



## Estimating the Number of Classes via Sample Coverage

Anne Chao, Shen-Ming Lee

*Journal of the American Statistical Association*, Volume 87, Issue 417 (Mar., 1992),  
210-217.

---

Your use of the JSTOR database indicates your acceptance of JSTOR's Terms and Conditions of Use. A copy of JSTOR's Terms and Conditions of Use is available at <http://www.jstor.org/about/terms.html>, by contacting JSTOR at [jstor-info@umich.edu](mailto:jstor-info@umich.edu), or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or otherwise used, in any form or by any means, except: (1) one stored electronic and one paper copy of any article solely for your personal, non-commercial use, or (2) with prior written permission of JSTOR and the publisher of the article or other text.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

*Journal of the American Statistical Association* is published by American Statistical Association. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/asa.html>.

---

*Journal of the American Statistical Association*  
©1992 American Statistical Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact [jstor-info@umich.edu](mailto:jstor-info@umich.edu).

©2001 JSTOR

# Estimating the Number of Classes via Sample Coverage

ANNE CHAO and SHEN-MING LEE\*

Assume that a random sample is drawn from a population with unknown number of classes and possibly unequal class probabilities. A nonparametric estimation technique is proposed to estimate the number of classes using the idea of sample coverage, which is defined as the sum of the cell probabilities of the observed classes. Since expected sample coverage can be well estimated, we were motivated to find its role in the estimation of the number of classes. This work generalizes the result of Esty to a nonparametric approach and extends Darroch and Ratcliff to incorporate the heterogeneity of the class probabilities. The coefficient of variation of the class sizes is shown to play an important role in the recommended estimation procedures. The performance of the proposed estimators is investigated by means of Monte Carlo simulations.

KEY WORDS: Heterogeneity; Multinomial; Number of species.

## 1. INTRODUCTION

A random sample of size  $n$  is taken from a population of elements belonging to  $N$  different classes. We assume that the classes are indexed by  $1, 2, \dots, N$ . For practical examples, see Goodman (1949), Craig (1953), Efron and Thisted (1976), Holst (1981), Esty (1982, 1985, 1986a), and Thisted and Efron (1987). Let  $p_i$  be the probability that any observation belongs to the  $i$ th class and  $X_i$  be the number of elements of the  $i$ th class observed in the sample ( $i = 1, 2, \dots, N$ ); then  $(X_1, X_2, \dots, X_N)$  is multinomially distributed. This is the most common model for species problems in biological applications. Let  $f_i$  be the number of classes that have exactly  $i$  elements in the sample, that is  $f_i = \sum_{j=1}^N I[X_j = i]$  ( $i = 1, 2, \dots, n$ ), where  $I[A]$  is the usual indicator function. Let  $D = \sum f_i$  be the number of distinct classes observed in the sample and sample size  $n = \sum if_i$ . The goal is to estimate  $N$  based on  $f_i$ .

If  $p_1 = p_2 = \dots = p_N = 1/N$  (the equally likely or equiprobable assumption), the problem reduces to an inference problem involving only one parameter  $N$ . Traditional estimation procedures have been discussed by many authors, including Lewontin and Prout (1956), Darroch (1958), Harris (1968), Johnson and Kotz (1977, pp. 136–139), Marchand and Schroeck (1982), and Holst (1981). The approximate maximum likelihood estimator (MLE) and minimum variance unbiased estimator (when the latter exists),  $\hat{N}_0$ , is the solution of the following equation:

$$D = N[1 - \exp(-n/N)], \quad (1.1)$$

with asymptotic variance

$$\text{var}(\hat{N}_0) = N/[e^{n/N} - (n/N) - 1]. \quad (1.2)$$

See Darroch (1958) and Harris (1968) for derivations.

In most practical applications, the equally likely assumption is not valid. Most previous authors adopted a parametric approach to handle heterogeneous populations

(i.e., unequal class probabilities). For example, Fisher, Corbet, and Williams (1943) assumed that for each species the number of elements observed in the sample follows a Poisson distribution and the Poisson parameter is assumed to have a gamma-type distribution. Many other papers on stochastic abundance models also make parametric assumptions; see, for example, Engen (1978) for a review.

The sample coverage of a random sample from a multinomial population is defined to be the sum of the probabilities of the observed classes. For an equiprobable population, the estimator [Eq. (2.3) in Sec. 2] proposed by Darroch and Ratcliff (1980) exactly used the idea of sample coverage. For heterogeneous populations, Esty (1985) was the first to apply the concept of sample coverage to estimate the number of classes in a parametric setup. The classes discussed by Esty are the different dies in minting. He assumed that the number of coins that each die produced follows a negative binomial distribution and obtained an estimator of the number of dies in terms of the sample coverage and the parameter of the negative binomial distribution.

Although the negative binomial is a useful model in numismatics, it may not be suitable in other fields. In Section 2 we propose a nonparametric estimation technique using the idea of sample coverage. The result of Esty (1985) then becomes a special case of the proposed method. In addition, our method extends Darroch and Ratcliff's estimator to handle heterogeneity of the class probabilities. Section 3 illustrates our method with an example and compares it with other estimates. Results of a simulation study are reported in Section 4 to show the general performance of our method.

## 2. SAMPLE COVERAGE AND PROPOSED ESTIMATORS

The sample coverage,  $C$ , is defined as the sum of the probabilities of the observed classes, that is,

$$C = \sum_{i=1}^N p_i I[X_i > 0]. \quad (2.1)$$

Note that  $C$  varies with the sample and is a random variable. A widely used "estimator" of  $C$  is

\* Anne Chao is Professor, Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043. Shen-Ming Lee is Associate Professor, Department of Statistics, Feng-Chia University, Tai-Chung, Taiwan 40724. This research was supported by the National Science Council of Taiwan under contract NSC 78-0208-M007-32. The authors thank the associate editor and the referee for carefully reading an early version of this article and for providing insightful comments and suggestions.

$$\hat{C} = 1 - f_1/n, \quad (2.2)$$

which was originally proposed by Turing (Good 1953) and has been discussed in Good (1953), Good and Toulmin (1956), Robbins (1968), Engen (1978), Esty (1982, 1983, 1986b). A variance estimator of  $\hat{C}$  and the construction of confidence intervals were given in Esty (1983).

If we estimate the number of classes without estimating the variation among the class probabilities, we usually can derive only a lower bound (Chao 1984, 1987), with the lower bound being achieved in the equiprobable case. Since the sample coverage can be well estimated in heterogeneous populations, we attempt to estimate the number of classes via a sample coverage approach that would take account of the variation among the class probabilities. This is our basic motivation.

In the equally likely case,  $C$  reduces to  $D/N$ . Hence a natural estimator of  $N$  in this case is

$$\hat{N}_1 = D/\hat{C}. \quad (2.3)$$

This estimator was first introduced by Darroch and Ratcliff (1980). They found that  $\hat{N}_1$  is asymptotically remarkably efficient in the equally likely case when compared to the usual MLE. Esty (1986a) also found that this estimator is superior to other estimators used in numismatics.

Suppose that the class sizes are iid with the following negative binomial distribution with parameters  $k$  ( $k > 0$ ) and  $r$  ( $0 < r < 1$ ),

$$P(\text{class size} = y) = \binom{y-1}{k-1} r^k (1-r)^{y-k}, \quad y = k, k+1, \dots \quad (2.4)$$

Esty (1985) obtained the following estimator:

$$\frac{D}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}} \frac{1}{k}. \quad (2.5)$$

Instead of estimating  $k$  from the sample, Esty (1985) suggested using  $k = 2$  because of its wide use in numismatics. The negative binomial distribution used in (2.4) has the alternative form  $\binom{k+y-1}{y} r^k (1-r)^y$  ( $y = 0, 1, \dots$ ). But it is easily seen that both forms lead to the same estimator following Esty's approach. The reason we adopt (2.4) is to avoid the case with class size being 0. Our following arguments basically treat  $(p_1, p_2, \dots, p_{N-1})$  as fixed parameters defined in an  $(N-1)$ -dimensional parameter space. We now state our main result, and the proof is given in the Appendix.

*Proposition 1.* Assume that a random sample of size  $n$  is drawn from a population of  $N$  cells with fixed cell probabilities  $\mathbf{p} = (p_1, p_2, \dots, p_N)$ ,  $\sum p_i = 1$ . Let  $p_1, p_2, \dots, p_N$  have mean  $\bar{p} = \sum_i p_i / N = 1/N$  and coefficient of variation (CV)  $\gamma = [\sum_i (p_i - \bar{p})^2 / N]^{1/2} / \bar{p}$ . ( $\gamma = 0$  is equivalent to that all  $p_i$ 's are equal.) Then we have

$$E(D)/E(C) = N - \frac{n(1-\bar{p})^{n-1}}{E(C)} \gamma^2 + R, \quad (2.6)$$

where

$$R = \frac{1}{E(C)} \sum_1^N \binom{n}{2} \bar{p}^2 (1-p_i^*)^{n-2} \left[ \frac{p_i - \bar{p}}{\bar{p}} \right]^3 - \frac{\theta}{E(C)} \sum_1^N \binom{n}{3} \bar{p}^3 (1-p_i^*)^{n-3} \left[ \frac{p_i - \bar{p}}{\bar{p}} \right]^4 \quad (2.7)$$

and  $p_i^* = \bar{p} + \theta(p_i - \bar{p})$ , for  $i = 1, 2, \dots, N$ ,  $0 < \theta < 1$ .

Observe that the remaining term  $R$  in (2.7) is a function of the third and fourth moments of  $p_i$ 's. See Table 1 for relevant numerical discussions.

We use Proposition 1 to derive an estimator of  $N$ . Since

$$E(f_1) = \sum_i np_i(1-p_i)^{n-1} \quad (2.8)$$

$$\approx n(1-\bar{p})^{n-1} - [2E(f_2) - 3E(f_3)]\gamma^2, \quad (2.9)$$

it follows from (2.6) and (2.9) that

$$N = \frac{E(D)}{E(C)} + \frac{E(f_1)}{E(C)} \gamma^2 + R^*, \quad (2.10)$$

where  $R^* \approx [2E(f_2) - 3E(f_3)]\gamma^4/E(C) - R$ . Extensive numerical results (see Table 1 for a selection) have suggested that  $R^*$  is usually negligible. Therefore, our main result will be based on the following approximation:

$$N \approx \frac{E(D)}{E(C)} + \frac{E(f_1)}{E(C)} \gamma^2. \quad (2.11)$$

To show how (2.11) works theoretically (i.e., to show  $R^*$  is negligible), we considered several combinations of  $N$  and  $n$  so that the sample coverage varies in a wide range. The cell probabilities were generated as the fractions of an iid realization from a negative binomial given in (2.4) with parameters  $k = 1$  and  $r = .04$ . For given  $p_i$ 's, all relevant expectations and CV in (2.11) can be obtained explicitly; see (A.1) and (A.2) in the Appendix and (2.8). Table 1 gives the average values of  $E(C)$ ,  $E(D)$ ,  $E(f_1)$ ,  $[E(D) + E(f_1)\gamma^2]/E(C)$  and true values of  $R^*$  over 50 different sets of  $p_i$ 's.

The results in Table 1 show that the approximation (2.11) works very well theoretically for all cases. Its behavior for other types of cell probabilities is also similar. In other words, the remaining term  $R^*$  in (2.10) is almost negligible and thus justifies the use of (2.11) numerically. Although its theoretical behavior seems to be satisfactory for all sample sizes, we do need sufficiently many data to estimate CV in practical situations. This will be addressed in the simulation section. We first discuss the estimation of CV. Good and Toulmin (1956) implied that

$$E \left[ \sum_{i=1}^n i(i-1)f_i / (n(n-1)) \right] = \sum_{i=1}^N p_i^2.$$

Thus

$$\gamma^2 = N \sum p_i^2 - 1 = N \sum_{i=1}^n i(i-1)E(f_i) / [n(n-1)] - 1.$$

To obtain an estimator of  $\gamma^2$ , we must replace  $N$  in the above by an initial estimate. From (2.3), we consider  $\hat{N}_1$

Table 1. Theoretical Behavior of Approximation (2.11) for Negative Binomial With  $k = 1$  and  $r = .04$

$N$	$n$	$E(D)$	$E(C)$	$E(f_i)$	(2.11) (right side)	$R^*$
100	10	9.2	.17	8.4	99.8	-.2
	20	16.9	.30	14.3	99.5	-.5
	50	33.9	.55	22.9	100.2	.2
	100	50.6	.75	25.4	99.7	-.3
	200	67.4	.89	22.6	100.3	.3
	400	81.0	.96	16.3	100.9	.9
200	10	9.6	.09	9.2	200.0	.0
	20	18.3	.17	16.7	199.7	-.3
	50	40.4	.35	32.6	199.5	-.5
	100	67.5	.55	45.5	200.8	.8
	200	101.0	.75	51.0	201.3	1.3
	400	134.5	.89	45.4	200.8	.8
400	10	9.8	.05	9.6	400.0	.0
	20	19.1	.09	18.2	400.4	.4
	50	44.6	.21	39.8	400.4	.4
	100	80.5	.35	64.8	399.3	-.7
	200	134.2	.55	90.1	401.8	1.8
	400	202.7	.75	102.1	399.4	-.6

an initial estimate; thus we have the following estimator of the nonnegative parameter  $\gamma^2$ :

$$\hat{\gamma}^2 = \max\left\{\hat{N}_1 \sum i(i-1)f_i / [n(n-1)] - 1, 0\right\}. \quad (2.12)$$

When the true value of CV is relatively large, we suggest using the following bias-corrected version of  $\hat{\gamma}^2$ :

$$\tilde{\gamma}^2 = \max\left\{\hat{\gamma}^2 \left\{1 + \frac{n(1-\hat{C})}{[n(n-1)\hat{C}]}\sum i(i-1)f_i\right\}, 0\right\}. \quad (2.13)$$

From (2.11), we then propose the following two estimators:

$$\hat{N}_2 = \frac{D}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}} \hat{\gamma}^2 \quad (2.14)$$

and

$$\hat{N}_3 = \frac{D}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}} \tilde{\gamma}^2, \quad (2.15)$$

where  $\hat{C}$  is defined by (2.2).

Since  $D/\hat{C}$  is an estimator under the equiprobable assumption, we see from (2.14) or (2.15) that it is biased low by the heterogeneity of the class probabilities. Formula (2.14) or (2.15) provides the magnitude of this bias. It depends on the sample coverage and the CV of the class probability distribution. For any fixed population the bias decreases when the sample coverage is increased, whereas for fixed sample coverage the bias increases as the CV becomes large.

An approximate variance of  $\hat{N}_2$  or  $\hat{N}_3$  can be obtained by using a standard asymptotic approach. Note that both  $\hat{N}_2$  and  $\hat{N}_3$  are functions of  $(f_1, f_2, \dots, f_n)$ , for example, if  $\hat{\gamma} > 0$

$$\begin{aligned} \hat{N}_2 &= \hat{N}_2(f_1, f_2, \dots, f_n) \\ &= \frac{\sum f_i}{1-f_1/n} + \frac{f_1}{1-f_1/n} \left[ \frac{\sum f_i}{1-f_1/n} \frac{\sum i(i-1)f_i}{n(n-1)} - 1 \right], \end{aligned} \quad (2.16)$$

where all the summations are from  $i = 1$  to  $i = n$ . Thus

$$\begin{aligned} \text{var}(\hat{N}_2) &\approx \sum_{k=1}^n A_k^2 \sum_{j=1}^N \alpha_j(k)[1 - \alpha_j(k)] \\ &+ \sum_{k=1}^n \sum_{s \neq k}^n \sum_{j=1}^N A_k A_s [-\alpha_j(k)\alpha_j(s)] \\ &+ \sum_{k=1}^n \sum_{s=1}^n A_k A_s \sum_{j=1}^N \sum_{m \neq j}^N [\beta_{j,m}(k, s) - \alpha_j(k)\alpha_m(s)], \end{aligned} \quad (2.17)$$

where  $A_k$  is the partial derivative of  $\hat{N}_2$  with respect to  $f_k$  evaluated at  $(Ef_1, \dots, Ef_n)$ ,  $\alpha_j(k) = \binom{n}{k} p_j^k (1-p_j)^{n-k}$ , and  $\beta_{j,m}(k, s) = \binom{ksn-n-k-s}{k, s} p_j^k p_m^s (1-p_j-p_m)^{n-k-s}$ . An approximate variance estimator will be developed later.

To connect our estimator with the work of Esty (1985), we now make an assumption on the generation of cell probabilities, which are regarded as random variables.

**Proposition 2.** Assume that the  $i$ th class consists of  $Y_i$  elements and the class sizes  $Y_1, Y_2, \dots, Y_N$  are iid random variables. Suppose that a random sample is drawn with replacement by selecting each element with the same probability, then the class probabilities  $P_i = Y_i / \sum_{i=1}^N Y_i$  ( $i = 1, 2, \dots, N$ ) satisfy that

$$E(P_i) = 1/N, \quad (2.18)$$

$$\text{cov}(P_i, P_j) = -\text{var}(P_i)/N + O(N^{-4}) \quad \text{for } i \neq j. \quad (2.19)$$

In particular, if  $Y_i, i = 1, 2, \dots, N$  are distributed as negative binomial given in (2.4), then

$$\text{var}(P_i) = (1-r)/[N^2k] + O(N^{-3}). \quad (2.20)$$

*Proof.* Equation (2.18) is obviously true. Equations (2.19) and (2.20) follow directly from that for  $i, j = 1, 2, \dots, N$  and  $i \neq j$ :

$$\text{var}(P_i) = N^{-2}(EY_i)^{-2} \text{var}(Y_i) + O(N^{-3}),$$

and

$$\text{cov}(P_i, P_j) = -N^{-3}(EY_i)^{-2} \text{var}(Y_i) + O(N^{-4}).$$

In this setup, it is clear that (2.11) holds conditioning on  $\mathbf{P} = (P_1, P_2, \dots, P_N)$ ; that is, replace expectation there by a conditional expectation. Under the assumption that the class sizes  $Y_i$  ( $i = 1, 2, \dots, N$ ) are negative-binomially distributed as given in (2.4), we have  $\gamma^2 \approx CV(P_i) = (1 - r)/k + O(N^{-1})$  by (2.20) from a conditional point of view. Esty (1985) also assumed that  $E(Y_i) \rightarrow \infty$ ; it is equivalent to letting  $r \rightarrow 0$ , since  $E(Y_i) = k/r$ . Thus  $\gamma^2 \rightarrow 1/k$ . Substituting  $\gamma^2 = 1/k$  into (2.11), we then have Esty's estimator given in (2.5).

We can also obtain the same conclusion using an unconditional approach.

*Proposition 3.* Assume that the cell probabilities  $(P_1, P_2, \dots, P_N)$  have a joint symmetric distribution with a common marginal distribution  $F(p)$  on  $(0, 1)$ . Let  $F(p)$  have mean  $p_0 = \int p dF(p)$  ( $p_0$  must be  $1/N$  in the multinomial model, but we use this general notation to indicate that the result holds for general  $p_0$ , and  $CV \gamma = [\int p^2 dF(p)/p_0^2 - 1]^{1/2}$ . Then

$$E(D)/E(C) = N - \frac{n(1 - p_0)^{n-1}}{E(C)} \gamma^2 + R_2,$$

where

$$R_2 = \frac{1}{E(C)} \int \binom{n}{2} p_0^2 (1 - p^*)^{n-2} \left[ \frac{p - p_0}{p_0} \right]^3 dF(p) - \frac{\theta}{E(C)} \int \binom{n}{3} p_0^3 (1 - p^*)^{n-3} \left[ \frac{p - p_0}{p_0} \right]^4 dF(p),$$

$p^* = p_0 + \theta(p - p_0)$  for  $0 < \theta < 1$ , and the expectation is unconditional.

Therefore, (2.11) still holds when the expectation is in an unconditional sense. Now it is clear that both conditional and unconditional approaches lead to the estimator  $\hat{N}_2$  and  $\hat{N}_3$  in (2.14) and (2.15). As discussed previously, if the class sizes  $Y_i$  ( $i = 1, 2, \dots, N$ ) are negative-binomially distributed, we have  $\gamma^2 = (1 - r)/k + O(N^{-1}) \rightarrow 1/k$  if  $E(Y_i) \rightarrow \infty$ . We then also have Esty's estimator (2.5).

We now discuss a variance estimator of the estimators (2.14) and (2.15). Here, we adopt an unconditional approach. Notice that unconditionally  $(f_0, f_1, \dots, f_n)$  is approximately multinomially distributed with parameter  $N$  and cell probability  $\int \binom{n}{i} p^i (1 - p)^{n-i} dF(p)$  ( $i = 0, 1, \dots, n$ ) and both  $\hat{N}_2$  and  $\hat{N}_3$  are functions of  $(f_1, f_2, \dots, f_n)$ ; thus the asymptotic normality and variance estimator can be obtained. In this unconditional approach, the sample size  $n = \sum if_i$  must be regarded as a random variable, and we write (if  $\hat{\gamma}^2 > 0$ ) (2.16) in a slightly different form:

$$\hat{N}_2 = \frac{\sum f_i}{1 - f_1/\sum if_i} + \frac{f_1}{1 - f_1/\sum if_i} \left[ \frac{\sum f_i}{1 - f_1/\sum if_i} \frac{\sum i(i - 1)f_i}{(\sum if_i)(\sum if_i - 1)} - 1 \right],$$

which then implies

$$\widehat{\text{var}}(\hat{N}_2) \approx \sum_{i=1}^n \sum_{j=1}^n \frac{\partial \hat{N}_2}{\partial f_i} \frac{\partial \hat{N}_2}{\partial f_j} \text{cov}(f_i, f_j), \tag{2.21}$$

where

$$\begin{aligned} \text{cov}(f_i, f_j) &= f_i(1 - f_i/\hat{N}_2) \quad \text{if } i = j \\ &= -f_i f_j / \hat{N}_2^2 \quad \text{if } i \neq j. \end{aligned}$$

Similarly, we can obtain variance estimators for  $\hat{N}_1$  and  $\hat{N}_3$ . The adequacy of this unconditional variance approximation will be numerically shown in the simulation studies of Section 4.

### 3. AN EXAMPLE

This data was described and analyzed in Holst (1981). Two hundred and four coins were found in a hoard of ancient coins. They were classified to different die types, and the purpose was to estimate the number of dies used in the minting process. The frequencies for the reverse side of the coins were 156 singletons, 19 doubletons, 2 triplets, and 1 quadruplet. Thus  $D = 178$  different dies were found in the hoard. For the obverse side, there were 141 different dies with frequencies in order (1-7) 102, 26, 8, 2, 1, 1, 1.

For the reverse side, Holst (1981) indicated that it is reasonable to assume that each die produced about the same number of coins. Hence he proposed  $\hat{N}_0 = 731$ , with a standard error (s.e.) estimate 130.6. Our estimator for the equiprobable case is  $\hat{N}_1 = 757$  (s.e. 143.5). Both results agree well. Here the sample coverage is estimated to be 23.5% with s.e. 4.2% based on a variance formula provided in Esty (1983). The CV estimates are  $\hat{\gamma} = .36$  and  $\tilde{\gamma} = .51$ ; thus we have  $\hat{N}_2 = 844$  (s.e. 186.6) and  $\hat{N}_3 = 932$  (s.e. 265.7).

For the obverse side, the sample coverage estimate is 50% (s.e. 5%) and the CV estimates are  $\hat{\gamma} = .69$  and  $\tilde{\gamma} = 1.0$ , which shows stronger evidence of heterogeneity. This also explains why the equally likely assumption is inappropriate for the obverse side, as concluded by Holst (1981). Hence we expect that both  $\hat{N}_0 = 258$  (s.e. 23.0) and  $\hat{N}_1 = 282$  (s.e. 32.9) severely underestimate the true population size. Our proposed estimates are  $\hat{N}_2 = 378$  (s.e. 65.2) and  $\hat{N}_3 = 480$  (s.e. 126.5). For both sides, the standard errors for the two proposed estimators are higher than the estimators assuming equiprobable condition. This seems unavoidable because of the estimation of CV. See numerical results in the next section.

### 4. SIMULATION STUDY

A simulation study was carried out to investigate the behavior of the proposed method. We studied the 48 combinations of:

1. Four types of populations—the equiprobable case ( $CV = 0$ ); negative binomial with  $k = 4, r = .04$  ( $CV = .4899$ );  $k = 2$  and  $r = .04$  ( $CV = 0.6928$ );  $k = 1$  and  $r = .04$  ( $CV = 0.9798$ ).
2. Three numbers of classes ( $N$ )—50, 100, and 200.
3. Four sample sizes ( $n$ )—50, 100, 200, and 400.

In the negative binomial cases, a random sample of size  $N$  was first generated and then the proportions were used as cell probabilities and fixed through the simulation.

For each combination, 200 data sets were generated. A preliminary study found that the results based on 200 trials are very close to those of 500 trials. Thus only 200 trials were run for each combination to save computing time. Then for each generated data set,  $\hat{N}_0$ - $\hat{N}_3$  and their standard error estimates were calculated. Finally, these 200 estimates and standard errors were averaged, and the sample standard error, as well as sample root mean squared error, (RMSE) was obtained. We present only the case  $N = 100$  in Tables 2-5. Conclusions for other values of  $N$  are generally similar. In each table, we also tabulate the average values of the number of observed classes ( $\bar{D}$ ), sample coverage ( $\bar{C}$ ), sample coverage estimate ( $\hat{\bar{C}}$ ), and  $CV$  estimates ( $\hat{\gamma}$  and  $\bar{\gamma}$ ).

Smaller sample sizes ( $n = 10$  or  $20$ ) were originally studied to investigate the estimator's behavior for low coverage cases. Unfortunately, in some trials, data were so sparse that all observed classes were singleton (i.e.,  $D = n = f_1$ ). Under such extreme cases, the usual MLE is  $\infty$ . Our estimate is also  $\infty$ , since the sample coverage estimate is 0. Hence it makes the point estimation comparison meaningless. This explains why we only considered sample size  $\geq 50$ , although Table 1 shows that, in theory, our method works for all  $n$ . It also indicates that alternative techniques are needed for very low coverage situations; see Esty (1982).

These tables show that in all cases,  $\hat{C}$  performs very well in estimating the sample coverage. This is exactly the motivation for attempting to estimate population size using the concept of sample coverage.

In the equally likely case (Table 2),  $CV = 0$ ,  $\hat{N}_1$  becomes the appropriate estimator in our approach.  $\hat{N}_0$  and  $\hat{N}_1$  behave similarly and work better than other estimates with respect to both bias and RMSE, since both are derived under equally likely assumption. This also numerically shows the conclusion of Darroch and Ratcliff (1980) that  $\hat{N}_1$  is highly efficient compared with the MLE. Our estimators  $\hat{N}_2$  and  $\hat{N}_3$  are expected to have a positive bias because the true  $CV$  is 0, whereas the estimates of it are slightly higher than 0. When  $n = 400$ , almost all classes were observed (the sample coverage tends to 1) and all estimators work well.

As expected, when  $CV > 0$ ,  $\hat{N}_0$  and  $\hat{N}_1$  consistently underestimate the true parameter, and the magnitude of the bias increases as the  $CV$  becomes large. When  $CV$  is high, the biases still exist even in the cases of high sample coverage (e.g.,  $n = 400$ ), whereas the proposed estimators are nearly unbiased.

The mean squared error (MSE) is the sum of variance and squared of bias. Theoretically, our approach involving estimation of the heterogeneity of class probabilities generally reduces bias but increases variance (because of the estimation of  $CV$ ). This is also clearly seen in Tables 3-5.  $\hat{N}_0$  and  $\hat{N}_1$  have the smallest standard errors but largest biases,  $\hat{N}_3$  has the smallest bias but largest standard error, and  $\hat{N}_2$  is somewhat in between. Based on the RMSE criterion, Table 3 shows that, for  $CV \approx .5$ ,  $\hat{N}_0$  and  $\hat{N}_1$  are still superior to  $\hat{N}_2$  or  $\hat{N}_3$  if  $n = 50$ ; for  $n = 100$ ,  $\hat{N}_0$ ,  $\hat{N}_1$ , and  $\hat{N}_2$  are generally comparable and all have smaller RMSE than  $\hat{N}_3$ ; for  $n > 100$ ,  $\hat{N}_2$  works best. In Table 4 ( $CV \approx .7$ ),  $\hat{N}_2$  outperforms all the others with respect to RMSE. In Table 5 ( $CV \approx 1$ ),  $\hat{N}_3$  becomes the choice except for  $n = 50$ .

Table 2. Comparison of Estimates for Equiprobable Case,  $N = 100^*$

Sample size $n$	Mean estimate	Mean bias	Mean estimated s.e.	Sample s.e.	Sample RMSE
50					
$\hat{N}_0$	107	7	30.0	33.5	34.3
$\hat{N}_1$	107	7	30.5	33.6	34.3
$\hat{N}_2$	113	13	37.9	39.3	41.5
$\hat{N}_3$	120	20	48.0	50.1	54.0
$\bar{D} = 39.37$		$\bar{C} = .395$	$\hat{\bar{C}} = .393$	$\hat{\gamma} = .131$	$\bar{\gamma} = .223$
100					
$\hat{N}_0$	101	1	12.1	12.1	12.2
$\hat{N}_1$	101	1	12.6	12.7	12.8
$\hat{N}_2$	104	4	16.2	15.4	16.0
$\hat{N}_3$	106	6	19.6	18.3	19.4
$\bar{D} = 63.28$		$\bar{C} = .634$	$\hat{\bar{C}} = .631$	$\hat{\gamma} = .110$	$\bar{\gamma} = .175$
200					
$\hat{N}_0$	101	1	4.8	4.6	4.6
$\hat{N}_1$	100	0	5.1	4.9	4.9
$\hat{N}_2$	101	1	6.3	5.5	5.6
$\hat{N}_3$	102	2	6.8	5.9	6.1
$\bar{D} = 86.77$		$\bar{C} = .866$	$\hat{\bar{C}} = .865$	$\hat{\gamma} = .083$	$\bar{\gamma} = .111$
400					
$\hat{N}_0$	100	0	1.4	1.4	1.4
$\hat{N}_1$	100	0	1.5	1.6	1.6
$\hat{N}_2$	100	0	1.6	1.6	1.6
$\hat{N}_3$	100	0	1.7	1.7	1.7
$\bar{D} = 98.16$		$\bar{C} = .982$	$\hat{\bar{C}} = .982$	$\hat{\gamma} = .059$	$\bar{\gamma} = .067$

\* $\bar{p} = .01$ ,  $CV = 0$ , 200 trials.

Table 3. Comparison of Estimates for Negative Binomial  $k = 4, r = .04, N = 100^*$

Sample size $n$	Mean estimate	Mean bias	Mean estimated s.e.	Sample s.e.	Sample RMSE
50					
$\hat{N}_0$	89	-11	22.4	27.7	29.8
$\hat{N}_1$	90	-10	23.4	28.8	30.6
$\hat{N}_2$	97	-3	30.6	41.1	41.2
$\hat{N}_3$	105	5	40.9	66.1	66.3
$\bar{D} = 37.69$		$\bar{C} = .450$	$\bar{C} = .444$	$\bar{\gamma} = .152$	$\bar{\gamma} = .259$
100					
$\hat{N}_0$	86	-14	9.2	10.4	17.4
$\hat{N}_1$	88	-12	10.3	11.3	16.5
$\hat{N}_2$	94	-6	14.5	15.4	16.6
$\hat{N}_3$	98	-2	19.0	19.9	20.0
$\bar{D} = 58.88$		$\bar{C} = .677$	$\bar{C} = .675$	$\bar{\gamma} = .215$	$\bar{\gamma} = .334$
200					
$\hat{N}_0$	90	-10	3.9	4.1	10.6
$\hat{N}_1$	93	-7	4.8	4.5	8.7
$\hat{N}_2$	97	-3	7.3	6.4	7.0
$\hat{N}_3$	99	-1	8.8	7.8	7.8
$\bar{D} = 80.32$		$\bar{C} = .869$	$\bar{C} = .869$	$\bar{\gamma} = .320$	$\bar{\gamma} = .424$
400					
$\hat{N}_0$	96	-4	1.3	2.3	5.0
$\hat{N}_1$	97	-3	2.0	2.5	3.7
$\hat{N}_2$	100	0	3.4	3.2	3.2
$\hat{N}_3$	101	1	3.8	3.5	3.5
$\bar{D} = 94.06$		$\bar{C} = .968$	$\bar{C} = .968$	$\bar{\gamma} = .430$	$\bar{\gamma} = .491$

\* $\bar{p} = .01, CV = .4899, 200$  trials.

Whether the reduction in the squared of bias can compensate for the increase in variance depends on  $CV, N$ , and  $n$ . Generally, consideration of  $CV$  is worthwhile for reducing MSE when  $CV$  is large and sample size is large (relatively to  $N$ ) enough to make stable estimation of  $CV$ . The general guidelines about how large the sample size should

be are still unclear to us. Our experience suggests that, for fixed  $N$ , the larger  $CV$  is, the smaller the  $n$  (or the sample coverage) needed. For example, in the case of  $N = 100$ , if  $CV \approx .5$ , sample coverage should be at least 70%; for  $CV \approx .7$  or 1, sample coverage 50% is enough. When  $CV$  is relatively small ( $\leq .5$  based on simulation experience),

Table 4. Comparison of Estimates for Negative Binomial  $k = 2, r = .04, N = 100^*$

Sample size $n$	Mean estimate	Mean bias	Mean estimated s.e.	Sample s.e.	Sample RMSE
50					
$\hat{N}_0$	76	-24	17.1	19.6	31.1
$\hat{N}_1$	77	-23	18.4	19.9	30.2
$\hat{N}_2$	84	-16	24.8	24.7	29.4
$\hat{N}_3$	91	-9	33.3	33.8	35.0
$\bar{D} = 36.12$		$\bar{C} = .496$	$\bar{C} = .487$	$\bar{\gamma} = .199$	$\bar{\gamma} = .333$
100					
$\hat{N}_0$	76	-24	7.4	9.1	25.9
$\hat{N}_1$	79	-21	8.6	10.0	23.7
$\hat{N}_2$	86	-14	13.3	13.9	19.9
$\hat{N}_3$	91	-9	18.1	18.4	20.4
$\bar{D} = 55.28$		$\bar{C} = .711$	$\bar{C} = .710$	$\bar{\gamma} = .309$	$\bar{\gamma} = .466$
200					
$\hat{N}_0$	81	-19	3.2	4.5	19.2
$\hat{N}_1$	85	-15	4.4	5.1	16.0
$\hat{N}_2$	93	-7	8.0	7.9	10.6
$\hat{N}_3$	97	-3	10.5	10.1	10.7
$\bar{D} = 74.30$		$\bar{C} = .877$	$\bar{C} = .877$	$\bar{\gamma} = .455$	$\bar{\gamma} = .610$
400					
$\hat{N}_0$	89	-11	1.1	2.8	10.9
$\hat{N}_1$	92	-8	2.2	3.1	8.6
$\hat{N}_2$	98	-2	4.9	4.7	5.0
$\hat{N}_3$	100	0	5.9	5.5	5.5
$\bar{D} = 88.42$		$\bar{C} = .961$	$\bar{C} = .961$	$\bar{\gamma} = .578$	$\bar{\gamma} = .682$

\* $\bar{p} = .01, CV = .6928, 200$  trials.

Table 5. Comparison of Estimates for Negative Binomial  $k = 1, r = .04, N = 100^*$

Sample size $n$	Mean estimate	Mean bias	Mean estimated s.e.	Sample s.e.	Sample RMSE
50					
$\hat{N}_0$	59	-41	11.4	13.7	42.9
$\hat{N}_1$	61	-39	13.0	14.4	41.2
$\hat{N}_2$	70	-30	19.4	22.3	37.6
$\hat{N}_3$	79	-21	29.0	38.1	43.7
$\bar{D} = 33.40$		$\bar{C} = .565$	$\bar{C} = .560$	$\bar{\gamma} = .265$	$\bar{\gamma} = .447$
100					
$\hat{N}_0$	63	-37	5.3	6.7	37.6
$\hat{N}_1$	67	-33	6.8	7.5	34.2
$\hat{N}_2$	76	-24	12.0	12.4	26.8
$\hat{N}_3$	83	-17	17.9	18.9	25.1
$\bar{D} = 49.96$		$\bar{C} = .761$	$\bar{C} = .754$	$\bar{\gamma} = .425$	$\bar{\gamma} = .640$
200					
$\hat{N}_0$	69	-31	2.2	4.5	31.0
$\hat{N}_1$	73	-27	3.7	5.3	27.1
$\hat{N}_2$	84	-16	8.3	9.2	18.3
$\hat{N}_3$	90	-10	12.1	12.6	16.2
$\bar{D} = 65.42$		$\bar{C} = .893$	$\bar{C} = .893$	$\bar{\gamma} = .586$	$\bar{\gamma} = .790$
400					
$\hat{N}_0$	80	-20	.74	3.3	20.7
$\hat{N}_1$	82	-18	2.1	3.8	18.0
$\hat{N}_2$	93	-7	6.4	7.0	9.8
$\hat{N}_3$	97	-3	8.9	9.2	9.6
$\bar{D} = 79.00$		$\bar{C} = .958$	$\bar{C} = .959$	$\bar{\gamma} = .735$	$\bar{\gamma} = .895$

\* $\bar{p} = .01, CV = .9798, 200$  trials.

the usual estimators assuming equiprobable condition are not seriously biased, so the improvement in bias is quite limited. Consequently, our method cannot reduce MSE unless  $n$  is sufficiently large. For large  $n$  and small  $CV$ , however, all estimators have similar performance. Thus there seems to be no advantage in estimating  $CV$  in these situations. If  $CV > .5$ , however, the use of the proposed estimation procedure to reduce MSE is warranted as long as the sample size is large enough, that is, when sample coverage is about 50% or more.

As for the choice between  $\hat{N}_2$  and  $\hat{N}_3$ , in terms of RMSE,  $\hat{N}_2$  seems to be superior to  $\hat{N}_3$  when  $CV$  is moderate, as in Tables 3 and 4 ( $CV \approx .5, .7$ ), but when  $CV$  is relatively large, as in Table 5 ( $CV \approx 1$ ),  $\hat{N}_3$  is preferable in the sense of having smallest bias and smallest RMSE among all estimators except for  $n = 50$ .

In conclusion, in addition to the case  $CV = 0$ , classical estimators assuming equiprobability are still appropriate when  $CV \leq .5$ . For  $CV > .5$  and data sufficiently large to allow stable estimation of  $CV$  (for example, the sample coverage exceeds 50% in this study), the proposed estimation procedure incorporating the  $CV$  term is suggested for practical use.

The standard error estimates for  $\hat{N}_1, \hat{N}_2,$  and  $\hat{N}_3$  calculated by the unconditional approach (column 4 in each table) are generally satisfactory compared with the sample standard errors (column 5 in each table).

It is clear from these results that the proposed estimators for nonequiprobable cases are generally biased downward because of the underestimation of  $CV$ . (The  $CV$  estimates are considerably lower than the true values unless  $n$  is large.) This is a principal source of error associated with the pro-

posed method. Whether it is possible to improve the estimation of  $CV$  and whether the improvement would lead to a better estimation of population size are still under investigation.

APPENDIX: PROOF OF PROPOSITION 1

Since for fixed  $\mathbf{p} = (p_1, p_2, \dots, p_N)$ ,

$$E(D) = N - \sum_{i=1}^N (1 - p_i)^n \tag{A.1}$$

and

$$E(C) = 1 - \sum_{i=1}^N p_i(1 - p_i)^n, \tag{A.2}$$

we can write that

$$E(D)/E(C) = N + g(\mathbf{p})/E(C),$$

where

$$g(\mathbf{p}) = N \sum_{i=1}^N p_i(1 - p_i)^n - \left[ \sum_{i=1}^N (1 - p_i)^n \right].$$

It is easy to prove that expanding  $g(\mathbf{p})$  with constraint  $\sum p_i = 1$  at  $\bar{\mathbf{p}} = (\bar{p}, \bar{p}, \dots, \bar{p}) = (1/N, \dots, 1/N)$  to the second order term is equivalent to expanding  $g(\mathbf{p})$  without the constraint  $\sum p_i = 1$ . For the latter expansion, we have

$$\begin{aligned} g(\bar{\mathbf{p}}) &= 0; \\ \left[ \frac{\partial g(\mathbf{p})}{\partial p_i} \right]_{\mathbf{p}=\bar{\mathbf{p}}} &= 0 \text{ for all } i = 1, 2, \dots, N; \\ \frac{\partial^3 g(\bar{\mathbf{p}})}{\partial p_i \partial p_j \partial p_k} &= 0. \end{aligned}$$

Thus (all indexes in the following proof run from 1 to  $N$ )

$$g(\mathbf{p}) = \frac{1}{2} \sum_i \left[ \frac{\partial^2 g(\mathbf{p})}{\partial p_i^2} \Big|_{\mathbf{p}=\mathbf{p}} \right] (p_i - \bar{p})^2 + \sum_{i < j} \sum \left[ \frac{\partial^2 g(\mathbf{p})}{\partial p_i \partial p_j} \Big|_{\mathbf{p}=\mathbf{p}} \right] (p_i - \bar{p})(p_j - \bar{p}) + R_1,$$

where  $R_1$  denotes the remaining term and

$$R_1 = \frac{1}{6} \sum_i \left[ \frac{\partial^3 g(\mathbf{p})}{\partial p_i^3} \Big|_{\mathbf{p}=\mathbf{p}^*} \right] (p_i - \bar{p})^3 + \sum_{i < j} \sum \left[ \frac{\partial^3 g(\mathbf{p})}{\partial p_i^2 \partial p_j} \Big|_{\mathbf{p}=\mathbf{p}^*} \right] (p_i - \bar{p})^2 (p_j - \bar{p}),$$

and  $\mathbf{p}^* = (p_1^*, p_2^*, \dots, p_N^*)$ ,  $p_i^* = \bar{p} + \theta(p_i - \bar{p})$  ( $i = 1, 2, \dots, N$ ;  $0 < \theta < 1$ ). Substituting all the derivatives and using that  $\sum_i p_i = \sum_i p_i^* = N\bar{p}$ ,  $\sum_{j \neq i} (p_j - \bar{p}) = \bar{p} - p_i$ , we then obtain (2.6) and (2.7) after rearrangement.

[Received September 1989. Revised March 1991.]

## REFERENCES

- Chao, A. (1984), "Nonparametric Estimation of the Number of the Classes in a Population," *Scandinavian Journal of Statistics*, 11, 265–270.
- (1987), "Estimating Population Size for Capture–Recapture Data With Unequal Catchability," *Biometrics*, 43, 783–791.
- Craig, C. C. (1953), "Use of Marked Specimens in Estimating Populations," *Biometrika*, 40, 170–176.
- Darroch, J. N. (1958), "The Multiple Recapture Census I: Estimation of a Closed Population," *Biometrika*, 45, 343–359.
- 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*, London: Chapman & Hall.
- Esty, W. W. (1982), "Confidence Intervals for the Coverage of Low Coverage Samples," *The Annals of Statistics*, 10, 190–196.
- (1983), "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample," *The Annals of Statistics*, 11, 905–912.
- (1985), "Estimation of the Number of Classes in a Population and the Coverage of a Sample," *Mathematical Scientist*, 10, 41–50.
- (1986a), "The Size of a Coinage," *Numismatic Chronicle*, 146, 185–215.
- (1986b), "The Efficiency of Good's Nonparametric Coverage Estimator," *The Annals of Statistics*, 14, 1257–1260.
- Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943), "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population," *Journal of Animal Ecology*, 12, 42–58.
- Good, I. J. (1953), "On the Population Frequencies of Species and the Estimation of Population Parameters," *Biometrika*, 40, 237–264.
- Good, I. J., and Toulmin, G. (1956), "The Number of New Species and the Increase of Population Coverage When a Sample Is Increased," *Biometrika*, 43, 45–63.
- Goodman, L. A. (1949), "On the Estimation of the Number of Classes in a Population," *Annals of Mathematical Statistics*, 20, 572–579.
- Harris, B. (1968), "Statistical Inference in the Classical Occupancy Problem Unbiased Estimation of the Number of Classes," *Journal of the American Statistical Association*, 63, 837–847.
- Holst, L. (1981), "Some Asymptotic Results for Incomplete Multinomial or Poisson Samples," *Scandinavian Journal of Statistics*, 8, 243–246.
- Johnson, N. L., and Kotz, S. (1977), *Urn Models and Their Applications: An Approach to Modern Discrete Probability Theory*, New York: John Wiley.
- Lewontin, R. C., and Prout, T. (1956), "Estimation of the Number of Different Classes in a Population," *Biometrika*, 12, 211–223.
- Marchand, J. P., and Schroeck, P. E. (1982), "On the Estimation of the Number of Equally Likely Classes in a Population," *Communications in Statistics, Part A—Theory and Methods*, 11, 1139–1146.
- Robbins, H. (1968), "Estimating the Total Probability of the Unobserved Outcomes of an Experiment," *Annals of Mathematical Statistics*, 39, 256–257.
- Thisted, R., and Efron, B. (1987), "Did Shakespeare Write a Newly-Discovered Poem?" *Biometrika*, 74, 445–455.