Environmental and Ecological Statistics **10,** 429–443, 2003

Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample

ANNE CHAO and TSUNG-JEN SHEN

Institute of Statistics, National Tsing Hua University, Hsin-Chu, TAIWAN 30043

Received December 2001; Revised July 2002

A biological community usually has a large number of species with relatively small abundances. When a random sample of individuals is selected and each individual is classified according to species identity, some rare species may not be discovered. This paper is concerned with the estimation of Shannon's index of diversity when the number of species and the species abundances are unknown. The traditional estimator that ignores the missing species underestimates when there is a non-negligible number of unseen species. We provide a different approach based on unequal probability sampling theory because species have different probabilities of being discovered in the sample. No parametric forms are assumed for the species abundances. The proposed estimation procedure combines the Horvitz–Thompson (1952) adjustment for missing species and the concept of sample coverage, which is used to properly estimate the relative abundances of species discovered in the sample. Simulation results show that the proposed estimator works well under various abundance models even when a relatively large fraction of the species is missing. Three real data sets, two from biology and the other one from numismatics, are given for illustration.

Keywords: biodiversity, entropy, Horvitz–Thompson estimator, jackknife, sample coverage, species, unequal probability sampling

1. Introduction

Assume that there are S species in a community and they are labeled from 1 to S. Denote the probabilities of species discovery (or relative abundance) by $(\pi_1, \pi_2, \ldots, \pi_S)$ where $\sum_{i=1}^{S} \pi_i = 1$. A widely used measure of biological diversity is Shannon's index of diversity defined by

$$H = -\sum_{i=1}^{S} \pi_i \log(\pi_i). \tag{1}$$

This index is also referred to as Shannon's information measure or entropy in the literature. Suppose a random sample of n individuals is taken with replacement from the community and each individual is classified correctly according to species identity. Let $X_i, i = 1, 2, \ldots, S$, be the number of individuals of the ith species observed in the sample, then $(X_1, X_2, \ldots, X_S)$ is a multinomial distribution with parameter $(n; \pi_1, \pi_2, \ldots, \pi_S)$, where $\sum_{i=1}^{S} X_i = n$. The missing species are those with zero frequency in the sample. A principal approach to the related inference problems is to adopt some parametric forms for $(\pi_1, \pi_2, \ldots, \pi_S)$; e.g., see Engen (1978), Magurran (1988), and Bunge and Fitzpatrick (1993) for a review of various models. In this paper, we consider a non-parametric approach in the sense that no parametric forms are assumed in our estimation procedure.

When the number of species is known and relatively small, a widely used estimator of H is the maximum likelihood estimator (MLE) given by

$$\hat{H}_{MLE} = -\sum_{i=1}^{S} \hat{\pi}_i \log(\hat{\pi}_i) = -\sum_{i=1}^{S} \frac{X_i}{n} \log\left(\frac{X_i}{n}\right), \tag{2}$$

where $\hat{\pi}_i = X_i/n$ (sample fraction or sample proportion) is the MLE of π_i. It is well known that the MLE is negatively biased (Basharin, 1959) and

$$E(\hat{H}_{MLE}) = H - \frac{S-1}{2n} + O(n^{-2}). \tag{3}$$

The above formula was derived under the assumption that S is known (Pielou, 1975) and a minimum requirement is $n > S$ (Hutcheson and Shenton, 1974). The bias could be removed for given n and S by adding $(S-1)/(2n)$ to the MLE. An alternative method of reducing bias is the jackknife methodology (Zahl, 1977), which will be discussed later in the Simulation Section.

In ecological applications, the true number of species is often unknown and some rare species may not be discovered in a sample of individuals because of the existence of many rare species. As indicated by Magurran (1988), a more substantial source of error for the MLE comes from a failure to include all species from the community in the sample. This error increases as the proportion of species discovered in the sample declines. See also Peet (1974) for more related discussion. In this case, a bias-corrected estimator can be obtained by substituting an estimated S in the bias formula, but this involves the estimation of the number of unseen species. Moreover, using an estimated S results in increased variance, as will be discussed in the Simulation Section. See Bunge and Fitzpatrick (1993) and Bunge *et al.* (1995) for history and developments of species estimation.

Norris and Pollock (1998) recently developed a non-parametric MLE approach under a mixed Poisson process for species sampling. They presented estimators for species richness and related parameters including the entropy. Their method will be applied to our examples in Section 4.

In this paper, we focus on situations in which the true total number of species is unknown and take unseen species into account in our estimation procedure. We provide a different approach to the estimation of diversity based on unequal probability sampling theory because species have different probabilities of being discovered in the sample. The proposed estimation procedure combines the Horvitz–Thompson estimator and the concept of sample coverage. The Horvitz–Thompson estimator is used for adjustment of missing species in an unequal probability sampling scheme. In order to properly estimate the relative abundance of the discovered species, the concept of sample coverage is used for adjustment for the sample fraction of unseen species. The Horvitz–Thompson estimator and the concept of sample coverage will be reviewed in Section 2.1. Our proposed estimator and its variance estimator are presented in Section 2.2. Simulation results are used to examine the performance of the proposed estimator in Section 3. In Section 4, three real data sets are analyzed for illustration. Some final concluding remarks and relevant discussion are made in Section 5.

2. Models and estimators

2.1 *Horvitz–Thompson estimator*

We first briefly review the Horvitz–Thompson estimator (Horvitz and Thompson, 1952) in an unequal probability sampling scheme; see also Thompson (1992). Consider a finite population with S units which are indexed from 1 to S. Here S is unknown. Let Y_i be a measurement associated with the ith unit and the purpose is to estimate the population total $\tau = \sum_{i=1}^{S} Y_i$. Consider any unequal probability sampling design and let λ_i be the probability that the ith unit is included in the sample. Assume this sample results in k distinct units ($k \leq S$), and their corresponding measurements are $(Y_{j_1}, Y_{j_2}, \ldots, Y_{j_k})$. An unbiased estimator for the population total, introduced by Horvitz–Thompson (1952), is

$$\hat{\tau}_{HT} = \sum_{m=1}^{k} \frac{Y_{j_m}}{\lambda_{j_m}} = \sum_{i=1}^{S} \frac{Y_i}{\lambda_i} I(A_i), \tag{4}$$

where A_i denotes the event that the ith unit is included in the sample and $I(A_i)$ is the usual indicator function (i.e., $I(A_i) = 1$ when the event A_i is true and $I(A_i) = 0$ otherwise). Note in Equation (4), the summation is over the distinct units in the sample and the missing units are not included. Therefore, the value of S is not involved in computing $\hat{\tau}_{HT}$, although S appears as the upper limit in the second summation of Equation (4). Any unit in the population may be selected several times, but each distinct unit of the sample is utilized only once. Each distinct unit in the summation is given a weight proportional inversely to the probability of that unit. In other words, the larger the probability of being included in the sample, the smaller the weight in the Horvitz–Thompson estimator. The weights are used for an adjustment of missing units. It is readily seen that the Horvitz–Thompson estimator is an unbiased estimator of the population total.

We now apply this estimator to the estimation of Shannon's entropy. Regard any biological community as a sampling population and regard each species in the community as a unit in an unequal probability sampling. Let the measurement associated with the ith species be $Y_i = -\pi_i \log(\pi_i)$, then the population total of the measurement is Shannon's index. When n individuals have been selected with replacement from the community, the probability of the ith species not being discovered in any individual is $1 - \pi_i$, thus it is not discovered in these n individuals is $(1 - \pi_i)^n$. Consequently, the probability of the ith species being included in the sample, λ_i, becomes $\lambda_i = 1 - (1 - \pi_i)^n$. Then a Horvitz–Thompson estimator based on (4) for given species abundance $(\pi_1, \pi_2, \ldots, \pi_S)$ is given by

$$\hat{H}_{HT} = -\sum_{i=1}^{S} \frac{\pi_i \log(\pi_i)}{1 - (1 - \pi_i)^n} I(A_i). \tag{5}$$

2.2 Proposed estimator

To obtain our proposed estimator, we need to substitute an adequate estimator for the relative species abundance π_i in Equation (5) for those discovered species. Without any loss of generality, assume the species counts for the k discovered species are $(X_1, X_2, \ldots, X_k)$. If we adopt the traditional MLE estimator $\hat{\pi}_i = X_i/n$, then $\sum_{i=1}^{k} \hat{\pi}_i = \sum_{i=1}^{S} \hat{\pi}_i I(X_i > 0) = 1$. This implies that any unseen species has zero probability of being discovered. Therefore, the missing species are ignored in the maximum likelihood approach. This is un-reasonable in many applications where rare species may exist. We will use the concept of sample coverage to modify the traditional sample proportion. The sample coverage is defined as

$$C = \sum_{i=1}^{S} \pi_i I[X_i > 0], \tag{6}$$

which represents the fraction of the total abundances of the discovered species. We can interpret $1 - C$ as the conditional (on data) probability of discovering a new species if an additional observation (i.e., individual) were to be taken. A well known estimator originally proposed by Turing (see Good, 1953) for this conditional probability is the proportion of singletons in the sample. This can be intuitively understood because a new species must be a singleton in the enlarged sample that includes the additional individual.

For notational convenience, define f_m as the number of species with m individuals in the sample, i.e., $f_m = \sum_{i=1}^{S} I(X_i = m)$, $m = 0, 1, 2, \ldots, n$. Note that f_0 denotes the number of missing species and we have $\sum_{m=0}^{n} f_m = S$, $\sum_{m=1}^{n} f_m = k$ and $\sum_{m=1}^{n} m f_m = n$. Using this notation, the sample coverage is estimated by $\hat{C} = 1 - f_1/n$. This estimator performs very well even in highly heterogeneous cases; e.g., see Esty (1986). Therefore, we have the following approximation

$$C = \sum_{i=1}^{S} \pi_i I[X_i > 0] \approx \hat{C} = 1 - f_1/n, \tag{7}$$

which also intuitively implies that the fraction of the abundances un-represented in the sample is approximately the proportion of singletons.

Under the model in which species count $\{(X_1, X_2, \ldots, X_S);\ X_i \geqq 0,\ i = 1, 2, \ldots, S\}$ is a multinomial distribution with parameter $(n; \pi_1, \pi_2, \ldots, \pi_S)$, the distribution of $\{(X_1, X_2, \ldots, X_k);\ X_i > 0,\ i = 1, 2, \ldots, k\}$ conditional on the k observed species is still multinomial but with parameter $(n; \pi_1^*, \pi_2^*, \ldots, \pi_k^*)$ where

$$\pi_i^* = \frac{\pi_i}{\sum_{j=1}^{S} \pi_j I(X_j > 0)} = \frac{\pi_i}{C}.$$

Therefore, conditional on there being unseen species, $\hat{\pi}_i = X_i/n$ is only a valid estimator for $\pi_i^* = \pi_i/C$, which gives a modified estimate $\tilde{\pi}_i = (X_i/n)\hat{C}$ for π_i. Ashbridge and Gouldie (2000) were the first to propose this modification and further applied it to the estimation of species richness. Note that for this modified estimator, the approximation in Equation (7) is satisfied because

$$\sum_{i=1}^{S} \tilde{\pi}_i \, I[X_i > 0] = \hat{C} \sum_{i=1}^{S} \frac{X_i}{n} \, I[X_i > 0] = \hat{C} = 1 - \frac{f_1}{n}.$$

Thus our proposed estimator that combines the Horvitz–Thompson adjustment and the concept of sample coverage is

$$\hat{H} = -\sum_{i=1}^{S} \frac{\tilde{\pi}_i \log(\tilde{\pi}_i)}{1 - (1 - \tilde{\pi}_i)^n} I(A_i) = -\sum_{i=1}^{S} \frac{\hat{C}\tilde{\pi}_i \log(\hat{C}\tilde{\pi}_i)}{1 - (1 - \hat{C}\tilde{\pi}_i)^n} I(A_i). \tag{8}$$

The proposed estimator is a function of the frequencies $(f_0, f_1, f_2, \ldots, f_n)$, which is approximately a multinomial distribution with parameter S and cell probabilities $(f_0/S, f_1/S, f_2/S, \ldots, f_n/S)$. Therefore, a variance estimator of the proposed estimator $\hat{H}$ can be obtained by using a standard asymptotic approach. That is, we have the following variance estimator

$$\hat{\text{var}}(\hat{H}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \hat{H}}{\partial f_i} \frac{\partial \hat{H}}{\partial f_j} \hat{\text{cov}}(f_i, f_j), \tag{9}$$

where $\hat{\text{cov}}(f_i, f_j) = f_i(1 - f_i/\hat{S})$ for $i = j$ and $\hat{\text{cov}}(f_i, f_j) = -f_i f_j/\hat{S}$ for $i \neq j$ and $\hat{S}$ denotes a proper estimator of the number of species. The estimation of S will be discussed below and the adequacy of this variance estimator will be shown in Section 3 by numerical results. When a reliable estimate for species richness is not obtainable, the jackknife technique (Efron and Tibshirani, 1993, p. 145), as suggested by a referee, provides a convenient variance estimator and confidence interval. However, our (unreported) simulation results show that the asymptotic variance estimator with a proper estimate for species richness is generally less biased than the jackknife method. Therefore, the asymptotic approach is used in this paper.

Since an estimator of the number of species is needed in the variance estimator and also in the bias-corrected MLE as will be discussed in later sections, we briefly address the estimation of species richness. A review on the topic is provided in Bunge and Fitzpatrick (1993). A computer program EstimateS which calculates various estimators of species richness is readily available from the website http://viceroy.eeb.uconn.edu/estimates. In this paper, we adopt the estimator based on the concept of sample coverage (see Chao and Lee, 1992; Chao *et al.*, 2000); see below. The authors have also developed a program, which will be discussed in Section 5.

In the sample coverage approach, abundant and rare species are treated separately. Abundant species are those having more than κ individuals in the sample; the observed rare species are those represented by only one, two, ..., and up to κ individuals in the sample. The estimation of the number of missing species is based entirely on the observed rare species because abundant species would be discovered in any sample anyway and thus they do not contain any information about the missing species. Let the total number of abundant and rare species in the sample be $S_{\text{abun}} = \sum_{i=\kappa+1}^{n} f_i = \sum_{i=1}^{S} I[X_i > \kappa]$ and $S_{\text{rare}} = \sum_{i=1}^{\kappa} f_i = \sum_{i=1}^{S} I[0 < X_i \leq \kappa]$. Then the estimator of the total number of species based on the estimated sample coverage $\hat{C}_{\text{rare}} = 1 - \sum_{i=1}^{k} if_i$ is Chao *et al.* (2000, Section 2)

$$\hat{S} = S_{\text{abun}} + \frac{S_{\text{rare}}}{\hat{C}_{\text{rare}}} + \frac{f_1}{\hat{C}_{\text{rare}}} \hat{\gamma}^2, \tag{10}$$

where

$$\hat{\gamma}^2 = \max\left\{\frac{S_{\text{rare}}}{\hat{C}} \frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{\left(\sum_{i=1}^{\kappa} if_i\right)\left(\sum_{i=1}^{\kappa} if_i - 1\right)} - 1, \quad 0\right\} \tag{11}$$

denotes the estimated squared coefficient of variation (CV) of the species abundance $\{\pi_1, \pi_2, \ldots, \pi_S\}$. The CV is defined as $\text{CV} = [\sum_{i=1}^{S} (\pi_i - \bar{\pi})^2/S]^{1/2}/\bar{\pi}$, where $\bar{\pi} = \sum_{i=1}^{S} \pi_i/S$. The value of CV characterizes the degree of heterogeneity among the species abundances. The CV is zero if and only if the species have equal abundance. The larger the CV, the greater the degree of heterogeneity. A value of the cut-off point, $\kappa = 10$, is adopted throughout the paper based on empirical experiences (Chao *et al.*, 1993).

It would be interesting to find under what circumstances we can conclude that there are no missing cells in a sample. Notice that when there are no singletons (i.e., $f_1 = 0$) and the sample size is large enough, the sample coverage is estimated from Equation (7) to be 1, which means that the probability of finding a new species in an additional draw of any individual is 0. Based on (10), the estimator of the total number of species is just the number of observed species. Consequently, if all species are represented by at least two individuals in the sample, then the species survey is complete. This is intuitively sensible to biologists and ecologists; see Colwell and Coddington (1994).

3. Simulation study

To examine the performance of the proposed estimator, we present in this section simulation results based on four types of abundance models. The number of species was fixed to be 100. Four abundance models for $(\pi_1, \pi_2, \ldots, \pi_{100})$ were considered and are given below, where c is a normalizing constant such that $\sum_{i=1}^{100} \pi_i = 1$. In Case 1, the relative abundances are proportional to a sample from a uniform distribution. In Case 2, the relative abundances are a random sample from a Dirichlet distribution with parameter 1, which is the well-known MacArthur's broken-stick model (MacArthur, 1957). Cases 3 and 4 consider a model where the abundances are fixed in the special form of a Zipf–Mandelbrot model (Zipf, 1965; Mandelbrot, 1977). All the simulation results are given in Tables 1 to 4. We considered three sample sizes ($n = 50$, 75, and 100). For each combination of abundance model and sample size, 5000 simulated data sets were generated.

Case 1. (Random uniform model) $\pi_i = ca_i$, where $(a_1, a_2, \ldots, a_{100})$ are a random sample from a uniform $(0, 1)$ distribution. (The average values of H and CV over the 5000 trials are 4.412 and 0.58 respectively.)

Case 2. (Broken-stick model) $\pi_i = ca_i$, where $(a_1, a_2, \ldots, a_{100})$ are a random sample from an exponential distribution. Or equivalently $(\pi_1, \pi_2, \ldots, \pi_{100})$ is a Dirichlet distribution with parameter 1. (The average values of H and CV over the 5000 trials are 4.188 and 1.00 respectively.)

Case 3. (Zipf–Mandelbrot model) $\pi_i = c/(i+2)$, $i = 1, 2, \ldots, 100$. $H = 4.088$ and $\text{CV} = 1.34$.

Case 4. (Zipf–Mandelbrot model) $\pi_i = c/i$, $i = 1, 2, \ldots, 100$. $H = 3.681$ and $\text{CV} = 2.25$.

For each generated data, the following four estimators and their estimated standard errors were calculated:

1. MLE: defined in Equation (2).
2. Bias-corrected MLE (MLE_bc): it is obtained by adding $(\hat{S}-1)/(2n)$ to the MLE, where $\hat{S}$ is the estimator given in (10).
3. Jackknife: the jackknife estimator proposed by Zahl (1977).
4. Proposed: our proposed estimator given in (8).

The estimated s.e. of the jackknife estimator was calculated using a pseudo-value approach; see Zahl (1977). For the other estimators, the estimated standard errors were obtained by an asymptotic method as described in Section 2.2. The associated 95% confidence interval for each estimator was constructed using the normal critical points. Although for the jackknife estimator the critical points based on the t-distribution were suggested in the literature, we used the normal critical points instead because the degrees of freedoms in our cases were large; thus the critical points for the normal and t-distribution are almost identical.

The non-parametric MLE proposed by Norris and Pollock (1998) was not included in our simulation comparison because the computation for a single data set in some cases took over an hour of computer time. This non-parametric MLE will be discussed for real data examples in the next section.

The resulting 5000 estimates and their s.e.'s were averaged to give the results in Tables 1–4 under the headings ''Average Estimate'', ''Average Bias'' and ''Average Estimated s.e.'' We also computed for the 5000 entropy estimates the sample s.e. and sample root mean squared error (RMSE); they are given under the headings ''Sample s.e.'' and ''Sample RMSE''. The percentage of data sets in which 95% confidence intervals cover

Table 1. Comparison of estimators, 5000 simulation trials $S=100$, random uniform model, CV = 0.58, $H=4.412$.

Size n (Species seen)	*Method*	*Average Estimate*	*Average Bias*	*Sample s.e.*	*Average Estimated s.e.*	*Sample RMSE*	*95% C. I. Coverage*
50	MLE	3.517	– 0.895	0.084	0.062	0.899	0.0
(37)	MLE_bc	4.398	– 0.014	0.346	0.324	0.346	90.0
	Jackknife	4.163	– 0.248	0.143	0.114	0.287	44.0
	Proposed	4.405	– 0.007	0.239	0.231	0.239*	93.5
75	MLE	3.759	– 0.652	0.075	0.057	0.657	0.0
(48)	MLE_bc	4.327	– 0.084	0.171	0.164	0.191	85.1
	Jackknife	4.297	– 0.115	0.120	0.099	0.166	75.5
	Proposed	4.409	– 0.002	0.158	0.150	0.158*	94.0
100	MLE	3.902	– 0.509	0.067	0.053	0.513	0.0
(57)	MLE_bc	4.329	– 0.082	0.115	0.110	0.142	82.6
	Jackknife	4.356	– 0.055	0.100	0.088	0.114*	87.7
	Proposed	4.416	0.005	0.119	0.114	0.119	94.3

*Denotes the smallest RMSE.

Table 2. Comparison of estimators, 5000 simulation trials $S = 100$, broken stick model, CV = 1.00, $H = 4.188$.

Size n (*Species seen*)	*Method*	*Average Estimate*	*Average Bias*	*Sample s.e.*	*Average Estimated s.e.*	*Sample RMSE*	*95% C. I. Coverage*
50	MLE	3.385	– 0.803	0.107	0.073	0.810	0.0
(34)	MLE_bc	4.098	– 0.090	0.290	0.275	0.303	85.9
	Jackknife	3.956	– 0.232	0.166	0.126	0.286	54.9
	Proposed	4.098	– 0.090	0.231	0.215	0.248*	89.5
75	MLE	3.596	– 0.592	0.098	0.068	0.600	0.0
(43)	MLE_bc	4.091	– 0.097	0.175	0.163	0.200	83.7
	Jackknife	4.063	– 0.125	0.141	0.109	0.189	75.7
	Proposed	4.118	– 0.070	0.166	0.154	0.180*	90.8
100	MLE	3.720	– 0.468	0.091	0.063	0.476	0.0
(50)	MLE_bc	4.109	– 0.079	0.134	0.122	0.155	85.3
	Jackknife	4.114	– 0.074	0.123	0.097	0.144*	84.8
	Proposed	4.136	– 0.052	0.136	0.126	0.145	92.5

*Denotes the smallest RMSE.

the true value is shown in the last column of each table. The average of the number of species seen in the sample is also listed in each table, and the proportion of discovered species is in the range of 27 to 57%.

As expected, the traditional MLE seriously underestimates in all cases. Although the MLE exhibits the smallest s.e., it has the largest RMSE due to a large bias. The maximum coverage probability for the 95% confidence interval is only 2%; thus almost none of the associated confidence intervals covered the true parameter. Therefore, the MLE cannot provide a reliable diversity estimate when there are unseen species.

The bias-corrected MLE largely removes the bias, especially when the value of CV is large. Tables 3 and 4 show that the bias-corrected MLE has the smallest magnitude of bias in the Zipf–Mandelbrot models. This bias-corrected estimator not only reduces bias but also improve the performance of its associated confidence intervals. The average of the coverage probabilities for 95% confidence intervals is about 90%. However, the substitution of an estimated number of species increases the variance so that the precision of the bias-corrected MLE is the lowest in the four estimators considered in this study. The resulting RMSE is larger than that of the proposed estimator; see below.

The jackknife estimator and the proposed estimator are generally comparable in bias and variance. If we restrict our comparison to these two estimators, then the jackknife estimator has smaller variance but larger bias whereas our proposed estimator has smaller bias but larger variance. The average coverage probabilities for the jackknife and the proposed estimators are respectively 71% and 93%. Thus the jackknife method generally produces an acceptable point estimate but the associated confidence interval has a much lower chance of containing the true parameter than the nominal level. The proposed estimator generally is preferable with respect the RMSE and the coverage probability of interval estimation. Even when there is a large fraction of missing species, our proposed

Table 3. Comparison of estimators, 5000 simulation trials $S = 100$, Zipf–Mandelbrot model, CV = 1.34, $H = 4.088$.

Size n (*Species seen*)	*Method*	*Average Estimate*	*Average Bias*	*Sample s.e.*	*Average Estimated s.e.*	*Sample RMSE*	*95% C. I. Coverage*
50	MLE	3.271	– 0.817	0.125	0.088	0.827	0.0
(32)	MLE_bc	4.036	– 0.052	0.325	0.341	0.329	90.1
	Jackknife	3.808	– 0.280	0.182	0.143	0.334	50.0
	Proposed	3.916	– 0.172	0.231	0.241	0.288*	87.0
75	MLE	3.470	– 0.618	0.114	0.083	0.629	0.0
(41)	MLE_bc	4.050	– 0.039	0.216	0.232	0.220	92.6
	Jackknife	3.920	– 0.168	0.155	0.126	0.229	71.2
	Proposed	3.967	– 0.121	0.172	0.191	0.211*	90.6
100	MLE	3.593	– 0.496	0.103	0.079	0.506	0.0
(48)	MLE_bc	4.060	– 0.029	0.165	0.186	0.167	95.1
	Jackknife	3.983	– 0.106	0.133	0.114	0.170	81.5
	Proposed	4.010	– 0.078	0.139	0.166	0.160*	94.8

*Denotes the smallest RMSE.

Table 4. Comparison of estimators, 5000 simulation trials $S = 100$, Zipf-Mandelbrot model, CV = 2.25, $H = 3.681$.

Size n (*Species seen*)	*Method*	*Average Estimate*	*Average Bias*	*Sample s.e.*	*Average Estimated s.e.*	*Sample RMSE*	*95% C. I. Coverage*
50	MLE	2.950	– 0.731	0.170	0.123	0.750	0.0
(27)	MLE_bc	3.707	0.026	0.399	0.472	0.400	95.9
	Jackknife	3.401	– 0.280	0.222	0.179	0.357	64.2
	Proposed	3.493	– 0.188	0.238	0.295	0.303*	92.5
75	MLE	3.116	– 0.565	0.152	0.115	0.585	0.2
(35)	MLE_bc	3.668	– 0.013	0.265	0.353	0.265	97.8
	Jackknife	3.499	– 0.182	0.190	0.158	0.263	76.1
	Proposed	3.571	– 0.110	0.191	0.259	0.221*	96.9
100	MLE	3.221	– 0.46	0.136	0.108	0.479	2.0
(41)	MLE_bc	3.668	– 0.012	0.202	0.305	0.202	99.3
	Jackknife	3.558	– 0.123	0.164	0.143	0.205	83.2
	Proposed	3.624	– 0.057	0.162	0.243	0.172*	99.1

*Denotes the smallest RMSE.

estimator can produce a reliable diversity index. The estimated s.e. formula given in (9) works well when the CV is not relatively large, but it overestimates when the CV is large as shown in Table 4.

4. Real data examples

4.1 *Tropical insect data (Janzen, 1973a, b) and bird data (Batten, 1976)*

Janzen (1973a, b) presented many valuable data sets on tropical foliage insects from sweep samples taken in 25 sites in Costa Rica and the Caribbean Islands. We select one set from Janzen's collection to illustrate our method. Table 5 gives the frequency counts for beetles collected respectively in day-time and night-time from the site referred to as ''Osa primary-hill, dry season, 1967'' in Janzen's paper.

For these two data sets, most species had only one, two or three individuals represented in the sample, and there were only a few abundant species. That is, the data information is concentrated on the lower-order capture frequencies. The estimated CV values for the day-time and night-time data are, respectively, 0.938 and 1.099 based on Formula (11) for the counts $\{f_1, f_2, \ldots, f_{10}\}$. These relatively high values of CV indicate that the community is highly heterogeneous in species abundances and any estimator that does not incorporate the heterogeneity would have severe negative bias for species richness.

Using (10) and (11), we obtain an estimate of 263 (s.e. 64.4) species for the day-time data, and an estimate of 269 (s.e. 69.7) species for the night-time data. This shows that a relatively large fraction of species has been missed in the samples. Substituting these two estimates, we can obtain the bias-corrected MLE. In Table 6, we list the MLE, bias-corrected MLE, jackknife estimate and the proposed estimate and their estimated s.e.'s. We also present the non-parametric MLE, which was calculated from a computer program

Table 5. Frequency counts for beetles data.

						Day-Time					
m	1	2	3	4	5	6	11			*Species seen*	*Individuals*
f_m	59	9	3	2	2	2	1			78	127
						Night-Time					
m	1	2	3	5	7	10	14	16	18	*Species seen*	*Individuals*
f_m	56	9	7	2	1	1	1	1	1	79	170

Table 6. Comparison of various estimates of Shannon's index of diversity for beetles data (estimated s.e. in parenthesis).

Estimate	*Day-Time*	*Night-Time*
MLE	4.08 (0.07)	3.83 (0.09)
Bias-corrected MLE	5.11 (0.38)	4.62 (0.38)
Jackknife	4.62 (0.11)	4.24 (0.12)
Non-parametric MLE	4.07	3.81
Proposed	4.70 (0.21)	4.30 (0.21)

Table 7. Frequency counts for birds data.

						Endemic Woodland							
m	1	2	3	5	6	11	16	21	25	26	35	*Species seen*	*Individuals*
f_m	4	3	5	1	1	1	1	1	1	1	1	20	170
						Conifer Plantation							
m	1	2	3	4	5	9	11	14	20	30	65	*Species seen*	*Individuals*
f_m	2	1	2	1	1	1	1	1	1	2	1	14	198

Table 8. Comparison of various estimates of Shannon's index of diversity for birds data (estimated s.e. in parenthesis).

Estimate	*Endemic Woodland*	*Conifer Plantation*
MLE	2.41 (0.07)	2.06 (0.07)
Bias-corrected MLE	2.47 (0.12)	2.09 (0.17)
Jackknife	2.48 (0.08)	2.10 (0.07)
Non-parametric MLE	2.40	2.04
Proposed	2.49 (0.10)	2.09 (0.16)

supplied by Dr James L. Norris. A bootstrap s.e. was suggested in Norris and Pollock (1998), but we were not able to do this because bootstrapping for the nonparametric MLE would be time-consuming.

Table 6 shows that the MLE and the non-parametric MLE are very close and yield the lowest estimate, but the bias-corrected MLE yields the highest estimate. The jackknife and the proposed estimates are in-between and the two estimates are higher than the MLE. The bias-corrected MLE has the lowest precision, as discussed in the simulation. All estimates imply that the diversity in the day-time is slightly higher than that in the night-time. Whether the observed difference is statistically significant is another interesting research topic; see Solow (1993). The 95% confidence intervals based on the jackknife are (4.40, 4.84) for the day-time data, and (4.00, 4.48) for the night-time data. The corresponding intervals based on the proposed estimator are respectively (4.29, 5.11) for the day-time data, and (3.89, 4.71) for the night-time data. Although our estimated s.e. is larger and thus the associated confidence interval is longer, the results in the Simulation Section lead us to expect that our intervals would have coverage probability closer to the nominal level. Same conclusion applies to the other two data analysis results as well.

We now investigate the behavior of all estimators when most species are found in the sample. Consider the interesting data sets originally reported by Batten (1976) and discussed in Magurran (1988, pp. 145–149). The purpose was to determine whether conifer plantations are less diverse than the endemic woodland in Ireland. The frequency counts of bird species in two woodland plots, a representative of the endemic woodland and a conifer plantation, are given in Table 7.

For both woodland plots, the data information is dominated by the abundant species. The estimated sample coverage is very high. Based on the estimator in (10), we conclude that there were two missing species for the endemic woodland, and only one species were undiscovered for the conifer plantation. Thus almost all species were discovered.

The five estimates of diversity for both plots are given in Table 8. The results exhibits little difference and all estimators are quite precise. For each estimator, the observed diversity is higher in the endemic woodland than in the conifer plantation. Solow (1993) applied a randomization procedure to assess the significance of difference in the observed MLE. Generally, in nearly complete species inventories and/or data information is concentrated on the high-order frequencies, all estimators yield approximately the same results and work equally well.

4.2 *Coin data (Holst, 1981)*

This data set was discussed in Holst (1981), Chao and Lee (1992), and Haas and Stokes (1998). Two hundred and four coins were found in a hoard of ancient coins. The coins were classified into different die types. For the obverse side, totally 141 types were identified and for the reverse side there were 178 types. The frequency count f_m in this example is interpreted as the number of die types with m coins in the sample. The data for the two sides are given in Table 9.

The MLE of Shannon's entropy is 4.80 for the obverse side and 5.13 for the reverse side. Holst (1981) and Chao and Lee (1992) concluded that for both sides there was a large proportion of classes un-discovered in the sample. Therefore, the MLE is likely to be negatively biased. Holst (1981) indicated that for the reverse side it is reasonable to

Table 9. Frequency counts for coin data.

				Obverse Side					
m	1	2	3	4	5	6	7	*Species seen*	*Individuals*
f_m	102	26	8	2	1	1	1	141	204
				Reverse Side					
m	1	2	3	4	5	6	7	*Species seen*	*Individuals*
f_m	156	19	2	1	0	0	0	178	204

Table 10. Comparison of various estimates of Shannon's index of diversity for coin data (estimated s.e. in parenthesis).

Estimate	*Obverse Side*	*Reverse Side*
MLE	4.80 (0.04)	5.13 (0.03)
Bias-corrected MLE	5.72 (0.19)	7.20 (0.49)
Jackknife	5.41 (0.07)	5.96 (0.05)
Non-parametric MLE	4.80	5.13
Proposed	5.56 (0.14)	6.63 (0.19)

assume that each die produced about the same number of coins whereas this assumption was inappropriate for the obverse side. The obverse side is more heterogeneous than the reverse side among the number of coins produced by dies. This can be seen because the estimated CV values for the obverse and reverse sides were 0.69 and 0.36 respectively. From (10), the total number of types for the obverse side is estimated to be 378 (s.e. 65) for the reverse side and 844 (s.e. 187) for the reverse side. The estimated number of dies for the reverse side is significantly higher than that for the observe side. This is understandable, because the designs on the obverse side (e.g., portrait of the sovereign for the data) are usually more complicated than those secondary designs on the reverse side. Substituting these two estimated numbers of dies, we get the bias-corrected estimates. In Table 10, we list all the five estimates for both sides. The MLE and non-parametric MLE yield identical results. The proposed estimate is much higher than the two MLE's for both sides. All estimates reveal that the reverse side is more diverse than the obverse side.

5. Concluding remarks and discussion

We have proposed a non-parametric method to estimate Shannon's index of diversity when there are unseen species in a sample. This approach combines the Horvitz–Thompson estimator and the concept of sample coverage to adjust for unseen species. The traditional maximum likelihood approach is valid only when the species survey is complete and all species are found in the sample. In such complete surveys, our proposed method and the previous estimators (the MLE, the bias-corrected MLE, jackknife estimator and the non-parametric MLE) work equally well. However, conducting a complete survey may require an extraordinary sampling effort because a large number of species with relatively small abundances may exist. When there is a non-negligible number of species missed in the sample as in the models considered in Section 3, our estimator is generally preferable to the previous estimators in terms of RMSE. The associated confidence interval constructed from our estimator also provides satisfactory coverage probability. The proposed method performs reasonably well even when a relatively large fraction of species is missing in the sample. Therefore, biodiversity can be estimated without expending much effort on searching for rare species.

Although we specifically deal with Shannon's index of diversity in this paper, the approach can be readily applied to other types of biological indices (Good, 1953; Smith and Grassle, 1977), e.g., the widely used Simpson's index of diversity. It follows from Smith and Grassle (1977) that both the MLE and minimum variance unbiased estimator (MVUE) exist for Simpson's index. Our (unreported) simulation results show that the proposed estimator for Simpson's index much improves the MLE, but the improvement over the MVUE is limited. Therefore, our method is more useful for situations in which no MVUE exists such as the entropy case considered in this paper.

A computer program SPADE (Species Prediction And Diversity Estimation), written in C language, that calculates various estimators for species richness and diversity indices may be obtained from the first author upon request and will be available soon on the website at http://chao.stat.nthu.edu.tw.

Acknowledgments

The authors would like to thank Dr James L. Norris for making available his computer program for calculating the non-parametric maximum likelihood estimator. All the non-parametric results in Section 4 were obtained by the program. Two reviewers kindly provided helpful comments and suggestions. The National Science Council of Taiwan is also acknowledged for its support for this research.

References

Ashbridge, J. and Goudie, I.B.J. (2000) Coverage-adjusted estimators for mark-recapture in heterogeneous populations. *Communications in Statistics-Simulation*, **29**, 1215–37.

Basharin, G.P. (1959) On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and Its Applications*, **4**, 333–6.

Batten, L.A. (1976) Bird communities of some Killarney woodlands. *Proceedings of the Royal Irish Academy*, **76**, 285–313.

Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**, 364–73.

Bunge, J., Fitzpatrick, M., and Handley, J. (1995) Comparison of three estimators of the number of species. *Journal of Applied Statistics*, **22**, 45–59.

Chao, A. and Lee, S.-M. (1992) Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, **87**, 210–17.

Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000) Estimating the number of shared species in two communities. *Statistica Sinica*, **10**, 227–46.

Chao, A., Ma, M.-C., and Yang, M.C.K. (1993) Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, **80**, 193–201.

Colwell, R.K. and Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society, London B*, **345**, 101–18.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Engen, S. (1978) *Stochastic Abundance Models*, Halsted Press, New York.

Esty, W. (1986) The efficiency of Good's nonparametric coverage estimator. *The Annals of Statistics*, **14**, 1257–60.

Good, I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–64.

Haas, P. and Stokes, L. (1998) Estimating the number of classes in a finite population. *Journal of the American Statistical Association*, **93**, 1475–87.

Holst, L. (1981) Some asymptotic results for incomplete multinomial or Poisson samples. *Scandinavian Journal of Statistics*, **8**, 243–6.

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–85.

Hutcheson, K. and Shenton, L.R. (1974) Some moments of an estimate of Shannon's measure of information. *Communications in Statistics*, **3**, 89–94.

Janzen, D.H. (1973a) Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions. *Ecology*, **54**, 659–86.

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

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

Magurran, A.E. (1988) *Ecological Diversity and Its Measurement*, Princeton, Princeton University Press, New Jersey.

Mandelbrot, B. (1977) *Fractals, Form, Chance and Dimension*, Freeman, San Francisco.

Norris III, J.L. and Pollock, K.H. (1998) Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics*, **5**, 391–402.

Peet, R.K. (1974) The measurement of species diversity. *Annual Review of Ecology and Systematics*, **5**, 285–307.

Pielou, E.C. (1975) *Ecological Diversity*, Wiley, New York.

Smith, W. and Grassle, J.F. (1977) Sampling properties of a family of diversity measures. *Biometrics*, **33**, 283–92.

Solow, A.R. (1993) A simple test for change in community structure. *Journal of Animal Ecology*, **62**, 191–3.

Thompson, S.K. (1992) *Sampling*, Wiley, New York.

Zahl, S. (1977) Jackknifing an index of diversity. *Ecology*, **58**, 907–13.

Zipf, G.K. (1965) *Human Behavior and Principle of Least Effort*, Addison-Wesley, New York.

Biographical sketches

Anne Chao is Professor, Institute of Statistics, National Tsing Hua University, Taiwan. Her interest in ecological and biological statistics began with analyzing birds' banding data in two estuaries of Taiwan. Since then, she has worked to enhance local people's appreciation of the diversity of natural environments. Specific methodological interests include species estimation and its applications, capture-recapture experiments and population size estimation, as well as related biological samplings.

Tsung–Jen Shen is a Ph.D. candidate in National Tsing Hua University, and is now working with Anne Chao on the estimation of various diversity indices. His area of research interest is species estimation and inferences in ecological statistics.