diversity

MDPI

Article

Comparing Nonparametric Estimators for the Number of Shared Species in Two Populations

Jack C. Yue [1,2,*], Murray K. Clayton [3] and Chi-Ruei Hung [1]

[1] Department of Statistics, National Chengchi University, Taipei 11605, Taiwan; terry89073@gmail.com
[2] Center for Fundamental Science, Kaohsiung Medical University, Kaohsiung 80378, Taiwan
[3] Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA; clayton@stat.wisc.edu
* Correspondence: csyue@nccu.edu.tw

Abstract: It is often of interest to biologists to evaluate whether two populations are alike with respect to a similarity index; assessing the numbers of shared species is one way to do this. In this study, we propose two Turing-type estimators for the probability of discovering new shared species and two jackknife-type estimators for the number of shared species in two populations. We use computer simulation and empirical data analysis to evaluate the proposed approach. The jackknife-type estimators provide stable and reliable estimates, for both the probability of discovering new shared species and the number of shared species. We also compare the jackknife-type estimates with that of using sample coverage to estimate the number of shared species. The estimate of using sample coverage has better performance in the case of even populations, while the jackknife-type estimates have smaller bias in the case of unbalanced populations. When combined with a stopping rule based on the probability of observing new shared species, confidence intervals based on the proposed jackknife-type estimators can provide better coverage probability for the true number of shared species. The jackknife-type estimates can provide coverage probability close to 0.95 in all examples.

Keywords: species diversity; number of share species; comparing populations; jackknife estimator; simulation

check for updates

Citation: Yue, J.C.; Clayton, M.K.; Hung, C.-R. Comparing Nonparametric Estimators for the Number of Shared Species in Two Populations. *Diversity* **2022**, *14*, 243. https://doi.org/10.3390/d14040243

Academic Editors: Michael Wink and Tsung-Jen Shen

Received: 14 January 2022
Accepted: 16 March 2022
Published: 26 March 2022

1. Introduction

Species diversity is a feature often used to compare populations. Among all measures, the number of species is a simple descriptor but its estimation is remarkably challenging. Indeed, there were over 550 papers on the topic as of 1991, as summarized by Bunge and Fitzpatrick [1]. Our primary interest in this paper is to study and evaluate the estimators of the number of shared species in two communities, borrowing ideas from the estimators of number of species in one population.

Good proposed an elegant idea for estimating the probability of discovering new species (Turing's estimator) [2], using only the information of species observed exactly once in the sample. Following Good's idea, Burnham and Overton applied a jackknife technique to obtain a nonparametric estimator of the number of species in one population based on the distribution of observed species frequency [3]. Chao and Lee proposed an alternative nonparametric estimator based on the concept of sample coverage [4], and Chao et al. later modified this estimator using the information of species appearing not more than 10 times in the sample [5].

The estimation of the number of shared species in two populations can be generalized from the species richness in one population. Using the information of sample coverage, Chao et al. proposed a nonparametric estimator of the number of shared species [6] and Chuang et al. developed three different types of jackknife estimators [7]. However, neither of these approaches takes advantage of jackknifing the sample and we don't know if there are enough observations to make the final decision. In a different approach, Yue and

Clayton modified Good's idea and proposed an estimator for the probability of observing new shared species in two populations [8]. They used this probability as an indicator to stop collecting more observations, which can lower overall study cost, in comparing species similarity between two populations. Therefore, in addition to developing two jackknife-type estimators for the number of shared species and comparing them to that by Chao et al. [6], we also evaluate if it is possible to use stopping indicator for estimating the number of shared species.

Note that, in addition to the proposed two jackknife-type estimators of the number of shared species in two populations, we also consider the feasibility of using the probability of observing new shared species as stopping rule. In the next section we briefly review the concept behind jackknife estimators, including Turing-type estimates of the probability of discovering new shared species. We then develop two nonparametric estimators for the number of shared species in two populations and discuss the variances of those estimators. We will use computer simulations and empirical analysis of varies data sets to evaluate the proposed approach.

2. Methodology

Suppose there are two populations and let $\overrightarrow{p} = (p_1, p_2, \ldots, p_s)$ and $\overrightarrow{q} = (q_1, q_2, \ldots, q_s)$ denote the species proportions of the two populations, where s is the number of distinct species in the pooled communities. In other words, if we randomly select a single sample, then the probabilities of observing the species i are p_i and q_i $(1 \leq i \leq s)$ in populations 1 and 2, respectively. Let s_0 be the number of shared species and, without loss of generality, let the species 1, 2, . . . , and s_0 be the shared species in both populations. Also, let $X_i(n)$ and $Y_i(n)$ denote the numbers of times of species i is observed based on n observations from each of populations 1 and 2, respectively, and let $s_0(n)$ denote the number of observed shared species from n (pairs of) observations.

The probability of observing a previously unseen species (which is listed) in a single sample draw from population 1 can be expressed as $u(n) = \sum\limits_i p_i \times I(X_i(n) = 0)$, where $I(\cdot)$ is indicator function [9]. The Turing estimate for the probability of discovering new species is based on the number of species appearing exactly once in the sample, i.e., $\widehat{u}(n) = \frac{g_1}{n}$ where $g_1 \equiv \sum\limits_i I(X_i(n) = 1)$ is the number of singletons [2]. However, Turing's estimate has a positive bias since $E(\widehat{u}(n)) = \sum\limits_i p_i(1-p_i)^{n-1}$ is larger than $E(u(n)) = \sum\limits_i p_i(1-p_i)^n$ [9].

The Turing-type estimator for the probability of discovering new shared species can be derived similarly. First, the probability of discovering new shared species after n observations is

$$v(n) = \sum_{i=1}^{s_0} p_i\, q_i \times I(X_i(n) = Y_i(n) = 0) + \sum_{i=1}^{s_0} (p_i \times I(X_i(n) = 0,\ Y_i(n) > 0) + q_i \times I(X_i(n) > 0,\ Y_i(n) = 0)) \tag{1}$$

where $(p_1, p_2, \ldots, p_s)$ and $(q_1, q_2, \ldots, q_s)$ are the species proportions of the two populations [8]. We propose two Turing-type estimators, denoted $v_1'(n)$ and $v_2'(n)$, based on Equation (1): the first is from [2] and the other is a direct extension from the one-population case. The first estimator is derived from $E(v(n))$, and $\frac{g_1}{n}$ is used to replace $u(n)$ as in Turing's estimate. Thus, $v_1'(n)$ can be expressed as

$$\begin{aligned} v_1'(n) = & \sum_{i=1}^{s} \tfrac{I(X_i(n)=1)}{n} + \sum_{i=1}^{s} \tfrac{I(Y_i(n)=1)}{n} + \sum_{i=1}^{s} \tfrac{I(X_i(n)=Y_i(n)=1)}{n} \\ & - \sum_{i=1}^{s} \tfrac{I(X_i(n)=0,Y_i(n)=1)}{n} - \sum_{i=1}^{s} \tfrac{I(X_i(n)=1,Y_i(n)=0)}{n}. \end{aligned} \tag{2}$$

Equation (2) is the probability that a shared new species occurs at the n^{th} sample point, given the sample statistics $X_i(n), Y_i(n)$ for $i = 1, 2, \cdots, s$. Since Turing's estimate has a positive bias, $v_1'(n)$ is also biased, as described in the Appendix of [8].

Another Turing-type estimator is to treat the two populations as two independent populations and then the two-population Turing's estimate is the sum of Turing's estimates from each population. Specifically, for the new shared species, we only consider the case where they are observed in one population but not yet observed in the other population. The estimator is expressed as

$$v_2'(n) = \sum_{i=1}^{s} \frac{I(X_i(n) = 1, Y_i(n) > 0)}{n} + \sum_{i=1}^{s} \frac{I(X_i(n) > 0, Y_i(n) = 1)}{n}. \tag{3}$$

The difference between $v_1'(n)$ and $v_2'(n)$ is $v_1'(n) - v_2'(n) = \sum_{i=1}^{s} \frac{I(X_i(n)=Y_i(n)=1)}{n}$, and thus $v_2'(n)$ has the potential to reduce the bias of $v_1'(n)$; in fact this will be shown to be the case in the next section.

We next develop jackknife-type estimators for the number of shared species similar to those used for the number of species [3]. For a single sample, their (first-order) jackknife estimate of the number of species in a single population is given by: $\widehat{s}_J = s_0^*(n) + \frac{n-1}{n}(f_1^*)$ where $s_0^*(n)$ is the number of observed species and f_1^* is the number of singletons. A similar idea can be applied to the case of two populations and we can use the number of species appearing once to develop the jackknife type estimate of number of shared species. Let f_{1+} (or f_{+1}) be the numbers of species appearing exactly once in the first (or second) population, which also appear at least once in the other population. Let f_{11} be the number of species appearing exactly once in both populations. Then, by analogy of using the singletons and the Equations (2) and (3), the jackknife-type estimators $\widehat{s}_J = s_0^*(n) + \frac{n-1}{n} \times$ (singleton) for the number of shared species can be expressed as $\widehat{s}_{J1} = s_0(n) + \frac{n-1}{n}(f_{1+} + f_{+1} + f_{11})$ and $\widehat{s}_{J2} = s_0(n) + \frac{n-1}{n}(f_{1+} + f_{+1})$. The derivation of these two estimators is outlined in Appendix A.

Using techniques similar to those used in the previous study [3], the jackknife-type estimators can also be expressed in the following form,

$$\widehat{s}_{J1} = s_0(n) + \frac{n-1}{n}\widehat{f}_1 = \sum_{i=1}^{n} a_i f_i \tag{4}$$

where

$$\begin{aligned} \widehat{f}_1 = \sum_{i=1}^{s} I(X_i(n) = 1) + \sum_{i=1}^{s} I(Y_i(n) = 1) + \sum_{i=1}^{s} I(X_i(n) = Y_i(n) = 1) \\ - \sum_{i=1}^{s} I(X_i(n) = 0, Y_i(n) = 1) - \sum_{i=1}^{s} I(X_i(n) = 1, Y_i(n) = 0) \end{aligned} \tag{5}$$

$a_1 = \frac{(n-1)(f_1+2f_{11})}{nf_1} + 1$, $a_2 = \cdots = a_n = 1$, and f_i is the number of species appearing exactly i times ($i \geq 1$) in either population.

One of the advantages of using the jackknife procedure is that the variance of the jackknife-type estimators can be derived easily. The variance of the first estimator is

$$Var(\widehat{s}_{J1}) = \sum_{i=1}^{n} a_i^2 f_i - \widehat{s}_{J1}. \tag{6}$$

The second estimator can also be expressed in a form similar to Equation (5):

$$\widehat{s}_{J2} = s_0(n) + \frac{n-1}{n}\widetilde{f}_1 = \sum_{i=1}^{n} b_i f_i \tag{7}$$

with variance

$$Var(\widehat{s}_{J2}) = \sum_{i=1}^{n} b_i^2 f_i - \widehat{s}_{J2} \tag{8}$$

where $b_1 = \frac{(n-1)(f_1+f_{11})}{nf_1} + 1$, $b_2 = \cdots = b_n = 1$. Since the difference between the two estimators from Equations (2) and (3) for the probability of discovering new shared species is $\sum_{i=1}^{s} \frac{I(X_i(n)=Y_i(n)=1)}{n}$, the difference between the two jackknife-type estimators from Equations (4) and (7) is $\frac{(n-1)f_{11}}{n}$.

Note that the jackknife-type estimators in Equations (4) and (7) are constructed similar to the form of jackknife estimator for one population, where the estimate of number of species is the sum of the number of observed species with $(n-1)/n$ multiplying the number of singletons in the sample. Interestingly, Chao's estimator for the number of shared species [6] also has the same form as Chao's estimator for the number of species in one population [4,5]. In particular, using a homogeneous population case as an example, Chao's estimator for the number of shared species can be expressed as $\widehat{s}_{Chao} = s_0(n) + \frac{s_{rare}(n)}{\widehat{C}}$, where $s_{rare}(n)$ is the number of observed rare shared species and $\widehat{C}$ is the estimate of sample coverage for the shared species. Using our notation, $s_{rare}(n) = \sum\limits_{i} I[0 < X_i(n),\ Y_i(n) \le 10]$ is the number of observed shared species appearing at most 10 times in both populations (i.e., rarely), and the sample coverage estimate is $\widehat{C} = \frac{\sum\limits_{i=1}^{s_{12}(n)} p_i^* \times q_i^* \times I[X_i(n)>0, Y_i(n)>0]}{\sum\limits_{i=1}^{s_{12}(n)} p_i^* \times q_i^*}$, with

$$p_i^* = \frac{p_i}{1 - \sum\limits_{i=1}^{s} \{p_i \times I[X_i(n) > 0, Y_i(n) > 10] + p_i \times I[X_i(n) > 10, Y_i(n) > 0]\}}$$

and

$$q_i^* = \frac{q_i}{1 - \sum\limits_{i=1}^{s} \{q_i \times I[X_i(n) > 0, Y_i(n) > 10] + q_i \times I[X_i(n) > 10, Y_i(n) > 0]\}}.$$

3. Simulation Studies

We first use computer simulation to evaluate the performance of $v_1'(n)$ and $v_2'(n)$, especially when used to form stopping rules that lead to estimates of the number of shared species, and compare three nonparametric estimators of the number of shared species in two populations: $\widehat{s}_{J1}$, $\widehat{s}_{J2}$, and Chao's estimate [6]. As pointed out in the previous study [8], the probability of observing new shared species can be used as a stopping indicator for sampling. We shall extend its role to develop the estimate for the number of shared species, and use the probability as a stopping indicator.

Similar to Yue and Clayton [8,10], we use geometric distributions to model the distribution of species within each population. That is, we assume that $p_i \propto \alpha^i$ and likewise for $q_i \propto \alpha^i$. In addition, we assume that the shared species are dominant in both populations [10,11]. We shall first evaluate the performance of estimators for the probability of discovering new shared species $v_1'(n)$ and $v_2'(n)$, using $v(n)$ as a benchmark. Note that the computer simulations conducted in this study are based on an Intel-based PC, using the statistical software R, version 2.12.0. All results are from 1000 simulation replications for each case.

Example 1. Suppose that the species proportions of the two populations follow geometric distributions and $p_i = q_i \propto \alpha^i$ with α = 0.9, 0.8, 0.7, and 0.6. Note that a larger α indicates a more even (or balanced) population structure, while a smaller α means that some species are dominant and the population structure is more unbalanced. Let the numbers of species in the two populations be 100, the number of shared species be 20 or 50, and the shared species are the most dominant species in each population. The results are each based on 1000 simulation runs.

Table 1 lists the probability and its estimates of discovering new shared species given that n observations are taken from each population and that the species proportions follow the geometric distributions stated above. As expected, the estimate $v'_1(n)$ has a larger bias, especially in the cases of smaller sample sizes. On the other hand, the estimate $v'_2(n)$ performs better in terms of bias for all cases and it is not influenced by the population structure (i.e., even or unbalanced). It seems that the deduction of $\sum_{i=1}^{s}\frac{I(X_i(n)=Y_i(n)=1)}{n}$ from $v'_1(n)$ is reasonable since $v'_1(n)$ has a positive bias [8], although it looks like $v'_2(n)$ could be under-biased from Equation (3). Nonetheless, based on these simulation results, it appears that the estimate $v'_2(n)$ is a better estimate for the probability of discovering new shared species.

Table 1. Probability of Discovering New Shared Species. Numbers of species in two populations are $s_1 = 100$ & $s_1 = 100$, the number of shared species is $s_0 = 20$ or $s_0 = 50$, and species proportions follow Geom(α).

n		α = 0.9			α = 0.8			α = 0.7			α = 0.6	
	$v(n)$	$v'_1(n)$	$v'_2(n)$	$v(n)$	$v'_1(n)$	$v'_2(n)$	$v(n)$	$v'_1(n)$	$v'_2(n)$	$v(n)$	$v'_1(n)$	$v'_2(n)$
						$s_0 = 20$						
100	0.04469	0.05077	0.04436	0.04196	0.05126	0.04144	0.02723	0.03558	0.02825	0.01931	0.02476	0.01966
200	0.00705	0.00730	0.00686	0.01804	0.02310	0.01900	0.01366	0.01841	0.01461	0.00979	0.01154	0.00935
500	0.00007	0.00006	0.00006	0.00408	0.00488	0.00425	0.00564	0.00686	0.00547	0.00408	0.00466	0.00373
1000	0	0	0	0.00068	0.00072	0.00068	0.00281	0.00331	0.00268	0.00191	0.00238	0.00191
1500	0	0	0	0.00011	0.00012	0.00012	0.00170	0.00198	0.00162	0.00133	0.00156	0.00126
2000	0	0	0	0.00003	0.00003	0.00003	0.00110	0.00140	0.00115	0.00102	0.00112	0.00091
3000	0	0	0	0	0	0	0.00061	0.00073	0.00062	0.00068	0.00079	0.00064
4000	0	0	0	0	0	0	0.00038	0.00044	0.00038	0.00049	0.00062	0.00049
5000	0	0	0	0	0	0	0.00026	0.00026	0.00023	0.00039	0.00050	0.00040
						$s_0 = 50$						
500	0.01816	0.02208	0.01784	0.00880	0.01111	0.00890	0.00550	0.00718	0.00574	0.00383	0.00487	0.00388
1000	0.00793	0.00941	0.00778	0.00461	0.00550	0.00441	0.00276	0.00339	0.00273	0.00197	0.00240	0.00192
1500	0.00433	0.00514	0.00437	0.00308	0.00379	0.00303	0.00185	0.00237	0.00191	0.00131	0.00155	0.00125
2000	0.00263	0.00303	0.00262	0.00225	0.00279	0.00222	0.00143	0.00167	0.00134	0.00095	0.00123	0.00098
3000	0.00110	0.00127	0.00114	0.00154	0.00186	0.00148	0.00093	0.00123	0.00097	0.00063	0.00086	0.00068
4000	0.00047	0.00055	0.00052	0.00110	0.00144	0.00115	0.00071	0.00086	0.00069	0.00048	0.00062	0.00049
5000	0.00024	0.00026	0.00024	0.00090	0.00113	0.00090	0.00056	0.00070	0.00056	0.00038	0.00049	0.00039

We shall continue the comparison of estimators for the number of shared species, despite the fact that the estimate $v'_1(n)$ is over-biased. Note that both the original and modified versions of Chao's estimates are considered in this study. However, we will only show the modified Chao's estimate (denoted as $\hat{s}_{C2}$ for the rest of this study) since it performs better than the original Chao's estimate. In the next example, we compare two jackknife-type and Chao's estimators for the number of shared species in two populations.

Example 2. We now consider the comparison of estimates for the number of shared species using the same settings as in Example 1 and show the averages and variances of estimates from 1000 simulation runs. In addition, we also include the case where the species proportions follow the Zipf's law, similar to that in [6]. We assume that $p_i = q_i \propto i^{\delta}$ with δ = 1, 1.5, and 2, and show only the averages of estimates. In general, more observations are required in the case of more unbalanced populations (i.e., smaller α and larger δ). To simplify the discussion, the cases where $p_i = q_i \propto \alpha^i$ with α = 0.9 and 0.7 will be used. The details of the simulation results can be found in Appendices B and C.

We first show the comparison of two jackknife-type and Chao's estimators for the number of shared species (Figures 1 and 2). In the even population case, Chao's estimate has the best performance for both s_0= 20 and 50. It converges much faster and does not have larger bias like the jackknife-type estimates. On the other hand, for the unbalanced population cases, the jackknife-type estimators (especially $\hat{s}_{J1}$) have a smaller bias, for both s_0= 20 or 50. But all estimators converge very slowly in the case of larger s_0 and unbalanced populations. It seems that, by analogy, the overbiased property of $v'_1(n)$ also carries over to the estimation of number of shared species in $\widehat{s}_{J1}$. In particular, since the behaviors of singletons can be very discrete in the cases of unbalanced populations, it is reasonable to be conservative and choose a slightly overbiased estimator.

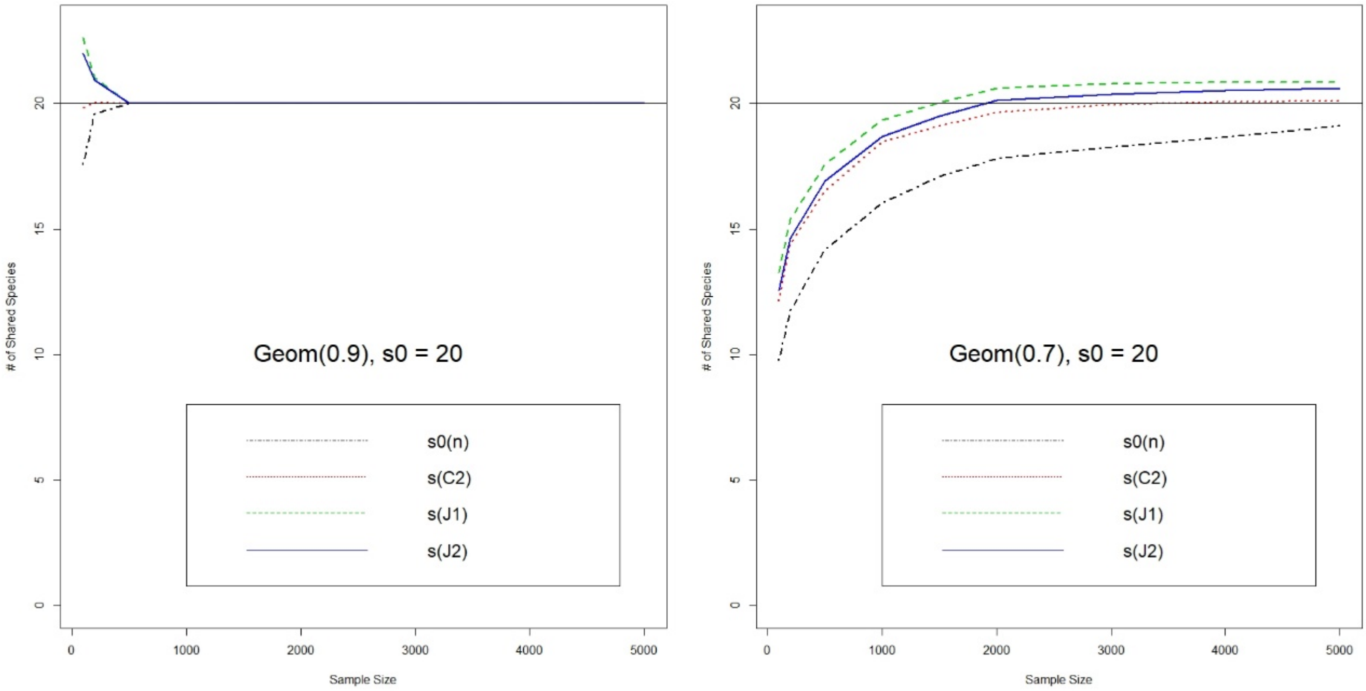

Figure 1. Estimates for the Number of Shared Species (Numbers of species in two populations are $s_1 = 100$ & $s_2 = 100$, and the number of shared species is $s_0 = 20$; $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

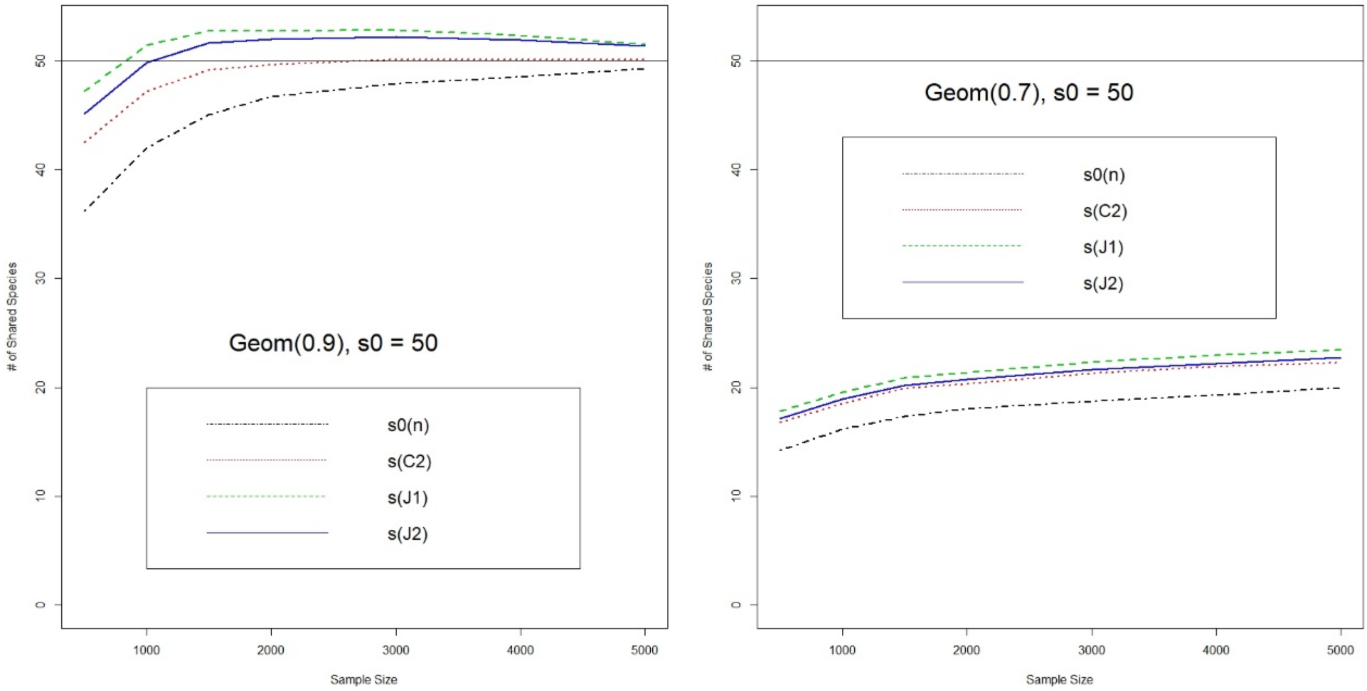

Figure 2. Estimates for the Number of Shared Species (Numbers of species in two populations are $s_1 = 100$ & $s_2 = 100$, and the number of shared species is $s_0 = 50$; $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

Note that, although we found that Chao's estimate performs well in the even population case, it can still produce undesirable results. For example, assume that the species proportions satisfy $p_i = (0.99)^i$ and $q_i = (0.9)^i$, and that the number of shared species is 80. Under this setting, there will be no observed rare shared species once the sample size is big enough. As shown in Table 2, we cannot compute Chao's estimate since all observed shared species appear more than 10 times. On the other hand, the jackknife-type estimators converge to the true number of shared species as the sample size increases.

Table 2. Estimates for the Number of Shared Species (Numbers of species in two populations are $s_1 = 100$ & $s_1 = 100$, the number of shared species is $s_0 = 80$, and species proportions are from Geom(0.99) and Geom(0.9); $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

n	$s_0(n)$	$\widehat{s}_{C2}$	$\widehat{s}_{J1}$	$\widehat{s}_{J2}$
500	42.68	51.40	53.74	53.42
1000	49.46	54.75	58.85	58.85
2000	55.87	55.97	64.88	64.88
3000	59.44	NA	68.21	68.21
5000	63.92	NA	72.36	72.36
8000	67.66	NA	75.57	75.57
10,000	69.57	NA	77.21	77.21
15,000	72.31	NA	78.96	78.96
20,000	74.18	NA	80.16	80.16

Note: Chao's estimates become N/A if the sample coverage = 0.

Next we compute the Monte Carlo variance of the two jackknife-type and Chao's estimators, and also the variance of jackknife-type estimators from Equations (6) and (8). Since all estimators converge to the true value fairly fast in the even population case (α = 0.8 & 0.9), we will focus on the case of α = 0.7. (Appendix B shows the details of simulation results for all cases $p_i = q_i \propto \alpha^i$ with α = 0.9, 0.8, 0.7, and 0.6). Figure 3 shows the sample variances of two jackknife-type and Chao's estimators from 1000 runs. On average, the jackknife-type estimators have smaller and smoother variances ($\widehat{s}_{J2}$ the smallest). The variance of Chao's estimate jumps up and down even when there are 2000 or more observations, which might indicate that Chao's estimate can still be unstable even when there are a lot of observations.

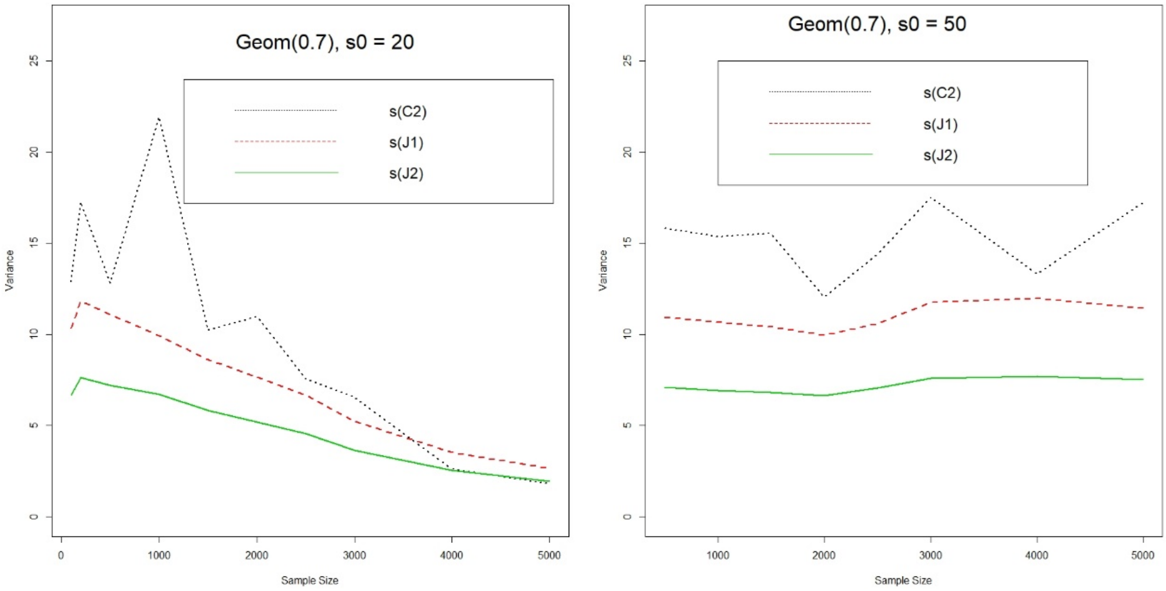

Figure 3. Variance of different Estimates for the Number of Shared Species ($J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, s_0(n): number of observed shared species).

We shall also check whether Equations (6) and (8) can provide reliable approximation to the variance of jackknife-type estimators, by using the sample variance from Monte Carlo simulation as the baseline. Figure 4 shows the variances from Equations (6) and (8) and those from Monte Carlo simulations which are marked with "Monte Carlo". Similar to the overbias in estimating the number of shared species, the variance of $\widehat{s}_{J1}$ from Equation (6) is always larger than that from Monte Carlo simulation. In contrast, the variance Equation (8) for $\widehat{s}_{J2}$ is a good approximation to that of Monte Carlo simulation. In any case, the variance formulae for the jackknife-type estimators provide fairly reliable approximations.

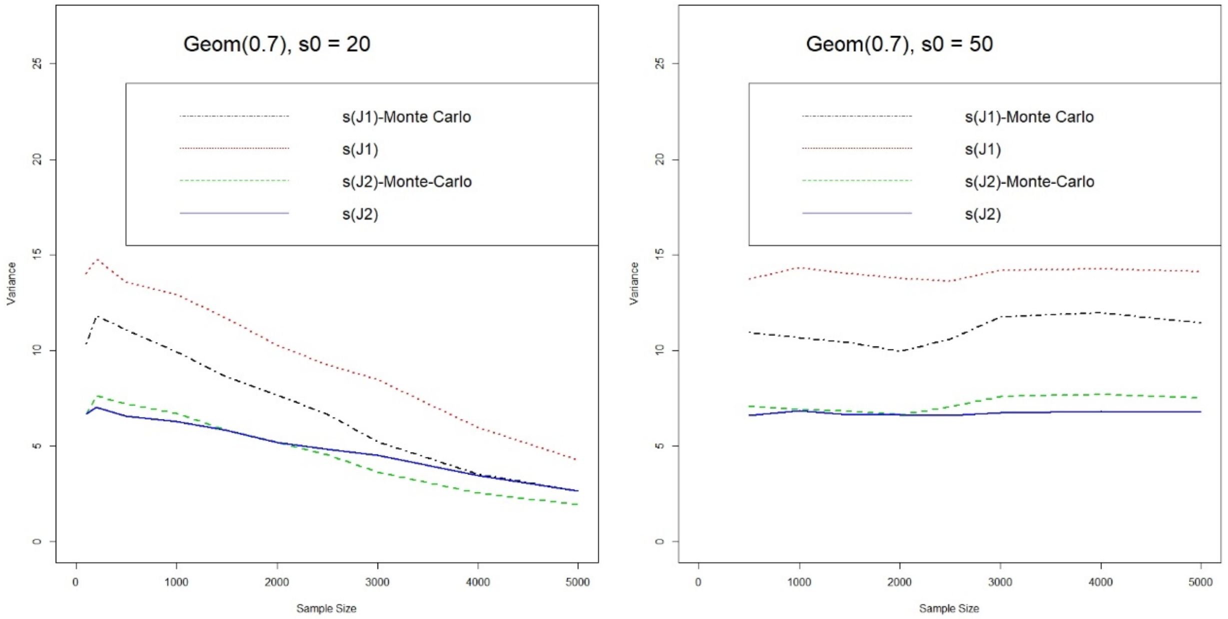

Figure 4. Variance Comparison (Sample vs. Monte Carlo) of Jackknife-type Estimates for the Number of Shared Species (*J*1 & *J*2: 1st & 2nd Jackknife estimates).

4. Empirical Studies

In addition to the simulations of the previous section, we also use empirical data to evaluate the three estimates of shared species. Four data sets are considered in this study: the first two are data on wild birds and on crabs [8], the third one is based on forest data, and the last one comes from Chinese literature. Also, we consider the case of sampling with replacement since there are finitely many observations in all data sets. In other words, we are using these data sets as representing the true populations, and our sampling emulates sampling from these populations.

Example 3. The Taiwan Bird data [11] contain two communities of wild birds consisting of 184 different species and 144,963 observations. There are 155 and 149 species in population 1 and 2, respectively, and 111 shared species (more than half are shared species). The shared species are dominant in each population, similar to the setting in the previous section. We therefore expect that the results of the jackknife-type estimates to be similar to those in the previous section.

Table 3 shows the estimates of the probability of discovering new shared species and the estimates of the number of shared species as a function of sample size. Moreover, we also calculate the coverage probability for the number of shared species; that is, the probability that the confidence interval $(\widehat{s} - 1.96 \times s.e., \widehat{s} + 1.96 \times s.e.)$ covers the true number of shared species. We expect this interval to behave approximately like a 95% confidence interval and so this coverage probability is intended to verify whether the estimate can be used in building confidence intervals. Note that $\widehat{s}$ is the estimate for the number of shared species, and its variance is calculated via 1000 simulation runs. Note that we can also use the variances via Equations (6) and (8) to compute the coverage for jackknife-type estimators (and the results of coverage probability are fairly close). However, the variance of Chao's estimator can only be computed via Monte Carlo simulation, and we shall compute the variances all based on simulation.

Table 3. Taiwan's Bird Data (Numbers of species in two populations are $s_1 = 155$ & $s_2 = 149$ and the number of shared species is $s_0 = 111$; $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

n	$v(n)$	$v_1'(n)$	$v_2'(n)$	$s_0(n)$	$\widehat{s}_{C2}$			$\widehat{s}_{J1}$			$\widehat{s}_{J2}$		
					Est.	s.e.	Prob	Est.	s.e.	Prob	Est.	s.e.	Prob
3000	0.07717	0.09651	0.07691	56.95	65.35	7.53	0.01	74.11	7.33	0.01	71.85	5.71	0.00
6000	0.04532	0.05594	0.04525	67.40	75.11	6.71	0.01	84.55	7.17	0.10	82.54	5.72	0.01
9000	0.03163	0.03872	0.03169	73.49	80.48	6.53	0.02	90.04	7.04	0.21	88.12	5.62	0.06
15,000	0.01924	0.02305	0.01919	80.54	86.72	5.81	0.03	96.39	6.86	0.42	94.60	5.50	0.23
24,000	0.01147	0.01354	0.01152	87.22	93.76	6.30	0.15	102.48	6.68	0.68	100.83	5.41	0.51
30,000	0.00879	0.01020	0.00878	90.08	96.14	5.84	0.24	105.30	6.68	0.76	103.64	5.40	0.66
36,000	0.00697	0.00803	0.00697	92.44	97.47	5.53	0.28	107.20	6.53	0.84	105.68	5.32	0.78
45,000	0.00517	0.00586	0.00516	95.48	100.12	4.85	0.35	109.71	6.36	0.92	108.30	5.23	0.88
51,000	0.00432	0.00489	0.00434	97.08	101.30	4.54	0.38	110.95	6.25	0.94	109.63	5.16	0.91

From the table we can see that, for the probability of discovering new shared species, $v_2'(n)$ again is a better estimate for small and large samples, and $v_1'(n)$ is always over-biased. The first jackknife-type estimate $\widehat{s}_{J1}$ of the number of shared species again is the largest among the three estimates, but, unlike the over-biasedness of $v_1'(n)$, it is still smaller than the true s when the sample drawn is large. Its variance decreases gradually as the sample size increases and becomes stable when the sample size is around 50,000, where the coverage probability is about 95%. The second jackknife-type estimate $\widehat{s}_{J2}$ has a similar behavior but it requires a larger sample to become stable.

Chao's estimate $\widehat{s}_{C2}$, on the other hand, does not reach the true number of shared species when the sample size is 51,000, and it might need considerably more samples to reach the true number. It seems that $\widehat{s}_{C2}$ is more conservative in estimating the number of shared species, and its coverage probability is too small even when there are 51,000 observations from each population (about 70% of the original sample size 144,963).

Example 4. The Panama Crab data [12] were collected in two coral communities at two locations in Panama. There are 55 and 50 species in populations 1 and 2, respectively, and 31 shared species, accounting for 74 different species and 5831 observations. Unlike the Taiwan Bird data, the shared species in the crab data are not so dominant and the number of shared species is less than half of the total species.

Among all the examples in these empirical analyses, the crab data have the smallest numbers of shared species and total observations. Because the smaller population in the crab data has about 1100 observations in total, we start with 110 observations from each population and consider only the case where the sample size is a multiple of 110 for computational simplicity (Table 4). Once again, $v_2'(n)$ is shown to be better than $v_1'(n)$ for estimating the probability of discovering new shared species, no matter what the sample size is. For the number of shared species, $\widehat{s}_{J1}$ has the largest averages and $\widehat{s}_{C2}$ is the smallest. Also, Chao's estimate performs the best in coverage probability.

Table 4. Panama's Crab Data (Numbers of species in two populations are $s_1 = 55$ & $s_2 = 50$ and the number of shared species is $s_0 = 31$; $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

n	$v(n)$	$v_1'(n)$	$v_2'(n)$	$s_0(n)$		$\widehat{s}_{C2}$			$\widehat{s}_{J1}$			$\widehat{s}_{J2}$	
					Est.	**s.e.**	**Prob**	**Est.**	**s.e.**	**Prob**	**Est.**	**s.e.**	**Prob**
110	0.03572	0.04486	0.03793	11.05	14.06	4.54	0.05	15.94	4.05	0.15	15.18	3.05	0.04
220	0.02167	0.02622	0.02185	14.03	18.65	9.85	0.97	19.77	4.47	0.33	18.81	3.30	0.12
330	0.01588	0.01866	0.01576	16.09	20.57	7.09	0.67	22.23	4.55	0.46	21.28	3.42	0.26
550	0.00995	0.01181	0.01003	18.90	22.58	6.66	0.84	25.38	4.65	0.66	24.40	3.52	0.51
1100	0.00495	0.00544	0.00484	22.70	26.23	4.09	0.78	28.68	4.21	0.82	28.02	3.39	0.77
1320	0.00390	0.00449	0.00402	23.72	27.73	4.69	0.95	29.65	4.16	0.84	29.02	3.38	0.83
1650	0.00315	0.00342	0.00307	24.71	28.63	4.91	0.97	30.36	4.04	0.88	29.78	3.30	0.86
1980	0.00236	0.00256	0.00236	25.67	29.03	4.96	0.98	30.73	3.67	0.85	30.34	3.14	0.85
2200	0.00207	0.00226	0.00209	26.17	29.36	4.47	0.97	31.14	3.66	0.89	30.75	3.11	0.88

The jackknife-type estimates never reached 90% of the coverage probability, although their estimates increase gradually and their variances are more stable. The reason why the jackknife-type estimates have smaller coverage probability is the variance, since the averages of $\widehat{s}_{C2}$ are smaller than those of $\widehat{s}_{J1}$ and $\widehat{s}_{J2}$ (and smaller than $s_0 = 31$). This matches the result that $\widehat{s}_{J2}$ has the smallest variance and smallest coverage probability. However, since $\widehat{s}_{J1}$ has a larger estimate of variance via Equation (6), $\widehat{s}_{J1}$ would have a better coverage probability if its variance were computed from Equation (6).

Example 5. Barro Colorado Island's Forest Data (We would like express our appreciation to Professor T.J. Shen, Department of Applied Mathematics, National Chung Hsing University, Taiwan, for providing this data set) are collected around the Gatun Lake area in Panama. The forest is separated into 4 regions (or populations): A, AB, D, and P. We choose regions A and AB in this study, containing 308 and 207 species, respectively. The reason for choosing this combination is that there are 207 shared species, i.e., AB can be treated as a sub-population of A, and the number of shared species in the two populations is equivalent to the number of species in AB. Also, the number of observations in region A is 242,083, much larger than that in region AB (5883).

Corresponding to region AB, the largest sample size considered is about two times its number of observations (12,000). As expected, $v_2'(n)$ is a good estimate of the probability for discovering new shared species and $v_1'(n)$ is always biased (Table 5). The jackknife-type estimates are fairly accurate estimates for the number of shared species, and they also have good coverage probabilities. Their variances decrease smoothly as the sample size increases. On the other hand, Chao's estimate grows slower, compared to of the jackknife-type estimates. Chao's estimate does not have a good coverage probability and it is likely that more observations are required.

Table 5. Barro Colorado Island's Forest Data (Numbers of species in two populations are $s_1 = 308$ & $s_2 = 207$ and the number of shared species is $s_0 = 207$; $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

n	$v(n)$	$v_1'(n)$	$v_2'(n)$	$s_0(n)$		$\widehat{s}_{C2}$			$\widehat{s}_{J1}$			$\widehat{s}_{J2}$	
					Est.	**s.e.**	**Prob**	**Est.**	**s.e.**	**Prob**	**Est.**	**s.e.**	**Prob**
600	0.06888	0.08510	0.06923	81.4	118.9	15.5	0.00	132.4	13.6	0.00	122.9	9.8	0.00
1200	0.03503	0.04244	0.03496	110.6	144.4	13.8	0.02	161.5	13.4	0.11	152.5	9.8	0.00
3000	0.01311	0.01547	0.01299	148.0	174.3	11.8	0.20	194.4	12.5	0.78	186.9	9.3	0.44
4500	0.00807	0.00935	0.00798	163.1	188.8	10.7	0.57	205.2	11.7	0.95	199.1	8.9	0.83
6000	0.00546	0.00621	0.00540	173.1	193.9	8.8	0.66	210.4	10.8	0.98	205.5	8.4	0.94
7500	0.00392	0.00424	0.00376	179.8	196.0	7.7	0.70	211.6	9.7	0.96	208.0	7.8	0.94
9000	0.00279	0.00316	0.00283	185.3	199.0	6.4	0.75	213.7	9.1	0.96	210.7	7.4	0.96
10,500	0.00208	0.00231	0.00211	188.9	199.9	5.4	0.73	213.2	8.2	0.98	211.0	6.8	0.98
12,000	0.00163	0.00175	0.00161	191.6	200.3	5.1	0.73	212.6	7.5	0.96	210.9	6.4	0.96

Example 6. The Chinese Novel Data contain two novels from Louis Cha Leung Yung, a famous Chinese writer. He has 10 famous historical novels, written between 1955 and 1972. The two novels chosen are "Fox of Snowy Mountain" (A) and "The Legendary Swordsman Enjoy Itinerant Life" (B) written in 1959 and 1967, respectively. We will treat different Chinese characters as different species. Then, there are 2591 and 3690 species in A and B, and 2457 shared species.

Novels A and B have about 110,000 and 420,000 characters (or observations). Thus, for computational efficiency, the sample size starts at 21,200 observations, about 20% of the observations in Novel A. We found that $v_2'(n)$ is a reliable estimate for the probability of discovering new shared species (Table 6). On the other hand, although $v_1'(n)$ is slightly over-biased, it is still a good estimate and is about 10% to 20% over-biased.

Table 6. Chinese Novel Data (Numbers of species in two populations are $s_1 = 2591$ & $s_2 = 3690$ and the number of shared species is $s_0 = 2457$; $J1$ & $J2$: 1st & 2nd Jackknife estimates, $C2$: Chao's estimate, $s_0(n)$: number of observed shared species).

n	$v(n)$	$v_1'(n)$	$v_2'(n)$	$s_0(n)$		$\hat{s}_{C2}$			$\hat{s}_{J1}$			$\hat{s}_{J2}$	
					Est.	s.e.	Prob	Est.	s.e.	Prob	Est.	s.e.	Prob
21,200	0.02421	0.02907	0.02402	1369.2	1681.6	41.2	0	1985.5	46.5	0	1878.4	34.0	0
42,400	0.01147	0.01354	0.01152	1721.3	1966.1	34.1	0	2295.5	43.4	0.04	2209.6	32.9	0
63,600	0.00697	0.00803	0.00697	1909.5	2100.9	27.2	0	2420.2	40.0	0.84	2352.8	31.1	0.09
74,200	0.00568	0.00649	0.00569	1976.8	2148.9	27.0	0	2458.4	38.3	0.96	2398.8	30.2	0.49
84,800	0.00471	0.00534	0.00472	2031.8	2187.5	25.0	0	2484.6	36.7	0.92	2432.3	29.4	0.86
95,400	0.00398	0.00446	0.00398	2078.0	2219.3	22.5	0	2503.8	35.3	0.78	2457.2	28.5	0.97
106,000	0.00341	0.00379	0.00340	2116.0	2245.6	22.3	0	2517.7	33.9	0.59	2476.3	27.7	0.90
127,200	0.00254	0.00281	0.00255	2178.8	2285.8	19.8	0	2536.0	31.5	0.26	2502.9	26.2	0.58
148,400	0.00197	0.00214	0.00198	2226.2	2312.3	17.2	0	2543.0	29.2	0.12	2516.6	24.7	0.28
169,600	0.00155	0.00168	0.00155	2262.7	2334.6	16.8	0	2546.9	27.4	0.05	2524.9	23.4	0.15

Neither Chao's estimate nor the jackknife-type estimates have desirable results in coverage probability. Unlike the previous three examples, the coverage probability does not stabilize as the sample size increases. The coverage probability of Chao's estimate is always 0, and those of the jackknife-type estimates decrease to 0 after reaching the maximum. It seems that the jackknife-type estimates can still provide useful information about the number of shared species, but the sample size is a very important factor. This result is similar to the optimal stopping for estimating the similarity index between two populations in Yue and Clayton [8]. Since it is not possible to sample all the individuals in the populations, knowing the appropriate time to stop sampling would be more feasible and cost efficient. Together with the probability of discovering new shared species $v_1'(n)$ and $v_2'(n)$, the jackknife-type estimators provide fairly accurate estimates to the number of shared species. For example, it seems that $v_1'(n) \leq 0.005$ or $v_2'(n) \leq 0.004$ is a possible candidate for stopping, where the coverage probability of jackknife-type estimators is around 0.95.

5. Conclusions

The rare species are often more important than dominant species in the estimation of the probability of discovering new species and the number of species in a population [13–15]. For example, two popular methods, Turing's and Chao's estimates, use the information on rare species for estimation of new species. The estimation of shared species in two populations can be directly extended from the methods used in one population. In this study, we establish jackknife-type estimates of shared species and compare it with that developed by Chao et al. [6].

First, we proposed a modified estimate for the probability for discovering new shared species in two populations, in order to reduce the bias of the estimate suggested by Yue and Clayton [8]. Then, based on these two estimates for discovering new shared species, we extended the jackknife-type estimate of Burnham and Overton [3] to obtain two estimates for the number of shared species in two populations. We compare these two jackknife-type estimates with that of Chao et al. [6]. Simulation studies and real examples confirm that the modified estimate $v_2'(n)$ has a smaller bias in estimating the probability of discovering new shared species, no matter what the sample size is.

For the number of shared species, the performance of estimates is influenced by the population structure and the sample size. In general, Chao's estimate has a smaller bias and converges to the true value much faster in the case of more even populations, and the jackknife estimates are better in the case of unbalanced populations (i.e., smaller α and larger δ). In the case of more even populations, all estimates are accurate even when there are not many observations. On the other hand, in the case of unbalanced populations, more observations are required and the jackknife-type estimates have a smaller bias. In addition, the variance of jackknife-type estimates can be approximated by the derived equations, which can be convenient in empirical analyses.

The coverage probability calculated in the real examples shows another difference between the jackknife and Chao's estimates. Applying a normal approximation for a 95% confidence interval, we evaluated the probability of covering the true number of shared species. Except for the Panama Crab data, Chao's estimate does not have coverage probability near 0.95. In contrast, both jackknife-type estimates can provide coverage probability close to 0.95 in all examples, provided that there are enough observations. Based on our experience, it seems that $v_1'(n) \leq 0.005$ (or $v_2'(n) \leq 0.004$) is a possible useful indicator for stopping sampling. When the sampling cost $c = 0.005$, the jackknife-type estimate $\widehat{s}_{J1}$ derived from $v_1'(n)$ in Yue and Clayton [8] has coverage probability close to 0.95 (except for the Panama Crab data). A similar result holds for another jackknife-type estimator $\widehat{s}_{J2}$. This is similar to the results in Yue and Clayton [8], although their interest is in the similarity index.

Note that we also conducted supplementary simulations to explore group sampling, group sampling with variable (i.e., random) numbers of observations, and sampling with one group observed sequentially and one group observed through a fixed sample. By and large the conclusions remain the same. It seems that the paired sampling represents the slowest incremental rate of accruing information and provides a useful baseline for examining the estimators.

As an alternative to our approach, using the sample coverage is another feasible approach for estimating species numbers, and there has been considerable success in using that for single populations. Among others, Chao and her colleagues have made important contributions to that topic [4,6]. However, addressing the sample coverage for estimating shared species requires a study separate from the work presented here.

Author Contributions: J.C.Y. proposed the general methodology, designed the simulation study and empirical analysis, and drafted the manuscript. M.K.C. designed the simulation study, interpreted the results, and helped to draft the manuscript. C.-R.H. conducted simulation studies and summarized the research results. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by a grant from the National Science Counsel in Taiwan, NSC 96-2118-M-004-003.

Institutional Review Board Statement: This study does not involve human participants, human tissue, or animal samples. The data used are open accessed and this study does not need ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: There are four data sets used in this study, all are open and accessible to the public.

Acknowledgments: The authors are grateful for the funding from the National Science Counsel in Taiwan, NSC 96-2118-M-004-003.

Conflicts of Interest: The authors declare that they have no competing interest.

Appendix A. Derivation of Jackknife-Type Estimators

The first estimator $\widehat{s}_{J1} = s_0(n) + \frac{n-1}{n}(f_{1+} + f_{+1} + f_{11})$ can be treated as a combination of jackknife-type and moment-type approaches. The jackknife-type estimate of the number of species for a population is the number of observed species plus $(n-1) \times u(n)$, where $u(n)$ is the probability of observing new species. Then, plugging into the Turing-type estimator for the probability of discovering new shared species $v_1'(n) = \frac{f_{1+}+f_{+1}+f_{11}}{n}$, we obtain the jackknife-type estimator $\widehat{s}_{J1}$ of the number of shared species in two populations.

The second estimator is based on jackknife technique noting that there are quite a few approaches to obtain jackknife estimators. Considering all possible combinations (i.e., permutations) is a natural choice, similar to Burnham and Overton [3]. There are two ways for counting possible combinations: one is pair-wise and the other is completely random. For the pairwise case, the observations are drawn in pairs, so ($X(i)$, $Y(i)$) are chosen together, where $X(i)$ and $Y(i)$ are the ith sample in the first and second populations, i.e., there are n possible jackknife subsamples if one pair of observations are omitted each time. For the completely random case, the observations are drawn randomly, so ($X(i)$, $Y(j)$) are chosen and it is possible $i \neq j$, which means that there are $n \times n = n^2$ possible jackknife subsamples if one observation is omitted from each population. Since the derivation of jackknife-type estimators are obtained via listing all possible combinations, we will only show the final results.

In the pair-wise case, depending on whether $X(i)$ and $Y(i)$ are both shared species and singletons, the jackknife-type estimator lies between two values

$$\text{Upper Bound} = s_0(n) + \frac{n-1}{n}(f_{1+} + f_{+1})$$

$$\text{Lower Bound} = s_0(n) + \frac{n-1}{n}(f_{1+} + f_{+1} - f_{11})$$

where f_{1+} is the number of species appearing exactly once in first population and at least once in the second population. The definition of f_{+1} is similar. The upper bound in the pair-wise case can be treated as a direct extension of the jackknife estimator of Burnham and Overton in the one population case, and thus we define the upper bound as the second jackknife-type estimator $\widehat{s}_{J2}$. The derivation of jackknife-type estimators in the completely random case is similar and the jackknife-type estimator equals $s_0(n) + \frac{n^2-1}{n^2}(f_{1+} + f_{+1} - \frac{2f_{11}}{n})$. Asymptotically, the jackknife-type estimator in the completely random case is very similar to those in the pair-wise case (closer to the upper bound).

In addition to the previous two jackknife-type estimators, it is also possible to derive other types of two-sample jackknife estimators. For example, Chuang et al. [7] used the jackknifing technique by Schechtman and Wang [16] and proposed a jackknife estimator $s_0(n) + \frac{n-1}{n}(f_{1+} + f_{+1}) - \frac{(n-1)^2}{n^2} f_{11}$.

We can see that these jackknife-type estimators have similar form, and only differ in how we weight the singletons. The differences would be more obvious in the case of small samples and are small if there are many observations. Still, there is another reason for choosing $\widehat{s}_{J1}$ and $\widehat{s}_{J2}$. The proposed estimators are based on the probability of discovering new shared species $v_1'(n)$ and $v_2'(n)$, and these probabilities can be used as stopping indicators. A detailed discussion of this can be seen in our empirical study (Section 4).

Appendix B. Estimates for the Number of Shared Species

Table A1. Numbers of species in two populations are $s_1 = 100$ & $s_2 = 100$, and the number of shared species $s_0 = 20$ or $s_0 = 50$ for the case of geometric distribution (*J*1 & *J*2: 1st & 2nd Jackknife estimates, *C*2: Chao's estimate, $s_0(n)$: number of observed shared species).

n	α = 0.9				α = 0.8				α = 0.7				α = 0.6			
	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$
$s_0 = 20$																
100	17.59	19.82	22.62	21.99	13.27	16.42	18.35	17.37	9.76	12.14	13.28	12.56	7.52	9.24	9.97	9.47
200	19.58	20.05	21.03	20.94	16.16	18.90	20.75	19.94	11.73	14.40	15.39	14.63	8.86	10.58	11.16	10.72
500	20.00	20.00	20.02	20.02	18.91	20.13	21.35	21.03	14.17	16.51	17.59	16.90	10.60	12.29	12.92	12.46
1000	20.00	20.00	20.00	20.00	19.79	20.13	20.51	20.47	16.03	18.49	19.34	18.70	12.00	13.71	14.38	13.91
1500	20.00	NA	20.00	20.00	19.97	20.06	20.15	20.15	17.08	19.11	20.04	19.51	12.76	14.51	15.10	14.64
2000	20.00	NA	20.00	20.00	19.99	20.02	20.04	20.04	17.81	19.64	20.62	20.12	13.29	14.93	15.53	15.11
3000	20.00	NA	20.00	20.00	20.00	20.00	20.01	20.01	18.26	19.96	20.81	20.39	13.82	15.81	16.36	15.84
4000	20.00	NA	20.00	20.00	20.00	20.00	20.00	20.00	18.66	20.07	20.86	20.52	14.10	15.79	16.48	16.01
5000	20.00	NA	20.00	20.00	20.00	20.00	20.00	20.00	19.11	20.09	20.86	20.62	14.70	16.44	17.17	16.68
$s_0 = 50$																
500	36.22	42.49	47.24	45.12	20.61	24.12	26.15	25.05	14.24	16.75	17.82	17.10	10.64	12.44	13.07	12.58
1000	42.06	47.21	51.46	49.83	23.66	27.13	29.16	28.07	16.17	18.55	19.56	18.90	12.00	13.70	14.40	13.91
1500	45.09	49.20	52.79	51.64	25.49	29.22	31.17	30.03	17.32	19.90	20.87	20.18	12.79	14.65	15.12	14.67
2000	46.75	49.63	52.81	51.99	26.87	30.44	32.44	31.32	18.03	20.29	21.36	20.71	13.40	15.18	15.85	15.36
3000	47.93	50.11	52.85	52.25	27.78	31.38	33.36	32.26	18.75	21.27	22.33	21.61	13.79	15.65	16.31	15.80
4000	48.54	50.14	52.34	51.95	28.55	31.98	34.12	32.99	19.26	21.92	22.94	22.18	14.21	16.03	16.77	16.25
5000	49.33	50.15	51.54	51.39	30.01	33.66	35.76	34.59	20.00	22.27	23.45	22.74	14.72	16.51	17.18	16.67

Note: Chao's estimates become N/A if the sample coverage = 0.

Table A2. Numbers of species in two populations are $s_1 = 100$ & $s_2 = 100$, and the number of shared species $s_{12} = 20$ or $s_0 = 50$ for the case of Zipf's law (*J*1 & *J*2: 1st & 2nd Jackknife estimates, *C*2: Chao's estimate, $s_0(n)$: number of observed shared species).

n	δ = 1				δ = 1.5				δ = 2			
	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$	$s_0(n)$	$\hat{s}_{C2}$	$\hat{s}_{J1}$	$\hat{s}_{J2}$
$s_0 = 20$												
500	15.00	29.25	17.97	17.97	11.00	19.03	15.95	16.94	8.00	14.00	8.99	8.99
1000	19.00	35.02	20.00	20.00	17.00	29.02	24.96	25.96	8.00	13.00	8.00	8.00
1500	20.00	27.00	20.00	20.00	20.00	33.01	22.00	22.00	14.00	23.00	16.99	16.99
2000	20.00	21.00	20.00	20.00	20.00	27.00	20.00	20.00	19.00	30.00	24.00	24.99
3000	20.00	20.00	20.00	20.00	20.00	24.00	20.00	20.00	19.00	27.00	21.00	21.00
4000	20.00	20.00	20.00	20.00	20.00	21.00	20.00	20.00	20.00	29.00	21.00	21.00
5000	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	28.00	20.00	20.00
$s_0 = 50$												
500	19.00	36.63	35.83	40.78	11.00	20.07	17.93	19.91	6.00	10.00	7.00	7.00
1000	34.00	64.50	60.87	67.83	22.00	40.04	37.92	42.90	10.00	17.00	15.97	17.96
1500	46.00	78.04	54.98	56.98	33.00	57.02	49.97	54.96	22.00	38.00	40.96	47.95
2000	50.00	82.00	51.00	51.00	42.00	73.00	53.99	56.99	25.00	43.00	31.99	33.99
3000	50.00	70.00	50.00	50.00	48.00	80.00	55.99	55.99	30.00	51.00	38.99	41.99
4000	50.00	60.00	50.00	50.00	49.00	79.00	52.00	52.00	31.00	50.00	39.00	41.00
5000	50.00	50.00	50.00	50.00	49.00	73.00	49.00	49.00	39.00	65.00	53.00	56.00

Appendix C. Variance of Estimates for the Number of Shared Species

Table A3. Numbers of species in two populations are $s_1 = 100$ & $s_2 = 100$, and the number of shared species $s_0 = 20$ or $s_0 = 50$ for the case of geometric distribution (*J*1 & *J*2: 1st & 2nd Jackknife estimates, *C*2: Chao's estimate, $s_0(n)$: number of observed shared species).

n	α = 0.9					α = 0.8					α = 0.7					α = 0.6				
	$\widehat{s}_{C2}$ **Sample**	$\widehat{s}_{J1}$		$\widehat{s}_{J2}$		$\widehat{s}_{C2}$ **Sample**	$\widehat{s}_{J1}$		$\widehat{s}_{J2}$		$\widehat{s}_{C2}$ **Sample**	$\widehat{s}_{J1}$		$\widehat{s}_{J2}$		$\widehat{s}_{C2}$ **Sample**	$\widehat{s}_{J1}$		$\widehat{s}_{J2}$	
		Sample	**Equation (6)**	**Sample**	**Equation (8)**		**Sample**	**Equation (6)**	**Sample**	**Equation (8)**		**Sample**	**Equation (6)**	**Sample**	**Equation (8)**		**Sample**	**Equation (6)**	**Sample**	**Equation (8)**
$s_0 = 20$																				
100	5.77	10.75	16.35	7.76	9.79	16.5	15.70	20.08	10.42	9.78	12.89	10.35	14.01	6.69	6.67	8.57	8.57	9.45	4.86	4.56
200	0.73	2.20	3.76	1.76	2.89	11.55	12.18	16.89	8.31	8.62	17.27	11.82	14.79	7.64	7.02	9.44	9.44	9.32	4.84	4.51
500	0.01	0.03	0.08	0.03	0.08	3.33	5.16	7.98	3.72	4.76	12.83	11.11	13.57	7.21	6.58	8.71	8.71	10.21	4.64	4.91
1000	0	0	0	0	0	0.57	1.05	1.73	0.86	1.33	21.94	9.93	12.93	6.70	6.28	10.48	10.48	9.91	4.78	4.78
1500	NA	0	0	0	0	0.17	0.24	0.31	0.21	0.30	10.26	8.62	11.69	5.83	5.83	10.96	10.96	9.59	5.11	4.61
2000	NA	0	0	0	0	0.03	0.06	0.12	0.06	0.12	10.99	7.67	10.28	5.18	5.20	8.89	8.89	9.36	4.94	4.58
3000	NA	0	0	0	0	0.01	0	0.03	0.01	0.03	7.56	6.69	9.27	4.56	4.85	17.26	17.26	9.63	5.15	4.69
4000	NA	0	0	0	0	0	0	0.01	0	0.01	6.55	5.22	8.48	3.63	4.54	8.22	8.22	10.23	4.64	4.86
5000	NA	0	0	0	0	0	0	0	0	0	2.61	3.55	5.98	2.55	3.45	7.77	7.77	9.18	4.70	4.49
$s_0 = 50$																				
500	29.93	32.19	43.50	21.70	20.92	19.63	16.40	22.71	10.55	10.83	15.83	10.96	13.73	7.11	6.60	10.33	10.33	9.84	5.17	4.65
1000	21.31	27.96	35.24	18.94	17.95	18.93	17.96	21.83	11.81	10.61	15.37	10.67	14.36	6.93	6.86	9.94	9.94	9.91	5.23	4.75
1500	15.98	19.16	27.31	13.34	14.54	23.59	19.64	22.96	12.85	10.85	15.56	10.42	14.04	6.83	6.64	17.62	17.62	10.04	4.92	4.81
2000	9.15	13.82	21.25	10.15	11.99	22.28	17.38	22.73	11.37	10.79	12.05	9.98	13.79	6.63	6.64	10.84	10.84	9.73	5.01	4.74
3000	3.48	7.01	11.24	5.15	7.30	20.05	17.00	22.63	11.08	10.78	17.54	11.75	14.22	7.59	6.75	9.14	9.14	11.01	5.06	5.17
4000	1.41	3.50	6.37	2.75	4.46	17.45	18.09	23.46	11.85	10.97	13.33	11.97	14.28	7.71	6.81	10.41	10.41	10.11	5.23	4.84
5000	0.67	2.10	3.34	1.72	2.53	18.37	17.60	22.54	11.49	10.75	17.25	11.45	14.15	7.53	6.79	9.15	9.15	9.80	5.16	4.70

Note: Chao's estimates become N/A if the sample coverage = 0.

References

1. Bunge, J.; Fitzpatrick, M. Estimating the Number of Species: A Review. *J. Am. Stat. Assoc.* **1993**, *88*, 364–373.
2. Good, I.J. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* **1953**, *40*, 237–264. [CrossRef]
3. Burnham, P.K.; Overton, S.W. Estimation of the Size of a Closed Population When Capture Probabilities Vary among Animals. *Biometrika* **1978**, *65*, 625–633. [CrossRef]
4. Chao, A.; Lee, S.-M. Estimating the Number of Classes via Sample Coverage. *J. Am. Stat. Assoc.* **1992**, *87*, 210–217. [CrossRef]
5. Chao, A.; Ma, M.; Yang, M.C.K. Stopping Rule and Estimation for Recapture Debugging with Unequal Detection Rates. *Biometrika* **1993**, *80*, 193–201. [CrossRef]
6. Chao, A.; Hwang, W.; Chen, Y.; Kuo, C. Estimating the Number of Shared Species in Two Communities. *Stat. Sin.* **2000**, *10*, 227–246.
7. Chuang, C.; Shen, T.; Hwang, W. Estimating the Number of Shared Species by a Jackknife Procedure. *Environ. Ecol. Stat.* **2015**, *22*, 759–778. [CrossRef]
8. Yue, J.C.; Clayton, M.K. Sequential Sampling in the Search for New Shared Species. *J. Stat. Plan. Inference* **2012**, *142*, 1031–1039. [CrossRef]
9. Rasmussen, S.L.; Starr, N. Optimal and Adaptive Stopping in the Search for New Species. *J. Am. Stat. Assoc.* **1979**, *74*, 661–667. [CrossRef]
10. Yue, J.C.; Clayton, M.K. An Overlap Measure based on Species Proportions. *Commun. Stat.-Theory Methods* **2005**, *34*, 2123–2131. [CrossRef]
11. Yue, J.C.; Clayton, M.K.; Lin, F. A Nonparametric Estimator of Species Overlap. *Biometrics* **2001**, *57*, 743–749. [CrossRef] [PubMed]
12. Smith, W.; Solow, A.R.; Preston, P.E. An Estimator of Species Overlap Using a Modified Beta-binomial Model. *Biometrics* **1996**, *52*, 1472–1477. [CrossRef]
13. Gaston, K. The Importance of Being Rare. *Nature* **2012**, *487*, 46–47. [CrossRef] [PubMed]
14. Mi, X.; Swenson, N.G.; Valencia, R.; Kress, W.J.; Erickson, D.L.; Pérez, A.J.; Ren, H.; Su, S.; Gunatilleke, N.; Gunatilleke, S.; et al. The Contribution of Rare Species to Community Phylogenetic Diversity across a Global Network of Forest Plots. *Am. Natuarlist* **2012**, *180*, 17–30. [CrossRef] [PubMed]
15. Schechtman, E.; Wang, S. Jackknifing Two-sample Statistics. *J. Stat. Plan. Inference* **2004**, *119*, 329–340. [CrossRef]
16. Shen, T.-J.; Chen, Y. Predicting the Number of Newly Discovered Rare Species: A Bayesian Weight Approach. *Conserv. Biol.* **2019**, *33*, 444–455. [CrossRef] [PubMed]