Statistica Sinica 3(1993), 19-34

GENERALIZED SAMPLE COVERAGE WITH AN APPLICATION TO CHINESE POEMS

M.-C. Ma and Anne Chao

National Tsing Hua University

Abstract: The sample coverage for a random sample is defined as the sum of the class probabilities of the observed classes in multinomial sampling for which only one class occurs in each independent observation. This study generalizes the concept of sample coverage to the case that multiple possibly dependent classes can occur for each observation. A consistent estimator for the generalized sample coverage and its mean squared error properties are developed. The resulting estimator is shown to be an approximate empirical Bayes estimator. A data set on Chinese poems is given for illustration. Results of a simulation study are reported to show the general performance of the proposed estimator and to suggest that the usual estimator, without considering the dependence among classes, may yield severe bias in some situations.

Key words and phrases: Sample coverage, discovering species, empirical Bayes.

1. Introduction

Each independent observation in multinomial sampling is classified to exactly one class (i.e., only one class can occur). The sample coverage of a random sample is defined as the sum of the class probabilities of the observed classes. An equivalent measure is one minus the sample coverage, which can be interpreted as the conditional probability of discovering a new species in an additional observation given the sample. An widely used "estimator" for this conditional probability is the proportion of the singletons originally proposed by Turing according to Good (1953). The estimator has been discussed in Good (1953), Good and Toulmin (1956), Harris (1959), Knott (1967), Robbins (1968), Engen (1978), Starr (1979), Chao (1981), Esty (1982, 1983, 1986), Cohen and Sackrowitz (1990) and Lo (1992). A variance estimator and the construction of confidence intervals are given in Esty (1983). There is a close relationship between the estimation of sample coverage and that of the number of classes. Refer to Darroch and Ratcliff (1980), Esty (1985) and Chao and Lee (1992) for details.

A seven-character quartet (chueh chü) 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 quartet 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 and the proportion of this sample covered is the main interest. In other words, if we were to select another seven-character quartet, what is the probability of finding at least a new character. Previous papers discussing literature studies, e.g. Efron and Thisted (1976), Thisted and Efron (1987) and McNeil (1973) have treated each single word in the sample as an independent observation and ignored the possible dependence among words. This may be reasonable for English literature, but for Chinese poems, many characters tend to occur together for "rhyming" or "symmetric" purposes. For examples, typical symmetric words: sun with moon, mountain with stream, spring with fall etc. Thus the dependence between characters occurring within a poem should be considered. In this work, we regard each selected seven-character quartet with 28 characters as an "observation", although each distinct Chinese character is still regarded as a class. Hence for each observation, at most 28 classes can occur (some characters may be repeated) and the dependence among characters occurring within an observation is allowed.

We present, in Section 2, a general model introduced in Chao and Lee (1990) to be applied to the Chinese poems problem and define generalized sample coverage. An "estimator" for the generalized sample coverage is proposed in Section 3. The mean squared error as well as other properties are derived. A simulation study is reported in Section 4 to show the general performance of the proposed estimator and to compare it with the usual estimator treating each character as an independent observation. In the final section, a data set on Chinese poems is given for illustration.

2. Generalized Sample Coverage

Assume that there are N classes and t independent observations, N is unknown. In our application, N denotes the number of distinct characters used for all poems, and a selected poem (seven-character quartet) is regarded as an observation. For each observation, at least one but at most n classes can occur, $n \leq N$. The sample space of each observation can be expressed as $S = \{(Z_1, Z_2, \ldots, Z_N) | Z_j = 0 \text{ or } 1,\ 0 < \sum Z_i \leq n\}$, where $Z_j = I[\text{the } j\text{th class occurs}]$ and I is the indicator function. Also let $\mathcal{P}_{Z_1 Z_2 \ldots Z_N}$ be the corresponding probability for the outcome $Z_1, Z_2, \ldots, Z_N$, $0 < \sum Z_i \leq n$. If $\sum Z_i > n$, $\mathcal{P}_{Z_1 Z_2 \ldots Z_N}$ is defined to be 0. For notational simplicity, let

$$P_w = \mathcal{P}_{Z_1 Z_2 \ldots Z_N}, \text{ where } w = \{i | Z_i = 1\}, \tag{2.1}$$

and w is a nonempty subset of $\{1, 2, \ldots, N\}$ with at most n elements.

Let the ith observation have outcome $(Z_{i1}, Z_{i2}, \ldots, Z_{iN})$ from the sample space S. Thus the t observations can be expressed as a $t \times N$ matrix (Z_{ij}), where

$$Z_{ij} = I\,[\text{the } j\text{th class is found in the } i\text{th observation}].$$

For example, with $N = 4$ and $n = 2$, the set of all possible outcomes for each observation is $S = \{(1000),(0100),(0010),(0001),(1100),(1010),(1001),(0110),(0101),(0011)\}$, where $\mathcal{P}_{1000} = P_{\{1\}}$ represents the probability that only class 1 is observed and the other classes are not observed. $\mathcal{P}_{1100} = P_{\{1,2\}}$ represents the probability that only classes 1 and 2 are observed and the other classes are not observed. A similar interpretation pertains to other outcomes. In the following, we shall drop the brace and write P_1 instead of $P_{\{1\}}$, P_{12} instead of $P_{\{1,2\}}$ etc. for simplicity.

Suppose $t = 2$ and we have a data matrix

$$\begin{bmatrix} Z_{11} & Z_{12} & Z_{13} & Z_{14} \\ Z_{21} & Z_{22} & Z_{23} & Z_{24} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

To define the sample coverage for this problem, first consider the following question: If we were to take an additional observation, what is the probability of discovering at least one new class. It is intuitively clear in the above example that this probability is $\bar{C} = \mathcal{P}_{0010} + \mathcal{P}_{0001} + \mathcal{P}_{1010} + \mathcal{P}_{1001} + \mathcal{P}_{0110} + \mathcal{P}_{0101} + \mathcal{P}_{0011} = P_3 + P_4 + P_{13} + P_{14} + P_{23} + P_{24} + P_{34}$, i.e., any observation which includes the class 3 or 4 will discover at least one new class. Note that for this example $P_{123} = P_{124} = P_{134} = P_{234} = P_{1234} = 0$. Consequently the sample coverage becomes $C = 1 - \bar{C} = \mathcal{P}_{100} + \mathcal{P}_{010} + \mathcal{P}_{110} = P_1 + P_2 + P_{12}$. Now we can define the generalized sample coverage in the following:

Definition 1. Let $\mathcal{A} = \{w | w \subseteq \{1, 2, \ldots, N\}, 0 < \#w \leq n\}$ where $\#w$ denotes the number of elements in w, be the collection of nonempty subset of $\{1, 2, \ldots, N\}$ with at most n elements. Let a random sample of t observations be taken and $X_{(k)} = \sum_{i=1}^{t} Z_{ik}$ be the number of occurrences of the kth class in t observations. We define the generalized sample coverage based on t observations as

$$\begin{aligned} C(= C_t) &= \sum_{w \in \mathcal{A}} P_w I[\text{all elements of } w \text{ occur in the sample}] \\ &= \sum_{w \in \mathcal{A}} P_w \Big\{ \prod_{k \in w} I\big[X_{(k)} \geq 1\big] \Big\}. \end{aligned} \tag{2.2}$$

The conditional probability of finding at least one new class in an additional observation given the sample is defined as

$$\bar{C}(= \bar{C}_t) = \sum_{w \in \mathcal{A}} P_w I[\text{at least one of the elements in } w \text{ does not occur}$$

in the sample]

$$= \sum_{w \in \mathcal{A}} P_w \Big\{ \max_{k \in w} I[X_{(k)} = 0] \Big\}. \tag{2.3}$$

We shall drop the use of subscript t in C_t and $\bar{C}_t$ wherever no confusion arises. Both C and $\bar{C}$ vary with the sample and $C + \bar{C} = 1$. In the special case that only one class occurs in each observation, C reduces to the total cell probabilities of the observed classes as defined in Good (1953) and Good and Toulmin (1956).

3. Estimator and Mean Squared Error

To derive an estimator for $E(\bar{C})$, the following definitions are needed: (All the summation of w, w', α and α' in the rest of this paper is over the class $\mathcal{A}$, unless otherwise stated.)

Definition 2. Define

$P_{(w)}$ = probability that all elements of w occur simultaneously

$$= \sum_{\{w' | w' \supseteq w\}} P_{w'};$$

$\bar{P}_{(w)}$ = probability that at least one element of w occurs

$$= \sum_{\{w' | w' \subseteq w\}} P_{(w')}(-1)^{\#w'-1};$$ (equivalently, $1 - \bar{P}_{(w)}$ is the probability that all elements of w do not occur);

$P^*_{(w|w')}$ = probability that all elements of w and at least one element of w' occur.

Definition 3. For a data matrix, we define for $w \in \mathcal{A}$ and $k = 1, 2, \ldots, n$,

$X_{(w)}$ = the number of observations in which all elements of w occur simultaneously

$$= \sum_{k=1}^{t} I[Z_{ki} = 1 \text{ for all } i \in w];$$

f_{k1} = the number of possible combinations of k classes that occur simultaneously and individually exactly once in t observations

$$= \sum_{1 \le i_1 < i_2 < \cdots < i_k \le N} \cdots \sum I\Big[X_{(i_1)} = X_{(i_2)} = \cdots = X_{(i_k)} = 1 \text{ and } X_{(i_1 i_2 \cdots i_k)} = 1\Big]$$

$$= \sum_{\{w | \#w = k\}} I\Big[X_{(w)} = 1 \text{ and } X_{(i)} = 1 \text{ for all } i \in w\Big].$$

Some properties of $\bar{C}$ and f_{k1} are given below:

Proposition 1.

$$\bar{C} = \sum_{w} P_{(w)} I\Big[X_{(i)} = 0 \ \forall \ i \in w\Big](-1)^{\#w-1}, \tag{3.1}$$

$$E(\bar{C}) = \sum_{w} P_{(w)}\Big(1 - \bar{P}_{(w)}\Big)^{t}(-1)^{\#w-1}, \tag{3.2}$$

$$E(f_{k1}) = \sum_{\{w|\#w=k\}} tP_{(w)}\Big(1 - \bar{P}_{(w)}\Big)^{t-1}. \tag{3.3}$$

Proof. From the definition (2.3) and

$$\begin{aligned}\max_{k\in w} I\Big[X_{(k)} = 0\Big] &= \sum_{k\in w} I\Big[X_{(k)} = 0\Big] - \sum\sum_{k,i\in w} I\Big[X_{(k)} = X_{(i)} = 0\Big] + \cdots \\ &\quad + (-1)^{\#w-1}\sum\sum\cdots\sum I\Big[X_{(k_1)} = \cdots = X_{(k_i)} = \cdots = 0\Big],\end{aligned}$$

(3.1) follows by noting that the coefficient of $I[X_{(i)} = 0 \ \forall \ i \in w]$ is $(-1)^{\#w-1}$ $\sum_{w'\supseteq w} P_{w'} = (-1)^{\#w-1}P_{(w)}$. Also, (3.2) and (3.3) follow immediately from (3.1) and the definition of f_{k1} respectively.

We now proceed to derive an estimator. Notice that if t is large,

$$\begin{aligned}E(\bar{C}) &= \sum_{k=1}^{n}\sum_{\{w|\#w=k\}} P_{(w)}\Big(1 - \bar{P}_{(w)}\Big)^{t}(-1)^{k-1} \\ &\approx \sum_{k=1}^{n}\sum_{\{w|\#w=k\}} P_{(w)}\Big(1 - \bar{P}_{(w)}\Big)^{t-1}(-1)^{k-1} \\ &= \sum_{k=1}^{n} E(f_{k1})(-1)^{k-1}/t.\end{aligned} \tag{3.4}$$

Thus an estimator of $E(\bar{C})$ is

$$\hat{\bar{C}}(=\hat{\bar{C}}_t) = \sum_{k=1}^{n} f_{k1}(-1)^{k-1}/t, \tag{3.5}$$

which is also used as a predictor of $\bar{C}$. If only one class is observed for each observation, $f_{k1} = 0$ for $k > 2$ and the estimator $\hat{\bar{C}}$ reduces to the proportion of singletons, which is Turing's formula given in Good (1953).

In the following, we obtain an alternative form for $\hat{\bar{C}}$:

Proposition 2. *Let Y_j $(= Y_j^t)$ be the number of singletons in the jth observation based on t observations, then*

$$\hat{\bar{C}}(= \hat{\bar{C}}_t) = \sum_{j=1}^{t} I(Y_j \geq 1)/t. \tag{3.6}$$

Proof. This follows from (3.5) and $f_{k1} = \sum_{j=1}^{t} \binom{Y_j}{k}$.

The proposed estimator given in (3.6) is simply the sample proportion of observations that contain singletons. This intuitively implies that only those observations that contain singletons are informative of the probability of discovering another singleton in an additional observation.

Let f_{k1}^{t+1} and Y_j^{t+1} be defined similarly to f_{k1} and Y_j except that they are based on $t+1$ observation if an additional observation were taken. It is then clear from (3.4) that the estimator

$$\hat{\bar{C}}_{t+1} = \sum_{k=1}^{n} \frac{f_{k1}^{t+1}(-1)^{k-1}}{t+1} = \sum_{j=1}^{t+1} \frac{I(Y_j^{t+1} \geq 1)}{t+1}$$

is an unbiased estimator of $E(\bar{C}_t)$.

We now extend the approach of Good (1953) to prove that $\hat{\bar{C}}$ is an approximate empirical Bayes estimators with respect to a uniform prior.

Proposition 3. *Let $\mathcal{E}$ be the event that there is at least one class which did not occur in the sample. The posterior mean of $\bar{C}$ given $\mathcal{E}$ under a uniform prior is then*

$$E(\bar{C}|\mathcal{E}) = \frac{\sum_{w} P_{(w)}\left[1 - \bar{P}_{(w)}\right]^t (-1)^{\#w-1}}{\sum_{w} \left[1 - \bar{P}_{(w)}\right]^t (-1)^{\#w-1}}. \tag{3.7}$$

Proof. For any $w \in \mathcal{A}$, it can be shown that

$$\begin{aligned} P\left(\mathcal{E}|\bar{C} = \bar{P}_{(w)}\right) &= P(\text{the set of unobserved classes is } w) \\ &= \sum_{w \subseteq w'} \left[1 - \bar{P}_{(w')}\right]^t (-1)^{\#w' - \#w}. \end{aligned}$$

Then for any $w \in \mathcal{A}$ the posterior probability that $\bar{C} = \bar{P}_{(w)}$ given $\mathcal{E}$ is

$$P(\bar{C} = \bar{P}_{(w)}|\mathcal{E}) = \frac{P\left(\mathcal{E}|\bar{C} = \bar{P}_{(w)}\right) P\left(\bar{C} = \bar{P}_{(w)}\right)}{\sum_{\alpha \in \mathcal{A}} P\left(\mathcal{E}|\bar{C} = \bar{P}_{(\alpha)}\right) P\left(\bar{C} = \bar{P}_{(\alpha)}\right)}.$$

Under a uniform prior that $P(\bar{C} = \bar{P}_{(w)})$ are equal for all $w \in \mathcal{A}$, we have

$$\begin{aligned}\sum_{\alpha}\sum_{\alpha\subseteq\alpha'}\left[1-\bar{P}_{(\alpha')}\right]^t(-1)^{\#\alpha'-\#\alpha} &= \sum_{\alpha'}\left[1-\bar{P}_{(\alpha')}\right]^t(-1)^{\#\alpha'}\left[\sum_{\alpha\subseteq\alpha'}(-1)^{\#\alpha}\right] \\ &= \sum_{\alpha'}\left[1-\bar{P}_{(\alpha')}\right]^t(-1)^{\#\alpha'-1}.\end{aligned}$$

It then follows that the posterior mean of $\bar{C}$ is

$$E(\bar{C}|\mathcal{E}) = \frac{\sum_{w}\bar{P}_{(w)}\sum_{w\subseteq w'}\left[1-\bar{P}_{(w')}\right]^t(-1)^{\#w'-\#w}}{\sum_{w}\left(1-\bar{P}_{(w)}\right)^t(-1)^{\#w-1}}. \tag{3.8}$$

Since it is easy to prove that $\sum_{\{w|w\subseteq w'\}}\bar{P}_{(w)}(-1)^{\#w-1} = P_{(w')}$, we have

$$\begin{aligned}&\sum_{w}\bar{P}_{(w)}\sum_{w\subseteq w'}\left[1-\bar{P}_{(w')}\right]^t(-1)^{\#w'-\#w} \\ =\ &\sum_{w'}\left[\sum_{w\subseteq w'}\bar{P}_{(w)}(-1)^{\#w-1}\right]\left[1-\bar{P}_{(w')}\right]^t(-1)^{\#w'-1} \\ =\ &\sum_{w'}P_{(w')}\left[1-\bar{P}_{(w')}\right]^t(-1)^{\#w'-1}.\end{aligned}$$

From Proposition 3, we can write

$$E(\bar{C}|\mathcal{E}) = \frac{E\left[\sum_i I\left(Y_i^{t+1}\geq 1\right)\right]/(t+1)}{E[I(\mathcal{E})]}$$

where Y_i^{t+1} is the number of singletons in the ith observation based on $t+1$ observations. If we were to take an additional observation, then $\sum_i I(Y_i^{t+1} \geq 1)/(t+1)$ is an empirical Bayes estimator. When t is large, our proposed estimator $\sum_i I(Y_i \geq 1)/t$ is thus an appoximate empirical Bayes estimator.

The mean squared error of the proposed estimator is then derived as follows:

Proposition 4. *The mean squared error of $\hat{\bar{C}}$ is given by*

$$\begin{aligned}&E(\hat{\bar{C}}-\bar{C})^2 \\ =\ &E(\hat{\bar{C}})/t - B/t\end{aligned}$$

$$+ \sum_{w\cap w'\neq\phi}\sum P_{(w)}P_{(w')}\Big[1-\bar{P}_{(w\cup w')}\Big]^t(-1)^{\#w+\#w'}$$

$$+ \sum_{w\cap w'=\phi}\sum P^*_{(w|w')}P^*_{(w'|w)}\Big[1-\bar{P}_{(w\cup w')}\Big]^{t-2}(-1)^{\#w+\#w'}$$

$$+ \sum_{w\cap w'=\phi}\sum P_{(w)}P_{(w')}\bar{P}^2_{(w\cup w')}\Big[1-\bar{P}_{(w\cup w')}\Big]^{t-2}(-1)^{\#w+\#w'}$$

$$- 2\sum_{w\cap w'=\phi}\sum P_{(w)}P^*_{(w'|w)}\bar{P}_{(w\cup w')}\Big[1-\bar{P}_{(w\cup w')}\Big]^{t-2}(-1)^{\#w+\#w'} \quad (3.9)$$

where

$$B=\sum_{w\cap w'=\phi}\sum\Big[P_{(w)}-P^*_{(w|w')}\Big]\Big[P_{(w')}-P^*_{(w'|w)}\Big]\Big[1-\bar{P}_{(w\cup w')}\Big]^{t-2}(-1)^{\#w+\#w'}. \quad (3.10)$$

Proof. From (3.6), we can show that

$$\begin{aligned}
E(\hat{\bar{C}}^2) &= t^{-2}E\Big\{\sum I(Y_i\geq 1)+\sum_{i\neq j}\sum I(Y_i\geq 1)I(Y_j\geq 1)\Big\}\\
&= \frac{E(\hat{\bar{C}})}{t}+t^{-2}E\sum_{i\neq j}\sum\left[\sum_{k=1}^{Y_i}\binom{Y_i}{k}(-1)^{k-1}\right]\left[\sum_{m=1}^{Y_j}\binom{Y_j}{m}(-1)^{m-1}\right]\\
&= \frac{E(\hat{\bar{C}})}{t}+t^{-2}\sum_{w\cap w'=\phi}\sum E\Big\{I\Big[X_{(i)}=1\ \forall\, i\in w\cup w', X_{(w)}=1, X_{(w')}=1,\\
&\qquad\qquad X_{(w\cup w')}=0\Big]\Big\}(-1)^{\#w+\#w'}\\
&= \frac{E(\hat{\bar{C}})}{t}+(1-\frac{1}{t})B. \qquad (3.11)
\end{aligned}$$

Also,

$$\begin{aligned}
E(\bar{C}^2) &= \sum_{w\cap w'\neq\phi}\sum P_{(w)}P_{(w')}\Big[1-\bar{P}_{(w\cup w')}\Big]^t(-1)^{\#w+\#w'}\\
&\quad+\sum_{w\cap w'=\phi}\sum P_{(w)}P_{(w')}\Big[1-\bar{P}_{(w\cup w')}\Big]^t(-1)^{\#w+\#w'}, \qquad (3.12)
\end{aligned}$$

and the cross product term is

$$-2E(\bar{C}\hat{\bar{C}})=-2\sum_{w\cap w'=\phi}\sum P_{(w)}\Big[P_{(w')}-P^*_{(w'|w)}\Big]\Big[1-\bar{P}_{(w\cup w')}\Big]^{t-1}(-1)^{\#w+\#w'}. \quad (3.13)$$

Combining the above three terms, we get (3.9).

Proposition 5. *The mean square error for any fixed n when t is large is given by*

$$E(\hat{\bar{C}} - \bar{C})^2 = \frac{E(\hat{\bar{C}})}{t} - \frac{B}{t} + \sum_{w \cap w' \neq \phi} \sum P_{(w)} P_{(w')} \left[1 - \bar{P}_{(w \cup w')}\right]^{t-2} (-1)^{\#w + \#w'}$$
$$+ \sum_{w \cap w' = \phi} \sum P^*_{(w|w')} P^*_{(w'|w)} \left[1 - \bar{P}_{(w \cup w')}\right]^{t-2} (-1)^{\#w + \#w'} + O(t^{-2}), \quad (3.14)$$

where B is defined in (3.10). *If we define M_{ij} for $i \neq j$ as the number of classes which appear only in the* ith *and* jth *observations (and not in others), an asymptotically unbiased estimator for the mean squared error of $\hat{\bar{C}}$ for large t and fixed n is then*

$$\frac{\hat{\bar{C}}(1 - \hat{\bar{C}})}{t - 1} + \sum_{i<j} \sum I\left[M_{ij} \geq 1, Y_i = 0, Y_j = 0\right] \Big/ \binom{t}{2}. \quad (3.15)$$

Proof. Since there are at most n classes observed for each observation, we can show that $\sum_w P_{(w)} \leq 2^n$. It then follows that

$$\sum_w \sum_{w'} P_{(w)} P_{(w')} \bar{P}^2_{(w \cup w')} \left[1 - \bar{P}_{(w \cup w')}\right]^{t-2} (-1)^{\#w + \#w'} = O(t^{-2}),$$

and

$$\sum_{w \cap w' \neq \phi} \sum P_{(w)} P_{(w')} \bar{P}_{(w \cup w')} \left[1 - \bar{P}_{(w \cup w')}\right]^{t-2} (-1)^{\#w + \#w'}$$
$$+ \sum_{w \cap w' = \phi} \sum P_{(w)} P^*_{(w'|w)} \bar{P}_{(w \cup w')} \left[1 - \bar{P}_{(w \cup w')}\right]^{t-2} (-1)^{\#w + \#w'} = O(t^{-2}),$$

(3.14) then follows. We now derive the estimator of the mean square error. From (3.11), an estimator for B is $t[\hat{\bar{C}}(\hat{\bar{C}} - 1/t)]/(t-1)$. Moreover, an unbiased estimator for $\sum_{w \cap w' \neq \phi} \sum P_{(w)} P_{(w')} [1 - \bar{P}_{(w \cup w')}]^{t-2} (-1)^{\#w + \#w'}$ is

$$\binom{t}{2}^{-1} \sum_{i<j} \sum \sum_{w \cap w' \neq \phi} \sum I[H(i, j, w, w')] (-1)^{\#w + \#w'},$$

where $H(i, j, w, w')$ denotes the event that all elements of w occur in the ith observation, all elements of w' occur in the jth observation, and none of the

elements of $w \cup w'$ occur in other observations. For given observations i and j, we have to enumerate how many combinations of w and w' satisfy the above specified condition. First fix the number of elements in $w \cap w'$, ℓ. There are, accordingly, $\binom{M_{ij}}{\ell}$ possibilities. Then there are $\binom{Y_i+M_{ij}-\ell}{k}$ choices for w and $\binom{Y_j+M_{ij}-\ell}{m}$ choices for w'. We finally have an alternate sum with respect to ℓ to remove all redundant pairs. The above estimator becomes

$$\binom{t}{2}^{-1} \sum\sum_{i<j} \sum_{\ell=1}^{M_{ij}} \left\{ \left[\sum_{k=0}^{Y_i+M_{ij}-\ell} \binom{Y_i + M_{ij} - \ell}{k} (-1)^k \right] \right.$$
$$\left. \times \left[\sum_{m=0}^{Y_j+M_{ij}-\ell} \binom{Y_j + M_{ij} - \ell}{m} (-1)^m \right] \right\} \binom{M_{ij}}{\ell} (-1)^{\ell-1}.$$

However, for $\ell < M_{ij}$ we have

$$\sum_{k=0}^{Y_i+M_{ij}-\ell} \binom{Y_i + M_{ij} - \ell}{k} (-1)^k = 0 = \sum_{m=0}^{Y_j+M_{ij}-\ell} \binom{Y_j + M_{ij} - \ell}{m} (-1)^m.$$

Thus only the terms for $\ell = M_{ij}$ are left and the above estimator then becomes

$$\binom{t}{2}^{-1} \sum\sum_{i<j} (-1)^{M_{ij}-1} [1 - I(Y_i > 0)][1 - I(Y_j > 0)]$$
$$= \binom{t}{2}^{-1} \sum\sum_{i<j} (-1)^{M_{ij}-1} I[M_{ij} \geq 1, Y_i = 0, Y_j = 0].$$

Similarly, define $G(i, j, w, w')$ as the event that all the elements of w and at least one element of w' occur in the ith observation, all elements of w' and at least one of w occur in the jth observation, and none of the elements of $w \cup w'$ occur in others. An unbiased estimator for $\sum_{w \cap w' = \phi} \sum P^*_{(w|w')} P^*_{(w'|w)} \left[1 - \bar{P}_{(w \cup w')}\right]^{t-2} (-1)^{\#w + \#w'}$ is

$$\binom{t}{2}^{-1} \sum\sum_{i<j} \sum_{w \cap w' = \phi} \sum I[G(i, j, w, w')] (-1)^{\#w + \#w'}$$
$$= \binom{t}{2}^{-1} \sum\sum_{i<j} \left((-1)^{M_{ij}} + 1 \right) I[M_{ij} \geq 2, Y_i = 0, Y_j = 0].$$

Combining the above three estimators, we then obtain (3.15).

For the special case that exactly one class can occur for each observation, the only possible value for M_{ij} is 1. It is obvious that $\sum\sum_{i<j} I[M_{ij} = 1, Y_i = 0, Y_j =$

$0] =$ number of doubletons; hence our variance estimator reduces to Esty's (1983) result.

The following proposition proves the consistency of the proposed estimator:

Proposition 6. *When* $N \to \infty$ *and* $t = t(N) \to \infty$ *such that* $t^{-2}E\sum\sum_{i<j} I\Big[M_{ij} \geq 1, Y_i = 0, Y_j = 0\Big] = o(1)$ *for any fixed* n, *then* $\hat{\bar{C}} - \bar{C} \xrightarrow{P} 0$.

Proof. It is easy to see that $B = O(1)$ and the conclusion follows directly from (3.14) and (3.15).

The condition in the above proposition can be easily shown to be valid in the special case that only one class is observed in each observation. We remark that a bias-corrected version of $\hat{\bar{C}}$ can be easily obtained by noting that

$$E(\hat{\bar{C}} - \bar{C}) = \sum_{w} \sum_{w' \subseteq w} P_{(w)} P_{(w')} \Big[1 - \bar{P}_{(w)}\Big]^{t-1} (-1)^{\#w + \#w'}.$$

An estimator for the bias is

$$\begin{aligned}
& \frac{1}{t(t-1)} \sum\sum_{i<j} \sum_{w} \sum_{w' \subseteq w} \Big\{ I[H(i,j,w,w')] + I[H(j,i,w,w')] \Big\} (-1)^{\#w + \#w'} \\
= \; & \frac{1}{t(t-1)} \sum\sum_{i<j} \sum_{\ell=1}^{M_{ij}} \left\{ \left[\sum_{k=0}^{Y_i + M_{ij} - \ell} \binom{Y_i + M_{ij} - \ell}{k} (-1)^k \right] \right. \\
& \left. + \left[\sum_{m=0}^{Y_j + M_{ij} - \ell} \binom{Y_j + M_{ij} - \ell}{m} (-1)^m \right] \right\} \binom{M_{ij}}{\ell} \\
= \; & \frac{1}{t(t-1)} \sum\sum_{i \neq j} \Big[I(Y_i = 0 = Y_j, M_{ij} \geq 1) + I(Y_i \geq 1, Y_j = 0, M_{ij} \geq 1) \Big].
\end{aligned}$$

Hence we obtain the bias-corrected estimator:

$$\hat{\bar{C}}_{bc} = \hat{\bar{C}} - \frac{1}{t(t-1)} \sum\sum_{i \neq j} \Big[I(Y_i = 0 = Y_j, M_{ij} \geq 1) + I(Y_i \geq 1, Y_j = 0, M_{ij} \geq 1) \Big]. \tag{3.16}$$

It is reduced to $\hat{\bar{C}}-$ (number of doubletons)$/\binom{t}{2}$ if only one class is observed for each observation.

Two intuitive explanations regarding the bias-corrected term are provided in the following: First, an approach similar to Lo (1992) is presented: Let $N_1^{t+1} = \sum_i I(Y_i^{t+1} \geq 1)$ be the number of samples with singletons based on $t+1$ samples. As mentioned before, $N_1^{t+1}/(t+1)$ is an unbiased estimator of $E(\bar{C}_t)$ and now we use N_1^t/t instead, where $N_1^t = \sum_i I(Y_i \geq 1)$. The purpose here is to evaluate the difference of $N_1^{t+1}/(t+1)$ and N_1^t/t. However, N_1^{t+1} is unobservable, so

we construct an "estimator" $\hat{N}_1^{t+1}$ as follows: the relation between N_1^{t+1} and N_1^t should be approximately the same as that of N_1^t and N_1^{t-1}, which can be obtained by successively deleting a single observation. If the ith observation is deleted, the following four cases may arise:

(1) if the ith observation contains singleton but no doubleton, we have

$$\hat{N}_1^{t+1} = N_1^t + 1;$$

(2) if the ith observation contains both a singleton and doubleton, we have

$$\hat{N}_1^{t+1} = N_1^t - \sum_{j \neq i} I\Big(Y_i \geq 1, Y_j = 0, M_{ij} \geq 1\Big) + 1;$$

(3) if the ith observation contains a doubleton but no singleton, we have

$$\hat{N}_1^{t+1} = N_1^t - \sum_{j \neq i} I(Y_i = 0, Y_j = 0, M_{ij} \geq 1);$$

(4) if the ith observation contains neither a singleton nor a doubleton, we have

$$\hat{N}_1^{t+1} = N_1^t.$$

Then an approximate bias is

$$\begin{aligned} & \frac{E\Big(\hat{N}_1^{t+1} | Y_1, Y_2, \ldots, Y_t, M_{ij},\ i, j = 1, 2, \ldots, t,\ i \neq j\Big)}{t+1} - \frac{N_1^t}{t} \\ = & \frac{-1}{(t+1)t} \sum\sum_{i \neq j} \Big[I(Y_i = 0 = Y_j, M_{ij} \geq 1) + I(Y_i \geq 1, Y_j = 0, M_{ij} \geq 1)\Big]. \end{aligned}$$

The second explanation is to consider a U-statistics for $E(\bar{C}_{t-2})$: Write

$$\begin{aligned} E(\hat{\bar{C}}_t - \bar{C}_t) &= E(\bar{C}_{t-1} - \bar{C}_t) \approx E(\bar{C}_{t-2} - \bar{C}_{t-1}) \\ &= E\Big[\frac{1}{t}\sum_{i=1}^{t} \frac{1}{t-1} \sum_{j \neq i} I\Big(Y_j^{(-i)} \geq 1\Big) - \frac{1}{t}\sum_{j=1}^{t} I(Y_j \geq 1)\Big], \quad (3.17) \end{aligned}$$

where $Y_j^{(-i)}$ is the number of singletons in the jth observation when the ith observation is deleted, $i \neq j$. It can be shown that

$$I\Big(Y_j^{(-i)} \geq 1\Big) = I(Y_j \geq 1) + I(Y_i = 0 = Y_j, M_{ij} \geq 1) + I(Y_i \geq 1, Y_j = 0, M_{ij} \geq 1).$$

Carrying out some algebra results in

$$\begin{aligned} E(\hat{\bar{C}}_t - \bar{C}_t) \approx\ & E\Big\{\frac{1}{t(t-1)} \sum\sum_{i \neq j} \Big[I(Y_i = 0 = Y_j, M_{ij} \geq 1) \\ & + I(Y_i \geq 1, Y_j = 0, M_{ij} \geq 1)\Big]\Big\}. \end{aligned}$$

This also provides an intuitive view of the bias-corrected term.

4. A Simulation Study

In this section, we compare our estimator with that of the usual approach without considering the possible dependence among characters occurring within a poem. If each character is treated as an independent observation, then an estimate for the probability of discovering a new character for an additional character is $f_{11}/(nt)$, the proportion of singletons. Therefore an estimate for the probability that there exists at least one new character for an additional poem with n characters is $\tilde{\bar{C}} = 1 - [1 - f_{11}/(nt)]^n$. We are then interested in the performance of $\tilde{\bar{C}}$ under the dependence situations.

We have carried out a limited simulation study to investigate the relative performance of $\hat{\bar{C}}$, bias-corrected $\hat{\bar{C}}_{bc}$, and $\tilde{\bar{C}}$. The number of classes N was fixed to be 200. Let $R_1, R_2, \ldots, R_{200}$ denote these 200 classes. For each observation, four classes ($n = 4$) were chosen in a way that two classes were randomly selected without replacement from R_{101},R_{102},...,R_{200} and the other two classes were a pair randomly selected from the 50 pairs (R_1, R_2),(R_3, R_4),...,(R_{99}, R_{100}). In other words, for $i = 1, 3, \ldots, 99$, both classes R_i and R_{i+1} always occurred together (if they occurred) in any observation. For each value of t (30 to 100 with an increment 10), 500 data sets were simulated. The average values of $\bar{C}$, $\hat{\bar{C}}$, $\hat{\bar{C}}_{bc}$ and $\tilde{\bar{C}}$ are shown in Table 1. Based on these 500 estimates, sample standard error as well as sample root mean squared error (RMSE) were also obtained. The average values of the estimated RMSE based on (3.15) for the proposed estimator are also given to check the adequacy of that formula.

Table 1 shows that the conventional estimator $\tilde{\bar{C}}$ ignoring the possible dependence severely overestimates in all cases. This implies that the usual sample coverage estimator has a severe negative bias. The proposed $\hat{\bar{C}}$ slightly overestimates and its bias-corrected form $\hat{\bar{C}}_{bc}$ is nearly unbiased. Both $\hat{\bar{C}}$ and $\hat{\bar{C}}_{bc}$ have comparable RMSE and generally perform better than the usual estimator $\tilde{\bar{C}}$ with respect to both bias and RMSE criteria.

The last column provides the averages of estimated RMSE based on (3.15). It seems that the theoretic formula is generally satisfactory compared with the sample RMSE if t is sufficiently large.

5. Application

Two hundred seven-character quartet ($t = 200$) were randomly selected from the collected work of Bai Juyi. Since each poem consists of 28 characters ($n = 28$), there is a total of 5600 characters in the data. The number of distinct characters is 1270 and the number of singletons is 539. Out of these 200 poems, there

are 181 poems with singletons and 19 without singletons. Thus the generalized sample coverage for this data is estimated to be 9.5% (19/200) and the conditional probability of finding at least one new character in the next quartet is 90.5%. If we treat each character as an independent observation and ignore the possible dependence within a quartet, then the probability of finding at least one new character is $1 - (1 - 539/5600)^{28} = 94.1\%$, and the sample coverage estimate is 5.9%. Based on the simulation given in the last section, we suspect that the probability 94.1% is likely to overestimate the true value. Out of the 19 poems without singletons, only one character "drinking a little" appears in both 17th and 24th poems and not in others. Therefore the estimated mean squared error for our estimate using (3.15) is $90.5\%(1 - 90.5\%)/199 + 1/\binom{200}{2} = (2.2\%)^2$. Thus if the asymptotic normality is valid, an approximate 95% confidence interval for the coverage of these 200 poems is 5.2% to 13.8% and 86.2% to 94.8% for the conditional probability of discovering at least a new character in the next additional quartet.

Acknowledgements

The authors are grateful to two referees for their insightful suggestions and comments, to Professor F.-Y. Shih for providing all the backgrounds of Chinese poems, to Mr. Y.-J. Yang for writing a program combining Chinese ET system and C Language to obtain all the frequencies needed in Section 5. This work is based on part of the Ph. D. Thesis of the first author under the supervision of the second author.

Table 1. Simulation results for comparing various estimators

$\bar{C}$: defined in (2.3);
$\hat{\bar{C}}$: defined in (3.5) and (3.6);
$\hat{\bar{C}}_{bc}$: bias-corrected estimator defined in (3.16);
$\tilde{\bar{C}}$: usual estimator without considering dependence.

t		average value	sample s.e.	sample RMSE	average of estimated RMSE
30	$\bar{C}$	.907	.013		
	$\hat{\bar{C}}$	.911	.057	.066	.057
	$\hat{\bar{C}}_{bc}$	.904	.061	.070	
	$\tilde{\bar{C}}$	.956	.026	.062	
40	$\bar{C}$	.831	.021		
	$\hat{\bar{C}}$	.837	.063	.080	.074
	$\hat{\bar{C}}_{bc}$	.828	.067	.083	
	$\tilde{\bar{C}}$	.905	.035	.092	
50	$\bar{C}$	.743	.031		
	$\hat{\bar{C}}$	.752	.064	.089	.084
	$\hat{\bar{C}}_{bc}$	.743	.067	.091	
	$\tilde{\bar{C}}$	.842	.044	.122	
60	$\bar{C}$	.655	.039		
	$\hat{\bar{C}}$	.661	.061	.094	.089
	$\hat{\bar{C}}_{bc}$	.652	.063	.095	
	$\tilde{\bar{C}}$	.761	.049	.135	
70	$\bar{C}$	.566	.045		
	$\hat{\bar{C}}$	.574	.054	.091	.091
	$\hat{\bar{C}}_{bc}$	.565	.055	.091	
	$\tilde{\bar{C}}$	.678	.051	.144	
80	$\bar{C}$	.487	.050		
	$\hat{\bar{C}}$	.493	.051	.093	.090
	$\hat{\bar{C}}_{bc}$	.485	.052	.093	
	$\tilde{\bar{C}}$	.594	.050	.141	
90	$\bar{C}$	.412	.052		
	$\hat{\bar{C}}$	.420	.051	.094	.087
	$\hat{\bar{C}}_{bc}$	.412	.052	.094	
	$\tilde{\bar{C}}$	.514	.052	.140	
100	$\bar{C}$	.345	.052		
	$\hat{\bar{C}}$	.354	.044	.086	.083
	$\hat{\bar{C}}_{bc}$	.348	.045	.086	
	$\tilde{\bar{C}}$	.441	.048	.131	

References

Chao, A. (1981). On estimating the probability of discovering a new species. *Ann. Statist.* **9**, 1339–1342. Correction **10**, 1331.

Chao, A. and Lee, S.-M. (1990). Estimating the number of unseen species with frequency counts. *Chinese J. Math.* **18**, 335–351.

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

Cohen A. and Sackrowitz, H. B. (1990). Admissibility of estimators of the probability of unobserved outcomes. *Ann. Inst. Statist. Math.* **42**, 623–636.

Darroch, J. N. and Ratcliff, D. (1980). A note on capture-recapture estimation. *Biometrics* **36**, 149–153.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

Engen, S. (1978). *Stochastic Abundance Models.* Chapman and Hall, London.

Esty, W. W. (1982). Confidence intervals for the coverage of low coverage samples. *Ann. Statist.* **10**, 190–196.

Esty, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *Ann. Statist.* **11**, 905–912.

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

Esty, W. W. (1986). The efficiency of Good's nonparametric coverage estimator. *Ann. Statist.* **14**, 1257–1260.

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

Good, I. J. and Toulmin, G. H. (1956). The number of new species and the increase of population coverage when a sample is increased. *Biometrika* **43**, 45–63.

Harris, B. (1959). Determining bounds on integrals with applications to cataloging problems. *Ann. Math. Statist.* **30**, 521–548.

Knott, M. (1967). Models for cataloging problems. *Ann. Math. Statist.* **38**, 1255–1260.

Lo, S.-H. (1992). From species problem to a general coverage problem via a new interpretation. *Ann. Statist.* **20**, 1094–1109.

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

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

Starr, N. (1979). Linear estimation of the probability of discovering a new species. *Ann. Statist.* **7**, 644–652.

Thisted, R., and Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74**, 445–455.

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

(Received July 1991; accepted June 1992)