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 Chao1,*

1Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan
2Master 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 hyper-diverse 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. Lamuntheang 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 Lamuntheang 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 $1, 2, \ldots, S$, 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 $i$th 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 $f_i$ (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, $f_i$ 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 $S_{\text{obs}} = \sum_{i=1}^{S} f_i$ 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 $E(f_1)$ 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:

$$E(f_k) = E\left[\sum_{i=1}^{S} I(X_i = k)\right]$$

$$= \sum_{i=1}^{S} \binom{n}{k} p_i^k (1 - p_i)^{n-k}, k = 0, 1, 2, \ldots, n. \quad (1a)$$

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

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

(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:

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

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

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

The equality in $(2a)$ holds when species abundances are homogeneous, that is, $p_1 = p_2 = \ldots = p_r$. 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:

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

$$\text{var}(\hat{S}_{\text{Chao1}}^*) = \frac{1}{4} \frac{(n - 1)^2}{n^2} \frac{f_1(2f_1 - 1)^2}{f_2} + \frac{1}{2} f_1(f_1 - 1) - \frac{1}{4} \frac{f_1^4}{\hat{S}_{\text{Chao1}}}.$$  

When sample size $n$ is sufficiently large and $S_{\text{Chao1}} > S_{\text{obs}}$, 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

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

where $R = \exp(1.96[1 + \text{var}(\hat{S}_{\text{Chao1}})/(\hat{S}_{\text{Chao1}} - S_{\text{obs}})^2]^{1/2})$. 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 $R^* = \exp(1.64[1 + \text{var}(\hat{S}_{\text{Chao1}})/(\hat{S}_{\text{Chao1}} - S_{\text{obs}})^2]^{1/2})$. 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 $\alpha_r = \sum_1^n \frac{n!}{r!(n - r)!}/f_r$, $r = 0, 1, \ldots$, 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 $\alpha_r$, $r=1,2,\ldots$, is not estimated by its sample frequency $r/n$, but rather by

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

(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):

\[
E \left[ \sum_{i=1}^{s} p_i I(X_i = r) \right] = \sum_{i=1}^{s} \left( \frac{n}{r} \right) p_i \gamma_i (1 - p_i)^{n-r} \\
= \sum_{i=1}^{s} \left( \frac{n}{r} \right) p_i^{r+1} (1 - p_i)^{n-(r+1)} \\
= \frac{r+1}{n-r} E(f_{r+1}) \approx \frac{r+1}{n} E(f_{r+1}).
\]

(5b)

If all the expected values in (5b) are approximated by their observed data, then the definition of \( \alpha_r \) 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 \(1 - \hat{\alpha}_0 f_0 = 1 - f_1/n\) (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:

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

(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 \( \alpha_r \):

\[
\hat{\alpha}_r = \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}}, \quad r = 1, 2, \ldots
\]

(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):

\[
\hat{C} = \begin{cases} 
1 - \frac{f_1}{f_2} \left( \frac{n-1}{n-1} f_1 + 2 f_{f_2} \right), & \text{if } f_2 > 0; \\
1 - \frac{f_1}{f_2} \left( \frac{n-1}{n-1} f_1 - 1 \right), & \text{if } f_2 = 0.
\end{cases}
\]

(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:

\[
|\text{bias}(\hat{S}_{\text{Chao}1})| = \frac{E(f_0) - (n-1)E(f_1)^2}{n} \approx \frac{2E(f_2)}{n(n-1)} - \frac{E(f_1)^2}{2E(f_2)/n(n-1)}.
\]

(8a)

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

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

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

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

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

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

(8b)
For the last two terms in the above formula, the Cauchy–Schwarz inequality yields:

\[
\left[ \sum_{i=1}^{S} p_i (1 - p_i)^{\alpha - 1} \right] \left[ \sum_{i=1}^{S} p_i^2 (1 - p_i)^{\alpha - 3} \right] \geq \left( \sum_{i=1}^{S} p_i^\alpha (1 - p_i)^{\alpha - 2} \right)^2 .
\]

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

\[
|\text{bias}(\hat{S}_{\text{Chao1}})| \approx \frac{1 - \alpha_3}{\alpha_3} \left[ \frac{1 - \alpha_4}{\alpha_4} - \frac{1 - \alpha_3}{\alpha_3} \right] \frac{2E(f_2)}{n(n - 1)}. \tag{9}
\]

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., \(\alpha_4\) is larger than \(\alpha_3\)). Applying the Good–Turing estimates in (7a) for \(\alpha_3\) and \(\alpha_4\) in (9), we then obtain a lower bound of species richness:

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

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:

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

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 \((f_1, f_2, \ldots, f_T)\) (such as the Chao1 or the iChao1 estimator). Using this approach, we derive a variance estimator:

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

where

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

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 \((f_1, f_2, f_3, \text{ and } f_4)\) 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 \((Z_i)\) 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 \(S_{\text{obs}}\) be the total number of species observed in the \(T\) sampling units, that is, only species with \(Z_i > 0\) contribute to \(S_{\text{obs}}\).

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 \(\theta_i\), which is defined as the chance of encountering at least one individual of the \(i\)th 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 (4a), 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 \((f_1, f_2, \ldots, f_T)\) 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:

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

\[ S_{\text{iChao2}} = S_{\text{Chao2}} + \frac{(T-3)}{4T} \frac{Q_3}{Q_4} \times \max \left( Q_1 - \frac{(T-3)}{2(T-1)} \frac{Q_2 Q_3}{Q_4}, 0 \right). \]

(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 \( \sum_{i=1}^{200} p_i = 1 \). 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 \((p_1, p_2, \ldots, p_5)\). 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) with parameter \( k = 4 \), \( r = 0.04 \), and \( p_i = c a_i \), 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 \( p_i = c a_i \), 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 \( p_i = c a_i \), 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 \((CV = 3.08)\).

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 \((S_{\text{obs}})\) 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 \((n = 200, 400, \text{and} 800)\) 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.” \((\text{columns 3 and 5 in Tables 1 and 2)}\). 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” \((\text{columns 4 and 6 in Tables 1 and 2).}\) 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 \((i.e., \text{low RMSE}; \text{see Walther and Moore, 2005).}\) 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 \((e.g., \text{Colwell and Coddington, 1994; Walther and Morand, 1998; Walther and Martin, 2001; Walther and Moore, 2005).}\)
An Improved Nonparametric Lower Bound of Species Richness

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 \( \text{iChao1} \) 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.

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

\( a \) 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.

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

\( a \) 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
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, 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.

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 \((f_1, f_2, f_3, f_4, f_5, f_6, f_{11}) = (56, 9, 7, 2, 1, 1, 1, 1)\), 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.

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 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.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Day-time</th>
<th>Night-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jackknife1</td>
<td>137.0 (10.9)</td>
<td>135 (10.6)</td>
</tr>
<tr>
<td>Jackknife2</td>
<td>187.0 (18.8)</td>
<td>182 (18.3)</td>
</tr>
<tr>
<td>LB</td>
<td>464.8 (770.7)</td>
<td>1343 (1436.9)</td>
</tr>
<tr>
<td>Chao1</td>
<td>271.4 (82.9)</td>
<td>253.2 (75.7)</td>
</tr>
<tr>
<td>iChao1</td>
<td>290.9 (89.1)</td>
<td>296.3 (81.3)</td>
</tr>
</tbody>
</table>
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 (Lamunthe and Böhm, 2011).

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Southern Namibia</th>
<th>Central Namibia</th>
<th>Etosha Pan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jackknife1</td>
<td>233.33 (12.38)</td>
<td>200.94 (11.23)</td>
<td>352.42 (15.18)</td>
</tr>
<tr>
<td>Jackknife2</td>
<td>283.66 (20.58)</td>
<td>238.66 (18.74)</td>
<td>427.08 (25.47)</td>
</tr>
<tr>
<td>LB</td>
<td>418.39 (183.66)</td>
<td>281.91 (105.69)</td>
<td>681.10 (249.54)</td>
</tr>
<tr>
<td>Chao2</td>
<td>270.26 (34.90)</td>
<td>216.50 (26.07)</td>
<td>402.21 (41.43)</td>
</tr>
<tr>
<td>iChao2</td>
<td>290.69 (38.46)</td>
<td>235.55 (33.74)</td>
<td>430.88 (46.83)</td>
</tr>
</tbody>
</table>

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 Böhm 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
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. 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.


*Received May 2013. Revised April 2014. Accepted April 2014.*
Supplementary Materials for “An improved non-parametric lower bound of species richness via a modified Good-Turing frequency formula” by Chun-Huo Chiu, Y. T. Wang, Bruno A. Walther and Anne Chao

All equation numbers refer to those in the main paper.

Web Appendix A: A summary of the results for the binomial-mixture and Poisson-mixture models

Web Appendix B: Variance formulas and confidence intervals

Web Appendix C: Simulation results and comparisons

Web Appendix D: Derivation details for the proposed improved lower bounds under sampling without replacement
Web Appendix A: A summary of the results for the binomial-mixture and Poisson-mixture models

(a) Binomial-mixture model

When species probabilities \((p_1, p_2, \ldots, p_S)\) are random parameters and follow an \(S\)-dimensional joint distribution with an unknown common marginal density \(h(p)\), then we have

\[
E(f_k) = S \int_0^1 \binom{n}{k} p^k (1 - p)^{n-k} h(p) dp, \quad k = 0, 1, 2, \ldots, n. \tag{A1}
\]

All the derivations and results are parallel to those in the main text. For example, the key step used to derive the Chao1 estimator is the Cauchy-Schwarz inequality (Eq. 1b in the main text) for “summation” terms. Such an inequality is also valid for “integral” terms:

\[
\left[ \int_0^1 (1 - p)^{p} h(p) dp \right] \left[ \int_0^1 p^2 (1 - p)^{n-p} h(p) dp \right] \geq \left[ \int_0^1 p (1 - p)^{n-p} h(p) dp \right]^2. \tag{A2}
\]

Then Eq. (A1) leads to the lower bound:

\[
E(f_0) \geq \frac{n-1}{n} \left[ E(f_1) \right]^2. \tag{A3}
\]

This equation (A3) is the same lower bound used to obtain the Chao1 estimator. Thus the Chao1 estimator is shown here to be also valid under a binomial-mixture model. Furthermore, throughout the derivation steps in the main text, we can replace all “summation” terms by the corresponding “integral” terms, and obtain exactly the same iChao1 estimator under a binomial-mixture model.

(b) Poisson-mixture model

In a Poisson-mixture model, we assume an arbitrary unknown mixing density \(g(\lambda)\) for the expected detection rate \(\lambda\) of any species so that we have

\[
E(f_k) = S \int_0^\infty \lambda^k \frac{\exp(-\lambda)}{k!} g(\lambda) d\lambda, \quad k = 0, 1, 2, \ldots. \tag{A4}
\]

A similar type of the Cauchy-Schwarz inequality under this model is:

\[
\left[ \int_0^\infty \exp(-\lambda) g(\lambda) d\lambda \right] \left[ \int_0^\infty \frac{\lambda^2 \exp(-\lambda)}{2} g(\lambda) d\lambda \right] \geq \left[ \int_0^\infty \lambda \exp(-\lambda) g(\lambda) d\lambda \right]. \tag{A5}
\]

Then Eqs. (A4) and (A5) lead to the following lower bound:

\[
E(f_0) \geq \frac{[E(f_1)]^2}{2E(f_2)}. \tag{A6}
\]

This equation (A6) is slightly different from the lower bound in Eq. (A3), where we have the term \((n-1)/n\). In Chao’s (1984) original derivation, the term \((n-1)/n\) was not considered. Here we provide a justification for the omission of the term \((n-1)/n\) even when \(n\) is not large.
In all derivation steps in the main text, we can replace the binomial term
\[ \sum_{i=1}^{S} \binom{n}{k} p_i^k (1-p_i)^{n-k} \] with the corresponding Poisson term
\[ S \int_{0}^{\infty} \lambda^k \frac{\exp(-\lambda)}{k!} g(\lambda) d\lambda. \] Note that all terms involving the size \( n \) in the main text will not appear under the Poisson-mixture models. Consequently, we directly obtain the same iChao1 estimator as that given in Eq. (10b).

REFERENCES

Web Appendix B. Variance formulas and confidence intervals

(1) Derivation of Eqs. (3a) and (3b)

Here we provide the derivation steps for the following approximate variance formula for the Chao1 estimator:

\[ \hat{S}_{\text{Chao1}} = S_{\text{obs}} + \frac{(n-1) f_1^2}{n 2f_2} \]

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

Under the models discussed in Section 2 of the main text, the abundance frequency counts \((f_0, f_1, \ldots, f_m)\), where \(m\) denotes the maximum frequency, follow approximately a multinomial distribution with a cell total \(S\) and cell probabilities \(\mu_0, \mu_1, \ldots, \mu_m\), where \(\mu_k = (Ef_k)/S\) for \(k = 0, 1, \ldots, m\). Then the moments of a multinomial distribution lead to the approximate variance-covariance formula:

\[ \text{cov}(f_i, f_j) = \begin{cases} \left( Ef_i \right)(1 - Ef_j)/S & i = j; \\ -(Ef_i)(Ef_j)/S & i \neq j. \end{cases} \]  

(B1)

The Chao1 estimator can be expressed as

\[ \hat{S}_{\text{Chao1}} = S_{\text{obs}} + \frac{(n-1) f_1^2}{n 2f_2} = S - f_0 + \frac{(n-1) f_1^2}{n 2f_2}. \]  

(B2)

For notational simplicity, we use a simple notation \(A = (n-1)/n\) and use \(\hat{S}\) instead of \(\hat{S}_{\text{Chao1}}\) in the following derivation. From (B2), we can regard the Chao1 estimator as a function of the abundance frequency counts \((f_0, f_1, \ldots, f_m)\) and apply a standard asymptotic approach to obtain:

\[ \var(\hat{S}) \approx \sum_{i=1}^{2} \sum_{j=0}^{2} \frac{\partial \hat{S}}{\partial f_i} \frac{\partial \hat{S}}{\partial f_j} \text{cov}(f_i, f_j) \]

\[ = \left( \frac{\partial \hat{S}}{\partial f_0} \right)^2 \var(f_0) + \left( \frac{\partial \hat{S}}{\partial f_1} \right)^2 \var(f_1) + \left( \frac{\partial \hat{S}}{\partial f_2} \right)^2 \var(f_2) + 2 \left( \frac{\partial h}{\partial \hat{f}_0} \right) \left( \frac{\partial h}{\partial \hat{f}_1} \right) \text{cov}(f_0, f_1) + 2 \left( \frac{\partial h}{\partial \hat{f}_0} \right) \left( \frac{\partial h}{\partial \hat{f}_2} \right) \text{cov}(f_0, f_2) + 2 \left( \frac{\partial h}{\partial \hat{f}_1} \right) \left( \frac{\partial h}{\partial \hat{f}_2} \right) \text{cov}(f_1, f_2), \]

where all the partial derivatives \(\frac{\partial \hat{S}}{\partial f_j}\) for \(j = 0, 1\) and 2 are evaluated at \((Ef_0, Ef_1, Ef_2)\).

Substituting

\[ \frac{\partial \hat{S}}{\partial f_0} = -1, \quad \frac{\partial \hat{S}}{\partial f_1} = A \frac{f_1}{f_2}, \quad \frac{\partial \hat{S}}{\partial f_2} = -A \left( \frac{f_1}{f_2} \right)^2 \]

and variance-covariance estimator (from Eq. B1)
we then obtain the approximate variance estimator
\[
\text{vá}r(\hat{S}) \approx f_0 \left(1 - \frac{f_0}{S}\right) + \left(\frac{f_1}{f_2}\right)^2 f_1 \left(1 - \frac{f_1}{S}\right) + 2A \frac{f_1}{f_2} f_0 f_1^2 S - A \left(\frac{f_1}{f_2}\right)^2 f_0 f_2 + A^2 \left(\frac{f_1}{f_2}\right)^3 f_1 f_2.
\]

When \( n \) is sufficiently large, we can replace \( f_0 \) and \( S \) in the above formula respectively by
\[
\hat{f}_0 = \frac{(n-1)}{n} \frac{f_1^2}{2f_2} \quad \text{and} \quad \hat{S} = S_{\text{obs}} + \frac{(n-1)}{n} \frac{f_1^2}{2f_2}.
\]

Then we obtain the approximate variance

<table>
<thead>
<tr>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{vá}r(\hat{S}) \approx f_2 \left[ 1 - \left(1 - \frac{1}{n}\right)^2 + \left(\frac{f_1}{f_2}\right)^4 + \left(\frac{f_1}{f_2}\right)^3 + \left(\frac{1}{n}\right) f_1^2 \right] ).</td>
</tr>
</tbody>
</table>

A similar derivation leads to the variance estimator of the bias-corrected estimator, as shown in Eq. (3b) of the main text.

(2) Derivation of the confidence intervals in Eq. (4) of the main text

A basic assumption in the construction of the confidence interval (C. I.) given in Eq. (4) of the main text is that \( \log(\hat{S} - S_{\text{obs}}) \) follows approximately a normal distribution with a mean \( \log(E_{f_0}) \) and a variance \( \sigma^2 \), i.e., \( \log(\hat{S} - S_{\text{obs}}) \sim N(\log(E_{f_0}), \sigma^2) \). This leads to a 95% C. I. of \( \log(E_{f_0}) \):

\[
[\log(\hat{S} - S_{\text{obs}}) - 1.96\times\sigma, \log(\hat{S} - S_{\text{obs}}) + 1.96\times\sigma],
\]

which gives an approximate 95% C. I. of \( E_{f_0} \):

\[
[(\hat{S} - S_{\text{obs}})/\exp(1.96\times\sigma), (\hat{S} - S_{\text{obs}})\times\exp(1.96\times\sigma)]. \tag{B3}
\]

The moment property of a log-normal distribution gives

\[
E(\hat{S} - S_{\text{obs}}) = \exp[\log(E_{f_0}) + \frac{1}{2} \sigma^2],
\]

\[
\text{var}(\hat{S} - S_{\text{obs}}) = \exp(2\log(E_{f_0}) + \sigma^2) \times (e^{\sigma^2} - 1).
\]

Then the squared coefficient of variation (CV) of the random variable \( \hat{S} - S_{\text{obs}} \) becomes

\[
CV^2 = \frac{\text{var}(\hat{S} - S_{\text{obs}})}{[E(\hat{S} - S_{\text{obs}})]^2} = \exp(\sigma^2 - 1). \tag{B4}
\]

A natural estimate of \( CV^2 \) is
\[
\frac{\text{var}(\hat{S} - S_{\text{obs}})}{(\hat{S} - S_{\text{obs}})^2} \approx \frac{\text{var}(\hat{S})}{(\hat{S} - S_{\text{obs}})^2}.
\]

Substituting this estimator into the formula \( \sigma^2 = \log(1 + CV^2) \) (which follows from Eq. B4), we then obtain an estimator of \( \sigma \). It then follows from (B3) that an approximate 95% C. I. of \( S \) is
\[
[S_{\text{obs}} + (\hat{S} - S_{\text{obs}})/R, \ S_{\text{obs}} + (\hat{S} - S_{\text{obs}}) \times R],
\]
where \( R = \exp\left\{ 1.96 \left[ \log(1 + \frac{\text{var}(\hat{S})}{(\hat{S} - S_{\text{obs}})^2}) \right]^{1/2} \right\} \).

(3) Comparisons of two approaches to construction confidence intervals

Approach 1: our suggested confidence interval of species richness is based on the assumption: \( \log(\hat{S} - S_{\text{obs}}) \sim N(\ \log(\ E_{\beta}) , \ \sigma^2) \); see Eq. (4) of the main text.

Approach 2: a confidence interval can also be constructed based on the assumption \( \log(\hat{S}) \sim N(\ \log(\hat{S}) , \ \sigma^2) \).

Web Table B1 compares the coverage probabilities of the 95% confidence intervals obtained from the two approaches based on the Chao1 and iChao1 estimators. In the homogeneous model (all species have the same abundances as shown in part (a) of Web Table B1), in which the estimated number of undetected species tends to be close to zero, Approach 2 produces confidence intervals with coverage probabilities being closer to the nominal level of 95% than those obtained from Approach 1. When heterogeneity exists among species abundances, in which case both the Chao1 and iChao1 estimators become lower bounds of species richness, Approach 1 is preferable to Approach 2 because all coverage probabilities based on Approach 1 are closer to 95% as shown in parts (b) and (c) in Web Table B1. Since species abundances in most applications are not homogeneous, we therefore recommend the use of Approach 1.
Web Table B1. Comparisons of coverage probabilities of 95% confidence intervals based on two approaches (see text for details) and two estimators (the Chao1 and iChao1 estimators).

(a) Results were calculated from 1000 simulated data sets from a homogeneous model, \( p_i = 1/S \) with \( S = 200 \) and \( CV= 0 \).

<table>
<thead>
<tr>
<th>Size n (species seen)</th>
<th>Estimator</th>
<th>Average Estimate</th>
<th>95% C.I. Coverage Approach 1</th>
<th>95% C.I. Coverage Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (127.19)</td>
<td>Chao1</td>
<td>204.16</td>
<td>0.95(^\dagger)</td>
<td>0.95(^\dagger)</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>206.74</td>
<td>0.92</td>
<td>0.93(^\dagger)</td>
</tr>
<tr>
<td>400 (173.47)</td>
<td>Chao1</td>
<td>200.87</td>
<td>0.94(^\dagger)</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>201.40</td>
<td>0.92</td>
<td>0.93(^\dagger)</td>
</tr>
<tr>
<td>800 (196.50)</td>
<td>Chao1</td>
<td>200.39</td>
<td>0.87</td>
<td>0.92(^\dagger)</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>200.55</td>
<td>0.86</td>
<td>0.93(^\dagger)</td>
</tr>
</tbody>
</table>

\(^\dagger\) Closer to 95% coverage between the two approaches.

(b) Results were calculated from 1000 simulated datasets from a log-normal (0, 1) model, with \( S = 200 \) and \( CV= 1.82 \).

<table>
<thead>
<tr>
<th>Size n (species seen)</th>
<th>Estimator</th>
<th>Average estimate</th>
<th>95% C.I. Coverage Approach 1</th>
<th>95% C.I. Coverage Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (91.47)</td>
<td>Chao1</td>
<td>159.16</td>
<td>0.69(^\dagger)</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>170.79</td>
<td>0.81(^\dagger)</td>
<td>0.74</td>
</tr>
<tr>
<td>400 (128.24)</td>
<td>Chao1</td>
<td>177.40</td>
<td>0.78(^\dagger)</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>187.22</td>
<td>0.92(^\dagger)</td>
<td>0.84</td>
</tr>
<tr>
<td>800 (160.85)</td>
<td>Chao1</td>
<td>188.59</td>
<td>0.88(^\dagger)</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>194.26</td>
<td>0.94(^\dagger)</td>
<td>0.85</td>
</tr>
</tbody>
</table>

\(^\dagger\) Closer to 95% coverage between the two approaches.

(c) Results were calculated from 1000 simulated datasets from a power decay model \( p_i = \frac{c}{i^{1/2}} \), with \( S = 200 \) and \( CV= 4.20 \).

<table>
<thead>
<tr>
<th>Size n (species seen)</th>
<th>Estimator</th>
<th>Average Estimate</th>
<th>95% C.I. Coverage Approach 1</th>
<th>95% C.I. Coverage Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chao1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>200 (59.84)</td>
<td>135.06</td>
<td>0.63†</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>147.03</td>
<td>0.80†</td>
<td>0.67</td>
</tr>
<tr>
<td>400 (88.09)</td>
<td>160.87</td>
<td>0.74†</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>172.79</td>
<td>0.88†</td>
<td>0.79</td>
</tr>
<tr>
<td>800 (123.63)</td>
<td>181.48</td>
<td>0.87†</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>194.70</td>
<td>0.94†</td>
<td>0.93</td>
</tr>
</tbody>
</table>

† Closer to 95% coverage between the two approaches.
Web Appendix C. Simulation results and comparisons

Web Table C1. Comparison of five species richness estimators based on 1000 simulation trials under a homogeneous model, \( p_i = 1/S \) with \( S = 200 \) and \( CV = 0 \). See Table 1 of the main text for the abbreviations of the estimators.

<table>
<thead>
<tr>
<th>Size ( n ) (species seen)</th>
<th>Estimator</th>
<th>Average Estimate</th>
<th>Sample s.e.</th>
<th>Average Estimated s.e.</th>
<th>Sample RMSE</th>
<th>95% C.I. Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (127.19)</td>
<td>Jackknife1</td>
<td>200.45*</td>
<td>10.46</td>
<td>12.14</td>
<td>10.46*</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>237.79</td>
<td>19.27</td>
<td>21.03</td>
<td>42.40</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>233.43</td>
<td>92.36</td>
<td>61.77</td>
<td>98.14</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>204.16</td>
<td>24.29</td>
<td>23.91</td>
<td>24.62</td>
<td>0.95†</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>206.74</td>
<td>30.16</td>
<td>30.51</td>
<td>30.96</td>
<td>0.92</td>
</tr>
<tr>
<td>400 (173.47)</td>
<td>Jackknife1</td>
<td>227.24</td>
<td>9.01</td>
<td>10.38</td>
<td>28.69</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>226.59</td>
<td>17.05</td>
<td>17.97</td>
<td>31.58</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>206.65</td>
<td>20.07</td>
<td>17.40</td>
<td>21.13</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>200.87*</td>
<td>10.12</td>
<td>9.86</td>
<td>10.14*</td>
<td>0.95†</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>201.40</td>
<td>12.36</td>
<td>11.94</td>
<td>12.81</td>
<td>0.92</td>
</tr>
<tr>
<td>800 (196.50)</td>
<td>Jackknife1</td>
<td>210.99</td>
<td>4.02</td>
<td>5.37</td>
<td>11.70</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>196.16</td>
<td>9.48</td>
<td>9.29</td>
<td>10.22</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>202.13</td>
<td>5.88</td>
<td>5.53</td>
<td>6.25</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>200.39*</td>
<td>2.91</td>
<td>2.91</td>
<td>2.93*</td>
<td>0.87†</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>200.55</td>
<td>3.39</td>
<td>3.40</td>
<td>3.50</td>
<td>0.86</td>
</tr>
</tbody>
</table>

* Denotes the smallest bias; # Denotes the smallest RMSE; † Closest to 95% coverage.
Web Table C2. Comparison of five species richness estimators based on 1000 simulation trials under a broken stick model, $p_i = c a_i$, where $(a_1, a_2, \ldots, a_{200})$ are a random sample from an exponential distribution. Or equivalently, $(p_1, p_2, \ldots, p_S)$ is a Dirichlet distribution with parameter 1 and CV = 0.96. See Table 1 of the main text for the abbreviations of the estimators.

<table>
<thead>
<tr>
<th>Size n (species seen)</th>
<th>Estimator</th>
<th>Average Estimate</th>
<th>Sample s.e.</th>
<th>Average Estimated s.e.</th>
<th>Sample RMSE</th>
<th>95% C.I. Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (101.20)</td>
<td>Jackknife1</td>
<td>154.13</td>
<td>9.80</td>
<td>10.11</td>
<td>46.90</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>178.49</td>
<td>16.53</td>
<td>17.51</td>
<td>27.12*</td>
<td>0.87†</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>196.92*</td>
<td>79.27</td>
<td>70.65</td>
<td>79.29</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>158.26</td>
<td>18.56</td>
<td>18.93</td>
<td>45.68</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>168.36</td>
<td>22.54</td>
<td>23.60</td>
<td>38.85</td>
<td>0.79</td>
</tr>
<tr>
<td>400 (136.15)</td>
<td>Jackknife1</td>
<td>179.32</td>
<td>9.05</td>
<td>9.27</td>
<td>22.57</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>192.04</td>
<td>16.06</td>
<td>16.06</td>
<td>17.92*</td>
<td>0.96†</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>201.11*</td>
<td>68.09</td>
<td>68.76</td>
<td>68.07</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>172.70</td>
<td>14.01</td>
<td>12.86</td>
<td>30.68</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>180.37</td>
<td>16.70</td>
<td>15.78</td>
<td>25.77</td>
<td>0.81</td>
</tr>
<tr>
<td>800 (161.23)</td>
<td>Jackknife1</td>
<td>190.50</td>
<td>7.48</td>
<td>7.65</td>
<td>12.09*</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>195.66</td>
<td>13.15</td>
<td>13.26</td>
<td>13.84</td>
<td>0.98†</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>201.92*</td>
<td>49.17</td>
<td>57.67</td>
<td>49.18</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>184.30</td>
<td>9.91</td>
<td>9.18</td>
<td>18.56</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>189.32</td>
<td>11.68</td>
<td>11.09</td>
<td>15.83</td>
<td>0.87</td>
</tr>
</tbody>
</table>

* Denotes the smallest bias; † Denotes the smallest RMSE; ‡ Closest to 95% coverage.
Web Table C3. Comparison of five species richness estimators based on 1000 simulation trials under a log-normal (0, 1) model, with $S = 200$ and $CV=1.82$. See Table 1 of the main text for the abbreviations of the estimators.

<table>
<thead>
<tr>
<th>Size $n$ (species seen)</th>
<th>Estimator</th>
<th>Average Estimate</th>
<th>Sample s.e.</th>
<th>Average Estimated s.e.</th>
<th>Sample RMSE</th>
<th>95% C.I. Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (91.47)</td>
<td>Jackknife1</td>
<td>142.25</td>
<td>10.81</td>
<td>10.06</td>
<td>58.75</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>172.54*</td>
<td>18.33</td>
<td>17.42</td>
<td>33.00*</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>240.93</td>
<td>158.29</td>
<td>128.87</td>
<td>163.34</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>159.16</td>
<td>28.20</td>
<td>25.81</td>
<td>49.62</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>170.79</td>
<td>33.38</td>
<td>31.86</td>
<td>44.33</td>
<td>0.81†</td>
</tr>
<tr>
<td>400 (128.24)</td>
<td>Jackknife1</td>
<td>180.72</td>
<td>9.66</td>
<td>10.23</td>
<td>21.56</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>203.49*</td>
<td>16.77</td>
<td>17.72</td>
<td>17.11*</td>
<td>0.94†</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>225.94</td>
<td>105.10</td>
<td>72.02</td>
<td>108.15</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>177.40</td>
<td>18.80</td>
<td>17.82</td>
<td>29.38</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>187.22</td>
<td>22.41</td>
<td>22.11</td>
<td>25.77</td>
<td>0.92</td>
</tr>
<tr>
<td>800 (160.85)</td>
<td>Jackknife1</td>
<td>201.71*</td>
<td>8.56</td>
<td>9.02</td>
<td>8.72*</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>210.67</td>
<td>15.74</td>
<td>15.62</td>
<td>19.00</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>210.18</td>
<td>40.11</td>
<td>36.35</td>
<td>41.34</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>188.59</td>
<td>11.99</td>
<td>11.23</td>
<td>16.55</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>194.26</td>
<td>14.61</td>
<td>13.89</td>
<td>15.68</td>
<td>0.94†</td>
</tr>
</tbody>
</table>

* Denotes the smallest bias; # Denotes the smallest RMSE; † Closest to 95% coverage.
Web Table C4. Comparison of five species richness estimators based on 1000 simulation trials under a Zipf-Mandelbrot model, $p_i = c/(i - 0.1)$, $i = 1, 2, \ldots, 200$ and CV = 3.08. See Table 1 of the main text for the abbreviations of the estimators.

<table>
<thead>
<tr>
<th>Size n (species seen)</th>
<th>Estimator</th>
<th>Average Estimate</th>
<th>Sample s.e.</th>
<th>Average Estimated s.e.</th>
<th>Sample RMSE</th>
<th>95% C.I. Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 (75.89)</td>
<td>Jackknife1</td>
<td>121.82</td>
<td>9.67</td>
<td>9.57</td>
<td>78.78</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>152.83</td>
<td>15.80</td>
<td>16.57</td>
<td>49.75</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>338.34</td>
<td>398.94</td>
<td>285.81</td>
<td>421.78</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>158.71</td>
<td>35.48</td>
<td>35.08</td>
<td>54.42</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>172.07*</td>
<td>40.62</td>
<td>42.62</td>
<td>49.26#</td>
<td>0.87†</td>
</tr>
<tr>
<td>400 (111.55)</td>
<td>Jackknife1</td>
<td>166.74</td>
<td>11.23</td>
<td>10.49</td>
<td>35.10</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>197.79*</td>
<td>18.65</td>
<td>18.17</td>
<td>18.76§</td>
<td>0.95†</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>247.78</td>
<td>141.89</td>
<td>108.40</td>
<td>149.55</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>178.79</td>
<td>27.14</td>
<td>24.32</td>
<td>34.41</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>190.75</td>
<td>31.96</td>
<td>29.64</td>
<td>33.23</td>
<td>0.95†</td>
</tr>
<tr>
<td>800 (150.13)</td>
<td>Jackknife1</td>
<td>203.71</td>
<td>10.09</td>
<td>10.34</td>
<td>10.74§</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Jackknife2</td>
<td>222.83</td>
<td>17.03</td>
<td>17.90</td>
<td>28.46</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>218.15</td>
<td>48.80</td>
<td>45.08</td>
<td>52.00</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Chao1</td>
<td>193.66</td>
<td>15.74</td>
<td>15.49</td>
<td>16.95</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>iChao1</td>
<td>200.91*</td>
<td>19.33</td>
<td>19.53</td>
<td>19.35</td>
<td>0.94†</td>
</tr>
</tbody>
</table>

* Denotes the smallest bias; § Denotes the smallest RMSE; † Closest to 95% coverage.
We compare the average biases for five species richness estimators: Chao1, iChao1, first-order jackknife (JK1), second-order jackknife (JK2), and the interpolated jackknife (I-JK) estimators using data sets generated from the same six models as used in the main text. All results are shown in Web Figure C1. We below summarize our findings:

(i) The homogeneous and negative binomial models (Panels a and b, in which CV values are low):
   
The I-JK estimator overestimates so that both the Chao1 and iChao1 estimators are superior to the I-JK estimator for all sample sizes. Compared with the JK1 and JK2 estimators, the I-JK estimator has a smaller positive bias when the sample size is relatively small. However, when the sample size is larger, the procedures proposed in Burnham and Overton (1978, 1979) select a worse estimator (JK1) instead of the better estimator (JK2). Therefore, the I-JK estimator does not necessarily perform better than the JK1 and JK2 estimators.

(ii) Other models (Panels c, d, e, and f, in which CV values are relatively high):
   
The I-JK estimator either improves over the JK1 and JK2 estimators or follows closely one of them. In all cases, the I-JK estimator exhibits the pattern typical of the JK1 and JK2 estimators: it underestimates when the sample size is relatively small, but then exceeds the true species richness and overestimates after this point. In some cases, these crossing points are not shown because they exceed the maximum sample sizes displayed in our figures. For each model, there is a limited range of sample sizes (near the crossing points) where the I-JK estimator is close to the true species richness and thus preferable to the Chao1 and iChao1 estimators. However, as discussed in the main text for the JK1 and JK2 estimators, this narrow range depends on the models used and is therefore not predictable. The I-JK estimator thus also exhibits a counter-intuitive pattern: its bias does not improve as sample size increases.
Web Figure C1. Comparison of the average biases of five species richness estimators: Chao1, iChao1, the first-order jackknife (JK1), the second-order jackknife (JK2), and the interpolated jackknife (I-JK) estimators for sample sizes between 100 and 800. The biases were obtained by averaging over 1000 data sets generated from six different species abundance models (see...
the main 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).

REFERENCES


Web Appendix D: Derivation details for the proposed improved lower bounds under sampling without replacement

The models and the proposed new lower bounds in the main text were derived under the assumption of sampling with replacement, in which individuals (or any sampling unit) can be repeatedly observed. As discussed in the main text (Section 7), some types of biological surveys rely on sampling without replacement which means that any individual (or any sampling unit such as plot or quadrat) can be surveyed at most once.

In this Appendix, we focus on the models and derivations for individual-based abundance data, and the results for incidence data will only be summarized. For abundance data, 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. Let $N_i$ (true species absolute abundance) be the unknown number of individuals of the $i$th 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.

Assume that a sample of $n$ individuals is taken from the community, with individuals being drawn without replacement. Let $S_{obs}, X_i$ (sample species frequency), $i = 1, 2, \ldots, S,$ and $f_k$ (sample abundance frequency counts), $k = 0, 1, \ldots, n$, be defined as those in the main text.

The sample fraction is defined as $q = n/N$, the ratio of the sample size over the known population size.

Following Chao and Lin (2012), we assume that the sample species frequencies $(X_1, X_2, \ldots, X_S)$ follow a generalized hypergeometric distribution:

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

Under this model, we have the following general expectation formula:

$$E(f_k) = \sum_{i=1}^{S} P(X_i = k) = \sum_{i=1}^{S} \binom{N_i}{k} \binom{N - N_i}{n-k} \binom{N}{n}.$$  

Chao and Lin (2012) derived a Chao1-type lower bound for the number of undetected species in the sample, which is a function of the sampling fraction and the number of singletons and doubletons. This lower bound was called $\hat{S}_{wor}$, where the subscript “wor” refers to “without replacement”, and has the following form:

$$\hat{S}_{wor} = S_{obs} + \frac{f_1^2}{n-1} \left( \frac{2 f_2 + q f_1}{1 - q} \right).$$

When the sample fraction $q$ tends to 0 (i.e. $n/N \to 0$), it reduces to the Chao1 lower bound for sampling with replacement; see (2b) in the main text.
Let the mean absolute abundance of those species that appeared \( r \) times in a sample of size \( n \) be \( F_r \). That is, \( F_r = \sum_{i=1}^{s} N_i I(X_i = r) / f_r \). Using the same approach as in equation (6) of the main text, we obtain the following derivation:

\[
E\left[ \sum_{i=1}^{s} \frac{N_i - r}{N - N_i - n + r + 1} I(X_i = r) \right] = \sum_{i=1}^{s} \frac{N_i - r}{N - N_i - n + r + 1} \binom{N - r}{n - r} = \sum_{i=1}^{s} \frac{r + 1}{n - r} \binom{N - r}{n - r} = r + 1 \frac{E(f_{r+1})}{n - r}.
\]

Also, we have

\[
E\left[ \sum_{i=1}^{s} \frac{N_i - r}{N - N_i - n + r + 1} I(X_i = r) \right] \approx \frac{F_r - r}{N - F_r - n + r + 1} E(f_r).
\]

Thus, a formula for the mean abundance \( F_r \) can be obtained as

\[
F_r = \frac{(r + 1)(N - n + r + 1)f_{r+1} + (n - r)rf_r}{(n - r)f_r + (r + 1)f_{r+1}}, \quad r = 0, 1, 2, \ldots
\]

(D1)

Our approach is to evaluate the magnitude of the bias associated with the Chao and Lin (2012) lower bound. The magnitude of the first-order bias is:

\[
\left| \text{bias}(\hat{S}_{\text{Chao}}) \right| = E(f_0) - \frac{n}{n - 1} \frac{\left[ E(f_1) \right]^2}{2E(f_2) + \frac{q}{1 - q} E(f_1)}
\]

\[
= \frac{n}{n - 1} \frac{2E(f_2) + \frac{q}{1 - q} E(f_1)}{E(f_0)} - \frac{\left[ E(f_1) \right]^2}{E(f_0)}.
\]

(D2)

We can now separately approximate each term in (D2). First, we have

\[
E(f_0) = \sum_{i=1}^{s} \frac{N - N_i - n + 1}{nN_i} E[I(X_i = 1)]
\]

\[
\approx \frac{N - F_r - n + 1}{nF_r} \sum_{i=1}^{s} \frac{nN_i}{N - N_i - n + 1} \binom{N - N_i}{n}.
\]

(D3)

Second, Chao and Lin (2012) showed that if \( N_i \) is much less than \( N \) for all \( i \), then we have
\[ \frac{n}{n-1}2E(f_x) + \frac{q}{1-q}E(f_1) \approx \sum_{i=1}^{s} \left( \frac{nN_i}{N-N_i-n+1} \right)^2 \left( \frac{N-N_i}{n} \right). \] (D4)

Note that for any species i that is observed three times, we have the following approximation:

\[ E[I(X_i = 3)] = \frac{n(n-1)(n-2)N_i(N_i-1)(N_i-2)}{6(N-N_i-n+1)(N-N_i-n+2)(N-N_i-n+3)} \left( \frac{N-N_i}{n} \right) \]

We then combine (D4) and (D5) to obtain

\[ \frac{n}{n-1}2E(f_x) + \frac{q}{1-q}E(f_1) \approx \sum_{i=1}^{s} \left( \frac{nN_i}{N-N_i-n+1} \right)^3 \left( \frac{N-N_i}{n} \right) \cdot 6E[I(X_i = 3)] \]

Third, we have the following approximation:

\[ E(f_1) \approx \sum_{i=1}^{s} \left( \frac{N-N_i-n+1}{nN_i} \right) \times 6E[I(X_i = 3)] \]

\[ \approx \left( \frac{N-F_1-n+1}{nF_1} \right)^2 \sum_{i=1}^{s} \left( \frac{nN_i}{N-N_i-n+1} \right)^3 \left( \frac{N-N_i}{n} \right). \] (D7)

It then follows from (D3), (D6) and (D7) that the numerator of (D2) becomes

\[ \frac{(N-F_1-n+1)(N-F_2-n+1)}{nF_1} \left( \frac{N-F_1-n+1}{nF_1} \right)^2 \times \sum_{i=1}^{s} \left( \frac{nN_i}{N-N_i-n+1} \right) \left( \frac{N-N_i}{n} \right)^3 \left( \frac{N-N_i}{n} \right). \] (D8)

For the last two terms above, the Cauchy-Schwarz inequality leads to the following:
The magnitude of the bias is hence bounded by

\[
\left| \text{bias}(\hat{S}_{wor}) \right| \geq \frac{N - F_1 - n + 1}{nf_3} \left( \frac{N - F_1 - n + 1}{nf_3} - \frac{N - F_1 - n + 1}{nf_3} \right) \left[ \frac{n}{n - 1} 2E(f_2) + \frac{q}{1-q} E(f_1) \right].
\]

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 absolute abundance than that of singletons (i.e., \( F_3 > F_1 \)). Using the estimators in (D1), we then obtain a lower bound of species richness:

\[
\hat{S}_{wor} + \frac{n(1-q)-3}{n} \frac{f_3}{4(1-q)f_4 + 3qf_3} \times \left( \frac{n}{n - 1} 2f_2 + \frac{q}{1-q} f_1 \right)
\]

\[
\times \max \left\{ \frac{n(1-q)-1}{n} \frac{f_1}{2(1-q)f_2 + qf_1} - \frac{n(1-q)-3}{n} \frac{f_3}{4(1-q)f_4 + 3qf_3}, 0 \right\}.
\]

When \( n \) is large enough, the above formula reduces to the following estimator which we refer to as \( \hat{S}_{wor1} \):

\[
\hat{S}_{wor1} = \hat{S}_{wor} + \frac{(1-q)f_3}{4(1-q)f_4 + 3qf_3} \times \left( 2f_2 + \frac{qf_1}{1-q} \right) \times \max \left\{ \frac{(1-q)f_1}{2(1-q)f_2 + qf_1} - \frac{(1-q)f_3}{4(1-q)f_4 + 3qf_3}, 0 \right\}.
\]

When \( q \) approaches 0, it approaches the proposed iChao1 estimator under sampling with replacement; see (10b) in the main text.

Using a parallel derivation for the incidence data of \( T \) sampling units randomly selected from \( T^* \) units in the study region, we can obtain the following improved lower bound:

\[
\hat{S}_{wor2} = \hat{S}_{wor} + \frac{T(1-q)-3}{T} \frac{Q_3}{4(1-q)Q_4 + 3qQ_3} \times \left( \frac{T}{T-1} 2Q_2 + \frac{q}{1-q} Q_1 \right)
\]

\[
\times \max \left\{ \frac{T(1-q)-1}{T} \frac{Q_1}{2(1-q)Q_2 + qQ_1} - \frac{T(1-q)-3}{T} \frac{Q_3}{4(1-q)Q_4 + 3qQ_3}, 0 \right\},
\]

where \( q = T/T^* \) is the sampling fraction. When \( q \) tends to 0, it approaches the proposed iChao2 estimator under sampling replacement; see (13) in the main text. Variance estimators of the two proposed estimators under sampling without replacement can be obtained directly by using the standard asymptotic method as in (11) of the main text.

REFERENCES