RESEARCH ARTICLE

Incidence-data-based species richness estimation via a Beta-Binomial model

Chun-Huo Chiu

Department of Agronomy, National Taiwan University, Taipei, Taiwan

Correspondence
Chun-Huo Chiu
Email: chchiu2017@ntu.edu.tw

Funding information
Taiwan Ministry of Science and Technology, Grant/Award Number: 109-2118-M-002-003-MY2

Handling Editor: Claudia Coleine

Abstract

1. Individual-based abundance data and sample-based incidence data are the two most widely used survey data formats to assess the species diversity in a target area, where the sample-based incidence data are more available and efficient for estimating species richness. For species individual with spatial aggregation, individual-unit-based random sampling scheme is difficult to implement, and quadrat-unit-based sampling scheme is more available to implement and more likely to fit the model assumption of random sampling. In addition, sample-based incidence data, without recording the number of individuals of a species and only recording the binary presence or absence of a species in the sampled unit, could considerably reduce the survey loading in the field.

2. In this study, according to sample-based incidence data and based on a beta-binomial model assumption, instead of using the maximum likelihood method, I used the moment method to derive the richness estimator. The proposed richness estimation method provides a lower bound estimator of species richness for beta-binomial models, in which the new method only uses the number of singletons, doubletons and tripletons in the sample to estimate undetected richness.

3. I evaluated the proposed estimator using simulated datasets generated from various species abundance models. For highly heterogeneous communities, the simulation results indicate that the proposed estimator could provide a more stable, less biased estimate and a more accurate 95% confidence interval of true richness compared to other traditional parametric-based estimators.

4. I also applied the proposed approach to real datasets for assessment and comparison with traditional estimators. The newly proposed richness estimator provides different information and conclusions from other estimators.

KEYWORDS
beta-binomial model, doubletons, sample-based incidence data, singletons, tripletons
1 | INTRODUCTION

Species richness, or the number of species in a particular area, is the simplest and most frequently used measurement of biodiversity. Understanding spatial and temporal patterns of species richness is essential to community ecology and conservation biology. However, because of sampling limitations, a complete species inventory is difficult to compile; many previous approaches in scientific literature were developed to address the underestimation of observed species richness associated with incomplete sampling data (Bunge et al., 2014; Bunge & Fitzpatrick, 1993; Chiu et al., 2014; Colwell & Coddington, 1994; Dorazio & Andrew Royle, 2003; Griffiths, 1973; Rocchetti et al., 2014; Wilson & Collins, 1992). There are two types of data typically collected in the field to assess richness diversity, namely individual-based abundance data, in which the sampling unit is the individual and the number of individuals (abundance) is recorded for each observed species, and sample-based incidence data, in which the sampling unit is a quadrat (or, e.g., a plot, a transect or a time interval) and only the binary presence or absence of species is recorded in each sampled unit (Colwell et al., 2012). Furthermore, there are generally two typical sampling schemes implemented to collect data, which are sampling with replacement and sampling without replacement. Depending on their data collection formats and sampling schemes, most previously derived estimators can be classified into parametric or nonparametric methods. Nonparametric methods yield robust statistical performance because no assumption is made related to the species detection probability; these methods are widely used by ecologists, and the lower bound estimator (Chao, 1984, 1987) and jackknife estimators (Burnham & Overton, 1978, 1979) are the most popular estimators. Parametric approaches assume the species detection probability is a random variable to reduce the number of parameters and applies the traditional statistical inference to estimate species richness (Chao & Bunge, 2002; Dorazio & Andrew Royle, 2003; Griffiths, 1973; Sanathanan, 1972, 1977). In general, parametric approaches usually work well when sample size is sufficiently large or the distribution of the species detection probability fits the model assumption. However, parametric approaches are often overlooked in ecological applications due to their unreliability, especially when data are sparse (Chao & Bunge, 2002).

For estimating species richness within a target area, sample-based incidence data that only the presence or absence of species is recorded in each sampling unit, without precisely recording the number of individuals for each species in the sampled plot, require less field effort than individual-based abundance data. In addition, for some individual organisms, especially micro-organisms, invertebrates or plants, it is hard to sample individuals randomly and independently due to their tiny size or highly spatial aggregated pattern. Therefore, sample-based incidence data are more practical compared to individual-based abundance data for estimating species richness. In this study, according to sample-based incidence data with a replacement sampling scheme, the proposed approach involved uses the parametric method to assume that the species detection probability is a random variable following a beta distribution. Instead of using the maximum likelihood (ML) method (Dorazio & Andrew Royle, 2003; Griffiths, 1973; Sanathanan, 1972, 1977), the traditional statistical moment method was applied to obtain the estimators of the parameters in the assumed beta-binomial model based on the rare species in the sample. The new richness estimator was shown as a nearly unbiased estimator under the homogeneous model and provided a parametric lower bound estimator in other discussed models. Compared to the ML method using an entire sample, the new estimation method only requires information about the rare species (i.e. the number of singletons, doubletons and tripletons) in the sample to estimate the number of undetected species. The simulation results indicated that the proposed approach reduced bias and root mean squared error (RMSE) more than traditional nonparametric estimators and the ML estimator (MLE), improving the accuracy of the coverage rate when the heterogeneity of species abundances was relatively high. Additionally, this method has the advantage of employing nonparametric approaches because only the number of rare species in the sample is required to estimate the number of undetected species, and information about the number of highly frequent species is not needed. Through this method, the field work required for field investigations can be greatly reduced.

The theoretical framework of the estimation method in our study is detailed in the following model and estimator section. In the simulation analysis section, we analyse the statistical performance of the proposed estimator approach in ecological species composition models by examining the bias, estimated standard error (SE), and coverage rate with a 95% confidence interval (CI). Afterwards, real data are used to illustrate the application of the newly proposed estimator and compare the new approach with other commonly used estimators. The final section provides a discussion of our estimation method.

2 | MATERIALS AND METHODS

2.1 | Model and estimator

We assume that there are $S$ species presented in the target area, which is an unknown parameter. In this study, the sampling unit is a plot, quadrat, or transect randomly sampled from a target region and only the absence or presence of a species is recorded in each sampled unit. If a sample contains $T$ sampling units or plots, and $Y_i$ is the number of plots that $i$th species observed or recorded in these $T$ plots, then $Y_i$ follows a binomial distribution with parameter $T$ and detection probability $x_i$. Here, the detection probability $x_i$ is determined not only by species population size, but also by many other biotic factors, such as individual colour, size, vocalizations and movement patterns.

To assess the heterogeneity of species detection probabilities, a mixed binomial model widely used in ecological studies was applied. I assumed that the count $Y_i$ follows a binomial distribution $\text{Binomial}(T, x_i)$, and $x_1, x_2, \ldots, x_S$ are independent and identically
distributed (i.i.d.) random variables with probability density function \(g(x)\). Herein, I presume that \(g(x)\) follows a beta distribution with two parameters \(\alpha, \beta\) ensuring species relative abundance distribution flexibility. Thus, I obtain the marginal distribution of the species count in the sample as follows:

\[
p_k = P(Y_i = k) = \int_0^1 P(X_i = k|x_i = x) \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} dx
\]

\[
= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \int_0^1 x^{\alpha - 1} (1 - x)^{\beta - 1} dx = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha) \Gamma(\beta)} \int_0^1 x^{\alpha + k - 1} (1 - x)^{\beta - 1} dx = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha) \Gamma(\beta)} \frac{\Gamma(\alpha + \beta + T)}{\Gamma(\alpha + T) \Gamma(\beta + T)}, \quad k = 0, 1, 2, \ldots, T. \tag{1}
\]

Let \(Q_k\) denote the number of species exactly observed in the \(k\) plots, \(Q_k = \sum_{i=1}^S I(Y_i = k), k = 0, 1, 2, \ldots\), where \(I(A)\) is an indicator function. \(I(A)\) equals to 1 if event \(A\) occurs and 0 if otherwise. Thus, \(S_{obs} = \sum_{k=1}^T Q_k\) is the observed richness in the sample and \(Q_0\) is the undetected richness. Based on the mixture model, the observed frequency counts \(\{Q_k; k \geq 1\}\) follow a multinomial distribution with the total sum \(S_{obs}\) and cell probabilities \(\left\{\frac{P_k}{1 - P_0}; k \geq 1\right\}\). The conditional likelihood function is expressed as:

\[
L(\alpha, \beta) = \frac{S_{obs}}{\prod_{k=1}^T Q_k} \prod_{k \geq 1} \left(\frac{P_k}{1 - P_0}\right)^{Q_k}. \tag{2}
\]

Then, the conditional MLE of \(\alpha\) and \(\beta\) can be obtained using an iterative numerical procedure (Santhanam, 1972, 1977). The mean detection probability of unobserved species can be obtained as:

\[
\hat{P}_0 = \frac{\Gamma(\hat{\alpha} + 1)}{\Gamma(\hat{\alpha} + T + 1)},
\]

Thus, based on Horvitz-Thompson theory, the MLE of richness can be obtained by:

\[
\hat{S}_{MLE} = \frac{S_{obs}}{1 - \hat{P}_0}. \tag{3}
\]

Next, use the MLE asymptotic approach to derive the variance estimator of the MLE (Equation 3). The variance decomposition formula can be applied to \(\hat{S}_{MLE}\) with respect to the observed number of species \(S_{obs}\) and expressed as:

\[
\text{Var}(\hat{S}_{MLE}) = E\left[\text{Var}(\hat{S}_{MLE} | S_{obs})\right] + \text{Var}\left[E(\hat{S}_{MLE} | S_{obs})\right]
\]

\[
\approx E(S_{obs}^2) \text{Var}(\hat{g}(\hat{\alpha}, \hat{\beta}) | S_{obs}) + \text{Var}(S_{obs}) E\left[\hat{g}(\hat{\alpha}, \hat{\beta}) | S_{obs}\right]^2,
\]

where \(\hat{g}(\hat{\alpha}, \hat{\beta}) = \frac{S_{obs}}{S_{obs} + 1}\) is a function of \(\hat{\alpha}\) and \(\hat{\beta}\). Additionally, \(S_{obs}\) approximately follows a binomial distribution with parameters \(S\) and \(E[S_{obs}]/S\). Therefore, \(\text{Var}(S_{obs})\) can be estimated by

\[
E\left[Q_k\right] = S \times P(Y=k) = S \sum_{k=0}^T \left\{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + k + 1)} \frac{\Gamma(\beta + T - k)}{\Gamma(\beta + T + 1)}\right\}.
\]

In this manuscript, we took another approach to estimate parameters of the beta-binomial model. On the basis of the ML method, the estimator of species richness uses all observed species in the sample, including frequently and rarely detected species, to estimate undetected richness. However, the MLE of richness is only robust when the model assumption is met. When the model assumption is violated, the MLE will become very unreliable (see Section 2.2 section). To modify the unreliability of MLE, we adopt a nonparametric concept. According to the Good–Turing frequency formula theory (Good, 1953, 2000) or the Cauchy–Schwarz inequality, the following inequality is held:

\[
\frac{TE[Q_0]}{E[Q_1]} \geq \frac{(T - 1)E[Q_1]}{2E[Q_2]} \geq \frac{(T - 2)E[Q_2]}{3E[Q_3]} \geq \frac{(T - 3)E[Q_3]}{4E[Q_4]} \geq \ldots
\]

This means that the rarer species could supply a more constrained range for undetected richness (i.e. rare species in a sample provide the most information about undetected species). Although some proposed nonparametric estimators use the entire sample to estimate undetected richness (Chao & Lee, 1992; Jiménez-Gamero & Puig, 2020; Puig & Kokonenj, 2018), however, for highly heterogeneous communities, Chao and Yang (1993) modified the estimator derived by Chao and Lee (1992) using the subset of the sample that consisted of only rare species to obtain a more robust estimator based on the same nonparametric approaches. That is why many widely used nonparametric estimating approaches use only rare species in the sample to estimate undetected richness (Burnham & Overton, 1978, 1979; Chao, 1984, 1987; Chiu et al., 2014).

Therefore, on the basis that rare species in the sample contain the most information on undetected richness, instead of using the ML method, I employed the statistical method of moments to derive a new richness estimator using only information on the rare species in a sample. Based on the marginal distribution of the species count in the sample, Equation 1, the expectation of the frequency counts can be expressed as:

\[
E[Q_k] = S \times P(Y=k) = S \sum_{k=0}^T \frac{\Gamma(\alpha + \beta) \Gamma(\alpha + k) \Gamma(\beta + T - k)}{\Gamma(\alpha) \Gamma(\beta) \Gamma(\alpha + \beta + T)}.
\]
We can now obtain the following expectations of undetected richness, singletons, doubletons and tripletons:

\[
E[Q_0] = S \left( \frac{T}{0} \right) \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \frac{\Gamma(\alpha + 1) \Gamma(\beta + T)}{\Gamma(\alpha + \beta + T)},
\]

(4.1)

\[
E[Q_1] = S \left( \frac{T}{1} \right) \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + \beta + T)} \frac{\Gamma(\alpha + 2) \Gamma(\beta + T - 1)}{\Gamma(\alpha + 2) \Gamma(\beta + T - 2)},
\]

(4.2)

\[
E[Q_2] = S \left( \frac{T}{2} \right) \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + \beta + T)} \frac{\Gamma(\alpha + 3) \Gamma(\beta + T - 2)}{\Gamma(\alpha + 3) \Gamma(\beta + T - 3)},
\]

(4.3)

\[
E[Q_3] = S \left( \frac{T}{3} \right) \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + \beta + T)} \frac{\Gamma(\alpha + 3) \Gamma(\beta + T - 2)}{\Gamma(\alpha + 3) \Gamma(\beta + T - 3)},
\]

(4.4)

When the number of the sampling unit (i.e. \(T\)) is sufficiently large, we have the following approximating equalities from Equation 4.1 to 4.4:

\[
E[Q_0] \approx \frac{\Gamma(\alpha + \beta + T)}{\Gamma(\alpha + \beta + 1)} \frac{1}{\Gamma(\beta + T - 1)} \approx \frac{1}{\alpha} \left( \beta + 1 \right),
\]

(5.1)

\[
E[Q_1] \approx \frac{\Gamma(\alpha + 1) \Gamma(\beta + 1)}{\Gamma(\alpha + 2) \Gamma(\beta + 1)} \approx \frac{2}{\alpha + 1} \left( \beta + 1 \right),
\]

(5.2)

\[
E[Q_2] \approx \frac{\Gamma(\alpha + 2) \Gamma(\beta + 1)}{\Gamma(\alpha + 3) \Gamma(\beta + 1)} \approx \frac{3}{\alpha + 2} \left( \beta + 1 \right),
\]

(5.3)

Then, from Equations 5.2 and 5.3, I formulate the following approximate equation:

\[
\frac{E[Q_1]E[Q_3]}{E[Q_2]^2} \approx \frac{\frac{\alpha + 1}{\alpha + 2}}{\alpha + 2} = \frac{2}{\alpha + 1} \left( 1 + \frac{1}{\alpha + 1} \right).
\]

Therefore, the parameter \(\alpha\) in the mixed binomial model can be approximately expressed as:

\[
\alpha \approx \frac{4E[Q_1]^2 - 3E[Q_1]E[Q_3]}{3E[Q_1]E[Q_3] - 2E[Q_2]^2}.
\]

Thus, the estimator of \(\alpha\) can be obtained as \(\hat{\alpha} = \frac{4Q_1^2 - 3Q_1Q_3}{3Q_1Q_3 - 2Q_2^2}\). To ensure \(\hat{\alpha} > 0\), the 3Q_1, Q_3 should be within the range of \(\left(2Q_2^2, 4Q_2^2\right)\). Next, according to Equations 5.1 and 5.2, we have the approximate relationship of \(E[Q_0], E[Q_1], E[Q_2]\) and shown as:

\[
\frac{E[Q_0]}{E[Q_1]} \approx \frac{1}{2\alpha}.
\]

Therefore, the expectation of undetected richness \(Q_0\) can be approximately presented as:

\[
E[Q_0] \approx \left( \frac{E[Q_1]^2}{2E[Q_2]} \left( 1 + \frac{1}{\alpha} \right) \right).
\]

(6)

Here, \(\frac{E[Q_1]^2}{2E[Q_2]}\) is the lower bound of \(E[Q_0]\) when \(T\) is large enough (Chao, 1987) and applies to all mixed binomial models. When expectations are replaced by the observed numbers, \(\hat{\alpha}\) is the widely used richness estimator (named as Chao2) in the literature (Chao, 1987).

Therefore, according to Equation 6, this implies \(-\frac{E[Q_1]^2}{2E[Q_2]} \leq \hat{\alpha} \leq 1\) and the assumption that the species detection rate follows a beta-binomial distribution. Since the Cauchy-Schwarz inequality can also show the inequality, \(\sum_{i=1}^{s} \frac{p_i(1-p_i)^{\beta-1} - \sum_{i=1}^{s} p_i(1-p_i)^{\beta-2}}{\left( \sum_{i=1}^{s} p_i^2 \right)^{\beta-2}} \geq \left( \sum_{i=1}^{s} p_i^2 \right)^{\beta-2}\).

This implies the inequality \(2E[Q_2] \leq 3E[Q_1]E[Q_3]\) when \(T\) is large enough. Then, according to the inequality above and Equation 6, the lower bound of absolute amplitude of bias of \(\frac{Q_1}{Q_2}\) is approximately equal to:

\[
\frac{E[Q_1]}{E[Q_2]} \approx \frac{3E[Q_1]E[Q_3] - 2E[Q_2]^2}{2E[Q_2] \left( 1 - \frac{2E[Q_2]^2}{3E[Q_1]E[Q_3]} \right)}.
\]

To obtain a more stable estimate and ensure \(\hat{\alpha} > 0\), which needs to meet the requirement of \(\frac{1}{2} \leq \frac{Q_1}{Q_2} \leq 1\), the newly proposed estimator is expressed as:

\[
\hat{S}_{\text{BB}} = S_{\text{obs}} + \hat{Q}_{0, \text{Chao2}} \left( 2 - \left( \frac{2Q_2^2}{3Q_1Q_3} \right)^\gamma \right),
\]

(7)

where \(\hat{Q}_{0, \text{Chao2}} = \left\{ \begin{array}{ll} \frac{T - 3\hat{Q}_1}{T} \quad \text{if } Q_2 > 0, \\ \frac{T - 3\hat{Q}_1}{T} \quad \text{if } Q_2 = 0 \end{array} \right. \) and the expression \(F^{-}\) is equal to \(\frac{1}{2} \left( \frac{1}{F} \right) \) if \(F \leq 1\) and \(1 + 1\) if otherwise. Here, \(Q_1\) or \(Q_3\) equal to zero will be replaced by 1 to make Equation 7 well defined. Although the newly proposed estimator (Equation 7) was derived based on the beta-binomial model assumption, it can also be interpreted as a bias-corrected estimator of Chao2 and be directly derived using the Good-Turing frequency formula (see Appendix S1 for details).

Since the newly proposed estimator is expressed in a simple closed form, without using the resampling method, the statistical delta method could be used to derive the estimator of variance. According to the asymptotic approximated distribution, in which
(Q₀, Q₁, ..., Qₗ) follows a multinomial distribution with a total size S and cell probabilities \( \frac{Q₀}{S}, \frac{Q₁}{S}, ..., \frac{Qₗ}{S} \) (Chao & Lee, 1992; Chao & Yang, 1993), the variance estimators of the aforementioned richness estimators (\( \hat{S} \)) can all be derived using the delta method. Thus, we have:

\[
\text{Var}(\hat{S}) \approx \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \frac{\partial \hat{S}}{\partial Q_i} \frac{\partial \hat{S}}{\partial Q_j} \text{cov}(Q_i, Q_j),
\]

where \( \text{cov}(Q_i, Q_j) = \begin{cases} Q_i \left( 1 - \frac{Q_i}{\hat{S}} \right) & \text{if } i = j \\ -Q_i/Q_j & \text{if } i \neq j \end{cases} \)

To derive the 95% CI of species richness and to ensure the lower bound of 95% CI of species richness is larger than observed richness, assume \( \hat{S} - \hat{S}_{obs} \) follows a log normal distribution (Chiu et al., 2014). Then the 95% CI of species richness is obtained as [\( \hat{S}_{obs} + S_{BB} \cdot \hat{S}_{obs} + (\hat{S} - \hat{S}_{obs}) \cdot R \)], where

\[
R = \exp \left\{ 1.96 \left[ \log \left( 1 + \frac{\text{Var}(\hat{S})}{(\hat{S} - \hat{S}_{obs})^2} \right) \right]^{1/2} \right\}
\]

From the process of derivation, the new richness estimator has the following properties: (i) Instead of using all species information in the sample as in other parametric approaches, the new estimator only uses the number of singletons, doubletons and tripletons to estimate undetected richness, and can be interpreted as a bias-corrected estimator of the Chao2 lower bound estimator. (ii) Although the newly proposed estimator derived based on a stochastic model can be classified as a parametric estimating method, it can also be classified as a nonparametric approach since the proposed estimator can be directly derived using the Good–Turing frequency formula without model assumption. (iii) The new estimator was shown as a nearly unbiased estimator in a homogeneous model (proved in Appendix S2) and would asymptotically converge to the true richness as the number of plots \( T \) increases.

| Table 1: Homogeneous model with mean detection probability of 0.05 and CV = 0 |  |  |  |  |  |
|---|---|---|---|---|---|
| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
| 10 (200.4) | \( \hat{S}_{\text{Chao2}} \) | 499.7 | −0.3⁴ | 67.3 | 70.2 | 67.2⁴ | 0.96⁴ |
|  | \( \hat{S}_{\text{SBB}} \) | 557.8 | 57.8 | 82.6 | 92.1 | 100.7 | 0.91 |
|  | \( \hat{S}_{\text{Jack1}} \) | 357 | −143 | 18.7 | 16.8 | 144.2 | 0 |
|  | \( \hat{S}_{\text{MLE}} \) | 9908.1 | 9408.1 | 65845.6 | 16928.2 | 66351.1 | 0.77 |
|  | \( \hat{S}_{\text{BB}} \) | 521.4 | 21.4 | 110 | 112.1 | 111.5 | 0.94⁴ |
| 20 (321.4) | \( \hat{S}_{\text{Chao2}} \) | 502 | 2⁴ | 32.5 | 34.9 | 32.5 | 0.96⁴ |
|  | \( \hat{S}_{\text{SBB}} \) | 516.8 | 16.8 | 43.5 | 46.9 | 46.6 | 0.92 |
|  | \( \hat{S}_{\text{Jack1}} \) | 509.9 | 9.9 | 17.8 | 18.9 | 20.4⁴ | 0.94⁴ |
|  | \( \hat{S}_{\text{MLE}} \) | 520 | 20 | 42.5 | 36.1 | 46.9 | 0.86 |
|  | \( \hat{S}_{\text{BB}} \) | 509.9 | 9.9 | 49.4 | 53.4 | 50.4 | 0.94⁴ |
| 40 (435.7) | \( \hat{S}_{\text{Chao2}} \) | 502 | 2⁴ | 16.4 | 15 | 16.5 | 0.94⁴ |
|  | \( \hat{S}_{\text{SBB}} \) | 506.3 | 6.3 | 20.5 | 18.7 | 21.4 | 0.92 |
|  | \( \hat{S}_{\text{Jack1}} \) | 571.9 | 71.9 | 14.4 | 16.3 | 73.3 | 0 |
|  | \( \hat{S}_{\text{MLE}} \) | 505.9 | 5.9 | 13.9 | 11.1 | 15⁴ | 0.87 |
|  | \( \hat{S}_{\text{BB}} \) | 504.9 | 4.9 | 24.2 | 20.7 | 24.7 | 0.94⁴ |
| 80 (491.7) | \( \hat{S}_{\text{Chao2}} \) | 500.2 | 0.2⁴ | 4 | 4.2 | 4 | 0.96⁴ |
|  | \( \hat{S}_{\text{SBB}} \) | 508.8 | 0.8 | 4.6 | 4.9 | 4.6 | 0.92 |
|  | \( \hat{S}_{\text{Jack1}} \) | 526.4 | 26.4 | 6.2 | 8.2 | 27.1 | 0 |
|  | \( \hat{S}_{\text{MLE}} \) | 500.6 | 0.6 | 2.9 | 3.2 | 3⁴ | 0.9 |
|  | \( \hat{S}_{\text{BB}} \) | 500.3 | 0.3 | 5 | 5.4 | 5 | 0.94⁴ |

⁴Denotes the least bias, lowest RMSE, and closest to 95% coverage.

Abbreviation: CI, confidence interval; CV, coefficient of variation; RMSE, root mean squared error; SE, standard error.
in other models (proved in Appendix S3). (iv) The estimator provides a lower bound estimator of species richness under the beta-binomial model, which is a relatively flexible ecological model that includes the broken stick and homogeneous models.

### 2.2 Simulation study

A simulation study was conducted to investigate the statistical behaviour of the new estimator and compare it with current estimators, namely the nonparametric-based Chao2 lower bound estimator (Chiu et al., 2014) and first-order jackknife estimator, and the parametric-based MLE. Where the ML estimates of parameters in the beta-binomial model can be efficiently calculated using the R function in the package 'extraDistir'.

(i) The Chao2 estimator ($\hat{S}_{\text{obs}} + \frac{T-1}{T} Q_1$, denoted as $\hat{S}_{\text{Chao}}$): the numbers of singletons and doubletons are used to estimate the number of undetected species.

(ii) The improved Chao2 estimator ($\hat{S}_{\text{obs}} + \frac{T-1}{T} Q_1 + \frac{T-3}{2T} Q_2$, denoted as $\hat{S}_{\text{Chao,impr}}$): compare to Chao2 estimator, this estimator uses additional information (namely, the numbers of tripletons and quadrupletons) to estimate the number of undetected species and correct the negative bias of Chao2 (Chiu et al., 2014).

(iii) The first-order jackknife estimator ($\hat{S}_{\text{obs}} + \frac{T-1}{T} Q_1$, denoted as $\hat{S}_{\text{Jack}}$): the number of singletons is used to estimate the number of undetected species.

(iv) The MLE (see Equation 3, denoted as $\hat{S}_{\text{MLE}}$): parameters are estimated by using ML method and estimate species richness using the Horvitz-Thompson theory.

(v) New proposed richness estimator (see Equation 7, denoted as $\hat{S}_{\text{BB}}$): the number of singletons, doubletons and tripletons are used to estimate the number of undetected species.

The simulation results were calculated from six widely used ecological species detection models. The number of species in each model was determined, where $c$ is a rescaling constant such that the maximum detection probability is a fixed constant. The coefficient of variation (CV) of the number of undetected species is also calculated to indicate the degree of heterogeneity of $\{x_1, x_2, \ldots, x_{\bar{n}}\}$.

### TABLE 2 Random uniform model with mean detection probability of 0.25 and CV = 0.58

| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
|---------------------------|-----------|------------------|------|-----------|----------------------|------------|---------------------|
| 10 (408.7)                | $\hat{S}_{\text{Chao}}$ | 452              | -48  | 13.2      | 12.2                  | 49.8       | 0.13                |
|                           | $\hat{S}_{\text{Chao}}$ | 462.2             | -37.8| 14.9      | 14.5                  | 40.6       | 0.43                |
|                           | $\hat{S}_{\text{Jack}}$ | 498.9             | -1.1a| 12.2      | 12.8                  | 12.2a      | 0.96a               |
|                           | $\hat{S}_{\text{MLE}}$ | 456.8             | -43.2| 12.8      | 12.8                  | 45.1       | 0.22                |
|                           | $\hat{S}_{\text{BB}}$  | 472.5             | -27.5| 20.5      | 20.2                  | 34.3       | 0.83                |
| 20 (451.1)                | $\hat{S}_{\text{Chao}}$ | 476.1             | -23.9| 11        | 9.5                   | 26.3       | 0.5                 |
|                           | $\hat{S}_{\text{Chao}}$ | 481.7             | -18.3| 12.6      | 11.1                  | 22.2       | 0.72                |
|                           | $\hat{S}_{\text{Jack}}$ | 499.4             | -0.6a| 9.7       | 9.6                   | 9.7b       | 0.91                |
|                           | $\hat{S}_{\text{MLE}}$ | 471.4             | -28.6| 7.7       | 6.3                   | 29.6       | 0.1                 |
|                           | $\hat{S}_{\text{BB}}$  | 488.8             | -11.2| 16.6      | 14.9                  | 20.0       | 0.93a               |
| 40 (475.6)                | $\hat{S}_{\text{Chao}}$ | 489.6             | -10.4| 8.3       | 7.1                   | 13.3       | 0.78                |
|                           | $\hat{S}_{\text{Chao}}$ | 492.7             | -7.3 | 9.5       | 8.6                   | 12         | 0.92                |
|                           | $\hat{S}_{\text{Jack}}$ | 501.9             | 1.9a | 7.3       | 7.1                   | 7.5b       | 0.92                |
|                           | $\hat{S}_{\text{MLE}}$ | 484.7             | -15.3| 4.7       | 3.7                   | 16         | 0.18                |
|                           | $\hat{S}_{\text{BB}}$  | 495.4             | -4.6 | 11.3      | 11                    | 12.2       | 0.93a               |
| 80 (489.3)                | $\hat{S}_{\text{Chao}}$ | 497               | -3  | 6.4       | 5.4                   | 7a         | 0.95a               |
|                           | $\hat{S}_{\text{Chao}}$ | 498.5             | -1.5 | 7.2       | 6.4                   | 7.3        | 0.94                |
|                           | $\hat{S}_{\text{Jack}}$ | 502.8             | 2.8  | 4.8       | 5.1                   | 5.6        | 0.84                |
|                           | $\hat{S}_{\text{MLE}}$ | 493.6             | -6.4 | 3.1       | 2.3                   | 7.1        | 0.6                 |
|                           | $\hat{S}_{\text{BB}}$  | 499.9             | -0.1a| 8.8       | 8                     | 8.8        | 0.96                |

*aDenotes the least bias, lowest RMSE, and closest to 95% coverage.

Abbreviations: CI, confidence interval; CV, coefficient of variation; RMSE, root mean squared error; SE, standard error.
Model 1: the homogeneous model (CV = 0): where \( x_i = 0.05, i = 1, 2, \ldots, S \), and there is no heterogeneity among species detection probabilities.

Model 2: the random uniform model (CV = 0.58): where \( x_i = ca_i, i = 1, 2, \ldots, S \), and \((a_1, a_2, \ldots, a_S)\) is a random sample from a uniform distribution with the parameters \((0, 1)\), and the maximum of \( x_i \) is set as 0.5.

Model 3: the broken stick model (CV = 1.03): where \( x_i = ca_i, i = 1, 2, \ldots, S \), and \((a_1, a_2, \ldots, a_S)\) is a random sample from an exponential distribution with a parameter of 1, and the maximum of \( x_i \) is set as 0.5. This model is commonly used in previous literature and equivalent to the Dirichlet distribution.

Model 4: the log-normal model (CV = 1.34): where \( x_i = ca_i, i = 1, 2, \ldots, S \), and \((a_1, a_2, \ldots, a_S)\) is a random sample from a log-normal distribution, and the maximum of \( x_i \) is set as 1.

Model 5: the Zipf–Mandelbrot model (CV = 1.84): where \( x_i = \left(\frac{c}{i+5}\right), i = 1, 2, \ldots, S \), is a commonly used model in ecology and linguistics (Magurran, 2004), and the maximum of \( x_i \) is set as 0.5.

Model 6: the Beta model (CV = 0.55): where \( x_i = ca_i, i = 1, 2, \ldots, S \), and \((a_1, a_2, \ldots, a_S)\) is a random sample from a beta distribution with the parameters \((2, 5)\) that meet our model assumption, and the maximum of \( x_i \) is set as 0.25.

The CV in these six models ranged from 0 to 1.84, covering most practical cases in real applications. Here, I considered four different quadrat numbers, 10, 20, 40, and 80, as different sampling sizes or efforts, and a total of 24 model-size combinational scenarios were generated. For each discussed estimator in each simulation scenario, the estimate and estimated standard error (SE) were averaged over 1000 simulated datasets to obtain the mean estimate and the mean estimated SE. The sample SE and RMSE over these 1000 estimates were also obtained. The percentage of the datasets in which the 95% CIs covered the true value is detailed in the tables (Tables 1–6). The average of observed richness over 1000 samples is also presented.

### RESULTS

#### 3.1 Simulation results

In general, a good species richness estimator should have as small bias and variance as possible (i.e. low RMSE). Moreover, the coverage rate of its associated 95% CI should be close to 0.95. As sample size increases, the estimator should also exhibit the following essential patterns: its bias, accuracy (as measured by RMSE), and coverage

| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
|---------------------------|-----------|------------------|------|-----------|---------------------|-------------|---------------------|
| 10 (210.4)                | \( \hat{\text{S}}_{\text{Chao2}} \) | 347.4            | -152.6 | 33.1      | 33.2                | 156.2       | 0.08                |
|                           | \( \hat{\text{S}}_{\text{Chao2}} \) | 374.8            | -125.2 | 39.7      | 40.7                | 131.3       | 0.31                |
|                           | \( \hat{\text{S}}_{\text{Jack}} \)  | 334.1            | -165.9 | 16.5      | 14.9                | 166.7       | 0                   |
|                           | \( \hat{\text{S}}_{\text{MLE}} \)  | 9333.7           | 8833.7 | 39,721.8  | 4047.7              | 40,595.2    | 0.92*               |
|                           | \( \hat{\text{S}}_{\text{BB}} \)   | 405.1            | -94.9  | 59.5      | 59.8                | 112.0                   | 0.80                |
| 20 (293.0)                | \( \hat{\text{S}}_{\text{Chao2}} \) | 390.4            | -109.6 | 24.3      | 23.4                | 112.0       | 0.08                |
|                           | \( \hat{\text{S}}_{\text{Chao2}} \) | 412.6            | -87.4  | 28.5      | 28.4                | 91.9        | 0.3                 |
|                           | \( \hat{\text{S}}_{\text{Jack}} \)  | 412.2            | -87.8  | 16        | 15                  | 89.2        | 0                   |
|                           | \( \hat{\text{S}}_{\text{MLE}} \)  | 555              | 55^a   | 99.7      | 110.6               | 113.6       | 0.98*               |
|                           | \( \hat{\text{S}}_{\text{BB}} \)   | 433.1            | -66.9  | 40.7      | 42.2                | 78.3        | 0.85                |
| 40 (366.1)                | \( \hat{\text{S}}_{\text{Chao2}} \) | 434.1            | -65.9  | 18.8      | 18.1                | 68.5        | 0.19                |
|                           | \( \hat{\text{S}}_{\text{Chao2}} \) | 449.9            | -50.1  | 22        | 21.7                | 54.7        | 0.54                |
|                           | \( \hat{\text{S}}_{\text{Jack}} \)  | 462.3            | -37.7  | 13.9      | 13.7                | 40.2        | 0.27                |
|                           | \( \hat{\text{S}}_{\text{MLE}} \)  | 530.6            | 30.6^a | 42.9      | 47.4                | 52.6        | 0.94*               |
|                           | \( \hat{\text{S}}_{\text{BB}} \)   | 464.7            | -35.3  | 30.6      | 31.4                | 46.7        | 0.89                |
| 80 (421.3)                | \( \hat{\text{S}}_{\text{Chao2}} \) | 462              | -38    | 13.9^a    | 13.1                | 40.4        | 0.4                 |
|                           | \( \hat{\text{S}}_{\text{Chao2}} \) | 471.2            | -28.8  | 16.2      | 15.9                | 33.1        | 0.7                 |
|                           | \( \hat{\text{S}}_{\text{Jack}} \)  | 486.6            | -13.4^a| 10.9      | 11.3                | 17.3        | 0.85                |
|                           | \( \hat{\text{S}}_{\text{MLE}} \)  | 516.8            | 16.8   | 19.2      | 23.6                | 25.5        | 0.92                |
|                           | \( \hat{\text{S}}_{\text{BB}} \)   | 487.6            | -24.4  | 22.1      | 21.4                | 25.3        | 0.93*               |

*Denotes the least bias, lowest RMSE, and closest to 95% coverage.

Abbreviations: CI, confidence interval; CV, coefficient of variation; RMSE, root mean squared error; SE, standard error.
rate of the CI should generally improve as sample size increases, and its estimates should finally converge to the true species richness when sample size is large enough. Using these criteria and according to the simulation results, we concluded the following findings:

1. In some cases (see Table 2), the jackknife estimator performed well in terms of bias and RMSE. However, as an estimator of species richness, the jackknife estimator violated the essential criteria that bias and accuracy (quantified by RMSE) should decrease with sample size (Tables 1 and 6). These results are consistent with those of Chiu et al. (2014). Of the discussed estimators, the jackknife estimator usually had the smallest SE, indicating favourable robustness; however, its coverage rate of CI is significantly lower than the claimed level, especially in cases with high heterogeneity (Tables 3–6).

2. Chao2 is almost unbiased in the homogeneous model (Table 1). However, its negative bias and CI coverage rate become worse as the heterogeneity of the species detection rates increased (Tables 2–6). Overall, Chao2 is advantageous as a richness estimator for its bias, RMSE and 95% CI coverage rate that perform better with increasing sample size (Tables 1–6) and tends to derive true species richness when sample size is sufficiently large. Compared with the newly proposed estimator, Chao2 had a smaller SE, larger bias and larger RMSE in the discussed hypothetical models (Tables 2–5).

3. The improved Chao2 (denoted as iChao2), a bias-corrected Chao2 estimator, fulfils the essential properties of a richness estimator (Tables 1–6). Compare to Chao2, iChao2 had a smaller bias, lower RMSE and more accurate 95% CI than Chao2 for most discussed hypothetical models (Tables 2–6). However, for the model with high CV (i.e. high heterogeneity), iChao2 still presented seriously negative bias and a low coverage rate of 95% CI when sample size is not large enough (Tables 3–5).

4. The MLE performs well in relation to bias, RMSE and 95% CI coverage rate when the hypothetical model is of low CV and sample size is large enough (Tables 1 and 2). However, when the model has a high CV, the MLE will tend to be unstable due to high variance (Tables 3–5). If beta distribution is assumed, the MLE performs well in terms of bias and RMSE (Table 6).

5. Generally, compared to other discussed estimators, the proposed estimator exhibits robust performance in most simulation cases.

### Table 4 Log-normal model with mean detection probability of 0.078 and CV = 1.34

| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
|---------------------------|-----------|------------------|------|-----------|---------------------|-------------|---------------------|
| 10 (211.8)                | \( \hat{S}_{\text{Chao}} \) | 362.7            | −137.3 | 37.4 | 36.9 | 142.3 | 0.18 |
|                           | \( \hat{S}_{\text{iChao}} \) | 392.9            | −107.1 | 43.2 | 43.7 | 115.4 | 0.46 |
|                           | \( \hat{S}_{\text{Jack}} \) | 336.3            | −163.7 | 16.2 | 15 | 164.5 | 0 |
|                           | \( \hat{S}_{\text{MLE}} \) | 151,528.7 | 151,028.7 | 32,361.1 | 52,827.3 | 154,439.9 | 0.03 |
|                           | \( \hat{S}_{\text{BB}} \) | 425.3 | −74.7 | 68.8 | 65.6 | 101.6 | 0.88 |
| 20 (299.5)                | \( \hat{S}_{\text{Chao}} \) | 419.9            | −80.1 | 27.7 | 27.7 | 84.7 | 0.38 |
|                           | \( \hat{S}_{\text{iChao}} \) | 447.3            | −52.7 | 32 | 33.3 | 61.6 | 0.71 |
|                           | \( \hat{S}_{\text{Jack}} \) | 431.2            | −68.8 | 16.8 | 15.8 | 70.8 | 0.03 |
|                           | \( \hat{S}_{\text{MLE}} \) | 128,063.8 | 127,563.8 | 59,954.4 | 34,760.8 | 140,886.8 | 0.01 |
|                           | \( \hat{S}_{\text{BB}} \) | 463.5 | −36.5 | 47.8 | 50.7 | 60.1 | 0.92 |
| 40 (381.8)                | \( \hat{S}_{\text{Chao}} \) | 458.5            | −41.5 | 19.1 | 19.1 | 45.7 | 0.55 |
|                           | \( \hat{S}_{\text{iChao}} \) | 476 | −24 | 23.3 | 23.5 | 33.4 | 0.88 |
|                           | \( \hat{S}_{\text{Jack}} \) | 491.6            | −8.4 | 13.3 | 14.6 | 15.7 | 0.96 |
|                           | \( \hat{S}_{\text{MLE}} \) | 5353.6 | 4853.6 | 21,439.2 | 1915.1 | 21,929.4 | 0.05 |
|                           | \( \hat{S}_{\text{BB}} \) | 485.9 | −14.1 | 33.8 | 32.4 | 36.6 | 0.92 |
| 80 (444.6)                | \( \hat{S}_{\text{Chao}} \) | 483.2            | −16.8 | 12.4 | 12 | 20.9 | 0.81 |
|                           | \( \hat{S}_{\text{iChao}} \) | 491.9            | −8.1 | 14.9 | 15 | 16.9 | 0.96 |
|                           | \( \hat{S}_{\text{Jack}} \) | 515.4            | 15.4 | 10.9 | 11.8 | 18.8 | 0.66 |
|                           | \( \hat{S}_{\text{MLE}} \) | 800 | 300 | 70.9 | 114.8 | 308.3 | 0 |
|                           | \( \hat{S}_{\text{BB}} \) | 497.6 | −2.4 | 19.5 | 19.5 | 19.6 | 0.94 |

*Denotes the least bias, lowest RMSE, and closest to 95% coverage.

Abbreviations: CI, confidence interval; CV, coefficient of variation; RMSE, root mean squared error; SE, standard error.
scenarios, with a reduced bias, a smaller RMSE and higher accuracy with regard to 95% CI coverage rate. This method adheres to the essential principle that the RMSE and bias should decrease with sample size. Except in the homogeneous model (Table 1), the new estimator reduces the negative bias and RMSE of Chao2 and iChao2 resulting from a small sample size or high heterogeneity of species detection rate (Tables 2–6). Although the SE of the newly proposed estimator is always larger than that of Chao2 and iChao2, its 95% CI coverage rate tends to be more highly accurate in most simulation scenarios (Tables 2–6).

6. The asymptotic approach used to estimate the variances of the discussed estimators performs well in most simulation cases except in some cases for the MLE (Tables 1–6).

### 3.2 | Real data analysis

#### 3.2.1 | Work example 1: Barro Colorado Island Forest dataset

Here, I used one real dataset of forest as a true community to evaluate and compare our estimator with other discussed estimators. The real forest dataset was from the 50 ha (1000 × 500 m) Barro Colorado Island (BCI) plot, Panama, censused in 1985 (Condit et al., 2012). In the plot, all individual trees and shrubs (≥ 1 cm diameter at breast height) were recorded and identified to species. There is a total of 238,018 individual trees belonging to 298 species in the BCI census dataset. Here, we fix the size of the quadrat as 20 × 20 m² and divide the BCI plot into 1250 quadrats. Then, we have the mean detection probability of species and CV of species detection probability as separately 0.175 and 1.36.

Four different numbers of sampling quadrats, namely 20, 40, 80 and 160, were implemented using the sampling with replacement scheme to investigate the effect of sampling effort on discussed richness estimators. For each sampling setting, 500 replicated samples were generated and only species presence–absence data in each quadrat were used for analyses. Then for each generated sample, we recorded the number of observed species and calculated the five discussed estimators: Chao2 estimator, iChao2 estimator, first-order Jackknife estimator, MLE and the newly proposed estimator. Finally, the estimate and their estimated SE were averaged over 500 simulated datasets to obtain the mean estimate and mean estimated SE. The sample SE, RMSE and coverage rate of 95% CI over these 500 estimates are also shown in Table 7.

#### Table 5 Zipf–Mandelbrot model with mean detection probability of 0.027 and CV = 1.84

| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
|---------------------------|-----------|------------------|------|-----------|----------------------|-------------|---------------------|
| 10 (93.9)                 | \(\hat{\text{S}}_{\text{Chao2}}\) | 265.6             | -234.4 | 74.6      | 62.6                 | 246         | 0.26                |
|                           | \(\hat{\text{S}}_{\text{iChao2}}\) | 287.2             | -212.8 | 80.6      | 68.1                 | 227.5       | 0.33                |
|                           | \(\hat{\text{S}}_{\text{Jack1}}\) | 164.1             | -335.9 | 15.6      | 11.2                 | 336.3       | 0                   |
|                           | \(\hat{\text{S}}_{\text{MLE}}\)   | 110,174.5         | 109,674.5 | 39,479.2 | 37,854.8             | 116,530.3   | 0.06                |
|                           | \(\hat{\text{S}}_{\text{BB}}\)    | 351.5             | -148.5  | 119.4     | 102.1                | 190.5       | 0.83                |
| 20 (152.4)                | \(\hat{\text{S}}_{\text{Chao2}}\) | 340.5             | -159.5  | 58        | 53.6                 | 169.7       | 0.36                |
|                           | \(\hat{\text{S}}_{\text{iChao2}}\) | 370.9             | -129.1  | 64        | 61.7                 | 144         | 0.54                |
|                           | \(\hat{\text{S}}_{\text{Jack1}}\) | 253.8             | -246.2  | 18.2      | 13.9                 | 246.9       | 0                   |
|                           | \(\hat{\text{S}}_{\text{MLE}}\)   | 131,739.5         | 131,239.5 | 17,327    | 30,317.1             | 132,372.7   | 0                   |
|                           | \(\hat{\text{S}}_{\text{BB}}\)    | 424.4             | -75.6   | 97.1      | 95.2                 | 123.1       | 0.91                |
| 40 (232.2)                | \(\hat{\text{S}}_{\text{Chao2}}\) | 409               | -91     | 43.5      | 42.3                 | 100.9       | 0.5                 |
|                           | \(\hat{\text{S}}_{\text{iChao2}}\) | 444.6             | -55.4   | 51.2      | 50.2                 | 75.4        | 0.83                |
|                           | \(\hat{\text{S}}_{\text{Jack1}}\) | 362.3             | -137.7  | 17.6      | 15.9                 | 138.8       | 0                   |
|                           | \(\hat{\text{S}}_{\text{MLE}}\)   | 144,342.6         | 143,842.6 | 10,640.6 | 33,798               | 144,233.7   | 0                   |
|                           | \(\hat{\text{S}}_{\text{BB}}\)    | 479.6             | -20.4   | 79.9      | 75.2                 | 82.5        | 0.92                |
| 80 (327.7)                | \(\hat{\text{S}}_{\text{Chao2}}\) | 461.1             | -38.9   | 30.2      | 29.5                 | 49.2        | 0.77                |
|                           | \(\hat{\text{S}}_{\text{iChao2}}\) | 491.5             | -8.5    | 36.4      | 36.6                 | 37.3        | 0.91                |
|                           | \(\hat{\text{S}}_{\text{Jack1}}\) | 468.9             | -31.1   | 17.7      | 16.7                 | 35.8        | 0.62                |
|                           | \(\hat{\text{S}}_{\text{MLE}}\)   | 152,341.7         | 151,841.7 | 6878.3    | 14,220.5             | 151,996.6   | 0                   |
|                           | \(\hat{\text{S}}_{\text{BB}}\)    | 499.1             | -0.9    | 53.4      | 54.1                 | 53.4        | 0.96                |

\(^a\)Denotes the least bias, lowest RMSE, and closest to 95% coverage.

Abbreviations: CI, confidence interval; CV, coefficient of variation; RMSE, root mean squared error; SE, standard error.
Methods in Ecology and Evolution

CHIU

CHIU

As with the simulation results, the results in Table 7 show that the number of observed species exhibits large negative bias, especially with lower sampling effort. The MLE is severely overestimating due to its unreliability (high variance) in the model with high CV. Although Chao2, iChao2 and first-order Jackknife estimator could reduce the bias of observed richness, the three estimators are still severely underestimating under low sampling effort especially in the small sample size. Compared with the newly proposed estimator, the SEs of Jackknife1, Chao2 and iChao2 are much lower in each sampling effort setting, and the coverage rates of CI of the two estimators are quite low, especially with lower sampling effort. The newly proposed estimator displays more robust performance in each sampling effort setting, which has lower bias and RMSE. Although the new estimator has higher SE in each sampling effort setting, the coverage rate of 95% CI is closer to 95% that is consistent with the results in simulation studies.

### Work example 2: Tropical ant data

I used another real dataset to illustrate the application of the proposed estimator. Here, tropical ant data collected at five elevations (50, 500, 1070, 1500 and 2000 m) in Costa Rica by Longino and Colwell (2011) were employed. Table 8 shows that the numbers of trapping samples at the five elevations were 599, 230, 150, 200 and 200, respectively; the reference sample size (T), observed species richness (S_{obs}), CV of empirical species detection rate and first 10 frequency counts (Q_1 \sim Q_{10}) are also noted. As Table 8 shows, the observed richness roughly decreases as elevation increases, and the observed richness in the elevation of 500 m is the highest. Table 8 also shows that the number of singletons is more than twice the number of doubletons at each elevation, except at the elevation of 2000 m, in which there are still many undetected species in each site.

In Table 9, the estimates and estimated SE of the various richness estimators are described. The results also demonstrated the presence of many undetected species at each elevation except the highest. As with the results in previous simulations, the MLE presents very unreliable estimates in the highly heterogeneous communities (high CV). Therefore, I ignore the ML richness estimates in the following discussions. The estimates of Chao2, the improved Chao2, the Jackknife estimator and the proposed estimator have the same rankings as observed richness along elevations, where the taxa of ants have the highest species richness at the elevation of 50 m and

---

**TABLE 6** Beta (2, 5) model with mean detection probability of 0.09 and CV = 0.55

| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
|---------------------------|-----------|------------------|------|-----------|---------------------|-------------|---------------------|
| 10 (276.8)                | \hat{S}_{Chao2} | 428.2            | -71.8| 33.5      | 31.9                | 79.2        | 0.5                 |
|                           | \hat{S}_{Chao2} | 454.2            | -45.8| 43.1      | 43.1                | 62.8        | 0.82                |
|                           | \hat{S}_{Jack1} | 437.2            | -62.8| 18.3      | 17                  | 65.4        | 0.1                 |
|                           | \hat{S}_{MLE}  | 488.1            | -11.9| 90.7      | 82.4                | 91.2        | 0.8                 |
|                           | \hat{S}_{IB}   | 457.5            | -42.5| 55.1      | 55.4                | 69.6        | 0.84                |
| 20 (374.6)                | \hat{S}_{Chao2} | 455.5            | -44.5| 19.3      | 18.5                | 48.4        | 0.49                |
|                           | \hat{S}_{Chao2} | 472.5            | -27.5| 24.3      | 24.1                | 36.7        | 0.84                |
|                           | \hat{S}_{Jack1} | 506              | 6    | 15.6      | 15.8                | 16.7        | 0.94                |
|                           | \hat{S}_{MLE}  | 477.8            | -22.2| 28        | 26.3                | 35.7        | 0.81                |
|                           | \hat{S}_{IB}   | 477.8            | -22.2| 31.8      | 31.7                | 38.8        | 0.88                |
| 40 (445.1)                | \hat{S}_{Chao2} | 480.2            | -19.8| 11.4      | 10.8                | 22.9        | 0.7                 |
|                           | \hat{S}_{Chao2} | 488.1            | -11.9| 13.6      | 13.5                | 18          | 0.91                |
|                           | \hat{S}_{Jack1} | 519              | 19   | 11.3      | 12                  | 22          | 0.58                |
|                           | \hat{S}_{MLE}  | 486.1            | -13.9| 11        | 10.3                | 17.7        | 0.8                 |
|                           | \hat{S}_{IB}   | 492.5            | -6.5 | 17.4      | 17.6                | 18.9        | 0.91                |
| 80 (479.5)                | \hat{S}_{Chao2} | 493.7            | -6.3 | 6.7       | 6.7                 | 9.2         | 0.94                |
|                           | \hat{S}_{Chao2} | 497              | -3   | 7.9       | 8.2                 | 8.4         | 0.96                |
|                           | \hat{S}_{Jack1} | 511.4            | 11.4 | 6.9       | 7.9                 | 13.3        | 0.52                |
|                           | \hat{S}_{MLE}  | 492.8            | -7.2 | 4.9       | 4.6                 | 8.7         | 0.81                |
|                           | \hat{S}_{IB}   | 498.9            | -1.1 | 9.8       | 10.3                | 9.9         | 0.96                |

*Denotes the least bias, lowest RMSE, and closest to 95% coverage.

Abbreviations: CI, confidence interval; CV, coefficient of variation; RMSE, root mean squared error; SE, standard error.
decrease with elevation. However, the Jackknife estimate showed that the species richness diversity at the five elevations were significantly different, while the Chao2, the improved Chao2 and the proposed estimator demonstrated that the richness at the elevations of 50 and 500 m were not significantly different. Besides that, only the newly proposed estimator showed that there is no significant difference between species richness at the elevations of 1070 and 1500 m at significant level 0.05.

The jackknife estimator had considerably lower estimates at each elevation, and based on results of the previous simulations, the proposed estimator may be prone to underestimation in the heterogeneous model when the sample size is not sufficiently large. As with the simulation results, the jackknife estimator had the smallest SE, but its estimate was neither more accurate nor more robust. Compared to the newly proposed estimator, Chao2 and iChao2 have lower estimates and SEs. As the simulation results show, when the sample is not sufficiently large and the heterogeneity of species composition is high, Chao2 and iChao2 with lower SEs and severely underestimated bias will cause low coverage probability of the 95% CI. In this dataset, the proposed estimator had the largest richness estimates and SE at

### Table 7: The statistical behaviours of the Chao2 ($\hat{S}_{\text{Chao2}}$), the improved Chao2 ($\hat{S}_{\text{iChao2}}$), the first-order jackknife estimator ($\hat{S}_{\text{Jack1}}$), the MLE ($\hat{S}_{\text{MLE}}$) and the proposed estimator ($\hat{S}_{\text{BB}}$)

| Size T (observed richness) | Estimator | Average estimate | Bias | Sample SE | Average estimated SE | Sample RMSE | 95% CI coverage rate |
|----------------------------|-----------|------------------|------|-----------|----------------------|-------------|---------------------|
| 20 (180.8)                 | $\hat{S}_{\text{Chao2}}$ | 214.8            | −83.2 | 14.9      | 13.8                 | 84.5       | 0.05                |
|                           | $\hat{S}_{\text{iChao2}}$ | 221.7            | −76.3 | 16.9      | 16                   | 78.2       | 0.13                |
|                           | $\hat{S}_{\text{Jack1}}$ | 222.4            | −75.6 | 9.4       | 8.9                  | 76.2       | 0                   |
|                           | $\hat{S}_{\text{MLE}}$ | 3412.9           | 3114.9 | 11,182.2  | 2211.6               | 11,597.1   | 0.83*              |
|                           | $\hat{S}_{\text{BB}}$ | 232.4            | −65.6  | 23.3      | 22.4                 | 69.6*      | 0.69                |
| 40 (207.2)                 | $\hat{S}_{\text{Chao2}}$ | 236.1            | −61.9 | 13.5      | 12.6                 | 63.3       | 0.13                |
|                           | $\hat{S}_{\text{iChao2}}$ | 242.1            | −55.9 | 15.4      | 14.7                 | 58         | 0.25                |
|                           | $\hat{S}_{\text{Jack1}}$ | 242.9            | −55.1 | 8.6       | 8.3                  | 55.8       | 0                   |
|                           | $\hat{S}_{\text{MLE}}$ | 578.4            | 280.4 | 272.1     | 376.1                | 390.5      | 0.63                |
|                           | $\hat{S}_{\text{BB}}$ | 254.7            | −43.3  | 20.7      | 20.1                 | 48.0*      | 0.75*               |
| 80 (229.2)                 | $\hat{S}_{\text{Chao2}}$ | 253.8            | −44.2 | 13.1      | 11.8                 | 46.1       | 0.32                |
|                           | $\hat{S}_{\text{iChao2}}$ | 258.7            | −39.3 | 15        | 13.7                 | 42.1       | 0.46                |
|                           | $\hat{S}_{\text{Jack1}}$ | 258.4            | −39.6 | 8.5       | 7.6                  | 40.5       | 0.02                |
|                           | $\hat{S}_{\text{MLE}}$ | 460.2            | 162.2 | 97        | 142.6                | 189        | 0.29                |
|                           | $\hat{S}_{\text{BB}}$ | 268.5*           | −29.5 | 19.6      | 18.7                 | 35.4*      | 0.83*               |
| 160 (248.0)                | $\hat{S}_{\text{Chao2}}$ | 273.2            | −24.8 | 13.9      | 12.8                 | 28.4       | 0.67                |
|                           | $\hat{S}_{\text{iChao2}}$ | 277.8            | −20.2 | 15.5      | 14.7                 | 25.5       | 0.77                |
|                           | $\hat{S}_{\text{Jack1}}$ | 273.9            | −24.1 | 7.8       | 7.1                  | 25.3       | 0.26                |
|                           | $\hat{S}_{\text{MLE}}$ | 418.7            | 120.7 | 44.4      | 77.1                 | 128.6      | 0.05                |
|                           | $\hat{S}_{\text{BB}}$ | 288.2            | −9.8   | 20.9      | 20.2                 | 23.1*      | 0.91*               |

*aDenotes the smallest bias, lowest RMSE, and closest to 95% coverage.

Abbreviations: CI, confidence interval; RMSE, root mean squared error; SE, standard error.

### Table 8: First 10 species frequency counts, number of plots, observed richness and the CV of empirical species detection rate at each elevation

| Site     | T | $S_{\text{obs}}$ | CV | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|----------|---|------------------|----|----|----|----|----|----|----|----|----|----|-----|
| h50m     | 599 | 228 | 1.87 | 49 | 23 | 18 | 14 | 9  | 10 | 4  | 8  | 6  | 2   |
| h500m    | 230 | 242 | 1.66 | 71 | 34 | 12 | 14 | 9  | 11 | 8  | 4  | 7  | 5   |
| h1070m   | 150 | 123 | 1.39 | 28 | 16 | 13 | 3  | 1  | 3  | 6  | 1  | 1  | 1   |
| h1500m   | 200 | 57  | 1.42 | 13 | 4  | 2  | 2  | 4  | 2  | 0  | 0  | 4  | 0   |
| h2000m   | 200 | 15  | 1.21 | 1  | 2  | 1  | 1  | 0  | 0  | 2  | 0  | 0  |     |

Abbreviation: CV, coefficient of variation.
TABLE 9 Richness estimates of the Chao2 ( $\hat{S}_{\text{Chao2}}$), the improved Chao2 ( $\hat{S}_{\text{iChao2}}$), the first-order jackknife estimator ( $\hat{S}_{\text{Jack}}$), the MLE ( $\hat{S}_{\text{MLE}}$) and newly proposed estimator ( $\hat{S}_{\text{BB}}$) and their SEs at the five elevations

| Estimator   | h50m | h500m | h1070m | h1500m | h2000m |
|-------------|------|-------|--------|--------|--------|
| $\hat{S}_{\text{obs}}$ | 228  | 242   | 123    | 57     | 15     |
| $\hat{S}_{\text{Chao2}}$ (SE) | 279.1 (19.8) | 314.8 (23.4) | 146.3 (12.2) | 77.0 (15.4) | 14.2 (0.7) |
| $\hat{S}_{\text{iChao2}}$ (SE) | 290.1 (22.3) | 326.8 (25.3) | 146.3 (12.4) | 79.7 (16.4) | 14.3 (1.1) |
| $\hat{S}_{\text{Jack}}$ (SE) | 276.0 (10.1) | 312.0 (12.0) | 150.0 (7.2) | 69.0 (5.8) | 15.0 (1.1) |
| $\hat{S}_{\text{MLE}}$ (SE) | 45.446 (3280) | 59.774 (4842) | 2372 (1476) | 296 (236) | 18.4 (6.4) |
| $\hat{S}_{\text{BB}}$ (SE) | 305.2 (28.8) | 321.9 (49.4) | 158.5 (18.9) | 87.5 (25.7) | 14.2 (0.9) |

Abbreviations: MLE, maximum likelihood estimator; SE, standard error.

each elevation, and based on the simulation results, could provide a more accurate 95% CI coverage rate. For these reasons, the results provided by the newly proposed estimator may more accurately reflect reality.

4 | DISCUSSION

According to the concept of the Good–Turing frequency formula, we derived a new richness estimator using the moment method, instead of the ML method, based on a beta-binomial model. The proposed estimator only uses information on rare species (i.e. the number of singletons, doubletons and tripletons) in the sample to estimate undetected richness. Therefore, precise recording of the number of observed abundant species in a field survey is unnecessary, which greatly reduces field sampling loading.

Compared to the widely used lower bound estimator (Chao2) and recently the bias-corrected Chao2 (iChao2), the proposed estimation method also provides an improved lower bound estimator under the beta-binomial model and reduces the biases and RMSEs (more accuracy) of Chao2 and iChao2 in most simulation scenarios. Although the frequently used Chao2 and jackknife estimators have smaller variances than does the proposed estimator, their 95% CI coverage rates are extremely low due to their greater bias, possibly resulting in erroneous conservation conclusions and consequent policy-making. The lower bias and higher SE of the newly proposed estimator ensure more accuracy in terms of 95% CI coverage of true species richness, as demonstrated in the simulation results tables.

According to statistical theory, random sampling is the fundamental assumption for deriving the estimator of species richness. In ecological studies, individual-based abundance data and sample-based incidence data are two common data types collected to assess the species richness of different taxa. In individual-based data, the sampling unit is the individual, which is randomly and independently sampled and identified to species. In sample-based incidence data, the sampling unit is not an individual, but a plot (or quadrat, net, trap, transect, timed survey, etc.), which is randomly and independently sampled from the target region, and only the incidence (presence or absence) of each species in the sampled unit is recorded. In the field, individuals of each species always present a strong or slight spatial aggregation pattern due to the biotic or abiotic effects that make individuals difficult to individually sample and count. In this case, sample-based incidence data, with their ease of collection and proper random sampling assumption, can be effectively adopted to estimate species richness. However, richness estimation based on sample-based incidence data is still a statistical challenge. Even if communities have the same richness and species abundance, different communities display specific spatial aggregation patterns of individuals that could lead to differing species detection rates for communities of different heterogeneities. Thus, a single statistical model is always insufficient in describing all the diverse ecological communities, so there is no unbiased richness estimator for all ecological communities. Nevertheless, in this study, although the newly proposed richness estimator was derived based on the constrained beta-binomial model through a parametric approach, the newly proposed estimator can be also derived as a bias corrected estimator of Chao2 using the Good–Turing frequency formula without model assumption (see Appendix S1). That can explain why the proposed estimator presents robust behaviour in all simulation scenarios, including homogeneous, random uniform, broken stick, log-normal, and Zipf–Mandelbrot models and a real forest community. As a result, I am confident that the proposed estimator can supply another candidate to estimate richness based on sample-based incidence data especially when sample size is not large enough. Species richness in the study is defined as the number of detectable species of the assemblage that is the target of estimation. Therefore, if a low species detection rate resulting from high spatial aggregation of individuals causes zero-inflation in the survey data, it may be challenging to accurately estimate species richness, due to the many almost undetectable species. In this case, the newly proposed estimator can supply a superior lower bound for species richness (see Tables 1–6).

The sample-based replacement sampling scheme is identical to a capture–recapture sampling scheme employed to estimate population size of a specific species. Therefore, the proposed estimator can also be used to estimate population size when only the heterogeneous detection rate of individuals is considered and without varying by time and without depending on the previous detection records.

ACKNOWLEDGEMENTS

I would like to thank the Editor (Dr. Aaron Ellison), the Associate Editor (Dr. Claudia Coleine) and two reviewers for providing insightful and thoughtful comments and suggestions. I am especially grateful for Reviewer 2’s comments regarding the detailed
correction and rigorousness of the article. This work is supported by the Taiwan Ministry of Science and Technology under Contract 109-2118-M-002-003-MY2.

CONFLICT OF INTEREST
There is no conflict of interest.

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13979.

DATA AVAILABILITY STATEMENT
R code for calculating the estimate and SE of discussed estimator (including Chao2, iChao2, MLE-based estimator and newly proposed estimator) is archived in Zenodo repository. https://doi.org/10.5281/zenodo.7015481 (Chiu, 2022). The BCI Plot data used in manuscript are available from the website: http://ctfs.si.edu/webatlas/datasets/bci/.

ORCID
Chun-Huo Chiu https://orcid.org/0000-0002-7096-2278

REFERENCES
Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: A review. Journal of the American Statistical Association, 88(421), 364–373.
Bunge, J., Willis, A., & Walsh, F. (2014). Estimating the number of species in microbial diversity studies. Annual Review of Statistics and Its Application, 1, 427–445.
Burnham, K. P., & 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., & 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., & Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. Biometrics, 58(3), 531–539.
Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. Journal of the American Statistical Association, 87(417), 210–217.
Chao, A., & Yang, M. C. (1993). Stopping rules and estimation for re-capture debugging with unequal failure rates. Biometrika, 80(1), 193–201.
Chiu, C. H. (2022). The R code for paper ‘incidence-data based species richness estimation via a Beta-binomial model’ in MEE. Zenodo, https://doi.org/10.5281/zenodo.7015481
Chiu, C. H., Wang, Y. T., Walther, B. A., & Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. Biometrics, 70(3), 671–682.

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology, 5(1), 3–21.
Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 345(1311), 101–118.
Condit, R., Lao, S., Pérez, R., Dolins, S. B., Foster, R., & Hubbell, S. (2012). [dataset] Barro Colorado Forest census plot data (Version 2012).
Cormack, R. M. (1989). Log-linear models for capture-recapture. Biometrics, 45, 395–413.
Dorazio, R. M., & Andrew Royle, J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics, 59(2), 351–364.
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(2), 101–111.
Griffiths, D. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, 29, 637–648.
Jiménez-Gamero, M. D., & Puig, P. (2020). A nonparametric method of estimation of the population size in capture-recapture experiments. Biometrical Journal, 62(4), 970–988.
Longino, J. T., & Colwell, R. K. (2011). Density compensation, species composition, and richness of ants on a neotropical elevational gradient. Ecosphere, 2(3), 1–20.
Magurran, A. E. (Ed.). (2004). Measuring biological diversity. Blackwell.
Puig, P., & Kokonendji, C. C. (2018). Non-parametric estimation of the number of zeros in truncated count distributions. Scandinavian Journal of Statistics, 45(2), 347–365.
Rocchetti, I., Alfó, M., & Böhning, D. (2014). A regression estimator for mixed binomial capture-recapture data. Journal of Statistical Planning and Inference, 145, 165–178.
Sanathanan, L. (1972). Estimating the size of a multinomial population. The Annals of Mathematical Statistics, 43, 142–152.
Sanathanan, L. (1977). Estimating the size of a truncated sample. Journal of the American Statistical Association, 72(359), 669–672.
Wilson, R. M. A., & Collins, M. F. (1992). Capture-recapture estimation with samples of size one using frequency data. Biometrika, 79, 543–553.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chiu, C.-H. (2022). Incidence-data-based species richness estimation via a Beta-Binomial model. Methods in Ecology and Evolution, 13, 2546–2558. https://doi.org/10.1111/2041-210X.13979