mathematics

MDPI

Article

A Richness Estimator Based on Integrated Data

Chun-Huo Chiu

Department of Agronomy, National Taiwan University, Taipei 10617, Taiwan; chchiu2017@ntu.edu.tw

Abstract: Species richness is a widely used measure for assessing the diversity of a particular area. However, observed richness often underestimates the true richness due to resource limitations, particularly in a small-sized sample or highly heterogeneous assemblage. To estimate the number of different species (species richness) present across several different sites (communities), researchers often use a combined collection of data (an integrated dataset). This dataset is created by collecting samples from each site individually and independently. However, the pooled sample of integrated data is no longer a random sample from the entire area, and the use of different sampling schemes results in different collected data formats. Consequently, employing a single sampling distribution to model the pooled sample becomes unfeasible, rendering existing richness estimators inadequate. This study provides a theoretical explanation for the applicability of Chao's lower bound estimator in assessing species richness across multiple sites based on the pooled sample. Additionally, a new non-parametric estimator is introduced, which adjusts the bias of Chao's lower bound estimator by leveraging the Good–Turing frequency formula. This proposed estimator only utilizes the richness of singletons, doubletons, and tripletons in the pooled sample to estimate undetected richness. Simulated datasets across various models are employed to demonstrate the statistical performance of the estimator, showcasing its ability to reduce the bias of observed richness and provide accurate 95% confidence intervals. Real datasets are also utilized to illustrate the practical application of the proposed approach.

Keywords: Chao's lower bound estimator; Good–Turing frequency formula; integrated data; singleton; doubleton; tripleton

MSC: 62G05

Citation: Chiu, C.-H. A Richness Estimator Based on Integrated Data. *Mathematics* **2023**, *11*, 3775. https://doi.org/10.3390/math11173775

Academic Editor: Manuel Alberto M. Ferreira

Received: 6 July 2023
Revised: 21 August 2023
Accepted: 1 September 2023
Published: 2 September 2023

1. Introduction

Species richness is the most commonly used quantitative diversity metric and is easily understood. The term "species" can be broadly defined to include biological species, software bugs, words in a book, genes, alleles, or other discrete entities, as reviewed in [1–3]. This article focuses on biological applications—specifically, the number of detectable species within a given area. However, due to constraints in resources or sampling, creating a complete species inventory for a target area is often unfeasible. Instead, a random sample, representing a small portion of the target area's size or community, is typically used to evaluate species diversity. In ecological studies, there are two main formats for assessing species diversity: individual-based abundance data and sample-based incidence data. Individual-based abundance data involve randomly sampling and identifying individual organisms to species and recording the frequency of species. Sample-based incidence data involve randomly sampling a plot, quadrat, trap, transect, or net from the target area and recording the presence or absence of species appearing in the sampled unit [4].

Since the true number of species in an area is the sum of the species observed in the sample and those not appearing in the sample, the observed richness in the sample always underestimates the true richness. Generally, the extent to which species are underestimated in a sample hinges on sampling efforts and sample completeness [5]. Accurately estimating the species richness of an assemblage remains a statistical challenge, particularly in

highly heterogeneous assemblages [6]. To address the negative bias of observed richness, numerous estimators have been proposed, leading to significant advancements in various disciplines (refer to the review papers [1,3,7,8] for detailed information).

Generally, richness estimators in the literature can be classified into three types: curve-fitting, parametric, and non-parametric approaches. Curve-fitting approaches utilize parametric curves to extrapolate species-accumulation or species-area curves, aiming to predict their asymptote as an estimate of species richness [9,10]. This method does not directly leverage the frequencies of common and rare species. Instead, it only anticipates the trajectory of the rising curve. Parametric approaches treat species composition as a random variable, adhering to a distribution with limited parameters [11–13]. This parameter reduction enables the application of standard traditional statistical inference procedures, such as the maximum likelihood method. A primary advantage of parametric methods is their simplicity. However, curve-fitting and parametric methods face challenges in selecting the appropriate parametric function or distribution. Models using different functions or distributions might fit the data similarly, resulting in vastly different estimates. Further, a well-fitting parametric model does not guarantee a satisfactory estimate of species richness. Non-parametric richness estimators, which do not make model assumptions about species detection probability or species composition, tend to be more robust and are often preferred by ecologists. In the realm of ecological research, Chao's lower bound estimators [14,15], which are rooted in the Cauchy–Schwarz inequality, are prominently used. Moreover, to address the bias in observed richness, jackknife-based estimators [16,17] were crafted. These work by consecutively excluding individuals from the data to analyze various sub-datasets. In addition, these non-parametric estimators do not require all the information on observed species; only rare species (singletons and doubletons) are used in the sample to estimate undetected richness. These estimators could show expected robust statistical behavior only when the sampling unit (i.e., an individual in abundance data and a plot in incidence data) is randomly sampled. However, due to resource constraints, a random sample is often only feasible in a limited area, not in a large-scale area.

In recent decades, monitoring species richness to reveal the impact of human activities on a large or global scale has become an increasingly urgent task [18–23]. However, estimating richness for large-scale areas (or across multiple sites) remains a statistical challenge, and no reliable estimator has been developed to date. In general, the collected datasets used to estimate the richness across multiple sites usually consist of the samples that are separately sampled from each site by implementing different sampling schemes. Therefore, this integrated dataset is composed of different kinds of data formats, including individual-based abundance data and sample-based incidence data. However, the widely-used rigorous estimators in the literature have their limitations due to their underlying theoretical assumptions, and they are not equipped to analyze this type of integrated data. Therefore, until now, no estimator has been specifically proposed for integrated data to estimate the richness across multiple sites.

In this article, I provide a theoretical interpretation of the applicability of Chao's lower bound estimator for estimating species richness based on a pooled sample of integrated data. Additionally, utilizing the Good–Turing frequency formula [24], I address the negative bias inherent in Chao's lower bound estimator and propose a bias-corrected alternative. The variance of the new estimator can be calculated through the asymptotic approach, and its 95% confidence interval can be obtained through logarithmic transformation. To evaluate the efficacy of this proposed estimator, three commonly used ecological models and two real datasets are utilized in simulation studies and illustrative examples. Based on simulation results from various scenarios of integrated data, both estimators significantly reduce the negative bias of observed richness, providing reliable lower bound estimates across various hypothetical models and exhibiting convergence towards the true richness as the sample size increases. Notably, the newly proposed bias-corrected estimator outperforms Chao's lower bound, exhibiting lower bias, lower root mean square error (RMSE), and a more

accurate 95% confidence interval (CI) for the true richness, particularly when dealing with small sample sizes or highly heterogeneous communities.

2. Materials and Methods

2.1. Sampling Distribution Model

Assume there are a total of S distinct species in the community of interest. In ecological studies, individual-based abundance data and sample-based incidence data are the most commonly collected data types in the assessment of richness diversity [4]. The sampling unit of individual-based abundance data is an individual independently sampled and identified to a species from the target area, the sampling unit of sample-based incidence data is a plot, quadrat, trap, transect, or net randomly sampled from the target area, and only the incidence (presence or absence) of species appearing in the selected plot is recorded.

For individual-based abundance data, assume n (a small fraction of community size) individuals are independently sampled by sampling with replacement or sampling without replacement. Let X_i be the number of individuals of species i counted in the sample. The species frequency or species abundance $(X_1, X_2, \ldots, X_S)$ could be assumed to follow a multinomial distribution with size n and probabilities $(p_1, p_2, \ldots, p_S)$, and the species frequency X_i follows a binomial distribution with parameters n and p_i:

$$X_i \sim Binomial(n, p_i),\ i = 1,\ 2, \ldots, S,$$

where p_i is the relative detection probability of species i. Let $f_k = \sum_{i=1}^{S} I(X_i = k)$ be the number of species that are observed exactly k times in the sample, $k = 1,\ 2, \ldots, n$. Therefore, $f_0,\ f_1,\ f_2$, and f_3 represent the undetected richness, singleton richness, doubleton richness, and tripleton richness in the abundance sample, respectively, where f_0 is an unknown parameter.

For sample-based incidence data, assume t sampling units are randomly sampled from the target area and only the incidence (presence or absence) of species in the sampled unit is recorded. Let Y_i be the number of units in which species i is detected in the t sampled units. Then, Y_i could be assumed to follow a binomial distribution with size t and probability $\pi_i,\ i = 1,\ 2,\ \ldots, S$:

$$Y_i \sim Binomial(t, \pi_i),\ i = 1,\ 2, \ldots, S,$$

where π_i is the detection probability of species i, which depends on the abundance, body size, and color of species i, as well as the investigator's capability. Let $Q_k = \sum_{i=1}^{S} I(Y_i = k)$ be the number of species that are detected in exactly k out of t sampling units, $k = 1,\ 2, \ldots t$. Therefore, $Q_0,\ Q_1,\ Q_2$, and Q_3 are the unseen richness, singleton richness, doubleton richness, and tripleton richness in the incidence sample, respectively, where Q_0 is an unknown parameter.

2.2. Richness Estimation for a Single Assemblage Using Integrated Data

Assuming that N samples are randomly collected from the target area through various sampling schemes, including individual-unit-based and sample-unit-based sampling methods, the integrated data comprise two formats (i.e., abundance data and incidence data), as commonly seen in ecological studies. To determine the richness of the assemblage, Chao1 and Chao2 [14,15], derived without model assumption on species detection rates, are the most commonly used estimators, which are briefly outlined below. In this context, I will not deeply explore jackknife-based estimators because they lack a theoretical basis for bias reduction in species richness estimation [25] and they exhibit inferior statistical performance compared to Chao's lower bound estimators [26].

2.2.1. Chao's Lower Bound Estimators

Based on Cauchy–Schwarz inequality, and without making any assumptions on species detection rates, Chao proposed lower bound estimators for richness in 1984 and

1987. These estimators were designed for individual-based abundance data and sample-based incidence data and are referred to as the Chao1 and Chao2 estimators, respectively. The Chao1 and Chao2 estimators are separately expressed as

$$\text{Chao1} = S_{obs} + \begin{cases} \frac{f_1^2}{2f_2} & if\ f_2 > 0 \\ \frac{f_1(f_1-1)}{2} & if\ f_2 = 0 \end{cases} \quad (1)$$

$$\text{Chao2} = S_{obs} + \begin{cases} \frac{t-1}{t}\frac{Q_1^2}{2Q_2} & if\ Q_2 > 0 \\ \frac{t-1}{t}\frac{Q_1(Q_1-1)}{2} & if\ Q_2 = 0 \end{cases}. \quad (2)$$

Chao's lower bound estimators only use the frequency counts of the two rarest species (i.e., the numbers of singleton and doubleton species) in the sample to estimate undetected richness.

On the basis of Cauchy–Schwarz inequality theory, Chao's lower bound estimators are unbiased when the detection rates of species are homogeneous (i.e., $p_i = \frac{1}{S}$ in Equation (1) or $\pi_i = c$ in Equation (2), for $i = 1, 2, \ldots, S$). In addition, according to the Good–Turing frequency formula [24], Chao et al. [27,28] show that Chao's lower bound estimators are nearly unbiased estimators only when rare species have approximately homogenous detection probabilities (or rates). Therefore, the degree of heterogeneity of the abundant species in the assemblage contains no information about the unbiasedness of Chao's lower bound estimators. When the detection rate of rare species is highly heterogeneous or the sample size is not large enough, in contrast to other parametric estimators, Chao1 or Chao2 can provide a lower bound and robust richness estimate [2,28]. However, Chao1 and Chao2 were separately derived based on different sampling models for abundance data and incidence data. Importantly, there is still no theoretical evidence or proof that Chao's lower bound estimator can be used to estimate species richness using a pooled sample of integrated data.

2.2.2. Extending Chao's Lower Bound Estimators for Integrated Data

Many richness estimators proposed in the literature, whether they are parametric or non-parametric, are designed for randomly sampled data. This means that the detection rate of a species for each random trial, such as a selected individual or plot, is assumed to be identical. These estimators assume that the underlying assumptions of the binomial distribution are met.

However, if N samples are separately collected using different sampling methods (e.g., sampling schemes, sampling efforts, plot sizes, or investigators) from the target area, the observed species count in the pooled sample no longer follows a binomial distribution. This violates the theoretical assumption of a random sample. This type of integrated data is often encountered in ecological studies, where individual-based abundance data and sample-based incidence data are collected from the same target area. While integrated data are commonly employed to estimate richness, no estimator has been rigorously designed for such data. In this section, I will theoretically illustrate how Chao's lower bound estimator can be modified to handle integrated data.

For individual-based abundance data, according to probability theory, when sample size n is sufficiently large and relative abundance (or detection probability) p is sufficiently small, the species frequency (X) follows a binomial distribution that converges to a Poisson distribution. This implies that the frequency (X) of rare species (i.e., p is sufficiently small) in the sample could approximate a Poisson distribution with mean $n\ p$ (i.e., $X \sim poi(np)$) for species with low detection rate. This convergence feature also applies to sample-based incidence data. When the number of plots t is large and the detection rate π tends to zero, the incidence count (Y) of rare species in the sample could approximate a Poisson distribution with mean $t\ \pi$ (i.e., $Y \sim poi(t\pi)$) for species with a low detection rate.

Without loss of generality, two random samples are collected from the target region through different sampling schemes, namely, individual-based sampling and plot-based sampling methods. These samples correspond to individual-based abundance data and sample-based incidence data, respectively. When the two sampled samples are pooled, the pooled species frequency $Z_i = X_i + Y_i$ represents the count of species i in the pooled sample. Here, Z_i is no longer a random variable following a binomial distribution.

Based on the convergence principle between the binomial and Poisson distributions discussed earlier, for species with low detection rates in the combined sampling scheme (i.e., small p_i and small π_i), the species abundance (Z_i) in the pooled sample approximately follows a Poisson distribution with a mean parameter $\lambda_i = np_i + t\pi_i$. For simplicity, let denote this mean parameter as $\lambda_i = md_i$, where m represents the unknown size of the pooled sample and d_i represents the detection rate of species i. Next, let $G_k = \sum_{i=1}^{S} I(Z_i = k)$ be the species frequency count, representing the number of species that are present exactly k times in the pooled sample. When k is small (e.g., k = 0, 1, 2 or 3) and the size of the pooled sample is sufficiently large, G_k is primarily contributed by the rare species, which approximately follow a Poisson distribution. Given a specific sampling scheme, all species in the region can be divided into a set of rare species, denoted as $\{s_{rare}\}$, and a set of abundant species, denoted as $\{s_{abun}\}$. Based on the existing convergence theory between the binomial distribution and the Poisson distribution, we have the approximation of the expectation of G_k for small k:

$$E[G_k] = E\left[\sum_{i=1}^{S} I(Z_i = k)\right] = \sum_{i\in\{S_{rare}\}} P(Z_i = k) + \sum_{i\in\{S_{abun}\}} P(Z_i = k).$$

When k is small, the probability that abundant species have a count of k tends to zero. Therefore, $\sum_{i\in\{S_{abun}\}} P(Z_i = k)$ is roughly equal to 0. We have

$$E[G_k] \approx \sum_{i\in\{S_{rare}\}} P(Z_i = k) + 0.$$

According to the convergence property between binomial and Poisson distribution for rare species, the following approximation is held:

$$E[G_k] \approx \sum_{i\in\{S_{rare}\}} P(Z_i = k) \approx \sum_{i\in\{S_{rare}\}} \frac{\lambda_i^k}{k!} e^{-\lambda_i} \approx \sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i}.$$

Therefore, we can derive the following four approximation equations for the expectation of undetected richness, singleton richness, doubleton richness, and tripleton richness, which represent the number of rare species in the pooled sample:

$$E[G_0] = E\left[\sum_{i=1}^{S} I(Z_i = 0)\right] = \sum_{i=1}^{S} P(Z_i = 0) \approx \sum_{i=1}^{S} e^{-\lambda_i} \tag{3a}$$

$$E[G_1] = E\left[\sum_{i=1}^{S} I(Z_i = 1)\right] = \sum_{i=1}^{S} P(Z_i = 1) \approx \sum_{i=1}^{S} \lambda_i\, e^{-\lambda_i} \tag{3b}$$

$$E[G_2] = E\left[\sum_{i=1}^{S} I(Z_i = 2)\right] = \sum_{i=1}^{S} P(Z_i = 2) \approx \sum_{i=1}^{S} \frac{{\lambda_i}^2}{2} e^{-\lambda_i} \tag{3c}$$

$$E[G_3] = E\left[\sum_{i=1}^{S} I(Z_i = 3)\right] = \sum_{i=1}^{S} P(Z_i = 3) \approx \sum_{i=1}^{S} \frac{{\lambda_i}^3}{6} e^{-\lambda_i}. \tag{3d}$$

It is worth emphasizing once again that these approximations are valid only for lower species frequency counts in the sample, under the condition that sample sizes (n and t) are sufficiently large. In Appendix A, I provide evidence that the aforementioned approximate equations hold by demonstrating their validity through numerical simulations.

Based on the Cauchy–Schwarz inequality, we have the following inequality:

$$\sum_{i=1}^{S} e^{-\lambda_i} \sum_{i=1}^{S} \lambda_i{}^2 e^{-\lambda_i} \geq \left(\sum_{i=1}^{S} \lambda_i \, e^{-\lambda_i} \right)^2. \tag{4}$$

According to Equations (3a) and (3b), Equation (4) is equivalent to $E[G_0]E[2G_2] \geq E[G_1]^2$. This inequality is also held when species detected mean abundance λ_i is assumed to be a random variable with probability density function $f(\lambda)$, expressed as

$$\left(S \int e^{-\lambda} \, f(\lambda) d\lambda \right) \left(S \int \lambda^2 e^{-\lambda} \, f(\lambda) d\lambda \right) \geq \left(S \int \lambda \, e^{-\lambda} \, f(\lambda) d\lambda \right)^2.$$

Therefore, we have the lower bound estimator of undetected richness $\hat{G}_0 = \frac{G_1^2}{2G_2}$, which uses the number of singletons and doubletons in the pooled sample to estimate undetected richness. Therefore, the proposed richness estimator could be interpreted as an extension of Chao1 or Chao2. It is denoted as Chao3, and the modified formula can be expressed as

$$\text{Chao3} = S_{obs} + \begin{cases} \frac{G_1^2}{2G_2} & if \; G_2 > 0 \\ \frac{G_1(G_1-1)}{2} & if \; G_2 = 0. \end{cases} \tag{5}$$

2.2.3. Modified Good–Turing Frequency Formula for Integrated Data

Before adjusting Chao's lower bound estimator for a more accurate estimator, it is essential first to introduce the Good–Turing frequency formula. Given a species abundance sample of size n collected randomly, let α_r symbolize the mean relative abundance of species that appear exactly r times in the sample, expressed as $\alpha_r = \sum_{i=1}^{S} p_i I(X_i = r)/f_r$. The Good–Turing frequency formula, designed to estimate α_r, is presented as $\hat{\alpha}_r = \frac{(r+1)f_{r+1}}{nf_r}; r = 1, 2, \ldots$.

This formula has its roots in the work of Alan Turing and I. J. Good during World War II. They collaborated on deciphering German ciphers and innovatively utilized this statistical method to estimate the true frequencies of rare code elements, including those undetected, based on observed frequencies in intercepted samples of Nazi code. Later, Good's papers in 1953 [29], and jointly with Toulmin in 1956 [24], shed light on Turing's wartime explorations concerning the frequency formula and other related research topics.

In ecological studies, sample coverage represents the proportion of the total number of individuals in a community that belong to the species represented in the sample. This is represented mathematically as $C = \sum_{i=1}^{S} p_i I(X_i > 0)$, providing a measure of the sample's completeness. Since sample coverage is equivalent to $C = 1 - \alpha_0 f_0$, it can be estimated as $1 - \hat{\alpha}_0 f_0 = 1 - f_1/n$ [24]. This provides insight into the proportion of individuals from sampled species. This metric helps ecologists determine how well their sample represents the underlying community and whether more sampling is needed.

The Good–Turing frequency formula also can be used to estimate the number of unobserved species in a sample, based on the intuitive concept that the mean relative abundance of unseen species should be no greater than the mean relative abundance of species observed once in the sample (i.e., $\alpha_0 \leq \alpha_1$). Upon employing the Good–Turing frequency formula to estimate α_0 and α_1, we arrive at the inequality $f_1/(nf_0) \leq 2f_2/(nf_1)$. Then, the lower bound estimator of undetected richness can be obtained, shown as $\hat{f}_0 = f_1^2/2f_2$, that is identical the Chao1 lower bound estimator initially derived via Cauchy–Schwarz inequality.

Based on the concept of the Good–Turing frequency formula, for the pooled sample from an integrated data, the mean detection rate of species which are present r times in the

pooled sample is denoted as $d_{(r)} = \sum_{i=1}^{S} d_i I(Z_i = r)/G_r$. That can be effectively estimated via a modified Good–Turing frequency formula, shown as

$$\hat{d}_{(r)} = \frac{(r+1)G_{r+1}}{mG_r}. \tag{6}$$

2.2.4. Modified Chao's Lower Bound Estimator for Integrated Data

According to the Cauchy–Schwarz inequality, we know that Chao3 is a lower bound estimator. Based on the Good–Turing frequency formula, Chao3 will be severely negatively biased when the rare species have a high degree of heterogeneity. In this section, the negative bias of Chao3 can be corrected based on the Good–Turing frequency formula [24]. The bias of Chao3 is approximately equal to

$$E\left[\frac{G_1^2}{2G_2}\right] - E[G_0] \approx \frac{\left(\sum_{i=1}^{S} md_i\, e^{-md_i}\right)^2}{2\sum_{i=1}^{S} \frac{(md_i)^2}{2} e^{-md_i}} - \sum_{i=1}^{S} e^{-md_i}.$$

Using the modified Good–Turing frequency formula (Equation (6)), we have the following approximate equations:

$$E[G_1] \approx \sum_{i=1}^{S} md_i\, e^{-md_i} = \sum_{i=1}^{S} \frac{2}{md_i} E[I(Z_i = 2)] \approx \frac{2}{md_{(2)}} E[G_2] \tag{7a}$$

$$E[G_0] \approx \sum_{i=1}^{S} e^{-md_i} = \sum_{i=1}^{S} \frac{1}{md_i} E[I(Z_i = 1)] \approx \frac{1}{md_{(1)}} E[G_1]. \tag{7b}$$

According to Equations (7a) and (7b), the bias of $G_1^2/(2G_2)$ can be approximately derived as

$$Bias_{Chao3} = E\left[\frac{G_1^2}{2G_2}\right] - E[G_0] \approx \left(\frac{1}{md_{(2)}} - \frac{1}{md_{(1)}}\right) E[G_1].$$

Therefore, the bias of Chao3 can be estimated by replacing $d_{(1)}$ and $d_{(2)}$ with $\hat{d}_{(1)}$ and $\hat{d}_{(2)}$, respectively. It is given as

$$\widehat{Bias}_{Chao3} = \frac{G_1^2}{2G_2}\left(\frac{2G_2^2}{3G_1G_3} - 1\right).$$

Then, we have the bias-corrected estimator of Chao3, expressed as

$$Chao3_{Adj} = S_{obs} + \frac{G_1^2}{2G_2}\left(2 - \left(\frac{2G_2^2}{3G_1G_3}\right)^{-}\right), \tag{8}$$

where $(A)^{-}$ equals 1 if $A \geq 1$ and A if $A < 1$. Here, as $G_3 = 0$ (or $G_1 = 0$), G_3 (or G_1) is replaced by 1 to make Equation (8) always well-defined. The mathematic form of the estimator shown in Equation (8) is identical to the parametric estimator proposed by Chiu [30,31] which was derived based on the beta-binomial mixture model for sample-based incidence data or based on the gamma-Poisson mixture model for individual-based abundance data. The new estimator can also be proved to be a lower bound of richness under the incidence-based beta-binomial mixture model or the abundance-based gamma-Poisson mixture model [30,31].

Since the $Chao3_{Adj}$ estimator utilizes the first three rarest species in the sample to estimate undetected richness, it can be applied to integrated data consisting of multiple samples randomly collected from the target area without adhering to a specific sampling model or scheme. For a comprehensive comparison, a table is provided in Appendix B. This table details the equations and symbols utilized in the proposed estimators, complete

with their definitions, origins (cited with references), and their statistical performances in estimating richness.

2.3. Estimating Richness across Multiple Assemblages Using Integrated Data

When there are N assemblages (sites), species sampling data are collected independently and separately from each site. The sampling data can be collected by either an individual-unit-based sampling method or a sample-unit-based sampling method. Let X_{ij} represent the number of sampling units (such as the number of individuals in individual-based abundance data or the number of plots in sample-based incidence data) of species i in the sample j, which is collected from the jth site. Here, i ranges from 1 to S for the S species, and j ranges from 1 to N for the N sites. If the sample size (i.e., the total number of individuals in abundance data or the total number of plots in incidence data) is sufficiently large in each sample, the counts (X_{ij}) of species with low detection rates will approximate a Poisson distribution. Then, the total count of species i in the pooled sample, denoted as $X_{i+} = \sum_{j=1}^{N} X_{ij}$, will approximate a Poisson distribution when the detection rate of species i in each site is uniformly low.

Let $G_k = \sum_{i=1}^{S} I(X_{i+} = k)$ be the number of species with a count of exactly k in the pooled sample. The approximate equations shown in Equations (3a)–(3d) are also applicable to the pooled sample of integrated data. Additionally, formulae for Chao3 and $\text{Chao3}_{\text{Adj}}$ can be derived to estimate species richness across multiple assemblages based on the Good–Turing frequency formula without making any specific model assumptions. Similarly, their variance estimators can be obtained using the asymptotic approach, and the 95% confidence interval (CI) of species richness can be derived by referring to the discussion surrounding Equation (9).

According to the derivation, the proposed richness estimator possesses the following properties: (i) when the samples are individually and randomly collected from each site, the sampled samples can be directly combined to estimate undetected richness, regardless of whether the data format in each sample is identical or not; (ii) the estimation of undetected richness is solely based on the frequency counts of the rarest species in the pooled sample; (iii) when the detection rates of rare species are homogeneous (including the homogeneous model as a special case) or the sample size is sufficiently large, the proposed estimators are nearly unbiased.

2.4. Estimation of the Variance for the Estimator

To derive the variance estimator for the proposed richness estimator, an asymptotic approach is employed. By defining G_{2+} as the total frequency count of species with a count of at least 3 in the sample (i.e., $G_{2+} = \sum_{k\geq 3} G_k$), the estimator Chao3 can be expressed as a function of (G_1, G_2, G_{2+}). The estimator of Chao3's variance could be obtained by the asymptotic approach in which (G_0, G_1, G_2, G_{2+}) approximate a multinomial distribution with parameters $\left(S, \frac{E[G_0]}{S}, \frac{E[G_1]}{S}, \frac{E[G_2]}{S}, \frac{E[G_{2+}]}{S}\right)$. Additionally, let G_{4+} be the total frequency count of species with a count of at least 4 in the sample (i.e., $G_{4+} = \sum_{k\geq 4} G_k$); then, $\text{Chao3}_{\text{Adj}}$ becomes a function of (G_1, G_2, G_3, G_{4+}). The estimator of $\text{Chao3}_{\text{Adj}}$'s variance can also be obtained using the asymptotic approach, where $(G_0, G_1, G_2, G_3, G_{4+})$ approximately follow a multinomial distribution with parameters $\left(S, \frac{E[G_0]}{S}, \frac{E[G_1]}{S}, \frac{E[G_2]}{S}, \frac{E[G_3]}{S}, \frac{E[G_{4+}]}{S}\right)$.

The variance estimator of the Chao3 or $\text{Chao3}_{\text{Adj}}$ can be derived via the delta method and is expressed as

$$\widehat{var}(\hat{S}) \approx \sum_i \sum_j \frac{\partial \hat{S}}{\partial G_i} \frac{\partial \hat{S}}{\partial G_j} \widehat{cov}(G_i, G_j),$$

where

$$\widehat{cov}(G, G_j) = \begin{cases} G_i(1 - G_i/\hat{S}) & if\ i = j \\ -G_i G_j/\hat{S} & if\ i \neq j \end{cases}.$$

To derive the 95% confidence interval (CI) of species richness and to ensure that the lower bound of the 95% CI of species richness is larger than the observed richness, assume $\hat{S} - S_{obs}$ follows a log-normal distribution [27,32]; then, the two-sided 95% CI of species richness is obtained as

$$\left[S_{obs} + \frac{\hat{S} - S_{obs}}{R}, S_{obs} + (\hat{S} - S_{obs})R\right], \text{where } R = exp\left\{1.96\left[\log\left(1 + \frac{Var(\hat{S})}{(\hat{S} - S_{obs})^2}\right)\right]^{\frac{1}{2}}\right\}. \quad (9)$$

When samples are randomly collected, Chao3 consistently provides a lower bound estimate of species richness. Similarly, the $\text{Chao3}_{\text{Adj}}$ also provides a lower bound estimate of species richness under the gamma-Poisson model or the beta-binomial model [30,31]. Therefore, in cases where the community exhibits high heterogeneity or the sample size is small, these two estimators can offer lower bound estimates and more informative one-sided 95% confidence intervals (CIs) of species richness, shown as

$$\left[S_{obs} + \frac{\hat{S} - S_{obs}}{R}, \infty\right], \text{where } R = exp\left\{1.65\left[\log\left(1 + \frac{Var(\hat{S})}{(\hat{S} - S_{obs})^2}\right)\right]^{\frac{1}{2}}\right\}.$$

3. Results

3.1. Hypothetical Species Composition Models for Simulation Study

A simulation study was conducted to examine the statistical behaviors of the new estimators. The study involved the use of three species abundance models to generate individual-based abundance data and three species detection models to generate sample-based incidence data. The number of species was kept constant at $S = 600$, and the simulated datasets were generated separately and independently using the following models.

3.1.1. Models for Individual-Based Abundance Sampling

The species detection probabilities (or species relative abundance) $(p_1, p_2, \ldots, p_S) = (ca_1, ca_2, \ldots, ca_S)$ in each model are provided below, where c is a normalizing constant such that $\sum_{i=1}^{S} p_i = 1$. The coefficient of variation (CV) of $(p_1, p_2, \ldots, p_S)$ is also presented to indicate the degree of heterogeneity of $(p_1, p_2, \ldots, p_S)$.

a. Abundance model 1, random uniform model (CV = 0.53), with $p_i = ca_i, i = 1, 2, \ldots, S$, where $(a_1, a_2, \ldots, a_S)$is a random sample from a uniform distribution.
b. Abundance model 2, broken-stick model (CV = 0.97), with $p_i = ca_i$, $i = 1, 2, \ldots, S$, where $(a_1, a_2, \ldots, a_S)$ is a random sample from an exponential distribution with parameter 1. This model is commonly used in the literature and is equivalent to the Dirichlet distribution.
c. Abundance model 3, log-normal model (CV = 1.56), with $p_i = ca_i$, $i = 1, 2, \ldots, S$, where $(a_1, a_2, \ldots, a_S)$ is a random sample from a log-normal distribution with parameters 0 and 1.

3.1.2. Models for Sample-Based Incidence Sampling

The species detection probabilities $(\pi_1, \pi_2, \ldots, \pi_S) = (ca_1, ca_2, \ldots, ca_S)$ in each model were determined, where c is a rescaling constant such that the maximum detection probability is a fixed at a constant value. The coefficient of variation (CV) of $(\pi_1, \pi_2, \ldots, \pi_S)$ is also calculated to indicate the degree of heterogeneity of $(\pi_1, \pi_2, \ldots, \pi_S)$.

d. Incidence model 1: the random uniform model (CV = 0.57), where $\pi_i = ca_i$, $i = 1, 2, \ldots, S$, and $(a_1, a_2, \ldots, a_S)$ is a random sample from a uniform distribution with parameters (0, 1), and scale c is used to control the maximum of $\{\pi_i, i = 1, 2, \ldots, S\}$.
e. Incidence model 2: the broken stick model (CV = 0.99), where $\pi_i = ca_i$, $i = 1, 2, \ldots, S$, and $(a_1, a_2, \ldots, a_S)$ is a random sample from an exponential distribution with parame-

ter 1, and scale c is used to control the maximum of $\{\pi_i,\ i = 1,\ 2, \ldots,\ S\}$. This model is commonly used in the literature and is equivalent to the Dirichlet distribution.

f. Incidence model 3: the log-normal model (CV = 1.23), where $\pi_i = ca_i,\ i = 1, 2, \ldots, S$, and $(a_1, a_2, \ldots, a_S)$ is a random sample from a log-normal distribution, and scale c is used to control the maximum of π_i.

The coefficient of variation (CV) in these six models ranged from 0 to 1.56, encompassing a wide range of values that encompass most practical cases in real-world applications.

In the simulation study, different sample sizes are considered to represent varying levels of sampling effort. For each simulation scenario, 1000 simulated datasets are generated. The estimates and their corresponding estimated standard errors (SE) are averaged across the 1000 simulated datasets to obtain the mean estimate and mean estimated SE. The sample SE and root mean square error (RMSE) are calculated based on the 1000 estimates to determine the sample SE and sample RMSE. The percentage of 95% confidence intervals (CIs) that cover the true value and the average observed richness are also calculated. All the simulation results are presented in Tables 1 and 2. For simplicity, the average estimates of the discussed estimators are plotted in Figures 1 and 2 to illustrate their statistical behavior as a function of sampling effort.

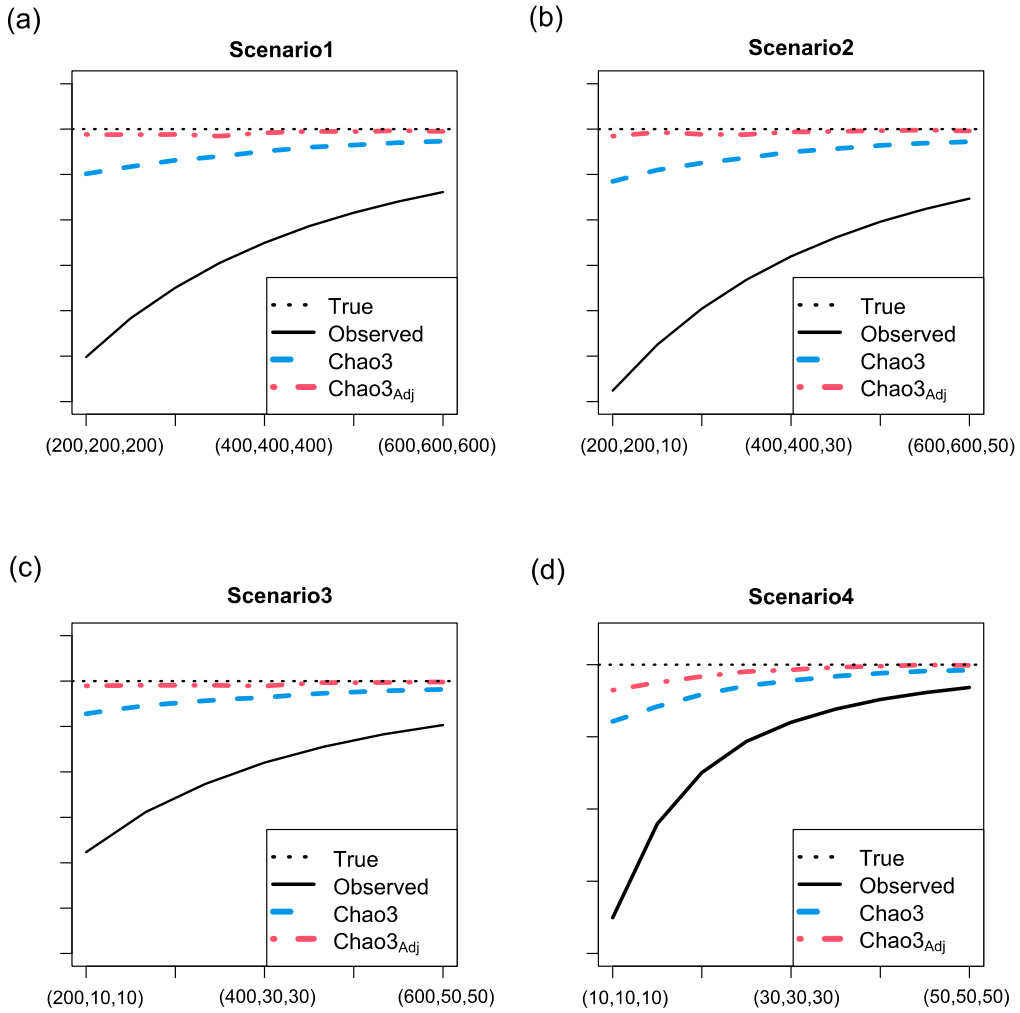

Figure 1. Plot of the true richness of one assemblage and the average richness estimates (including observed richness, Chao3, and the adjusted Chao3 ($\text{Chao3}_{\text{Adj}}$)) as a function of the sampling effort for four different scenarios. Each scenario involves different combinations of abundance and incidence data sets: (**a**) three abundance data sets; (**b**) two abundance data sets and one incidence data set (**c**) one abundance data and two incidence data sets; and (**d**) three incidence data sets.

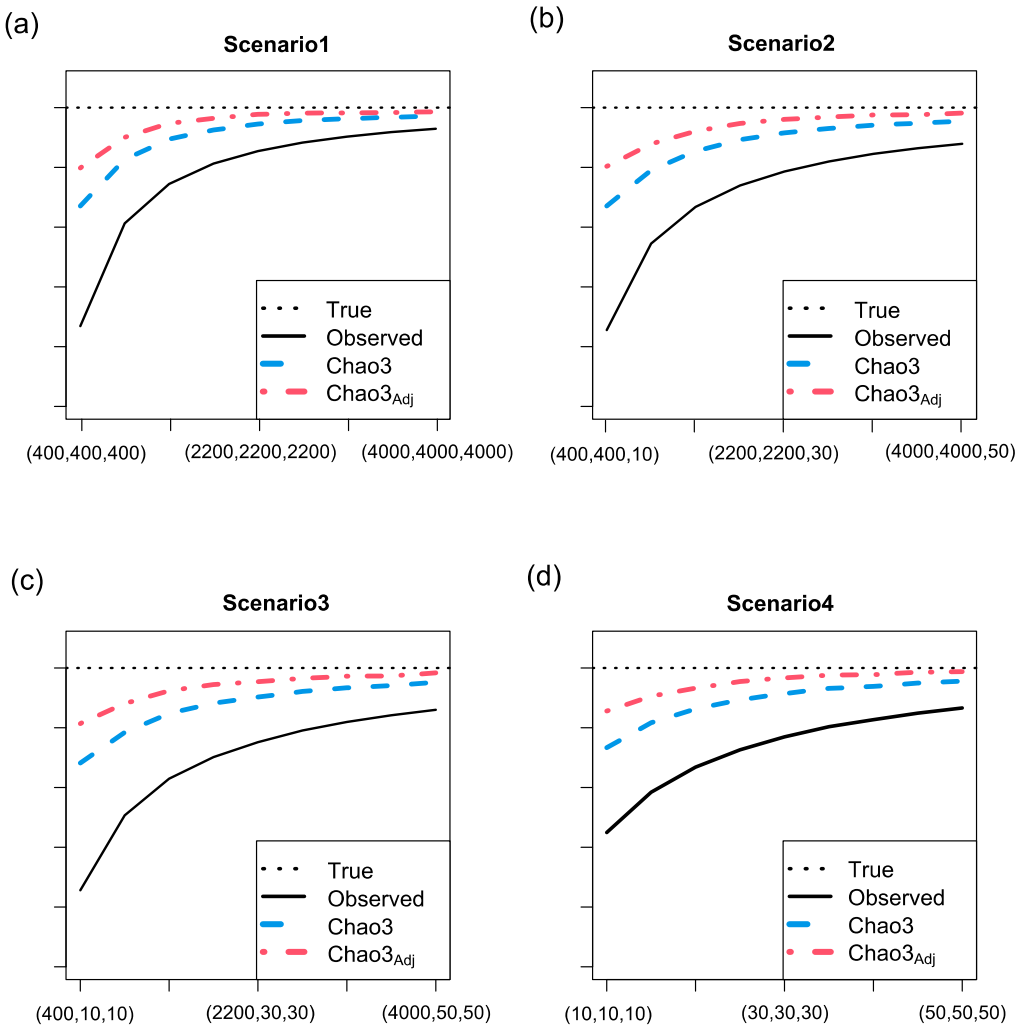

Figure 2. Plot of the true richness of multiple assemblages and the average richness estimates (including observed richness, Chao3, and the adjusted Chao3 ($Chao3_{Adj}$)) as a function of the sampling effort for four different scenarios. Each scenario involves different combinations of abundance and incidence data sets: (**a**) three abundance data sets; (**b**) two abundance data sets and one incidence data set; (**c**) one abundance data set and two incidence data sets; and (**d**) three incidence data sets.

Table 1. The statistical behavior of Chao3 and $Chao3_{Adj}$ are analyzed in four scenarios to estimate the richness of one assemblage.

Size (Observed Richness)	Estimator	Average Estimate	Bias	Sample SE	Average Estimated SE	Sample RMSE	95% CI Coverage Rate
			Scenerio1				
200, 200, 200 (352.3)	Chao3	557.3	−42.7	37.6	38.7	56.9 †	0.834
	$Chao3_{Adj}$	601.6	1.6 †	71.6	68.2	71.5	0.856 †
400, 400, 400 (480.0)	Chao3	582.1	−17.9	21.4	20.9	27.8 †	0.882
	$Chao3_{Adj}$	601.6	1.6 †	35.7	34.1	35.7	0.87 †
600, 600, 600 (536.2)	Chao3	590.4	−9.6	14.3	13.4	17.2 †	0.91
	$Chao3_{Adj}$	599.9	−0.1†	22.1	21.5	22.1	0.95 †

Table 1. *Cont.*

Size (Observed Richness)	Estimator	Average Estimate	Bias	Sample SE	Average Estimated SE	Sample RMSE	95% CI Coverage Rate
			Scenerio2				
200, 200, 10 (307.9)	Chao3	548.8	−51.2	50.1	46.8	71.6 †	0.804
	$Chao3_{Adj}$	598.2	−1.8 †	93.2	82.1	93.1	0.890 †
400, 400, 20 (441)	Chao3	571	−29	26.3	25.3	39.1 †	0.8
	$Chao3_{Adj}$	596.3	−3.7 †	45.4	41.9	45.5	0.91 †
600, 600, 40 (513.4)	Chao3	584.9	−15.1	16.9	16.2	22.6 †	0.858
	$Chao3_{Adj}$	598.8	−1.2 †	27.4	26.4	27.4	0.932 †
			Scenerio3				
200, 10, 10 (330.8)	Chao3	547.9	−52.1	41	41.4	66.3 †	0.786
	$Chao3_{Adj}$	587.1	−12.9 †	74.5	70.7	75.5	0.872 †
400, 20, 20 (460.2)	Chao3	572.1	−27.9	22.6	22.4	35.9 †	0.814
	$Chao3_{Adj}$	593.6	−6.4 †	39.6	36.9	40	0.902 †
600, 40, 40 (538.5)	Chao3	589.3	−10.7	13.2	13.1	17 †	0.902
	$Chao3_{Adj}$	599.8	−0.2 †	20.7	21.3	20.6	0.95 †
			Scenerio4				
10, 10, 10 (445.0)	Chao3	570.1	−29.9	24.7	24.2	38.8 †	0.808
	$Chao3_{Adj}$	589.4	−10.6 †	41.8	38.2	43.1	0.88 †
20, 20, 20 (540.5)	Chao3	585	−15	11.8	11.8	19.1 †	0.826
	$Chao3_{Adj}$	594	−6 †	18.5	19.2	19.4	0.961 †
40, 40, 40 (583.9)	Chao3	596.3	−3.7	6.1	5.7	7.1 †	0.954
	$Chao3_{Adj}$	599.6	−0.4 †	9.1	10.5	9.1	0.952 †

Note: data in Scenerio1, Scenerio2, Scenerio3, and Scenerio4 are separately composed of three abundance data, two abundance data and one incidence datum, one abundance datum and two incidence data, and three incidence data, respectively. † denotes the least bias, lowest RMSE, and closest to 95% coverage. Abbreviations: SE, standard error; RMSE, root mean square error; CI, confidence interval.

Table 2. The statistical behavior of Chao3 and adjusted Chao3 ($Chao3_{Adj}$) were analyzed in four scenarios to estimate the richness of multiple assemblages.

Sizes (Observed Richness)	Estimator	Average Estimate	Bias	Sample SE	Average Estimated SE	Sample RMSE	95% CI Coverage Rate
			Scenerio1				
500, 500, 500 (457)	Chao3	544.8	−55.2	21.7	20	59.3	0.42
	$Chao3_{Adj}$	572.9	−27.1 †	38.2	35.5	46.8 †	0.892 †
1000, 1000, 1000 (529.4)	Chao3	573.1	−26.9	13.5	12.8	30.1	0.6
	$Chao3_{Adj}$	587.3	−12.7 †	22.7	22.1	26 †	0.906 †
2000, 2000, 2000 (569.5)	Chao3	589.5	−10.5	9.3	8.3	14.1 †	0.806
	$Chao3_{Adj}$	596.3	−3.7 †	15.2	14.5	15.6	0.972 †
			Scenerio2				
500, 500, 10 (430.7)	Chao3	518	−82	20.1	20	84.4	0.13
	$Chao3_{Adj}$	545.7	−54.3 †	34.7	35.4	64.4 †	0.786 †
1000, 1000, 20 (502.5)	Chao3	553.2	−46.8	15.2	14.5	49.2	0.306
	$Chao3_{Adj}$	572.4	−27.6 †	26.4	25.4	38.1 †	0.858 †
2000, 2000, 40 (547.9)	Chao3	580.2	−19.8	12.9	11.7	23.7	0.694
	$Chao3_{Adj}$	593	−7 †	22.1	20.6	23.2 †	0.942 †

Table 2. *Cont.*

Sizes (Observed Richness)	Estimator	Average Estimate	Bias	Sample SE	Average Estimated SE	Sample RMSE	95% CI Coverage Rate
			Scenerio3				
500, 10, 10 (452.6)	Chao3	540.9	−59.1	21.1	19.9	62.7	0.32
	$Chao3_{Adj}$	567.2	−32.8 †	36.8	34.9	49.3 †	0.862 †
1000, 20, 20 (524.9)	Chao3	570.4	−29.6	13.8	13.1	32.6	0.57
	$Chao3_{Adj}$	584.4	−15.6 †	23.4	22.6	28 †	0.912 †
2000, 40, 40 (566.4)	Chao3	587.4	−12.6	9.1	8.5	15.5	0.782
	$Chao3_{Adj}$	594.4	−5.6 †	14.5	14.6	15.4 †	0.984 †
			Scenerio4				
10, 10, 10 (492.4)	Chao3	553.4	−46.6	18	16.6	50	0.41
	$Chao3_{Adj}$	575.3	−24.7 †	31.6	29.3	40.1 †	0.886 †
20, 20, 20 (541.8)	Chao3	576.1	−23.9	12.7	11.6	27	0.642
	$Chao3_{Adj}$	587.3	−12.7 †	21.3	20.1	24.8 †	0.920 †
40, 40, 40 (571.8)	Chao3	588.3	−11.7	7.9	7.6	14.1	0.806
	$Chao3_{Adj}$	594.2	−5.8†	12.8	13.4	14†	0.976 †

Note: data in Scenerio1, Scenerio2, Scenerio3, and Scenerio4 are separately composed of three abundance data, two abundance data and one incidence datum, one abundance datum and two incidence data, and three incidence data, respectively. † denotes the least bias, lowest RMSE, and closest to 95% coverage. Abbreviations: SE, standard error; RMSE, root mean square error; CI, confidence interval.

3.2. Simulation Results for Richness Estimation in a Single Assemblage

In this case, the integrated dataset consists of three random samples that are independently collected from the same assemblage. Each sample is simulated separately based on one of the three discussed abundance/incidence models, representing different sampling situations or methods. Different sample sizes are considered to indicate varying levels of sampling efforts, ranging from n = 200 to 600, with an increment of 50 for abundance data, and t = 10 to 50 with an increment of 5 for incidence data.

Four different scenarios are examined, including:

a. Three abundance models: random uniform, broken-stick, and log-normal;
b. Two abundance models: random uniform and broken-stick; one incidence model: log-normal;
c. One abundance model: random uniform; two incidence models: broken-stick and log-normal;
d. Three incidence models: random uniform, broken-stick, and log-normal.

The simulation results for these four scenarios are presented separately in Figure 1a–d and Table 1.

To estimate the richness of a single assemblage based on integrated data, as shown in Figure 1 and Table 1, under various scenarios of integrated datasets, both Chao3 and $Chao3_{Adj}$ can effectively reduce the negative bias of observed richness, and their bias and RMSEs decrease as sample size increases. When the sample size is small, both Chao3 and $Chao3_{Adj}$ provide a lower bound for the true richness, and they approach the true richness as sampling increases. The estimator of variance derived via the asymptotic method could perform well in all simulation scenarios (shown in Table 1).

Compared to Chao3, $Chao3_{Adj}$ offers a nearly unbiased and resilient estimate (with reduced bias and RMSE) and a more accurate 95% CI in every simulation scenario (as illustrated in Figure 1 and Table 1), even if the sample size is small.

3.3. Simulation Results for Richness Estimation across Multiple Assemblages

To evaluate the statistical behavior of the discussed estimators for richness estimation based on integrated data, three assemblages (sites) are assumed here. The integrated data consist of three samples that are collected separately from each site. It is assumed that the three sites comprise S = 600 species, with each site containing 300 species. There are varying numbers of shared species and unique species in each site.

Different sample sizes are considered to represent different sampling efforts, ranging from n = 400 to 4000, with an increment of 450 for abundance data, and t = 10 to 50, with an increment of 5 for incidence data. Four different scenarios are examined:

a. Three abundance models: random uniform, broken-stick, and log-normal;
b. Two abundance models: random uniform and broken-stick; one incidence model: log-normal;
c. One abundance model: random uniform; two incidence models: broken-stick and log-normal;
d. Three incidence models: random uniform, broken-stick, and log-normal.

The simulation results for these four scenarios are presented separately in Figure 2a,d and Table 2.

To assess richness across multiple assemblages using integrated data, both Figure 2 and Table 2 indicate that Chao3 and $\text{Chao3}_{\text{Adj}}$ effectively mitigate the negative bias of observed richness in each discussed integrated data scenario. As the sample size increases, the bias and RMSE for these two estimators decline. With limited sample sizes, Chao3 and $\text{Chao3}_{\text{Adj}}$ act as lower bounds for true richness. As sample size increases, these estimators converge toward true richness. The variance estimator, derived through the asymptotic method, consistently performs well across all simulated scenarios, as corroborated by Table 2. While $\text{Chao3}_{\text{Adj}}$ exhibits higher variance compared to Chao3, it yields more precise and consistent estimates, demonstrating less bias and RMSE. This ensures a more dependable 95% CI across all simulation scenarios, as emphasized in both Figure 2 and Table 2.

3.4. Remarks of Simulation Results

Undoubtedly, for a fixed sample size, a superior species richness estimator should exhibit lower bias and variance (i.e., low RMSE). Additionally, the coverage rate of its associated 95% confidence interval should be close to 0.95. As the sample size increases, an effective estimator should exhibit the following key characteristics: its bias should decrease; its accuracy (measured by RMSE) should enhance; and the coverage rate of its confidence interval should generally become better, ultimately approximating the true species richness when the sample size is adequately expansive. Based on these criteria, the following findings can be concluded from the simulation results:

a. In all simulation scenarios presented in Tables 1 and 2 and Figures 1 and 2, both Chao3 and $\text{Chao3}_{\text{Adj}}$ consistently provide robust lower bound estimates in all hypothetical models, and they tend to approach the true richness as the sample size increases;
b. Both Chao3 and $\text{Chao3}_{\text{Adj}}$ exhibit the essential statistical behaviors: their bias and RMSE decrease, resulting in more accurate 95% confidence intervals as the sample size increases (Tables 1 and 2);
c. The estimators of the discussed estimators' variance, derived using the asymptotic approach, perform well across all simulation scenarios (Tables 1 and 2);
d. Compared to Chao3, $\text{Chao3}_{\text{Adj}}$ exhibits lower bias, larger standard errors, and more accurate 95% confidence intervals for the true richness in all simulation scenarios (Tables 1 and 2);
e. When samples are directly collected from the entire region (Table 1), $\text{Chao3}_{\text{Adj}}$ has higher RMSEs compared to Chao3; however, when samples are separately collected from each local area within the target region (Table 2), $\text{Chao3}_{\text{Adj}}$ demonstrates lower RMSEs.

These findings collectively demonstrate the favorable performance of $Chao3_{Adj}$ in terms of bias, standard error, and accuracy of the 95% confidence interval, particularly when samples are collected from each site within the target area.

3.5. Using Datasets as True Assemblages

I utilized two biological survey datasets, representing true assemblages and generated separate datasets, from these two assemblages. In each dataset, the observed species relative abundance was considered as the true species relative abundance or detection probability. A sample of size n (or t) was then generated through sampling with replacement to create the sampling dataset. Different sample sizes were considered to indicate varying levels of sampling efforts.

The average estimate and other relevant statistics obtained using the 1000 generated datasets, as a function of sample size, are depicted in Figures 3 and 4 and Table A3 (refer to Appendix C for detailed information). These evaluations aimed to assess the statistical behaviors of the discussed richness estimators across four different sampling scenarios: three abundance data; two abundance data and one incidence datum; one abundance datum and two incidence data; and three incidence data.

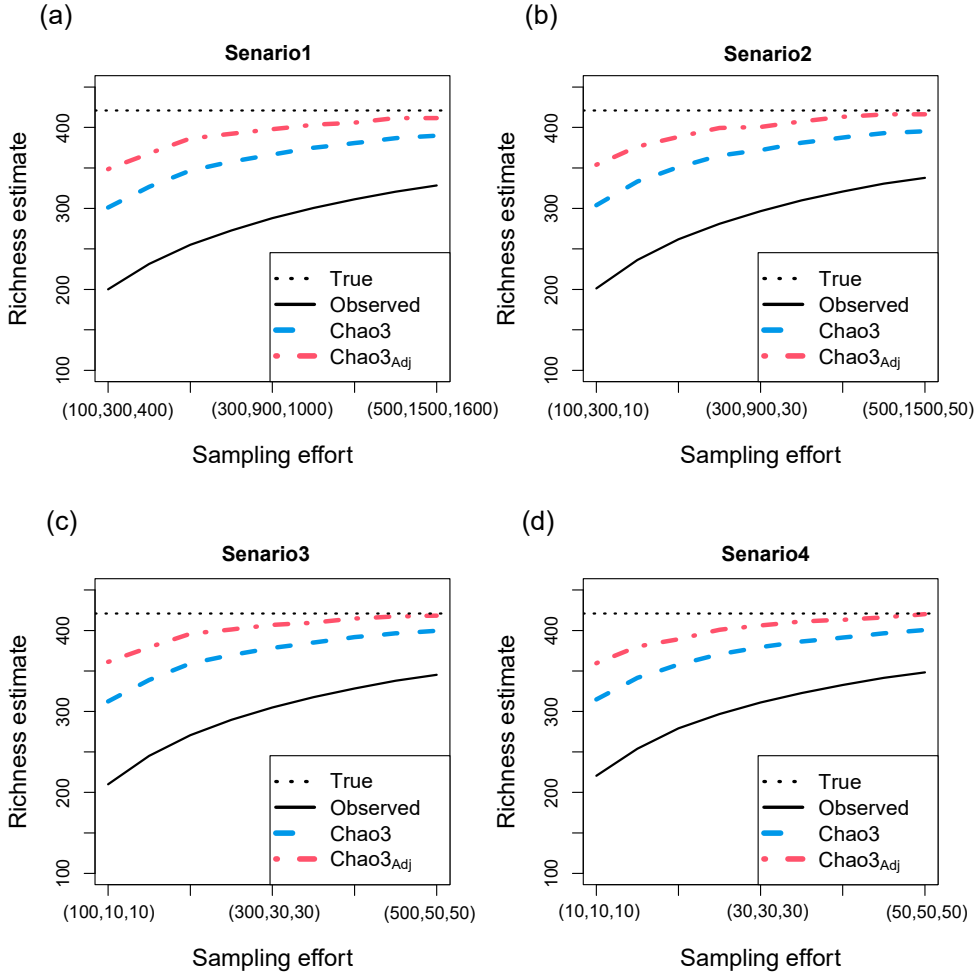

Figure 3. Plot of the true number of moth species and the average richness estimates (including observed richness, Chao3, and the adjusted Chao3 ($Chao3_{Adj}$)) as a function of the sampling effort for four different scenarios. Each scenario involves different combinations of abundance and incidence data: (**a**) three abundance data; (**b**) two abundance data and one incidence datum; (**c**) one abundance datum and two incidence data; and (**d**) three incidence data.

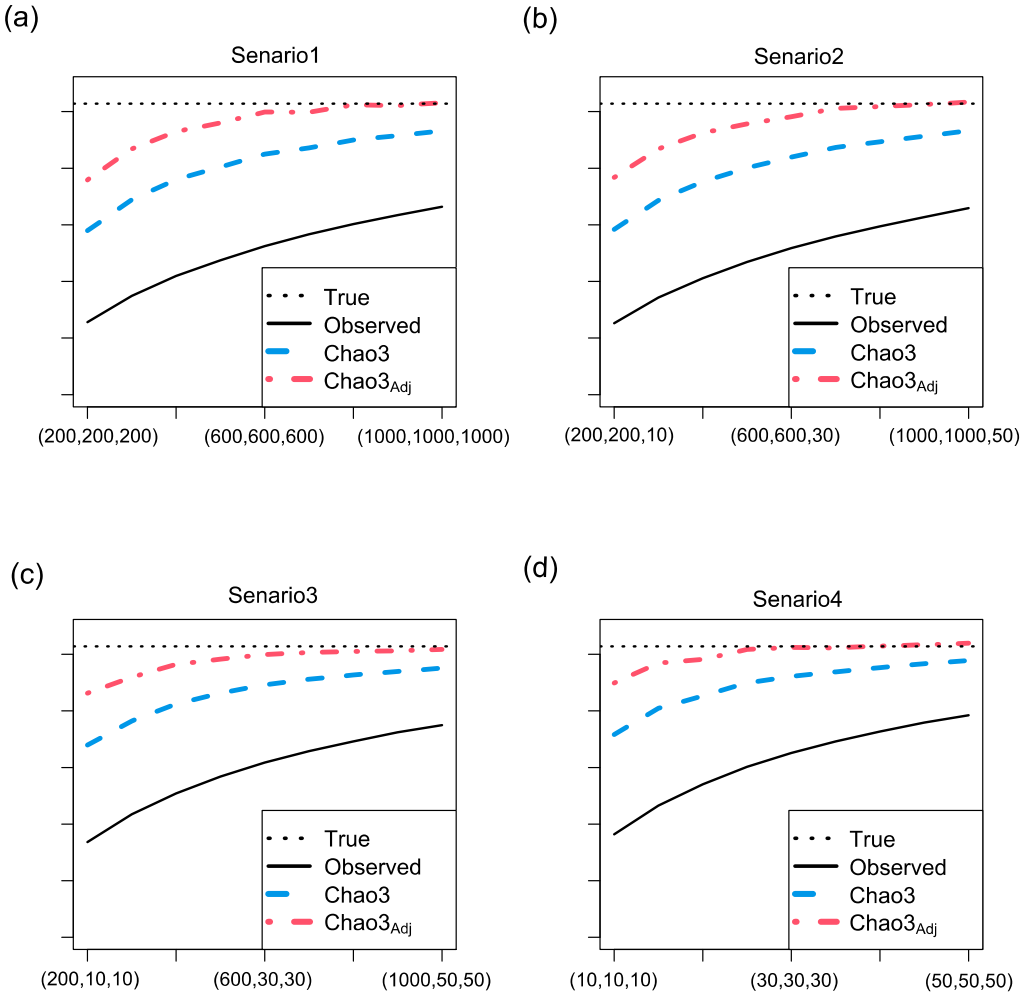

Figure 4. Plot of the true number of xylobiont beetle species and three average richness estimates (including observed richness, Chao3, and the adjusted Chao3 ($Chao3_{Adj}$), as a function of the sampling effort for four different scenarios. Each scenario includes different combinations of abundance and incidence data sets: (**a**) three abundance data sets; (**b**) two abundance data sets and one incidence data set; (**c**) one abundance data set and two incidence data sets; and (**d**) three incidence data sets.

3.5.1. Moth Species Data

Moth species data were collected in the Golfo Dulce region of the Costa Rican rainforest from July to October 2014 [33]. The target region was divided into three types of forest: creek forest; slope forest; and ridge forest. Light traps were set up at 18 sites, with six replicates within each forest type. Further details can be found in [33].

Table 3 presents a summary of the data, including the sample size, observed richness, and the first five species frequency counts for each forest type. In the pooled sample, a total of 421 species were recorded, with 115, 285, and 356 species observed in the creek, slope, and ridge forests, respectively. In this case, the survey datasets are considered as the true assemblages. The proportion of species in the sample is assumed to represent the species' relative abundance for generating individual-based abundance data, while the ratio between species abundance and the maximum abundance is considered as the species' detection probability for generating sample-based incidence data. Consequently, each type of forest has its corresponding abundance model and incidence model. Four different scenarios are examined:

a. Three abundance models: creek, slope, and ridge;
b. Two abundance models: creek and slope; one incidence model: ridge;
c. One abundance model: creek; two incidence models: slope and ridge;

d. Three incidence models: creek, slope, and ridge.

The simulation results for each scenario are depicted separately in Figure 3a–d and Table A3 (Appendix C).

Table 3. The summary of three moth samples separately collected from creek, slope, and ridge habitats in the Costa Rica rain forest [33], and three beetle samples separately collected from Quercus robur, Tilia cordata, and Fraxinus excelsior tree species in the Leipzig floodplain forest [34].

Moth Species Data							
Habitat	Sample Size	Observed Richness	Sample CV	f_1	f_2	f_3	f_{4+}
creek	461	115	1.86	54	22	9	32
slope	2382	285	2.40	92	47	23	123
ridge	3710	356	2.68	94	59	31	172
Beetle Species Data							
Tree Species	Sample Size	Observed Richness	Sample CV	f_1	f_2	f_3	f_{4+}
Quercus robur	2205	174	2.72	74	29	10	61
Tilia cordata	1737	198	2.37	92	27	16	63
Fraxinus excelsior	1797	184	3.30	77	31	11	65

Abbreviations: CV, coefficient of variation; f_1, f_2, f_3, and f_{4+} are, respectively, the singleton richness, doubleton richness, tripleton richness, and the number of species observed more than three times in the sample.

3.5.2. Xylobiont Beetle Species Data

The second real dataset comprises xylobiont beetle species data collected from the Leipzig floodplain forest in 2016 [34]. The beetle species data were collected separately from three dominant tree species in the Leipzig floodplain forest area: Quercus robur (QR); Tilia cordata (TC); and Fraxinus excelsior (FE). Table 3 provides information on the sample size, observed richness, and the first three species frequency counts for each tree species. In total, 307 beetle species were observed, with 174, 198, and 184 species recorded in QR, TC, and FE tree species, respectively.

In this case, the survey datasets are treated as the true assemblages, and the species abundance/incidence model is constructed using the same method discussed earlier for each tree species. Four different scenarios are considered, including:

a. Three abundance models: QR, TC, and FE;
b. Two abundance models: QR and TC; one incidence model: FE;
c. One abundance model: QR; two incidence models: TC and FE;
d. Three incidence models: QR, TC, and FE.

The simulation results for each scenario are presented separately in Figure 4a,d and Table A4 (Appendix C).

Abbreviations: CV, coefficient of variation; f_1, f_2, f_3, and f_{4+} are, respectively, the singleton richness, doubleton richness, tripleton richness, and the number of species observed more than three times in the sample.

The results of the analysis, as depicted in Figures 3 and 4 and the Appendix C in Appendix C, demonstrate that both Chao3 and $\text{Chao3}_{\text{Adj}}$ effectively reduce the underestimation of observed richness. From a theoretical standpoint, and considering the results of the simulation study, it is expected that $\text{Chao3}_{\text{Adj}}$ exhibits lower bias compared to Chao3, particularly when there is high heterogeneity as indicated by a high coefficient of variation (CV). Furthermore, the Appendix C in Appendix C confirm that $\text{Chao3}_{\text{Adj}}$ has lower bias, higher standard error, and lower root mean square error (RMSE). The higher estimated standard error of $\text{Chao3}_{\text{Adj}}$ compared to Chao3 suggests that the former estimator may provide a more accurate 95% confidence interval for true richness. This observation aligns with the findings of the simulation study presented in Tables 1 and 2.

Overall, the results support the notion that $\text{Chao3}_{\text{Adj}}$ has less bias and performs better in terms of standard error and RMSE, indicating its potential to provide more accurate

estimates and confidence intervals for true richness, particularly in scenarios with higher heterogeneity.

4. Discussion and Conclusions

Species richness is the most commonly used diversity metric in ecological research. Numerous methodologies for estimating total species richness in a given area have been explored in the scholarly literature. These methods can be broadly categorized as either parametric or non-parametric estimators. Parametric estimators leverage assumptions about species compositions and necessitate complex computational processes to resolve likelihood functions. Furthermore, these estimators often encounter convergence issues during iterative numerical procedures or yield high variance when sample sizes are small. As such, they are less suited to small sample sizes and seldom applied in ecological studies. In contrast, non-parametric estimators, which do not impose assumptions on species composition and feature simple, closed formulae, tend to be more robust in various simulation cases. Consequently, they have gained widespread adoption in ecological studies. However, parametric and non-parametric approaches are derived assuming the sample is collected randomly from the target area, whereas in the abundance sample, the number of individuals belonging to a specific species, or in the incidence sample, the number of plots where a species is detected, both adhere to a binomial distribution.

Estimating species richness for a large-scale area or multiple assemblages poses a statistical challenge due to the difficulty of obtaining a random sample from the entire region. Typically, integrated data collected for assessing species richness in such cases consist of multiple samples that are individually sampled from each assemblage or local-area. Additionally, these samples may employ different sampling schemes or strategies. Therefore, the detection probability of a species may vary across the samples, and the data format (individual-based abundance data or sample-based incidence data) can differ among the samples. As a result, the pooled sample of the integrated data cannot be considered a random sample from the entire region, even though each individual sample is randomly collected from its respective local area or assemblage. Consequently, the pooled sample from the integrated data cannot be modeled using a traditional sampling distribution. Additionally, no estimator has been previously developed in the literature to estimate richness based on integrated data.

In this context, richness estimators that rely only on the frequency counts of rare species in the sample have been theoretically demonstrated to be applicable to the pooled sample, as long as the samples are randomly collected and the sample size is not excessively small. While numerous non-parametric techniques are grounded on the frequency tallies of infrequent species, such as the widely-adopted jackknife estimators, these often contravene essential standards, where bias and root mean squared error (RMSE) ought to diminish with the increasing sample size. Additionally, they are not consistently reliable, especially with limited data or when the assemblage is highly heterogenous [26,30,35]. Hence, this manuscript does not delve into these estimators; instead, it focuses on the widely used Chao's lower bound estimator, which utilizes the numbers of singletons and doubletons to estimate undetected richness and provides a reabiable estimate when sample size is not sufficiently large. This is the primary approach discussed in this text. In this research, a lower bound estimator (Chao3) and its bias-corrected estimator ($\text{Chao3}_{\text{Adj}}$) are theoretically proven to be suitable for estimating richness in multiple assemblages based on the pooled sample from integrated data. Chao3 derived using Cauchy–Schwarz inequality provides a lower bound richness estimate, while $\text{Chao3}_{\text{Adj}}$ corrects the bias of Chao3 based on the Good–Turing frequency formula.

Since a single statistical model cannot accurately fit all ecological communities, there is no existing richness estimator that is uniformly unbiased for all such communities. Therefore, the development of a more robust estimator becomes an essential issue in estimating species richness. In this case, an estimator should be designed with functions such that both its bias and accuracy (quantified by RMSE), the two most crucial properties

for an estimator, decrease as the sample size increases. Additionally, the coverage rate of the 95% confidence interval should approach 0.95 as the sample size expands. Based on these critical criteria, I arrived at the following conclusions from our simulation results. In all simulated scenarios, the observed richness in the samples was significantly underestimated, particularly when the sample size was small or when the species composition of the community was highly heterogeneous. Simulation results demonstrate that both Chao3 and $\text{Chao3}_{\text{Adj}}$ could correct the severe negative bias of observed richness, and their bias and RMSEs decreased as the sample size increased across all models discussed. These two estimators provide lower bound estimates in all hypothetical models and tend to converge to the true richness as the sample size increases. This implies that both proposed estimators can be used to estimate regional richness based on the pooled sample from integrated data, which aligns with the theoretical findings. Notably, when the sample size is small or the community exhibits high heterogeneity, Chao3 presents a significant negative bias, and its 95% confidence interval (CI) coverage rate is generally much lower than 0.95. In this case, $\text{Chao3}_{\text{Adj}}$ outperforms Chao3 with lower bias, lower root mean square error (RMSE), higher standard error (s.e.), and a more accurate 95% CI for true richness. This indicates that $\text{Chao3}_{\text{Adj}}$ tends to be more stable and less susceptible to the challenges mentioned compared to the traditional Chao's lower bound estimator.

In the text, all proposed estimators are based on the assumption that each sampling unit is collected independently. When individuals in the sample are not sampled independently, and individuals of the same species are more likely to be sampled, the proposed estimator can be severely negatively biased. Therefore, when individuals of a species exhibit spatial aggregation patterns, it becomes challenging to collect them independently and individually. In such cases, the individual-based sampling method may not be practical to implement. In these situations, it is recommended to use the sample-based incidence sampling method for collecting data to assess species richness. This method could approximately ensure that the sampling units are sampled independently to align with the underlying model assumptions.

The proposed estimators depend solely on data related to rare species, namely, the count of singletons, doubletons, and tripletons in the aggregated sample, in order to estimate unobserved richness. Compared to more complex computations such as the maximum likelihood method, these methods offer a computational advantage as they provide estimates more simply, and their user-friendliness is emphasized by their straightforward formulae. From a practical standpoint, another significant benefit is that these estimators eliminate the need for detailed tracking of the count of abundant species observed during field surveys, thereby considerably reducing the field sampling burden.

In summary, while the newly introduced estimators show promising results in the hypothetical models and two real datasets, their applicability still necessitates further validation using a broader range of real datasets in the future.

Funding: This research was funded by National Science and Technology Council (Taiwan), grant number 111-2118-M-002-002-MY2.

Data Availability Statement: All R codes used in this paper are archived on Zenodo: https://doi.org/10.5281/zenodo.8118860 (accessed on 6 July 2023). The moth species dataset used in this paper is archived on the website: https://doi.org/10.5061/dryad.783p8m2 (accessed on 6 July 2023). The xylobiont beetle species dataset used in this paper is archived on the website: https://doi.org/10.5061/dryad.d7wm37q0g (accessed on 6 July 2023).

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Numerical Study to Show That the Expectation of Frequency Counts of Rare Species in the Pooled Sample Is Approximately Identical to the Probability Sum of the Poisson Distribution

Assume that there are S(=400) species contained in the target area, when both abundance data and incidence data are collected from the same area, species abundance (X_i)

follows a binomial distribution with size n and probability p_i (i.e., $X_i \sim Binomial(n,\ p_i)$), where $\sum_{i=1}^{S} p_i = 1$. Additionaly, species incidence count (Y_i) follows a binomial distribution with size t and probability π_i (i.e., $Y_i \sim Binomial(t, \pi_i)$). Let the species frequency in the pooled sample be $Z_i = X_i + Y_i$; then, $G_k = \sum_{i=1}^{S} I(Z_i = k)$ is the count of species frequency in the pooled sample.

Here, I implement a simulation study to show that the following equation (Equation (A1)) is approximately held when sample size (n) and (t) is large and k is small (i.e., $k = 0,\ 1,\ 2,\ 3$):

$$E[G_k] = \sum_{i=1}^{S} P(Z_i = k) \approx \sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i},\ where\ \lambda_i = np_i + t\pi_i. \tag{A1}$$

a. For the individual-based abundance sampling model, the species detection probabilities (or species relative abundance) $p_i = ca_i\ and\ a_i \sim U(0,\ 1),\ i = 1,\ \ldots,\ S$, where c is a normalizing constant such that $\sum_{i=1}^{S} p_i = 1$;
b. For the sample-based incidence model, the species detection probabilities $\pi_i \sim U(0.05, 0.2), i = 1,\ 2, \ldots, 50\ and\ \pi_i \sim U(0.8, 1),\ i = 51,\ 52,\ \ldots, 100$.

In the simulation study, different sample sizes are understood to indicate different sampling efforts. For each simulation scenario, 1000 simulated datasets are generated; then, G_k is averaged over the 1000 simulated datasets to estimate $E[G_k]$. The table below shows that the equation (Equation (A1)) could be roughly held for $k = 0,\ 1,\ 2,\ 3$ when the sample sizes (n and t) are sufficiently large.

Table A1. Expectation of the count for the first five rare species frequencies.

Sample Size		$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$n = 100, t = 10$	$\sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i}$	59.4	65.1	43	23.4	14.3	13.6
	$E[G_k]$	55.6	66.9	44.8	22	8	3.2
$n = 200, t = 20$	$\sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i}$	21.9	39.9	42.8	35.8	25.7	16.3
	$E[G_k]$	20.2	39.2	43.3	37.4	27.1	17
$n = 300, t = 30$	$\sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i}$	9.4	22.4	30.5	32	29.3	24.4
	$E[G_k]$	8.5	21.5	30	32.5	29.9	25.7
$n = 400, t = 40$	$\sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i}$	4.4	12.8	20.5	24.7	25.6	24.3
	$E[G_k]$	4	12	20	24.5	25.4	24.8
$n = 500, t = 50$	$\sum_{i=1}^{S} \frac{\lambda_i^k}{k!} e^{-\lambda_i}$	2.2	7.4	13.7	18.4	20.7	21.2
	$E[G_k]$	1.9	6.9	13.1	18.1	20.6	21.2

The simulation results show that the expectations of the first three frequency counts (i.e., $G_0,\ G_1,\ G_2,\ G_3$) are roughly identical to the probability sum of the Poisson distribution with mean $\lambda_i = np_i + t\pi_i,\ i = 1,\ 2,\ \ldots, S$.

Appendix B. The Summary of the Statistical Properties of the Richness Estimators Discussed in the Text

Table A2. The summary of the statistical properties of the richness estimators discussed in the text.

Available Data and Notation	Richness Estimator	Pluses and Minuses
Individual-based abundance data: sampling unit is an individual randomly selected from target assemblage and identified as species. S_{obs} : the observed richness; f_1 : the singleton richness in the sample; f_2 : the doubleton richness in the sample; f_3 : the tripleton richness in the sample.	Chao1 [6] $S_{obs} + \begin{cases} \frac{f_1^2}{2f_2} & if\ f_2 > 0 \\ \frac{f_1(f_1-1)}{2} & if\ f_2 = 0 \end{cases}$	1. A lower bound estimator of richness for all species composition models; 2. A nearly unbiased estimator when rare species are homogeneous; 3. Has severely negative bias when community is highly heterogeneous.
	Chao1$_{\text{Adj}}$ [19] $S_{obs} + \frac{f_1^2}{2f_2}\left(2 - \frac{2f_2^2}{3f_1f_3}\right)$	1. A lower bound estimator of richness for gamma-Poisson models; 2. Compared to Chao1, Chao1$_{\text{Adj}}$ has less bias, higher variance, lower RMSE, and a more accurate coverage rate of 95% confidence interval.
Sample-based incidence data: sampling unit is a quadrat or plot, and only the incidence of species appearing in the selected plot is recorded. S_{obs} : the observed richness; Q_1 : the singleton richness in the sample; Q_2 : the doubleton richness in the sample; Q_3 : the tripleton richness in the sample; t : the number of selected plot.	Chao2 [7] $S_{obs} + \begin{cases} \frac{t-1}{t}\frac{Q_1^2}{2Q_2} & if\ Q_2 > 0 \\ \frac{t-1}{t}\frac{Q_1(Q_1-1)}{2} & if\ Q_2 = 0 \end{cases}$	1. A lower bound estimator of richness for all species composition model; 2. A nearly unbiased estimator when rare species are homogeneous; 3. Has severely negative bias when community is highly heterogeneous.
	Chao2$_{\text{Adj}}$ [20] $S_{obs} + \frac{t-1}{t}\frac{Q_1^2}{2Q_2}\left(2 - \frac{2Q_2^2}{3Q_1Q_3}\right)$	1. A lower bound estimator of richness for beta-binomial models; 2. Compared to Chao2, Chao2$_{\text{Adj}}$ has less bias, higher variance, lower RMSE, and has more accurate coverage rate of 95% confidence interval.
Pooled sample of integrated data: we directly pool the individual-based abundance data and sample-based incidence data as a new sample. S_{obs} : the observed richness; G_1 : the singleton richness in pooled sample; G_2 : the doubleton richness in the pooled sample; G_3 : the tripleton richness in the pooled sample.	Chao3 (New proposed) $S_{obs} + \begin{cases} \frac{G_1^2}{2G_2} & if\ G_2 > 0 \\ \frac{G_1(G_1-1)}{2} & if\ G_2 = 0 \end{cases}$	1. Chao3 is available for pooled sample of integrated data; 2. A lower bound estimator of richness when sample size is large enough; 3. Has severely negative bias when community is highly heterogeneous.
	Chao3$_{\text{Adj}}$ (New proposed) $S_{obs} + \frac{G_1^2}{2G_2}\left(2 - \frac{2G_2^2}{3G_1G_3}\right)$	1. Compared to Chiao3, Chao3$_{\text{Adj}}$ has less bias, higher variance, and lower RMSE; 2. Has more accurate coverage rate of 95% confidence interval.

Appendix C

Table A3. The statistical behavior of Chao3 and $Chao3_{Adj}$ were analyzed in four scenarios to estimate the number of the moth species (richness = 421) [22].

Size (Observed Richness)	Estimator	Average Estimate	Bias	Sample SE	Average Estimated SE	Sample RMSE	95% CI Coverage Rate
			Scenerio1				
100, 300, 400 (200.0)	Chao3	301.6	−119.4	30.4	29.4	123.2	0.158
	$Chao3_{Adj}$	350.7	−70.3 †	57.8	56.8	91 †	0.85 †
300, 900, 1000 (288.2)	Chao3	367.6	−53.4	22.1	22.4	57.8	0.49
	$Chao3_{Adj}$	399.1	−21.9 †	39.6	41.7	45.2 †	0.926 †
500, 1500, 1600 (328.2)	Chao3	391.9	−29.1	19.8	18.7	35.2 †	0.754
	$Chao3_{Adj}$	416.4	−4.6 †	36.1	34.1	36.3	0.943 †
			Scenerio2				
100, 300, 10 (201.2)	Chao3	303.7	−117.3	30.1	29.5	121.1	0.18
	$Chao3_{Adj}$	353.4	−67.6 †	58.3	56.9	89.2 †	0.846 †
300, 900, 30 (297.3)	Chao3	333	−88	30.5	27.2	93.1	0.294
	$Chao3_{Adj}$	377.2	−43.8 †	57.9	51.9	72.6 †	0.891 †
500, 1500, 50 (338.4)	Chao3	351.8	−69.2	26	25.1	73.9	0.408
	$Chao3_{Adj}$	389	−32 †	49	47.2	58.5 †	0.931 †
			Scenerio3				
100, 10, 10 (200)	Chao3	310.2	−110.8	29.4	28.8	114.7	0.186
	$Chao3_{Adj}$	357.1	−63.9 †	56.3	54.8	85.1 †	0.846 †
300, 30, 30 (288.2)	Chao3	339.2	−81.8	28.6	26.5	86.7	0.32
	$Chao3_{Adj}$	380	−41 †	54.7	49.6	68.3 †	0.882 †
500, 50, 50 (328.2)	Chao3	357.2	−63.8	24.5	24.3	68.3	0.4
	$Chao3_{Adj}$	393.7	−27.3 †	46.5	45.9	53.9 †	0.942 †
			Scenerio4				
10, 10, 10 (221.1)	Chao3	313.3	−107.7	28.7	26.7	111.4	0.168
	$Chao3_{Adj}$	356.3	−64.7 †	54.6	50.9	84.6 †	0.842 †
30, 30, 30 (311.1)	Chao3	341.3	−79.7	25.4	25	83.7	0.286
	$Chao3_{Adj}$	379.5	−41.5 †	47.6	47	63.1 †	0.911 †
50, 50, 50 (348.0)	Chao3	359.3	−61.7	23.6	23.2	66	0.428
	$Chao3_{Adj}$	393.7	−27.3 †	44.4	43.3	52.1 †	0.941 †

Note: data in Scenerio1, Scenerio2, Scenerio3 and Scenerio4 are separately composed by three abundance data, two abundance data and one incidence datum, one abundance datum and two incidence data, and three incidence data, respectively. † denotes the least bias, lowest RMSE, and closest to 95% coverage. Abbreviations: SE, standard error; RMSE, root mean square error; CI, confidence interval.

Table A4. The statistical behavior of Chao3 and $Chao3_{Adj}$ were analyzed in four scenarios to estimate the number of the beetle species (richness = 207) in Leipzig floodplain forest [23].

Size (Observed Richness)	Estimator	Average Estimate	Bias	Sample SE	Average Estimated SE	Sample RMSE	95% CI Coverage Rate
			Scenerio1				
200, 200, 200 (132.2)	Chao3	212.3	−94.7	29.9	29.1	99.3	0.321
	$Chao3_{Adj}$	253.7	−53.3 †	59	56.2	79.5 †	0.856 †
600, 600, 600 (196.4)	Chao3	233.7	−73.3	28.4	27.4	78.6	0.428
	$Chao3_{Adj}$	271.6	−35.4 †	55.4	52.5	65.8 †	0.912 †
1000, 1000, 1000 (228.1)	Chao3	250	−57	28.3	26.5	63.6	0.561
	$Chao3_{Adj}$	287.6	−19.4 †	54.4	50.5	57.7 †	0.928 †
			Scenerio2				
200, 200, 10 (115.7)	Chao3	199.1	−107.9	34.6	31.8	113.3	0.287
	$Chao3_{Adj}$	245.9	−61.1 †	69.2	62.3	92.3 †	0.841 †
600, 600, 30 (181.3)	Chao3	221.3	−85.7	30.3	29.5	91	0.386
	$Chao3_{Adj}$	263.1	−43.9 †	59.7	56.9	74.1 †	0.892 †
1000, 1000, 50 (215.8)	Chao3	238.3	−68.7	30.3	28.3	75.1	0.497
	$Chao3_{Adj}$	278.8	−28.2 †	59.7	54	66 †	0.912 †
			Scenerio3				
200, 10, 10 (127.5)	Chao3	209.3	−97.7	30.4	30	102.3	0.334
	$Chao3_{Adj}$	252.2	−54.8 †	59.4	58.5	80.8 †	0.855 †
600, 30, 30 (195.4)	Chao3	232	−75	29.1	27.8	80.5	0.43
	$Chao3_{Adj}$	269.9	−37.1 †	57.1	52.9	68.1 †	0.885 †
1000, 50, 50 (227.8)	Chao3	245.5	−61.5	26.4	25.3	66.9	0.504
	$Chao3_{Adj}$	278.4	−28.6 †	50.4	47.7	57.9 †	0.929 †
			Scenerio4				
10, 10, 10 (141.1)	Chao3	228.5	−78.5	33.2	31	85.2	0.453
	$Chao3_{Adj}$	274.3	−32.7 †	66.1	60.3	73.7 †	0.882 †
30, 30, 30 (213.0)	Chao3	251.5	−55.5	27.9	28.5	62.2	0.613
	$Chao3_{Adj}$	291.2	−15.8 †	55.1	54.3	57.3 †	0.916 †
50, 50, 50 (246.2)	Chao3	264.4	−42.6	27.8	25.6	50.8 †	0.685
	$Chao3_{Adj}$	297.5	−9.5 †	52.7	47.7	53.6	0.922 †

Note: The data sets in Scenerio1, Scenerio2, Scenerio3, and Scenerio4 are separately composed of three abundance data sets, two abundance data sets and one incidence data set, one abundance data set and two incidence data sets, and three incidence data sets. † denotes the least bias, lowest RMSE, and closest to 95% coverage. Abbreviations: SE, standard error; RMSE, root mean square error; CI, confidence interval.

References

1. Bunge, J.; Fitzpatrick, M. Estimating the number of species: A review. *J. Am. Stat. Assoc.* **1993**, *88*, 364–373.
2. Colwell, R.K.; Coddington, J.A. Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **1994**, *345*, 101–118. [PubMed]
3. Chao, A.; Chiu, C.-H. Species richness: Estimation and comparison. *Wiley StatsRef Stat. Ref. Online* **2016**, *1*, 26.
4. Colwell, R.K.; Chao, A.; Gotelli, N.J.; Lin, S.-Y.; Mao, C.X.; Chazdon, R.L.; Longino, J.T. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* **2012**, *5*, 3–21. [CrossRef]
5. Hortal, J.; Borges, P.A.; Gaspar, C. Evaluating the performance of species richness estimators: Sensitivity to sample grain size. *J. Anim. Ecol.* **2006**, *75*, 274–287. [CrossRef]
6. Bunge, J.; Willis, A.; Walsh, F. Estimating the number of species in microbial diversity studies. *Annu. Rev. Stat. Appl.* **2014**, *1*, 427–445. [CrossRef]
7. Wilson, R.M.; Collins, M.F. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* **1992**, *79*, 543–553. [CrossRef]

8. Chao, A. Species estimation and applications. In *Encyclopedia of Statistical Sciences*; Balakrishnan, N., Read, C.B., Vidakovic, B., Eds.; Wiley: New York, NY, USA, 2005; pp. 7907–7916.
9. Flather, C.H. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *J. Biogeogr.* **1996**, *23*, 155–168. [CrossRef]
10. Ter Steege, H.; Prado, P.I.; Lima, R.A.d.; Pos, E.; de Souza Coelho, L.; de Andrade Lima Filho, D.; Salomão, R.P.; Amaral, I.L.; de Almeida Matos, F.D.; Castilho, C.V. Biased-corrected richness estimates for the Amazonian tree flora. *Sci. Rep.* **2020**, *10*, 10130. [CrossRef]
11. Chao, A.; Bunge, J. Estimating the number of species in a stochastic abundance model. *Biometrics* **2002**, *58*, 531–539. [CrossRef]
12. Sanathanan, L. Estimating the size of a multinomial population. *Ann. Math. Stat.* **1972**, 142–152. [CrossRef]
13. Sanathanan, L. Estimating the size of a truncated sample. *J. Am. Stat. Assoc.* **1977**, *72*, 669–672. [CrossRef]
14. Chao, A. Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **1984**, *11*, 265–270.
15. Chao, A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **1987**, *43*, 783–791. [CrossRef] [PubMed]
16. Burnham, K.P.; Overton, W.S. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **1978**, *65*, 625–633. [CrossRef]
17. Burnham, K.P.; Overton, W.S. Robust estimation of population size when capture probabilities vary among animals. *Ecology* **1979**, *60*, 927–936. [CrossRef]
18. Bellard, C.; Bertelsmeier, C.; Leadley, P.; Thuiller, W.; Courchamp, F. Impacts of climate change on the future of biodiversity. *Ecol. Lett.* **2012**, *15*, 365–377. [CrossRef]
19. Cardinale, B.J.; Duffy, J.E.; Gonzalez, A.; Hooper, D.U.; Perrings, C.; Venail, P.; Narwani, A.; Mace, G.M.; Tilman, D.; Wardle, D.A. Biodiversity loss and its impact on humanity. *Nature* **2012**, *486*, 59–67. [CrossRef]
20. Cavicchioli, R.; Ripple, W.J.; Timmis, K.N.; Azam, F.; Bakken, L.R.; Baylis, M.; Behrenfeld, M.J.; Boetius, A.; Boyd, P.W.; Classen, A.T. Scientists' warning to humanity: Microorganisms and climate change. *Nat. Rev. Microbiol.* **2019**, *17*, 569–586. [CrossRef] [PubMed]
21. Delgado-Baquerizo, M.; Maestre, F.T.; Reich, P.B.; Jeffries, T.C.; Gaitan, J.J.; Encinar, D.; Berdugo, M.; Campbell, C.D.; Singh, B.K. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat. Commun.* **2016**, *7*, 1–8. [CrossRef] [PubMed]
22. Stork, N.E. How many species of insects and other terrestrial arthropods are there on Earth? *Ann. Rev. Entomol.* **2018**, *63*, 31–45. [CrossRef] [PubMed]
23. Louca, S.; Mazel, F.; Doebeli, M.; Parfrey, L.W. A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol.* **2019**, *17*, e3000106. [CrossRef]
24. Good, I.J.; Toulmin, G. The Number of New Species and the Increase of Population Coverage When a Sample Is Increased. *Biometrika* **1956**, *43*, 45–63. [CrossRef]
25. Cormack, R.M. Log-linear models for capture-recapture. *Biometrics* **1989**, *45*, 395–413. [CrossRef]
26. Chiu, C.H.; Wang, Y.T.; Walther, B.A.; Chao, A. An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics* **2014**, *70*, 671–682. [CrossRef]
27. Chao, A.; Chiu, C.H.; Colwell, R.K.; Magnago, L.F.S.; Chazdon, R.L.; Gotelli, N.J. Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. *Ecology* **2017**, *98*, 2914–2929. [CrossRef] [PubMed]
28. Chao, A.; Colwell, R.K. Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT Stat. Oper. Res. Trans.* **2017**, *41*, 3–54.
29. Good, I.J. The population frequencies of species and the estimation of population parameters. *Biometrika* **1953**, *40*, 237–264. [CrossRef]
30. Chiu, C.-H. A more reliable species richness estimator based on the Gamma–Poisson model. *PeerJ* **2023**, *11*, e14540. [CrossRef]
31. Chiu, C.H. Incidence-data-based species richness estimation via a Beta-Binomial model. *Methods Ecol. Evol.* **2022**, *13*, 2546–2558. [CrossRef]
32. Chao, A.; Lee, S.-M. Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* **1992**, *87*, 210–217. [CrossRef]
33. Rabl, D.; Gottsberger, B.; Brehm, G.; Hofhansl, F.; Fiedler, K. Moth assemblages in Costa Rica rain forest mirror small-scale topographic heterogeneity. *Biotropica* **2020**, *52*, 288–301. [CrossRef]
34. Haack, N.; Grimm-Seyfarth, A.; Schlegel, M.; Wirth, C.; Bernhard, D.; Brunk, I.; Henle, K. Patterns of richness across forest beetle communities—A methodological comparison of observed and estimated species numbers. *Ecol. Evol.* **2021**, *11*, 626–635. [CrossRef] [PubMed]
35. Chiu, C.H. A species richness estimator for sample-based incidence data sampled without replacement. *Methods Ecol. Evol.* **2023**, 1–12. [CrossRef]