the iChao1 estimates that the minimum species richness for the day-time data is 291 with a one-sided 95% lower confidence bound of 188, and the corresponding minimum species richness for the night-time data is 297 with a one-sided 95% lower confidence bound of 199; see Section 2 for the formula for constructing a one-sided confidence interval.

6.2. *Ciliates Data*

Foissner, Agatha, and Berger (2002) took 51 soil samples from three areas of Namibia and recorded the detection or nondetection of soil ciliate species in each sample. Detailed sampling locations, procedures, and species identifications were described in Foissner et al. (2002). In short, 331 species were detected within three sampled areas, namely the Southern Namibia Desert (15 samples, 154 species), the Central Namibia Desert (17 samples, 136 species), and the Etosha Pan Desert (19 samples, 234 species). For simplicity, we will refer to these three areas as Southern Namibia, Central Namibia, and Etosha Pan. The first four incidence frequency counts (Q_1, Q_2, Q_3, Q_4) for these three areas were (85, 29, 14, 9), (69, 28, 13, 4), and (125, 44, 26, 14), respectively. The purpose here is to compare and rank the species richness of these three areas.

The estimated sample coverage for these three areas is 76.6%, 82.1%, and 77.6%, respectively, and the corresponding estimated CVs are 3.00, 3.39, and 2.55. These high estimated CVs imply that the three communities are highly heterogeneous in species detection probabilities. Since these data are incidence data, we report in Table 4 various estimates (along with their estimated s.e.) including the Chao2 estimator given in (12b), the iChao2 estimator given in (13), the LB estimate and the two jackknife estimators.

Table 4 shows that the LB estimate for the incidence data is substantially larger than the other four estimates and is also associated with a large variation. The first-order jackknife estimator is lower than the other three estimates, but the second-order jackknife estimate is very close to the proposed iChao2 estimate. For each of the three areas, the iChao2 estimate is consistently higher than the corresponding Chao2 estimate. Nevertheless, all methods show a consistent species richness ranking of the three areas: The species richness of the Etosha Pan was significantly higher than that of Southern Namibia, which was significantly higher than that of Central Namibia. The proposed iChao2 estimates demonstrate that the extent of under-estimation of the observed number of species for the three areas is at least 47%, 42%, and 46%, respectively.

7. Conclusion and Discussion

In highly heterogeneous communities, the traditional Chao1 lower bound for abundance data and the Chao2 lower bound for incidence data unavoidably have large negative bias when sample sizes are not sufficiently large. For both data types, we propose an improved lower bound called iChao1 and iChao2 estimator, respectively. Although our derivations focus on the model that the species abundances (for abundance data) or species detection probabilities (for incidence data) are fixed parameters, all the proposed estimators are also valid under a binomial-mixture model. For abundance data, a classical model is the Poisson-gamma mixture model as presented by Fisher et al. (1943). With a little modification in the model formulation, we can further show that the proposed iChao1 estimator is also valid under a Poisson-gamma model (Wang, 2010); see Web Appendix A for details. A worthwhile future research topic is to extend this work to the mixed power series distributions as discussed in Böhning et al. (2013).

Our simulations show that the iChao1 estimator removes a large portion of the negative bias which was associated with the traditional Chao1 and Chao2 estimators. Furthermore, the new estimators have good accuracy and coverage probability for the associated confidence intervals. These new estimators always reduce bias over the traditional estimators and improve accuracy and confidence interval coverages when the heterogeneity of species abundances is relatively high. These new estimators will be featured in the Program SPADE (Species Prediction And Diversity Estimation, http://chao.stat.nthu.edu.tw/softwareCE.html) following the publication of this article.

The proposed new estimators were derived under the assumption of sampling with replacement, in which individuals (or any other sampling unit) can be repeatedly observed. However, in some surveys, sampling is done without replacement. This type of 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. Chao and Lin (2012) derived Chao1 type and Chao2 type estimators under sampling without replacement. Parallel derivations to those developed in Section 5 lead to the corresponding improved estimators under sampling without replacement. We summarize the results in Web Appendix D.

For abundance data and the Chao1 estimator, only the numbers of singletons and doubletons are used to estimate the number of undetected species. For the iChao1 estimator,

we use additional information (tripletons and quadrupletons). Thus, greater sampling effort is needed to collect this additional information. However, the payback is that we can have a less biased and more accurate nonparametric species richness estimator. Our simulation results revealed that the improvement is warranted, especially for highly heterogeneous communities. Thus, we suggest expending more sampling effort to reach species frequencies of up to the fourth frequency, for both abundance and incidence data.

8. Supplementary Materials

Web Appendices and Web Tables referenced in Sections 2, 5–7 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank an Associate Editor and two reviewers for carefully reading an earlier version and providing insightful and helpful comments and suggestions, which substantially improved the article. We also acknowledge that one reviewer added a very relevant reference. This work was supported by the Taiwan National Science Council under Projects 100-2118-M007-006 (for C.-H.C. and A.C.) and 101-2811-M007-088 (for Y.T.W.). C.-H.C. is supported by a postdoctoral fellowship from National Tsing Hua University, Taiwan. B.A.W. acknowledges financial support from Taipei Medical University's SEED grant.

References

Böhning, D., Baksh, M. F., Lerdsuwansri, R., and Gallagher, J. (2013). The use of the ratio-plot in capture-recapture estimation. *Journal of Computational and Graphical Statistics* **22**, 133–155.

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

Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**, 625–633.

Burnham, K. P. and Overton, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60**, 927–936.

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. and Chiu, C.-H. (2014). Estimation of species richness and shared species richness. To appear as an entry in *Handbook of Methods and Applications of Statistics in the Atmospheric and Earth Sciences*. NY: Wiley.

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

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

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

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

Chiarucci, A., Enright, N. J., Perry, G. L. W., Miller, B. P., and Lamont, B. B. (2003). Performance of nonparametric species richness estimators in a high diversity plant community. *Diversity and Distributions* **9**, 283–295.

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

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45**, 395–413.

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.

Foissner, W., Agatha, S., and Berger, H. (2002). Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha region and the Namib Desert. *Denisia* **5**, 1–1459.

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

Good, I. J. (2000). Turing's anticipation of empirical bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation* **66**, 101–111.

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

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

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

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

Lanumteang, K. and Böhning, D. (2011). An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics and Data Analysis* **55**, 2302–2311.

MacArthur, R. H. (1957). On the relative abundances of bird species. *Proceedings of the National Academy of Science of the United States of America* **43**, 193–295.

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

Novotny, V. and Basset, Y. (2000). Rare species in communities of tropical insect herbivores: Pondering the mystery of singletons. *Oikos* **89**, 564–572.

Palmer, M. W. (1991). Estimating species richness: The second-order jackknife reconsidered. *Ecology* **72**, 1512–1513.

Robbins, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics* **39**, 256–257.

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

Walther, B. A. and Martin, J.-L. (2001). Species richness estimation of bird communities: How to control for sampling effort? *Ibis* **143**, 413–419.

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

Walther, B. A. and Morand, S. (1998). Comparative performance of species richness estimation methods. *Parasitology* **116**, 395–405.

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

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

Received May 2013. *Revised April* 2014. *Accepted April* 2014.