Statistica Sinica **6**(1996), 403-418

ESTIMATING THE NUMBER OF SPECIES VIA A MARTINGALE ESTIMATING FUNCTION

Anne Chao, Paul Yip* and Huey-Shyan Lin

*National Tsing Hua University and University of Hong Kong**

Abstract. A martingale estimating function is proposed to estimate the number of species under a multinomial model with possibly unequal cell probabilities. This approach provides a class of estimators including the maximum likelihood estimator for the equiprobable case and the nonparametric sample coverage estimator (Chao and Lee (1992)) for the non-equiprobable case. Consistency of the proposed estimators is discussed. A simulation study investigates the behavior of the proposed procedure. A data set on Chinese poems is given for illustration.

Key words and phrases: Number of classes, multinomial, heterogeneity, sample coverage, zero mean martingale.

1. Introduction

Estimating the number of species in a population is a classical problem in biological applications. Bunge and Fitzpatrick (1993) provided a review of various models and approaches. They also compiled an extended bibliography on this topic with over 550 references.

We focus on the most common multinomial model: Consider a population which consists of N unknown distinct species. We search this population by selecting one element at a time, noting its species identity and returning it to the population. A search is called an n-stage search if n selections are made. Imagine that the species are labeled $1, \ldots, N$ in any arbitrary fashion. In practical species search studies, any sampling element is identified by its species classification rather than the labeling order. This labeling is just for convenience of mathematical treatment. Let p_i be the probability that the selected one belongs to the ith species and let X_{ik} be the number of elements of the ith species in a k-stage search, then $(X_{1k}, \ldots, X_{Nk})$ is multinomially distributed for all $k = 1, \ldots, n$. Our aim is to estimate N, the number of distinct species, after n selections are made.

As indicated in Bunge and Fitzpatrick (1993), there are three principal frequentist procedures: In the first approach, one postulates a parametric functional form for the multinomial cell probabilities (e.g., McNeil (1973)). In the second approach, one approximates the distribution of the cell probabilities by

a parametric probability density function (e.g., Sichel (1986)). In the third nonparametric sample coverage approach, one estimates the number of species via the estimation of sample coverage (e.g., Chao and Lee (1992)). The coefficient of variation of the cell probabilities is shown to play an important role in measuring the heterogeneity of the population in the sample coverage approach.

In this paper, we present an alternative nonparametric technique using a martingale estimating function (Godambe (1985)) via the notion of sample coverage. This approach provides a class of estimators including the maximum likelihood estimator for the equiprobable case and the sample coverage estimator proposed by Chao and Lee (1992) for the non-equiprobable case. A history on the application of sample coverage to species and population size estimation can be found in Chao and Lee (1992) and references therein. The use of a martingale estimating function for related capture-recapture models is given in Becker (1984), Becker and Heyde (1990), Yip (1989, 1991) and Yip, Fong and Wilson (1993). However, the previous martingale estimating function approaches only deal with the equiprobable case. This paper extends it to incorporate the heterogeneity of species probabilities.

In Section 2, we present the martingale estimating function approach via the idea of sample coverage. A class of estimators is derived and the bootstrap method is proposed to obtain a variance estimator. In Section 3, results of a simulation are reported to assess the performance of the proposed procedure. A data set on the poems of Chinese poet Bai Juyi is given in Section 4 for illustration.

2. Martingale Estimating Function

A zero-mean-martingale (ZMM) in discrete time is a stochastic process $\{\mathcal{M}_k : k = 1, 2, \ldots\}$ such that $E(\mathcal{M}_1) = 0$, and for all $k = 1, 2, \ldots$ we have $E|\mathcal{M}_k| < \infty$ and $E(\mathcal{M}_{k+j} \mid \mathcal{F}_k) = \mathcal{M}_k$ for $j = 0, 1, \ldots$, where $\mathcal{F}_k$ denotes the σ-field generated by the search process of a k-stage search, i.e., $\mathcal{F}_k = \sigma\{X_{1i}, \ldots, X_{Ni};\ i = 1, \ldots, k\}$.

Let $f_{ik} = \sum_{j=1}^{N} I[X_{jk} = i]$ be the number of classes that have exactly i elements in a k-stage search, where $I(\cdot)$ is the indicator function. Denote the number of distinct species in a k-stage search by D_k, $D_k = \sum_{i=1}^{k} f_{ik} = \sum_{j=1}^{N} I[X_{jk} > 0]$. Further, let $m_k = I[\text{the } k\text{th selection is a discovered species}]$ and $u_k = I[\text{the } k\text{th selection is an undiscovered species}]$. Thus $u_k + m_k = 1$ for all k. Define the sample coverage of a k-stage search, C_k, as

$$C_k = \sum_{i=1}^{N} p_i I[\text{the } \ i\text{th species has already been discovered in a } \ k\text{-stage search}]. \tag{2.1}$$

Let $A_{i,k}$ be the event that the ith species is not captured in samples $1, \ldots, k-1$ but captured in sample k and $B_{i,k}$ be the event that the ith species is not captured

in samples $1, \ldots, k$. Then

$$E(u_k \mid \mathcal{F}_{k-1}) = E\Big\{\sum_{i=1}^{N} I(A_{i,k}) \mid \mathcal{F}_{k-1}\Big\} = \sum_{i=1}^{N} p_i I[B_{i,k-1}].$$

It follows from definition (2.1) that for $k = 1, \ldots, n$ ($C_0 \equiv 0$)

$$E(u_k \mid \mathcal{F}_{k-1}) = (1 - C_{k-1}). \tag{2.2}$$

Similarly, we have

$$E(m_k \mid \mathcal{F}_{k-1}) = C_{k-1}. \tag{2.3}$$

We can further show that

$$\begin{aligned}\mathrm{Var}(u_k \mid \mathcal{F}_{k-1}) &= \mathrm{Var}\Big\{\sum_{i=1}^{N} I(A_{i,k}) \mid \mathcal{F}_{k-1}\Big\} \\ &= \sum_{i=1}^{N} p_i(1-p_i) I(B_{i,k-1}) + \sum\sum_{i \neq j} (-p_i p_j) I(B_{i,k-1}) I(B_{j,k-1}) \\ &= C_{k-1}(1 - C_{k-1}). \end{aligned} \tag{2.4}$$

Similarly, we have

$$\mathrm{Var}(m_k \mid \mathcal{F}_{k-1}) = C_{k-1}(1 - C_{k-1}), \tag{2.5}$$

and

$$\mathrm{Cov}(u_k, m_k \mid \mathcal{F}_{k-1}) = -C_{k-1}(1 - C_{k-1}). \tag{2.6}$$

Based on (2.2) and (2.3), we can construct the following martingale difference:

$$\begin{aligned}\mathcal{D}_k &= NC_{k-1}[u_k - (1 - C_{k-1})] - N(1 - C_{k-1})[m_k - C_{k-1}] \\ &= (NC_{k-1})u_k - N(1 - C_{k-1})m_k.\end{aligned}$$

Thus the process $\mathcal{M} = \{\mathcal{M}_k : k = 1, 2, \ldots\}$ where $\mathcal{M}_k = \sum_1^k \mathcal{D}_i$ is a ZMM. To obtain a general martingale estimating function for N, we "integrate" a bounded predictable weight function w_{k-1} with respect to $\mathcal{M}$. Then the estimating function becomes

$$\mathcal{M}_n^* = \sum_{k=1}^{n} w_{k-1}\mathcal{D}_k = \sum_{k=1}^{n} w_{k-1}[(NC_{k-1})u_k - N(1 - C_{k-1})m_k] \tag{2.7}$$

$$= \sum_{k=1}^{n} w_{k-1}[(NC_{k-1}) - Nm_k]. \tag{2.8}$$

In this paper, we discuss two possible weight functions: $w_{k-1} = 1$ for all k and the optimal weight function suggested in Godambe (1985). He showed that the optimal weight function for a given martingale difference $\mathcal{D}_k$ is given by

$$w^*_{k-1} = \frac{E[\partial \mathcal{D}_k / \partial N \mid \mathcal{F}_{k-1}]}{E[\mathcal{D}_k^2 \mid \mathcal{F}_{k-1}]} \ . \tag{2.9}$$

Since N is an integer, instead of taking the derivative with respect to N we compute the first difference of $\mathcal{D}_k$. Here "optimal" is in the sense of giving the tightest asymptotic confidence interval in the class of estimating function $\mathcal{M}^*_n$.

For the special case that all p_i's are equal, i.e., $p_1 = p_2 = \cdots = p_N = 1/N$, we have $NC_{k-1} = D_{k-1}$ ($D_0 \equiv 0$) and $N(1-C_{k-1}) = N - D_{k-1}$. Hence the estimating function (2.7) reduces to $\sum\ w_{k-1}[D_{k-1}u_k - (N - D_{k-1})m_k]$, which is similar to that considered by Yip (1989, 1991), Yip, Fong and Wilson (1993), Becker (1984) and Becker and Heyde (1990) for capture-recapture models. Equating the above estimating function to its mean gives the estimator $\sum\ w_{k-1}D_{k-1} / \sum\ w_{k-1}m_k$. If $w_{k-1} = 1$, the above reduces to

$$\hat{N}_0 = \sum_{k=1}^{n}\ D_{k-1} \Big/ \sum_{k=1}^{n}\ m_k. \tag{2.10}$$

This is equivalent to the Schnabel estimator for capture-recapture studies if each selection is regarded as a trapping sample (Schnabel (1938) or Seber (1982)). Note that the Schnabel estimator depends on the sequential ordering of the selections. In other words, the Schnabel estimator is not invariant to permutation of selections because of conditioning on the previous history sequentially.

In case all of the p_i's are equal, the maximum likelihood estimator (MLE) $\hat{N}_{0,mle}$ is the solution of the following equation (Darroch (1958))

$$\sum_{i=0}^{D_n - 1} (N - i)^{-1} = n \ / \ N \ . \tag{2.11}$$

As N, $n \to \infty$ such that $n/N \to \lambda > 0$, the asymptotic variance of the MLE is

$$\mathrm{Var}(\hat{N}_{0,mle}) = N \ /[\exp(\lambda) - \lambda - 1]. \tag{2.12}$$

Note that the MLE is invariant to permutation of selections.

It follows from (2.3)-(2.6) and (2.9) that the optimal weight in this case is $w^*_{k-1} = 1/(N - D_{k-1})$. Accordingly, the optimal estimator corresponding to the optimal weight is the solution for the following equation

$$\sum_{1}^{n}[D_{k-1} - Nm_k]/(N - D_{k-1}) = 0.$$

It is easy to see the above equation is equivalent to (2.11); hence the optimal weight martingale estimator is exactly the MLE in the equiprobable case. Following a similar derivation as in Yip (1989, 1991), an estimated standard error for $\hat{N}_0$ and $\hat{N}_{0,mle}$ can be obtained by substituting appropriate weight and estimate respectively in the following:

$$\Big\{\sum_{k=1}^{n} w_{k-1}^2 D_{k-1}(N - D_{k-1})\Big\}^{1/2} \Big/ \sum_{k=1}^{n} w_{k-1} m_k. \tag{2.13}$$

In most practical applications, the equally-likely assumption is invalid. To relax this assumption, the nuisance parameters $p_1, \ldots, p_N$ can be modeled by the following two arguments:
(1) the fixed-effects model: $p_1, \ldots, p_N$ are regarded as fixed parameters. The essential relevant parameters are the mean $\bar{p} = \sum_{i=1}^{N} p_i/N = 1/N$ and the coefficient of variation (CV) $\gamma = [\sum_{i=1}^{N}(p_i - \bar{p})^2/N]^{1/2}/\bar{p}$.
(2) the random-effects model: $p_1, \ldots, p_N$ are a random sample from a N-dimensional variable $(P_1, \ldots, P_N)$ with the constraint $\sum P_i = 1$. Assume that $(P_1, \ldots, P_N)$ has a symmetric joint CDF with a common marginal $F(p)$ on (0,1). Then $F(p)$ has mean $\bar{p} = \int p\, dF(p) = 1/N$ by the common marginal assumption. Define the CV of $F(p)$ as $\gamma = [\int (p - \bar{p})^2 dF(p)]^{1/2}/\bar{p}$.

Both approaches will lead to exactly the same estimator. The basic motivation for handling the heterogeneous case is the following: instead of estimating N directly, we estimate it via the estimation of NC_{k-1}. The identity $NC_{k-1} = D_{k-1}$ is no longer valid when the p_i's are not equal. From (2.8), if an "estimator" $\widehat{NC}_{k-1}$ of NC_{k-1} can be found such that the magnitude of the term

$$\sum_{k=1}^{n} w_{k-1}[(NC_{k-1}) - (\widehat{NC}_{k-1})]$$

is negligible, then our estimating function for the heterogeneous case becomes

$$\mathcal{M}_n^{**} = \sum_{k=1}^{n} w_{k-1}\mathcal{D}_k^* = \sum_{k=1}^{n} w_{k-1}[(\widehat{NC}_{k-1}) - Nm_k], \tag{2.14}$$

where $\mathcal{D}_k^* = (\widehat{NC}_{k-1}) - Nm_k$. Here $\sum \mathcal{D}_i^*$ is approximately a ZMM and (2.14) gives the following estimator:

$$\sum_{k=1}^{n} w_{k-1}(\widehat{NC}_{k-1}) \Big/ \sum_{k=1}^{n} w_{k-1} m_k. \tag{2.15}$$

When the estimator $\widehat{NC}_{k-1}$ is $\mathcal{F}_{k-1}$-measurable, an asymptotic standard error is

$$N\Big\{\sum_{k=1}^{n} w_{k-1}^2[C_{k-1}(1 - C_{k-1})]\Big\}^{1/2} \Big/ \sum_{k=1}^{n} w_{k-1} m_k. \tag{2.16}$$

Our procedure is to find an estimator for $E(NC_{k-1})$ and subsequently use it as an estimator of NC_{k-1}. It can be shown that under both types of model

$$E(NC_{k-1}) = E(D_{k-1}) + \gamma^2 E(f_{1,k-1}) + R,$$

where $f_{10} \equiv 0$ and R is the remainder term in the expansion. Under the fixed-effects (random-effects) model, the expectation is in a conditional (unconditional) sense. The magnitude of the remainder term R is generally negligible compared to N. For example, if $(P_1, \ldots, P_N)$ has a symmetric Dirichlet distribution, we can theoretically show that R/N tends to 0 as N, $n \to \infty$ such that $n/N \to \lambda > 0$. (See Lin, Chao and Lee (1993) for a proof). Thus R will be ignored and we have

$$\widehat{NC}_{k-1} = D_{k-1} + \hat{\gamma}^2 f_{1,k-1}, \tag{2.17}$$

where $\hat{\gamma}$ is a CV estimator. We adopt the estimator of the CV based on $k-1$ selections, $\hat{\gamma}^2_{k-1}$, from Chao and Lee (1992) where

$$\hat{\gamma}^2_m = \max\left\{ \frac{D_m \sum_{i=1}^m i(i-1) f_{im}}{m(m-1)[1 - f_{1m}/m]} - 1,\ 0 \right\}. \tag{2.18}$$

It follows from (2.15), (2.17) and (2.18) that an estimator for N when $w_k = 1$ is given by

$$\hat{N} = \sum_{k=k_0}^{n} [D_{k-1} + \hat{\gamma}^2_{k-1} f_{1,k-1}] \Big/ \sum_{k=k_0}^{n} m_k. \tag{2.19}$$

All the summation starts with an initial time k_0 since we need sufficient observations to get a stable estimate of the CV. In the simulation of Section 3, we chose $k_0 = \sum_{i=1}^{10} i f_{in}/2$ since those species with more than 10 occurrences are treated separately as will be described later. The estimation of the CV is the most difficult part in this procedure. Alternatively, we estimate this parameter after all observations have been obtained. That is, consider the following estimator with a modified CV estimate:

$$\begin{aligned} \tilde{N} &= \sum_{k=1}^{n} [D_{k-1} + \hat{\gamma}^2_n f_{1,k-1}] \Big/ \sum_{k=1}^{n} m_k \\ &= \hat{N}_0 + \hat{\gamma}^2_n \sum_{k=1}^{n} f_{1,k-1} \Big/ \sum_{k=1}^{n} m_k. \end{aligned} \tag{2.20}$$

The optimal weight can be shown to be approximately equal to $w^*_{k-1} = 1/(1 - C_{k-1})$. An estimated weight $\hat{w}^*_{k-1}$ is obtained by substituting $\hat{C}_{k-1} = 1 - f_{1,k-1}/(k-1)$; (see Good (1953) and Robbins (1968)). Thus, we have two estimators associated with the optimal weight:

$$\hat{N}_w = \sum_{k=k_0}^{n} \hat{w}^*_{k-1} [D_{k-1} + \hat{\gamma}^2_{k-1} f_{1,k-1}] \Big/ \sum_{k=k_0}^{n} \hat{w}^*_{k-1} m_k \tag{2.21}$$

and

$$\tilde{N}_w = \sum_{k=1}^{n} \hat{w}^*_{k-1}[D_{k-1} + \hat{\gamma}_n^2 f_{1,k-1}] \Big/ \sum_{k=1}^{n} \hat{w}^*_{k-1} m_k. \tag{2.22}$$

Simulation results showed that the estimates of the asymptotic standard error given in (2.16) usually underestimate the true standard error in the non-equiprobable case. We adopt a bootstrap method to obtain a variance estimator. First, note that in many situations, $(f_{0n}, f_{1n}, \ldots, f_{nn})$ is approximately multinomially distributed with parameter N and cell probabilities $N^{-1}\sum_{i=1}^{N}\binom{n}{k}p_i^k(1-p_i)^{n-k}$ or $\binom{n}{k}\int p^k(1-p)^{n-k}dF(p)$, $k = 0, 1, \ldots, n$ for the fixed-effects and random-effects models respectively. The argument is similar to that given in the Appendix of Darroch et al. (1993). Note here we only have $\sum_{i=1}^{n} iE(f_{in}) = n$ instead of $\sum_{i=1}^{n} if_{in} = n$ in this approximation. In the case of $\hat{N}$, a bootstrap replication $(f^*_{0n}, f^*_{1n}, \ldots, f^*_{nn})$ under both models is then generated from a multinomial distribution with parameter $\hat{N}$ and estimated cell probabilities $f_{kn}/\hat{N}$, $k = 0, 1, \ldots, n$. However, the occurrence history of each species is needed to calculate the martingale estimator. For each of the species appearing i times, $i = 1, \ldots, n$, we randomly selected i observations from $\{1, 2, \ldots, n^*\}$ without repetition, where $n^* = \sum_{i=1}^{n} if^*_{in}$, indicating this species is discovered in the selected observations and not in others. That is, we randomly choose one sequential order from all possible permutations as the searching order.

For each set of the data, we generate B replications and B bootstrap estimates $\hat{N}^*_i$ can then be obtained, $i = 1, \ldots, B$. The bootstrap variance of $\hat{N}$ is simply the sample variance of $\hat{N}^*_i$, $i = 1, \ldots, B$, i.e.

$$\hat{\mathrm{Var}}(\hat{N}) = \Big[\sum_{i=1}^{B}(\hat{N}^*_i)^2 - (\sum_{i=1}^{B}\hat{N}^*_i)^2/B\Big] \Big/ (B-1).$$

We now show the sample coverage estimator proposed in Chao and Lee (1992) can be regarded as a special case of our approach in the following sense: Suppose we can extend the search one additional stage, i.e., the $(n+1)$th selection, after an n-stage search has been made. If we consider only the martingale difference $\mathcal{D}^*_{n+1}$ in (2.14), then an estimator becomes

$$\widehat{NC}_n / \; m_{n+1}, \tag{2.23}$$

where $m_{n+1} = I$[the additional observation is a discovered species]. Further, note that

$$E(m_{n+1}) = \sum_{i=1}^{N} p_i[1 - (1-p_i)^n] = E(C_n).$$

Replacing m_{n+1} in (2.23) by an estimator $\hat{C}_n$ of $E(C_n)$ and substituting $\widehat{NC}_n = D_n + \hat{\gamma}_n^2 f_{1n}$ (see Equation (2.17)), we then have exactly the sample coverage estimator (Chao and Lee (1992))

$$\hat{N}_s = \frac{D_n}{\hat{C}_n} + \frac{f_{1n}}{\hat{C}_n}\,\hat{\gamma}_n^2, \tag{2.24}$$

which is invariant to permutation of selections. Thus, the sample coverage estimator can be regarded as a martingale estimator conditioned on all data.

Simulation results have suggested that the above procedure generally produces reasonable estimates if the CV is not too large. When the CV is large, it implies long frequency data. Hence, a modified procedure is recommended as follows: Since the species with large class probabilities are discovered many times, they may be ignored from a practical point of view; i.e. we only consider those species with no more than κ occurrences. A suitable value of κ might be 10 as suggested in Chao, Ma and Yang (1993). The number of species that have occurred more than κ times is then added to the resulting estimate. That is, we only concentrate on a subset of species so that the CV for these species is smaller than that of the original one. In the simulation of Section 3, we treat species more than 10 times ($\kappa = 10$) separately and apply our procedure to only those species appearing up to 10 times.

A theoretical justification on the use of sample coverage and the martingale estimator is the consistency property based on an extended result of Chen (1980, 1981a, 1981b). (See Lin, Chao and Lee (1993) for details.) We merely summarize the conclusion as follows: Under a multinomial model, if the species probabilities $(P_1, \ldots, P_N)$ follow a symmetric Dirichlet distribution with parameter α, then $\gamma^2 \approx 1/\alpha$. As N, $n \to \infty$ such that $n/N \to \lambda > 0$, then
(1) if the CV can be consistently estimated by an estimator $\hat{\gamma}$, say, then the sample coverage estimator is consistent, i.e., we have

$$N^{-1}\Big[\frac{D_n}{\hat{C}_n} + \frac{f_{1n}}{\hat{C}_n}\,\hat{\gamma}^2\Big] \to 1 \ \text{ in probability;}$$

(2) under the same condition, the martingale estimator is consistent. That is, for any initial starting time k_0, we can show that

$$N^{-1}\Big\{\sum_{k=k_0}^{n}[D_{k-1} + \hat{\gamma}^2 f_{1,k-1}]\Big/\sum_{k=k_0}^{n} m_k\Big\} \to 1 \ \text{ in probability.}$$

It is clear that the CV plays the most important role in the consistency property. If the CV is known or can be consistently estimated, we are able to estimate the number of "invisible" species consistently. Otherwise, consistency becomes an

unattainable ideal because "there is nearly always a good chance that there are a large number of extremely rare species", as I. J. Good pointed out in Bunge and Fitzpatrick (1993).

3. Simulation Study

A simulation study was carried out to compare the relative merits of the martingale and sample coverage estimators. Comparisons of sample coverage estimator and other estimators are given in Chao and Lee (1992). We have focused on heterogeneous populations, since for the equi-probable case, the martingale estimators are reduced to the traditional estimators, which have already been studied by Chao and Lee (1992) and many others. We only present the simulation results for the fixed-effects model. Conclusions for the random-effects trials are generally similar. The following cases are reported: (the number of species N was fixed to be 100, the constant c in the first five cases is a normalizing constant such that $\sum\ p_i = 1$). The first five cases are in a form of Zipf's law which is widely prevalent in natural frequency data. For the other cases, the proportions of a set of 100 negative binomial variables were fixed through the simulation and used as the cell probabilities based on numismatics applications (see Esty (1985)).

Case 1. (CV = 2.25). $p_i = c/i,\ i = 1, 2, \ldots, 100.$
Case 2. (CV = 1.34). $p_i = c/(i+2),\ i = 1, 2, \ldots, 100.$
Case 3. (CV = 0.99). $p_i = c/(i+5),\ i = 1, 2, \ldots, 100.$
Case 4. (CV = 0.54). $p_i = c/(i+20),\ i = 1, 2, \ldots, 100.$
Case 5. (CV = 0.32). $p_i = c/(i+50),\ i = 1, 2, \ldots, 100.$
Case 6. (CV = 0.93). $p_i = X_i / \sum_{j=1}^{100} X_j$, where $X_1, X_2, \ldots, X_{100}$ are realizations from negative binomial (1, 0.04).
Case 7. (CV = 0.68). $p_i = X_i / \sum_{j=1}^{100} X_j$, where $X_1,\ X_2, \ldots, X_{100}$ are realizations from negative binomial (2, 0.04).
Case 8. (CV = 0.43). $p_i = X_i / \sum_{j=1}^{100} X_j$, where $X_1,\ X_2, \ldots, X_{100}$ are realizations from negative binomial (4, 0.04).

For each case and fixed sample size (100 and 200), 200 data sets were generated. For each generated data set, four martingale estimators ($\hat{N}$, $\tilde{N}$, $\hat{N}_w$ and $\tilde{N}_w$) and the sample coverage estimator $\hat{N}_s$ proposed by Chao and Lee (1992) as well as their estimated s.e.'s were calculated. If there were species appearing more than 10 times, we treated them separately. The initial value k_0 in calculating $\hat{N}$ and $\hat{N}_w$ was taken as $\sum_{i=1}^{10} if_{in}/2$. To get s.e. estimates for the martingale estimators, 200 bootstrap replications were used for each generated data set. The s.e. estimator for the sample coverage estimator was provided in Equation (2.21) of Chao and Lee (1992). Based on the 200 simulated data sets, the sample s.e.

as well as sample root mean squared error (RMSE) for each estimator were then obtained. All the results are given in Table 1.

Table 1. Simulation results for comparing estimates, 200 runs;
$\hat{N}$: martingale estimator, see (2.19);
$\tilde{N}$: martingale estimator using a modified CV estimator, see (2.20);
$\hat{N}_w$: optimal weighted martingale estimator, see (2.21);
$\tilde{N}_w$: optimal weighted martingale estimator using a modified CV estimator, see (2.22);
$\hat{N}_s$: sample coverage estimator proposed in Chao and Lee (1992), see (2.24).

n	method	estimate	bias	estimated s.e.	sample s.e.	sample RMSE
Case 1. ($p_i = 1/i$, CV= 2.25)						
100	$\hat{N}$	78	−22	21.0	18.0	28.7
	$\tilde{N}$	78	−22	19.1	15.3	26.5
	$\hat{N}_w$	79	−21	21.5	17.7	27.7
	$\tilde{N}_w$	80	−20	19.4	15.8	25.2
	$\hat{N}_s$	89	−11	22.5	20.8	23.6
200	$\hat{N}$	92	−8	14.7	13.4	15.9
	$\tilde{N}$	91	−9	12.6	12.2	14.9
	$\hat{N}_w$	92	−8	14.5	13.3	15.2
	$\tilde{N}_w$	93	−7	13.7	12.2	13.9
	$\hat{N}_s$	100	0	16.3	14.8	14.8
Case 2. ($p_i = 1/(i+2)$, CV= 1.34)						
100	$\hat{N}$	86	−14	18.6	17.2	22.4
	$\tilde{N}$	85	−15	17.0	16.2	22.0
	$\hat{N}_w$	87	−13	18.8	17.2	21.7
	$\tilde{N}_w$	88	−12	17.3	15.4	19.9
	$\hat{N}_s$	96	−4	20.4	20.7	21.1
200	$\hat{N}$	95	−5	11.8	11.6	12.6
	$\tilde{N}$	94	−6	11.0	10.8	12.3
	$\hat{N}_w$	96	−4	11.6	11.6	12.3
	$\tilde{N}_w$	96	−4	11.0	11.0	11.7
	$\hat{N}_s$	201	1	13.0	13.3	13.3

Table 1. (Continued)

n	method	estimate	bias	estimated s.e.	sample s.e.	sample RMSE
Case 3. ($p_i = 1/(i+5)$, CV= 0.99)						
100	$\hat{N}$	87	−13	17.0	17.2	21.5
	$\tilde{N}$	88	−12	15.4	16.6	20.7
	$\hat{N}_w$	88	−12	16.8	17.3	20.9
	$\tilde{N}_w$	90	−10	15.6	16.6	19.6
	$\hat{N}_s$	95	−5	17.9	19.4	20.1
200	$\hat{N}$	97	−3	10.3	10.0	10.4
	$\tilde{N}$	96	−4	9.7	8.9	9.6
	$\hat{N}_w$	98	−2	10.1	9.7	10.0
	$\tilde{N}_w$	98	−2	9.6	8.8	9.0
	$\hat{N}_s$	101	1	11.2	10.4	10.5
Case 4. ($p_i = 1/(i+20)$, CV= 0.54)						
100	$\hat{N}$	95	−5	15.9	17.2	17.8
	$\tilde{N}$	95	−5	14.5	14.7	15.5
	$\hat{N}_w$	96	−4	16.2	16.9	17.4
	$\tilde{N}_w$	96	−4	14.7	14.8	15.4
	$\hat{N}_s$	98	−2	16.1	17.1	17.1
200	$\hat{N}$	98	−2	7.6	7.8	8.1
	$\tilde{N}$	97	−3	7.3	7.1	7.6
	$\hat{N}_w$	98	−2	7.4	7.6	7.9
	$\tilde{N}_w$	98	−2	7.0	7.1	7.4
	$\hat{N}_s$	99	−1	7.9	8.2	8.2
Case 5. ($p_i = 1/(i+50)$, CV= 0.32)						
100	$\hat{N}$	101	1	16.3	16.4	16.4
	$\tilde{N}$	99	−1	14.4	13.9	13.9
	$\hat{N}_w$	102	2	16.4	16.3	16.5
	$\tilde{N}_w$	100	0	14.6	14.1	14.1
	$\hat{N}_s$	101	1	15.4	15.3	15.3
200	$\hat{N}$	99	−1	6.8	7.1	7.1
	$\tilde{N}$	99	−1	6.6	6.8	6.8
	$\hat{N}_w$	100	0	6.7	6.9	6.9
	$\tilde{N}_w$	100	0	6.2	6.4	6.4
	$\hat{N}_s$	100	0	6.9	7.2	7.2

Table 1. (Continued)

n	method	estimate	bias	estimated s.e.	sample s.e.	sample RMSE
Case 6. (class size $\sim$ NB$(1, 0.04)$, CV$= 0.93$)						
100	$\hat{N}$	76	-24	11.9	11.8	26.9
	$\tilde{N}$	75	-25	11.2	10.2	26.9
	$\hat{N}_w$	77	-23	11.9	11.7	26.3
	$\tilde{N}_w$	76	-24	11.1	10.0	25.7
	$\hat{N}_s$	79	-21	12.7	11.8	23.7
200	$\hat{N}$	83	-17	7.5	8.2	18.8
	$\tilde{N}$	84	-16	7.3	8.4	18.3
	$\hat{N}_w$	84	-16	7.4	8.1	18.1
	$\tilde{N}_w$	85	-15	7.1	8.1	17.2
	$\hat{N}_s$	88	-12	8.3	9.0	15.3
Case 7. (class size $\sim$ NB$(2, 0.04)$, CV$= 0.63$)						
100	$\hat{N}$	86	-14	13.2	12.2	18.8
	$\tilde{N}$	85	-15	12.0	10.9	18.5
	$\hat{N}_w$	87	-13	13.1	12.0	18.0
	$\tilde{N}_w$	86	-14	12.0	10.8	17.6
	$\hat{N}_s$	88	-12	13.1	12.1	17.4
200	$\hat{N}$	91	-9	6.8	7.5	11.8
	$\tilde{N}$	90	-10	6.6	6.8	12.1
	$\hat{N}_w$	91	-9	6.7	7.4	11.4
	$\tilde{N}_w$	91	-9	6.4	6.9	11.2
	$\hat{N}_s$	93	-7	7.2	8.0	10.7
Case 8. (class size $\sim$ NB$(4, 0.04)$, CV$= 0.42$)						
100	$\hat{N}$	94	-6	15.1	15.4	16.6
	$\tilde{N}$	94	-6	13.4	13.2	14.5
	$\hat{N}_w$	95	-5	15.0	15.3	16.3
	$\tilde{N}_w$	94	-6	13.5	13.0	14.1
	$\hat{N}_s$	95	-5	14.6	14.1	14.9
200	$\hat{N}$	98	-2	6.8	7.2	7.5
	$\tilde{N}$	97	-3	6.5	7.3	8.1
	$\hat{N}_w$	98	-2	6.6	7.2	7.5
	$\tilde{N}_w$	97	-3	6.2	7.0	7.4
	$\hat{N}_s$	98	-2	6.9	7.8	8.1

It is clear from Table 1 that all four martingale estimates are very close. As expected, the martingale estimator using an estimated optimal weight has smaller s.e. than the constant weight estimator, but the improvement is not significant. The martingale estimator $\tilde{N}(\tilde{N}_w)$ using a modified CV estimator has smaller RMSE than the estimator $\hat{N}(\hat{N}_w)$ using an adaptive CV estimator. As far as the RMSE is concerned, the martingale and sample coverage estimators are generally comparable. Note that the sample coverage estimator has smaller bias whereas the martingale estimators have smaller s.e.'s. It seems that $\tilde{N}_w$ generally has the smallest RMSE for the trials of Zipf's law. In all cases, the sample coverage estimate is higher than the four types of martingale estimates. The bootstrap s.e. estimates are generally satisfactory since they are close to the sample s.e.'s in most cases.

Table 2. Frequencies of Chinese poem data ($n = 2800$, $D_n = 857$).

i	f_{in}	i	f_{in}	i	f_{in}	i	f_{in}	i	f_{in}
1	393	7	17	13	6	19	1	26	1
2	171	8	16	14	3	20	2	27	1
3	89	9	3	15	1	21	1	30	1
4	51	10	10	16	3	23	1	32	1
5	25	11	9	17	6	24	2	34	1
6	29	12	7	18	4	25	1	56	1

4. Chinese Poem Data

A seven-character quartet is a Chinese poem of 28 characters which are divided into four parts with seven characters in each part. In a study of the seven-character quartets of China's most popular poet of the Tan'g Dynasty, Bai Juyi, 200 seven-character quartets were randomly selected from Bai's collected work. See Ma and Chao (1993) for a discussion on this data set. In this application, the species are distinct characters. Here we use the first 100 selected poems for illustration. Totally there were $n = 2800$ Chinese characters and $D_n = 857$ for distinct ones. The plot of D_k, $k = 1, 2, \ldots, 2800$ is presented in Figure 1. The frequencies for this data are listed in Table 2. Our aim is to estimate the number of distinct characters that had been used in Bai's collected quartets.

Usually the long frequency data as given in Table 2 implies large variation on the class probabilities, and consequently a large CV. The CV estimate $\hat{\gamma}_n = 1.43$ for all frequencies is quite large, which shows strong evidence of heterogeneity among the classes. If we wrongly assume that all the classes are equally-likely, the Schnabel estimate given in (2.10) is 910 with s.e. 18 using (2.13) and the MLE is 896 with s.e. 7 using (2.12). Based on many previous simulations, the estimates derived from the equiprobable assumption are generally biased downwards in

the heterogeneous situation. Hence these two estimates are likely to have severe negative biases.

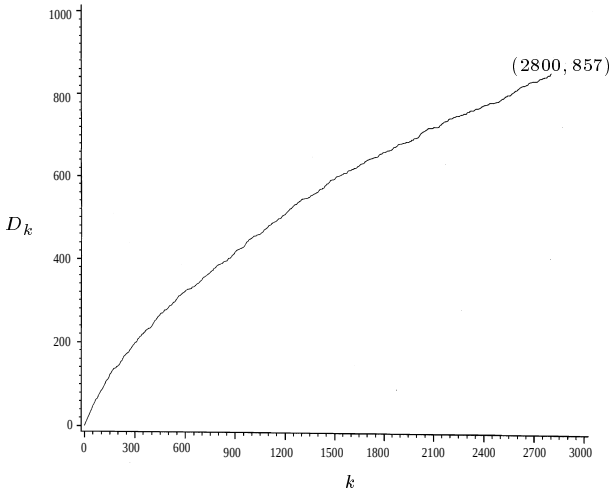

Figure 1. Plot of D_k with respect to k for Chinese Poem Data

Table 3. Analysis results for Chinese poem data

Equi-probable:		
$\hat{N}_0$	910	(s.e. = 18)
$\hat{N}_{0,mle}$	896	(s.e. = 7)
Martingale (weight = 1):		
$\hat{N}$	1297	(s.e. = 44)
$\tilde{N}$	1216	(s.e. = 41)
Martingale (optimal weight):		
$\hat{N}_w$	1311	(s.e. = 46)
$\tilde{N}_w$	1264	(s.e. = 46)
Sample coverage:		
$\hat{N}_s$	1372	(s.e. = 57)

As suggested in Section 3, we treat high and low frequency separately to obtain both the sample coverage and martingale estimates. A restriction is imposed on the subset with frequency no more than only 10 times, and the CV estimate reduces to $\hat{\gamma}_n = 0.78$. The classes appearing more than 10 times are then added to the resulting estimate. All the estimates and their s.e.'s are given in Table 3. The sample coverage estimate 1372 is slightly higher than the four martingale estimates, which is consistent with the findings in the simulation. The sample coverage estimator has also slightly higher estimated s.e. as expected. For a con-

stant weight, the martingale estimate and bootstrap s.e. based on 200 replications are $\hat{N} = 1297$ (s.e. $= 44$) and $\tilde{N} = 1216$ (s.e. $= 41$); for the estimated optimal weight, we have $\hat{N}_w = 1311$ (s.e. $= 46$) and $\tilde{N}_w = 1264$ (s.e. $= 46$). The initial value for the estimators $\hat{N}$ and $\hat{N}_w$ is taken as $k_0 = \sum_{i=1}^{10} if_{in}/2 = 940$. All the four martingale estimates are very close. Thus, we conclude the total number of distinct characters is around 1300 with an estimated s.e. around 50 in Bai's collected quartets. This number is much less than that of the most commonly used Chinese characters, 5000, as adopted in the Chinese E−TEN software system. The statistical result here may provide interesting evidence for Bai's honor as "the most popular poet in the Tan'g Dynasty" and "the poet of ordinary people". A program (written in C Language) which calculates all the proposed martingale and sample coverage estimates as well as their s.e.'s is available upon request.

Acknowledgements

The authors thank a referee for helpful comments which have led to some clarifications of the presentation. Part of the research was done while the second author (P. Y.) visited National Tsing Hua University. Research was supported by the National Science Council of Taiwan under Contracts NSC-83-0208-M007-006.

References

Becker, N. G. (1984). Estimating population size from capture-recapture experiments in continuous time. Austral. J. Statist. **26**, 1-7.

Becker, N. G. and Heyde, C. C. (1990) Estimating population size from multiple recapture experiments. Stochastic Process. Appl. **36**, 77-83.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. J. Amer. Statist. Assoc. **88**, 364-373.

Chao, A. and Lee, S. M. (1992). Estimating the number of classes via sample coverage. J. Amer. Statist. Assoc. **87**, 210-217.

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.

Chen, W.-C. (1980). On the weak form of Zipf's law. J. Appl. Probab. **17**, 611-622.

Chen, W.-C. (1981a). Limit theorems for general size distributions. J. Appl. Probab. **18**, 139-147.

Chen, W.-C. (1981b). Some local limit theorems in the symmetric Dirichlet-multinomial urn models. Ann. Inst. Statist. Math. **33**, 405-415.

Darroch, J. N. (1958). The multiple-recapture census I. Estimation of a closed population. Biometrika **45**, 343-359.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. J. Amer. Statist. Assoc. **88**, 1137-1148.

Esty, W. W. (1985). Estimation of the number of classes in a population and the coverage of a sample. Math. Scientist **10**, 41-50.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. Biometrika **72**, 419-428.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika **40**, 237-264.

Lin, H.-S., Chao, A. and Lee, S.-M. (1993). Consistency of an estimator of the number of species-in memory of Professor Wen-Chen Chen. J. Chinese Statist. Assoc. **31**, 253-270 (in Chinese).

Ma, M.-C. and Chao, A. (1993). Generalized sample coverage with an application to Chinese poems. Statist. Sinica **3**, 19-34.

McNeil, D. R. (1973). Estimating an author's vocabulary. J. Amer. Statist. Assoc. **68**, 92-96.

Robbins, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. Ann. Math. Statist. **39**, 256-257.

Sichel, H. S. (1986). Parameter estimation for a word frequency distribution based on occupancy theory. Comm. Statist., Theory Methods **15**, 935-949.

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. Amer. Math. Monthly **45**, 348-352.

Seber, G. A. F. (1982). The Estimation of Animal Abundance and Related Parameters, 2nd edition. Griffin, London.

Yip, P. (1989). An inference procedure for a capture and recapture experiment with time-dependent capture probabilities. Biometrics **45**, 471-479.

Yip, P. (1991). A martingale estimating equation for a capture-recapture experiment in discrete time. Biometrics **47**, 1081-1088.

Yip, P. S. F., Fong, D. Y. T. and Wilson, K. (1993). Estimating population size by recapture sampling via estimating function. Stochastic Models **9**, 179-193.

Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan 30043.

Department of statistics, University of Hong Kong, Hong Kong.

(Received February 1994; accepted April 1995)