Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Population size estimation for capture-recapture models with applications to epidemiological data

P. K. Tsay [a] [b] & A. Chao [b]

[a] Department of Public Health , Chang Gung University , Taiwan

[b] Institute of Statistics , National Tsing Hua University , Taiwan

Published online: 02 Aug 2010.

To cite this article: P. K. Tsay & A. Chao (2001) Population size estimation for capture-recapture models with applications to epidemiological data, Journal of Applied Statistics, 28:1, 25-36, DOI: 10.1080/02664760120011572

To link to this article: http://dx.doi.org/10.1080/02664760120011572

Journal of Applied Statistics, Vol. 28, No. 1, 2001, 25–36

Population size estimation for capture–recapture models with applications to epidemiological data

P. K. TSAY[1,2] **& A. CHAO**[2], [1]*Department of Public Health, Chang Gung University, Taiwan and* [2]*Institute of Statistics, National Tsing Hua University, Taiwan*

ABSTRACT *The capture–recapture method is applied to estimate the population size of a target population based on ascertainment data in epidemiological applications. We generalize the three-list case of Chao & Tsay (1998) to situations where more than three lists are available. An estimation procedure is presented using the concept of sample coverage, which can be interpreted as a measure of overlap information among multiple list records. When there is enough overlap, an estimator of the total population size is proposed. The bootstrap method is used to construct a variance estimator and confidence interval. If the overlap rate is relatively low, then the population size cannot be precisely estimated and thus only a lower (upper) bound is proposed for positively (negatively) dependent lists. The proposed method is applied to two data sets, one with a high and one with a low overlap rate.*

1 Introduction

In a typical capture–recapture experiment, the target population is sampled several times (or over a certain number of occasions) using traps or nets. We assume that over the experiment period the population size is a constant, which is our parameter of interest. For each trapping sample, a unique tag or mark is attached to a first-capture, whereas for a recapture its tag number is recorded. Biologists and ecologists have long recognized that the recapture information (i.e. overlap information) collected by marking or tagging can be used to estimate the number missing from all the samples. Seber (1982, 1986, 1992) and Schwarz & Seber (1999) provided comprehensive reviews on models of estimating animal abundance and on capture–recapture models in particular.

Correspondence: Anne Chao, Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043. E-mail: chao@stat.nthu.edu.tw

ISSN 0266-4763 print; 1360-0532 online/01/010025-12
DOI: 10.1080/02664760120011572

The purpose of many epidemiological surveillance studies is to estimate the size of a population characterized by a rare trait by merging several existing but incomplete lists of the target population. In this paper, we focus on the applications of the capture–recapture method to such epidemiological studies. Hence, 'capture in a sample' in a multiple capture–recapture study corresponds to 'being recorded or identified in a list', and 'capture probability' becomes 'ascertainment probability' or 'probability of being identified'. A 'sample' in this paper also refers to a list, source or a record.

The capture–recapture technique originally developed for animal studies has been applied to human populations under the term 'multiple-record system' for an arbitrary number of lists, and under the term 'dual-system' especially for two lists. The earliest references to such applications can be traced back to Sekar & Deming (1949) for two samples, Wittes & Sidel (1968) for three samples, Wittes (1974) for four samples and Fienberg (1972) for five samples. Overviews of the topic are provided in Hook & Regal (1995) and the International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995a, b). A short encyclopaedia article is given in Chao (1998).

The traditional approach assumes that all individuals have the same 'capture' probabilities for each sample. This equal-catchability assumption implies that the samples are independent. Dependence among samples leads to a bias for any estimator derived under independence. This bias is called 'correlation bias' in the census undercount context; see Darroch *et al.* (1993). The causes of dependence may derive from two sources: local (list) dependence within each individual and heterogeneity among individuals. These two types of dependencies are usually confounded and cannot be easily disentangled in a data analysis. According to the various ways of modelling dependence, there are three distinct approaches, as described below.

(1) Ecological models: this approach specifies various forms of capture probabilities based on empirical investigations of animal ecology. Three sources of variations are considered: time, behavioural response and heterogeneity. Local independence arises from an individual's behavioural response and heterogeneity among individuals causes another type of dependence. See Pollock (1991) and Chao (2000) for a review.

(2) Loglinear models: in the loglinear approach, the data are regarded as a form of an incomplete 2^t contingency table (t is the number of lists) for which the cell corresponding to those individuals unlisted by all samples is missing. Then various loglinear models are fitted to the observed cells and the chosen model is projected on to the unobserved cell by assuming that there is no t-factor interaction. The two types of dependencies can be modelled by including some specific interaction or common interaction effects in the loglinear models. See Fienberg (1972), Cormack (1989), Agresti (1994) and IWGDMF (1995a, b).

(3) Sample coverage approach: Chao & Tsay (1998) proposed a sample coverage approach, which can deal with the two types of dependencies, to evaluate the undercount for a special type of census problem. This approach aims to quantify the overlap information by using sample coverage and to measure the dependence by using the coefficient of covariation (CCV) parameters. The estimated overlap rate and estimated CCVs are then used to estimate the population size; see Sections 2 and 3 for details.

Chao & Tsay (1998) mainly discussed the cases of two and three lists because these are the two commonly encountered situations in census undercount applications. The overlap rate for the census problem is generally high. However, in epidemiological applications, more than three lists are usually recorded and the overlap rate is often relatively low. In this paper, we extend the three-list case of Chao & Tsay (1998) to handle the data with more than three lists. When there is enough overlap, an estimation procedure of the total population size is proposed. If the overlap rate is relatively low, then the population size cannot be precisely estimated and thus only a lower (upper) bound is proposed for positively (negatively) dependent lists.

In Section 2, we review the sample coverage approach to capture–recapture proposed in Chao & Tsay (1998). The approach encompasses many existing models as special cases. The extension of Chao & Tsay (1998) to the general case with more than three lists is treated in Section 3. In Section 4, we illustrate the applications by using two data sets. One data set with low overlap information concerns infection by the acute hepatitis A virus in an outbreak at a college in northern Taiwan, and the other set with high overlap rate is ascertainment data on diabetes discussed in the literature. Some remarks about the limitations of the capture–recapture method are discussed in Section 5.

2 Sample coverage approach

Assume that the true population size is N and the individuals are indexed by $1, 2, \ldots, N$. Individuals are assumed to act independently. There are t samples and they are indexed by $1, 2, \ldots, t$. Our resulting estimator is invariant to the ordering of the lists. The ascertainment data for all individuals can be conveniently expressed by an $N \times t$ matrix $X = (X_{ij})$, where $X_{ij} = I$ [the ith individual is listed in the jth sample], and I[A] is the usual indicator function for the event A.

Most ascertainment data for all identified individuals are aggregated into a categorical form. Let

$$Z_{s_1 s_2 \ldots s_t} = \sum_{i=1}^{N} I[X_{i1} = s_1, X_{i2} = s_2, \ldots, X_{it} = s_t] \tag{1}$$

be the number of individuals with capture history $s_1, s_2, \ldots, s_t$, where $s_j = 0$ denotes absence in sample j and $s_j = 1$ denotes presence in sample j. For example, when $t = 3$, there are seven observed cells $Z_{100}, Z_{010}, Z_{001}, Z_{110}, Z_{101}, Z_{011}$ and Z_{111} where Z_{100} is the number of individuals listed in sample 1 only, Z_{011} is the number of individuals listed in samples 2 and 3 but not in sample 1. A similar interpretation pertains to other capture histories. The missing cell Z_{000} denotes the number unidentified from all lists. When we add over a sample, the subscript corresponding to that sample is replaced by a '+' sign. For example, $Z_{+11} = Z_{011} + Z_{111}$ and $Z_{++1} = Z_{001} + Z_{011} + Z_{101} + Z_{111}$, etc. Let n_j, $j = 1, 2, \ldots, t$ be the number of individuals listed in sample j. Thus for $t = 3$, we have $n_1 = Z_{1++}$, $n_2 = Z_{+1+}$, and $n_3 = Z_{++1}$. The notation has a direct extension to data with more than three samples.

The sample coverage approach aims to model the dependence by 'coefficient of covariation' (CCV) parameters. We can use either a fixed-effect or a random-effect model. Throughout this paper, we shall adopt a random-effect model and assume that there is no local dependence but there is heterogeneity, unless otherwise stated.

Let $P_{ij} = P(X_{ij} = 1)$ be the capture probability of the ith individual in the jth sample. The model assumes that $\{(P_{i1}, P_{i2}, \ldots, P_{it}),\ i = 1, 2, \ldots, N\}$ are a random sample from a t-dimensional distribution $F_{P_1, P_2, \ldots, P_t}(p_1, p_2, \ldots, p_t)$. The CCV for samples j and k is defined as

$$\gamma_{jk} = \frac{E[(P_j - \mu_j)(P_k - \mu_k)]}{\mu_j \mu_k} = \frac{\mathrm{cov}(P_j, P_k)}{\mu_j \mu_k} = \frac{E(P_j P_k)}{\mu_j \mu_k} - 1 \tag{2}$$

where $\mu_j = E(P_j)$ denotes the average capture probability for the jth sample.

The magnitude of γ_{jk} measures the degree of dependence of samples j and k. If the two samples are independent, then $\gamma_{jk} = 0$. The two samples are positively (negatively) dependent if $\gamma_{jk} > 0 (\gamma_{jk} < 0)$. The CCV for an arbitrary number of m samples under a random-effect model is defined as $(m = 2, 3, \ldots, t)$

$$\gamma_{k_1 k_2 \ldots k_m} = \frac{E[(P_{k_1} - \mu_{k_1})(P_{k_2} - \mu_{k_2}) \ldots (P_{k_m} - \mu_{k_m})]}{\mu_{k_1} \mu_{k_2} \ldots \mu_{k_m}} \tag{3}$$

In the three-sample case, there are, in total, eight parameters (i.e., $N, \mu_1, \mu_2, \mu_3, \gamma_{12}, \gamma_{13}, \gamma_{23}, \gamma_{123}$) whereas there are only seven observable cells. In the independent case, all CCVs $(\gamma_{12}, \gamma_{13}, \gamma_{23}, \gamma_{123})$ vanish.

The above model includes many models commonly used in both animal and human populations. For example, it includes the following ecological model:

$$P(X_{ij} = 1) = h_i e_j \tag{4}$$

where $(e_1, e_2, \ldots, e_t)$ and $(h_1, h_2, \ldots, h_N)$ denote, respectively, the time-varying and heterogeneity effects. This model is referred to as model M_{th} in the literature. In the special case that $e_1 = e_2 = \ldots = e_t = 1$, it reduces to model M_h. Fisher, Corbet and Williams in their 1943 pioneering paper assumed that the heterogeneity effects $(h_1, h_2, \ldots, h_N)$ are a random sample from a gamma distribution with density $f_H(h) = \beta^\alpha h^{\alpha - 1} \exp(-\beta h)/\Gamma(\alpha)$. Under model M_{th} and the gamma assumption, we have $\mu_j = (\alpha/\beta) e_j$, *and*

$$\gamma_{k_1 k_2 \ldots k_m} = \frac{E[H - E(H)]^m}{[E(H)]^m} = \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} \frac{\Gamma(\alpha + j)}{\alpha^j \Gamma(\alpha)} \tag{5}$$

Note that the CCV for any two samples becomes the square of the coefficient of variation (CV) of $(h_1, h_2, \ldots, h_N)$ and thus any two samples are positively correlated. The parameter CV has been used to measure the degree of heterogeneity in many ecological models (Lee & Chao, 1994). In the special case of gamma heterogeneity, we have $\gamma_{jk} = 1/\alpha$. It follows from the moment property of the gamma distribution (e.g. see Bowman & Shenton, 1988, p. 5) that the CCVs for any number of samples are positive and are functions of α only. Specifically, we have $\gamma_{jkm} = 2/\alpha^2$ for any three samples, and $\gamma_{jkmn} = 3(\alpha + 2)/\alpha^3$ for any four samples.

Our model also includes the Rasch (1961) model as a special case. The Rasch model assumes that $P(X_{ij} = 1) = w_i a_j/(1 + w_i a_j)$, where w_i denotes the heterogeneity effect of the ith individual and a_j denotes the sample effect (or time effect) of the jth list. The Rasch model has been extensively used for human populations, especially in educational testing and psychological statistics. Sanathanan (1972) was the first to introduce this model to the context of population size estimation. She adopted a normal random-effect for $(w_1, w_2, \ldots, w_N)$. For a Rasch model, it can be theoretically verified that the CCV for any two samples is positive, but the

CCVs for more than two samples are not necessarily positive, see Chao & Tsay (1998, Table 3).

We remark that when both types of dependencies exist, a fixed-effect model is needed. The general definition for the CCV becomes

$$\gamma_{k_1 k_2 \ldots k_m} = \frac{1}{N} \sum_{i=1}^{N} \frac{E\left[(X_{ik_1} - \mu_{k_1})(X_{ik_2} - \mu_{k_2}) \ldots (X_{ik_m} - \mu_{k_m})\right]}{\mu_{k_1} \mu_{k_2} \ldots \mu_{k_m}}$$

where $\mu_j = \sum_{i=1}^{N} E(X_{ij})/N$. It is readily seen that definition (3), under a random-effect model, is the limiting value of this general definition. The derivation with proper modifications in the following section is still valid for a fixed-effect model with two types of dependencies.

3 Estimators of population size

The concept of sample coverage comes originally from I. J. Good and A. M. Turing (Good, 1953) and it has been applied to the species estimation for heterogeneous communities (Bunge & Fitzpatrick, 1993). This concept has been modified for the general multiple-sample case by Chao & Tsay (1998). In such a case, the sample coverage can be interpreted as a measure of the overlap fraction. As discussed in the Introduction, the overlap information among lists plays an important role in estimating the number of missing animals. The basic idea is that the sample coverage can be well estimated even in the presence of two sources of dependencies. Thus, an estimate of population size can be derived via the relationship between the population size and the sample coverage and CCVs. Chao & Tsay (1998) mainly dealt with two and three-sample cases. We now extend their analysis to an arbitrary number of lists.

There are now t populations and the capture probabilities of the kth population $\prod_k$ are $\{P_{ik}, i = 1, 2, \ldots, N\}, k = 1, 2, \ldots, t$. Let L_j denote 'list j' and define the sample coverage of the combination list of the other $t-1$ lists omitting the kth list with respect to $\prod_k$ as

$$\begin{aligned} C_{\prod_k}\left(\bigcup_{j \neq k} L_j\right) &= \frac{\sum_{i=1}^{N} P_{ik} I[\sum_{j \neq k}^{t} X_{ij} > 0]}{\sum_{i=1}^{N} P_{ik}} \\ &= 1 - \frac{\sum_{i=1}^{N} P_{ik} I[\sum_{j \neq k}^{t} X_{ij} = 0]}{\sum_{i=1}^{N} P_{ik}} \end{aligned} \tag{6}$$

We remark here that when both types of dependencies exist, the probabilities in $\prod_k$ are replaced by $\{E(X_{ik} \mid X_{ij}, j \neq k), i = 1, 2, \ldots, N\}$ and P_{ik} in (6) is replaced by $E(X_{ik} \mid X_{ij}, j \neq k)$.

From equation (6), we define the sample coverage as

$$C = \frac{1}{t} \sum_{k=1}^{t} C_{\prod_k}\left(\bigcup_{j \neq k} L_j\right) \tag{7}$$

Let S_n be the number of individuals that are listed in sample n only (i.e. singletons); that is, for $t = 4$, $S_1 = Z_{1000}$, $S_2 = Z_{0100}$, $S_3 = Z_{0010}$ and $S_4 = Z_{0001}$. Taking the expectation of C, we have

$$E(C) \approx 1 - \frac{1}{t} \sum_{k=1}^{t} \left[\frac{E[P_k \prod_{j \neq k}(1 - P_j)]}{E(P_k)}\right] = 1 - \frac{1}{t} \sum_{k=1}^{t} \frac{E(S_k)}{E(n_k)} \tag{8}$$

where $P_1, P_2, \ldots, P_t$ are defined prior to equation (2). It follows from (8) that an estimator of sample coverage is

$$\hat{C} = 1 - \frac{1}{t} \sum_{k=1}^{t} \frac{S_k}{n_k} \tag{9}$$

For $t = 4$, the estimator reduces to

$$\hat{C} = 1 - \frac{1}{4} \left(\frac{Z_{1000}}{n_1} + \frac{Z_{0100}}{n_2} + \frac{Z_{0010}}{n_3} + \frac{Z_{0001}}{n_4} \right)$$

which is one minus the average of the fraction of singletons. Obviously, singletons cannot contain any overlap information. Therefore, the sample coverage can be interpreted as a measure of overlap. When neither dependence exists, it follows from the definition of C that $C = D/N$, where D denotes the average of overlapped cases, i.e.

$$D = \frac{1}{t} \sum_{k=1}^{t} \left\{ \sum_{i=1}^{N} I\left[\sum_{j \neq k} X_{ij} > 0 \right] \right\} = M - \frac{1}{t} \sum_{k=1}^{t} S_k \tag{10}$$

Here, M denotes the number of different cases identified in at least one of the lists. Hence, in the independent case, a valid estimator is

$$\hat{N}_0 = D/\hat{C} \tag{11}$$

which is a ratio of overlapped cases to an estimated overlap fraction. From (10), we have

$$E(D) = N - \frac{1}{t} \sum_{k=1}^{t} E\left[\prod_{j \neq k} (1 - P_j) \right] \tag{12}$$

Define $H(i,j) = Z_{k_1 k_2 \ldots k_t} I[k_i = 1,\ k_j = +,\ k_n = 0 \text{ for } n \neq i, j]$, and $A(i,j) = H(i,j) + H(j,i)$. For example, $A(1,2) = Z_{1+00} + Z_{+100}$, $A(2,3) = Z_{01+0} + Z_{0+10}$. To obtain the proposed estimator, we need the following three propositions. The reader is referred to Tsay (1997) for proofs.

Proposition 1: Based on (8) and (12), an expansion of $E(D)/E(C)$ gives

$$N = \frac{E(D)}{E(C)} + \frac{1}{tE(C)} \sum_{i=1}^{t} \sum_{j=i+1}^{t} E[A(i,j)]\, \gamma_{ij} + \frac{R}{tE(C)}$$

where R denotes the remainder term and is given by

$$\frac{R}{N} = \sum_{j=1}^{t} \left[\sum_{\substack{k_1 \\ k_1 \neq j}} \Psi_{jk_1} \gamma_{jk_1} - \sum_{\substack{k_1 < k_2 \\ k_1, k_2 \neq j}} \sum \Phi_{jk_1k_2} \gamma_{jk_1k_2} + \ldots + \sum_{\substack{k_1 < \ldots < k_{t-1} \\ k_1, \ldots, k_{t-1} \neq j}} \sum \ldots \sum (-1)^t \Phi_{jk_1 \ldots k_{t-1}} \gamma_{jk_1k_2 \ldots k_{t-1}} \right]$$

and

$$\Psi_{jk_1} = \mu_{k_1} \left[\prod_{r \neq k_1, j} (1 - \mu_r) \right] - N^{-1} E[H(k_1, j)]$$

and

$$\Phi_{jk_1\ldots k_m} = \mu_{k_1}\mu_{k_2}\ldots\mu_{k_m}\left[\prod_{r\neq k_1,\ldots,k_m,j}(1-\mu_r)\right]$$

Proposition 2: The remainder term R in Proposition 1 vanishes under the following situations:

- Neither local dependence nor heterogeneity exists.
- There is no local dependence but heterogeneity only exists in one sample.
- There is no heterogeneity but local dependence occurs only in two samples.
- The underlying model is a heterogeneous random-effect model M_{th} described in equation (4), where $(h_1, h_2, \ldots, h_N)$ are a random sample from a gamma distribution with density $\beta^{\alpha}h^{\alpha-1}\exp(-\beta h)/\Gamma(\alpha)$.

Since the gamma distribution can cover a wide range of types of heterogeneity, as shown in Fisher *et al.* (1943), this is our main motivation for ignoring the remainder term R. In Proposition 1, if R is ignored and the expectations are replaced by the observed or estimated values, then we have the following proposed estimating equation:

$$N = \frac{D}{\hat{C}} + \frac{1}{t\hat{C}}\sum\sum_{i<j}A(i,j)\gamma_{ij} \tag{13}$$

Let $B(i,j) = Z_{k_1k_2\ldots k_t}I[k_i = k_j = 1, k_n = +$ for $n \neq i,j]$. For example, $B(1,2) = Z_{11++}$ and $B(2,3) = Z_{+11+}$. Using the relationship that $\gamma_{ij} = NB(i,j)/[n_in_j] - 1$ and substituting it into (13), an estimator for the t-sample case can be derived as

$$\hat{N} = \left[\frac{D}{\hat{C}} - \frac{1}{t\hat{C}}\sum\sum_{i<j}A(i,j)\right]\left\{1 - \frac{1}{t\hat{C}}\sum\sum_{i<j}\frac{A(i,j)B(i,j)}{n_in_j}\right\}^{-1} \tag{14}$$

A large sample property of the above estimator is given as follows.

Proposition 3: As N is large, we have

$$\frac{\hat{N}}{N} \xrightarrow{P} 1 - \frac{R/N}{W + R/N}$$

where R/N is given in Proposition 1 and

$$W = \sum_{j=1}^{t}\left[\sum_{\substack{k_1<k_2\\k_1,k_2\neq j}}\sum\mu_{k_1}\mu_{k_2}\sigma_{k_1k_2} + \ldots + \sum_{\substack{k_1<\ldots<k_{t-1}\\k_1,\ldots,k_{t-1}\neq j}}\sum\ldots\sum(-1)^t(t-1)\mu_{k_1}\ldots\mu_{k_{t-1}}\sigma_{k_1\ldots k_{t-1}}\right]$$

where $\sigma_{k_1k_2\ldots k_m} = E(P_{k_1}P_{k_2}\ldots P_{k_m})/[\mu_{k_1}\mu_{k_2}\ldots\mu_{k_m}]$. Therefore, in the four situations in Proposition 2, the proposed estimator (14) is consistent.

A variance estimator and confidence interval associated with the estimator $\hat{N}$ can be constructed by using a non-parametric bootstrap procedure (Efron & Tibshirani, 1993). The reader is referred to Chao & Tsay (1998) for other relevant references. As defined in Section 2, let $\Omega = \{\omega \mid \omega = s_1s_2\ldots s_t\}$ denote all capture histories. The capture counts, $\{Z_\omega \mid \omega\in\Omega\}$, as shown in Darroch *et al.* (1993), are approximately distributed as a multinomial distribution in many situations. A bootstrap sample $\{Z^{\star}_{\omega} \mid \omega\in\Omega\}$ is generated from a multinomial distribution with cell

total $\hat{N}$ and cell probabilities $Z_\omega/\hat{N}$ for any observable cell, and $1 - \sum_{w \neq (00\ldots0)} Z_\omega/\hat{N}$ for the missing cell. Then based on the generated observed data $\{Z^\star_\omega \mid \omega \in \Omega \backslash (00\ldots0)\}$, a bootstrap estimate is obtained. After B replications, the bootstrap variance estimator of $\hat{N}$ is simply the sample variance of those B bootstrap estimates. A log-transformation (Chao, 1989) can then be used to construct an associated confidence interval based on the estimated bootstrap variance.

Previous simulation studies (Chao *et al.*, 1996; Chao & Tsay, 1998) have shown that the proposed estimator $\hat{N}$ performs well if the sample coverage is relatively high. However, when the sample coverage is low, i.e. there are relatively many singletons, we feel the data do not contain sufficient information to estimate the number of missing. In this case, the undercount and population size cannot be estimated precisely due to insufficient overlap. Consequently, a large standard error is usually associated with the proposed estimator. Previous studies for $N = 200$ have suggested that the estimated sample coverage should be at least 55% to ensure that the estimate is reliable. For other population sizes, a practical data-dependent guideline can be determined from the estimated bootstrap s.e. associated with the estimator. That is, the estimate is unavoidably useless if the estimated s.e. becomes unacceptable (say, it exceeds one-third of the population size estimate). When this happens, we suggest the following one-step estimator, which is obtained by one iteration step for N from equation (13). The one-step estimator turns out to be

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{t\hat{C}} \sum\sum_{i<j} A(i,j)\,\hat{\gamma}_{ij} \tag{15}$$

where

$$\hat{\gamma}_{ij} = \frac{B(i,j)}{n_i n_j}\left[\frac{D}{\hat{C}} + \frac{1}{t\hat{C}} \sum\sum_{r<s} A(r,s)\left(\frac{D}{\hat{C}}\frac{B(r,s)}{n_r n_s} - 1\right)\right] - 1 \tag{16}$$

Based on (13), the one-step estimator can usually be regarded as a lower (upper) bound for positively (negatively) dependent samples. A standard error estimate and confidence interval can be similarly constructed by the bootstrap procedure as described before.

4 Examples

4.1 Hepatitis A virus data (3-sample, low overlap)

The purpose of this study was to estimate the number of people who were infected by the hepatitis A virus (HAV) in an outbreak in northern Taiwan. The outbreak occurred from April to July 1995, mainly in and around a technical college with approximately 5000 students. Our data are restricted to those records of students of that college. In total, 271 cases were reported from the following three sources. (1) P-list: records based on a serum test taken by the Institute of Preventive Medicine, Department of Health of Taiwan. There were 135 ascertained cases. (2) Q-list: records provided by the National Quarantine Service. The records included 122 cases reported by doctors in local hospitals. (3) E-list: records based on questionnaires conducted by epidemiologists. Cases were either confirmed by serum test or identified by symptom combinations. A total of 126 cases were in the E-list.

TABLE 1. Data on hepatitis A virus and diabetes

Hepatitis A List			
P	Q	E	Data
1	1	1	$Z_{111} = 28$
1	1	0	$Z_{110} = 21$
1	0	1	$Z_{101} = 17$
1	0	0	$Z_{100} = 69$
0	1	1	$Z_{011} = 18$
0	1	0	$Z_{010} = 55$
0	0	1	$Z_{001} = 63$
0	0	0	—

Diabetes list				
1	2	3	4	Data
1	1	1	1	$Z_{1111} = 58$
1	1	1	0	$Z_{1110} = 157$
1	1	0	1	$Z_{1101} = 18$
1	1	0	0	$Z_{1100} = 104$
1	0	1	1	$Z_{1011} = 46$
1	0	1	0	$Z_{1010} = 650$
1	0	0	1	$Z_{1001} = 12$
1	0	0	0	$Z_{1000} = 709$
0	1	1	1	$Z_{0111} = 14$
0	1	1	0	$Z_{0110} = 20$
0	1	0	1	$Z_{0101} = 7$
0	1	0	0	$Z_{0100} = 74$
0	0	1	1	$Z_{0011} = 8$
0	0	1	0	$Z_{0010} = 182$
0	0	0	1	$Z_{0001} = 10$
0	0	0	0	—

The data are given in Table 1. Several loglinear models were fitted to these data. Except for the saturated model, the loglinear models, which do not take into account heterogeneity (i.e. models (PE, Q), (QE, P), (PQ, E), (PQ, QE), (PQ, PE) and (QE, PE)), do not fit the data well. The quasi-symmetric and three partial quasi-symmetric models (see Darroch *et al.*, 1993), which incorporate heterogeneity effects, fit the data well. The four adequate models produce approximately the same estimates, 1300, with an approximate estimated s.e. of 520. This relatively large estimated s.e. shows the data are actually insufficient to model the heterogeneity effect.

We now illustrate our estimation procedure. From equations (9) and (10), the sample coverage is estimated to be $\hat{C} = 51.27\%$ and the average of the overlapped cases is $D = 208.67$. If we incorrectly assume that the three samples are independent, then from (11) an estimate would give $\hat{N}_0 = D/\hat{C} = 407$, which is slightly larger than the estimate of 388 based on an independent loglinear model. The estimator (14) yields an estimate of $\hat{N} = 971$, but a large estimated bootstrap s.e. 925 renders the estimate useless. The estimated s.e. was calculated by using a bootstrap method based on 1000 replications. We feel these data, with an estimated sample coverage of 51%, do not contain enough information to correct for undercount. The proposed one-step estimator in (15) gives an estimate of $\hat{N}_1 = 508$ with an estimated

s.e. of 40 using 1000 bootstrap replications, which subsequently produce a 95% confidence interval of (407, 591).

We remark that, in December 1995, the National Quarantine Service of Taiwan conducted a screen serum test for the HAV antibody for all students at the college at which the outbreak of the HAV occurred. After suitable adjustments, they have concluded that the final figure of the number infected is about 545. Thus, this example presents a very valuable data set with the advantage of a known true parameter. Our estimator $\hat{N}_1$ provides a satisfactory lower bound.

4.2 Diabetes data (4-sample, high overlap)

The data given in Table 1 on diabetes were collected and discussed in Bruno *et al.* (1994) and IWGDMF (1995a). The purpose of collecting these data was to estimate the number of diabetes patients in a community in Italy based on the following four lists: diabetic clinic and/or family physician visits (list-1, 1754 cases), hospital discharges (list-2, 452 cases), prescriptions (list-3, 1135 cases) and regent strips and insulin syringes (list-4, 173 cases). In total, 2069 cases were identified. Bruno *et al.* (1994) found that the loglinear model (12, 13, 23, 24, 34) fits the data well and obtained an estimate of 2771 with a 95% confidence interval of (2492, 3051). When they further stratified for the pattern of treatment (dietary control, hypoglycaemia agents and insulin), an estimate of 2586 was obtained with a 95% interval of (2341, 2830). IWGDMF (1995a) analysed the data by including heterogeneity terms to several proper loglinear models and selected the final model by Akaike Information Criterion. They obtained an estimate of 2834 based on the stratified data.

The sample coverage for these data is estimated to be 80.35%. Since the coverage estimate is relatively high, we would prefer the use of $\hat{N}$ given in (14). The proposed $\hat{N} = 2609$ (s.e. $= 81$). The s.e. estimate was obtained using 1000 bootstrap replications. The corresponding 95% confidence interval for $\hat{N}$ is (2472, 2792). If the data are analysed within each stratum, we have the sum of the three estimates as $\hat{N} = 2559$, which is very close to our unstratified results.

5 Discussion

We have presented in this paper a sample coverage approach to estimate the size of a target population in epidemiological applications. This approach uses the overlap information (i.e. estimated sample coverage) and dependence measures (i.e. CCV estimates) to correct for undercount.

There are several basic assumptions that should be fulfilled for the capture–recapture methods. In addition to the closure assumption (i.e. the population size remains unchanged in the study period), a basic assumption is an explicit definition or interpretation of the 'target population'. Gutteridge & Collin (1994), in a prevalence study of physical disability, reported that two sources might have different interpretations of disability and its severity. Thus, the 'target population' for two sources might become inconsistent. One of the other basic assumptions is that all identification 'marks' should be correctly recorded and matched. Although in most epidemiological studies this assumption can obviously be fulfilled, in reality it might be an impediment in developing countries, as indicated by Black & McLarty (1994).

An implicit assumption is that the joint 'capture' probability for any individual in *all* lists should be positive, so that overlap information can be obtained. In other words, any individual unascertained in any list is purely a 'random zero' (missing due to small chance), not a 'structure zero' (missing due to impossibility). If some cases are systematically missed by one or more sources, precluding overlap information being obtained, then those 'uncatchable' individuals cannot be included in our target population. An extreme example of this situation can be seen where the first list collected cases from a certain area whereas the other list collected cases from another disjoint area. There is no way to collect source intersection, and hence, capture–recapture models are not applicable to estimate the total number of cases in the area involved. The estimating target is actually the size of those jointly 'ascertainable' or 'catchable' individuals by *all* sources.

We reiterate that a serious limitation of the capture–recapture methods is that sufficiently high overlap information is required to obtain a precise population size estimate and to model dependencies among samples. Coull & Agresti (1999) demonstrated that the likelihoods under some random-effect models for sparse information might become quite flat and the resulting estimates might become unstable; the associated confidence intervals also tend to be wide. In the HAV example discussed in Section 4.1, we have also shown in a sample coverage approach that a large variation might be associated with the resulting estimator due to insufficient overlap.

An interactive *S-plus* CARE (for capture–recapture) program that calculates the proposed estimators and their associated s.e., as well as confidence intervals up to six lists, is available from the second author upon request. It will also be available soon from the website at http://www.stat.nthu.edu.tw/~chao/.

Acknowledgements

The authors thank the reviewer for helpful comments, which were very useful in revising the manuscript. This work is based on part of the Ph.D. Thesis of the first author under the supervision of the second author and was supported by the National Science Council of Taiwan under contract NSC 87–2118-M007-010.

REFERENCES

AGRESTI, A. (1994) Simple capture–recapture models permitting unequal catchability and variable sampling effort, *Biometrics*, 50, pp. 494–500.

BLACK, J. F. P. & MCLARTY, D. G. (1994) Capture–recapture technique: difficult to use in developing countries (Letter), *British Medical Journal*, 308, p. 531.

BOWMAN, K. O. & SHENTON, L. R. (1988) *Properties of Estimators for the Gamma Distribution* (New York, Marcel Dekker).

BRUNO, G. B., BIGGERI, A., LAPORTE, R. E., MCCARTY, D., MERLETTI, F. & PAGONO, G. (1994) Application of capture–recapture to count diabetes, *Diabetes Care*, 17, pp. 548–556.

BUNGE, J. & FITZPATRICK, M. (1993) Estimating the number of species: a review, *Journal of the American Statistical Association*, 88, pp. 364–373.

CHAO, A. (1989) Estimating population size for sparse data in capture–recapture experiments, *Biometrics*, 45, pp. 427–438.

CHAO, A. (1998) Capture–recapture. In *Encyclopedia of Biostatistics*, P. ARMITAGE and T. COLTON (eds) (New York, Wiley) pp. 482–486.

CHAO, A. (2000) An overview of closed capture–recapture models, to appear in *Journal of Agricultural, Biological and Environmental Statistics*.

Chao, A. & Tsay, P. K. (1998) A sample coverage approach to multiple-system estimation with application to census undercount, *Journal of the American Statistical Association*, 93, pp. 283–293.

Chao, A., Tsay, P. K., Shau, W.-Y. & Chao, D.-Y. (1996) Population size estimation for capture–recapture models with applications to epidemiological data, *Proceedings of Biometrics Section, American Statistical Association*, pp. 108–117.

Cormack, R. M. (1989) Log-linear models for capture–recapture, *Biometrics*, 45, pp. 395–413.

Coull, B. A. & Agresti, A. (1999) The use of mixed logit models to reflect heterogeneity in capture–recapture studies, *Biometrics*, 55, pp. 294–301.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. & Junker, B. W. (1993) A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability, *Journal of the American Statistical Association*, 88, pp. 1137–1148.

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

Fienberg, S. E. (1972) The multiple recapture census for closed populations and incomplete 2^k contingency tables, *Biometrika*, 59, pp. 591–603.

Fisher, R. A., Corbet, A. S. & 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, pp. 42–58.

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

Gutteridge, W. & Collin, C. (1994) Capture–recapture technique: quick and cheap (Letter), *British Medical Journal*, 308, p. 531.

Hook, E. B. & Regal, R. R. (1995) Capture–recapture methods in epidemiology: methods and limitations, *Epidemiological Reviews*, 17, pp. 243–264.

International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995a) Capture–recapture and multiple-record systems estimation I: history and theoretical development, *American Journal of Epidemiology*, 142, pp. 1047–1058.

International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995b) Capture–recapture and multiple-record systems estimation II: applications in human diseases, *American Journal of Epidemiology*, 142, pp. 1059–1068.

Lee, S.-M. & Chao, A. (1994) Estimating population size via sample coverage for closed capture–recapture models, *Biometrics*, 50, pp. 88–97.

Pollock, K. H. (1991) Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future, *Journal of the American Statistical Association*, 86, pp. 225–238.

Rasch, G. (1961) On general laws and the meaning of measurement in psychology. *In:* J. Neyman (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 321–333 (Berkeley, University of California Press).

Sanathanan, L. (1972) Models and estimation methods in visual scanning experiments, *Technometrics*, 14, pp. 813–829.

Schwarz, C. J. & Seber, G. A. F. (1999) A review of estimating animal abundance III, *Statistical Science*, 14, pp. 427–456.

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

Seber, G. A. F. (1986) A review of estimating animal abundance, *Biometrics*, 42, pp. 267–292.

Seber, G. A. F. (1992) A review of estimating animal abundance II, *International Statistical Review*, 60, pp. 129–166.

Sekar, C. & Deming W. E. (1949) On a method of estimating birth and death rates and the extent of registration, *Journal of the American Statistical Association*, 44, pp. 101–115.

Tsay, P. K. (1997) Comparison of log-linear models and sample coverage methods in capture–recapture experiments, Ph.D. Thesis, National Tsing Hua University, Taiwan.

Wittes, J. T. (1974) Applications of a multinomial capture–recapture method to epidemiological data, *Journal of the American Statistical Association*, 69, pp. 93–97.

Wittes, J. T. & Sidel, V. W. (1968) A generalization of the simple capture–recapture model with applications to epidemiological research, *Journal of Chronic Diseases*, 21, pp. 287–301.