

TUTORIAL IN BIOSTATISTICS

The applications of capture-recapture models to epidemiological data

Anne Chao^{1,*}, P. K. Tsay¹, Sheng-Hsiang Lin¹, Wen-Yi Shau² and Day-Yu Chao³

¹*Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan*

²*Graduate Institute of Clinical Medicine, National Taiwan University, Taipei, Taiwan*

³*Graduate Institute of Epidemiology, National Taiwan University, Taipei, Taiwan*

SUMMARY

Capture-recapture methodology, originally developed for estimating demographic parameters of animal populations, has been applied to human populations. This tutorial reviews various closed capture-recapture models which are applicable to ascertainment data for estimating the size of a target population based on several incomplete lists of individuals. Most epidemiological approaches merging different lists and eliminating duplicate cases are likely to be biased downwards. That is, the final merged list misses those who are in the population but were not ascertained in any of the lists. If there are no matching errors, then the duplicate information collected from a capture-recapture experiment can be used to estimate the number of missed under proper assumptions. Three approaches and their associated estimation procedures are introduced: ecological models; log-linear models, and the sample coverage approach. Each approach has its unique way of incorporating two types of source dependencies: local (list) dependence and dependence due to heterogeneity. An interactive program, CARE (for capture-recapture) developed by the authors is demonstrated using four real data sets. One set of data deals with infection by the acute hepatitis A virus in an outbreak in Taiwan; the other three sets are ascertainment data on diabetes, spina bifida and infants' congenital anomaly discussed in the literature. These data sets provide examples to show the usefulness of the capture-recapture method in correcting for under-ascertainment. The limitations of the methodology and some cautionary remarks are also discussed. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

The purpose of many epidemiological surveillance studies is to estimate the size of a population by merging several incomplete lists of names in the target population. Some examples are as follows:

1. An outbreak of the hepatitis A virus (HAV) occurred in and around a college in northern Taiwan from April to July 1995 [1]. Cases of students in that college were

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

Contract/grant sponsor: National Science Council of Taiwan; contract/grant numbers: NSC 87-2118-M007-101, NSC89-2118-M007-006

Received February 2000

Accepted April 2001

ascertained by three sources: (i) P-list, records based on a serum test taken by the Institute of Preventive Medicine, Department of Health of Taiwan – there were 135 identified cases; (ii) Q-list, local hospital records reported by the National Quarantine Service – 122 cases were found; (iii) E-list, records collected by epidemiologists – there were 126 cases. Merging the three lists by eliminating duplicate records resulted in 271 ascertained cases. This data set has the advantage of a known true number of infected because a screen serological check for all students was conducted after the three surveys. In Section 5, we use this data set to show the need of correction for undercount.

2. Hook *et al.* [2] and Regal and Hook [3] presented a data set on spina bifida collected in New York State from 1969–1974. Three lists were collected from birth certificates, death certificates and medical rehabilitation files. There were 513, 207 and 188, cases, respectively, on the three lists and in total 626 ascertained cases. A method of estimating the number of missed cases was discussed by the above authors to assess the completeness of the survey and to accurately estimate the prevalence rate.
3. Bruno *et al.* [4] collected a data set on diabetes in a community in Italy based on the following four records: diabetic clinic and/or family physician visits (1754 cases); hospital discharges (452 cases); prescriptions (1135 cases), and purchases of reagent strips and insulin syringes (173 cases). A total of 2069 cases were identified. Despite the active identification, Bruno *et al.* concluded that there were still some people who could not be identified. The purpose was then to estimate the number of missing diabetes patients and to adjust for undercount.
4. Wittes *et al.* [5] and Fienberg [6] analysed a multiple lists data in an attempt to estimate the number of infants born with a specific congenital anomaly in Massachusetts during a fixed time interval. Five distinct types of sources were considered: obstetric records (183 cases); other hospital records (215 cases); list maintained by state Department of Public Health (36 cases); list maintained by state Department of Mental Health (263 cases), and records by special schools (252 cases). The total number of cases identified was 537. Previous studies [5–7] have shown that inferences regarding the number of missing cases can be made under proper assumptions.

These four data sets will be discussed in Section 5 as illustrations. As indicated by the International Working Group of Disease Monitoring and Forecasting (IWGDMF) [8, 9], similar examples arise in various disease categories such as cancer, drug use, infectious diseases, injuries and others.

The adjustment of undercount has an analogue in the biological sciences: estimating the number of unseen animals in a closed population. Here a closed population means that there is no birth, death or migration so that the population size is a constant during the study period. The estimation of population size is a classical problem and has been extensively discussed in the literature. Although doing complete census counts for some clustered animal populations, especially in colonies, is not impossible, biologists have long realized that complete enumeration is nearly an unattainable ideal for most mobile populations and that proper adjustment for undercount is needed.

Since most animals cannot be drawn like balls in an urn or numbers on a list, traditional random sampling techniques are not easily applicable to biological surveys. To tackle the above undercount problem, special types of sampling schemes have been developed.

Capture-recapture sampling has been widely used to adjust for undercount in the biological sciences [10–13]. It would be unnecessary and almost impossible to count every animal in a closed population in order to obtain an accurate estimate of population size. The recapture information (that is, source-overlap information or source intersection) collected by marking or tagging can be used to estimate the number of missing under proper assumptions.

In contrast, epidemiologists have attempted to enumerate all relevant cases to obtain the prevalence rates for various diseases. Some studies based on health records did ascertain almost all patients [14]. However, as Hook and Regal [15] indicated, most prevalence surveys merging several records of lists are likely to miss some cases and thus be biased to underestimate. There is relatively little literature in the health sciences on the assessment of the completeness of these types of surveys or on the adjustment for under-ascertainment. Therefore, as commented by LaPorte *et al.* [16], people know more about the number of animals than the count of diseases. In this tutorial, we focus on the applications of the closed capture-recapture methods to such epidemiological studies. In the same way that ecologists and biologists count animals, we introduce in this paper the use of capture-recapture models to count human populations.

In Section 2, the capture-recapture technique and its adaptation for use in human populations are reviewed. In Section 3, the general data structure and the concept of two types of dependencies are formulated. Section 4 introduces three closed capture-recapture models which are applicable to ascertainment data in epidemiology. A program developed by the authors is demonstrated to estimate population size in Section 5 using the four data sets described in the beginning of this section. Some remarks about the limitations of capture-recapture methods are discussed in Section 6.

2. CAPTURE-RECAPTURE

2.1. *Animal populations*

In an animal capture-recapture experiment, traps or nets are placed in the study area and the population is sampled several times. At the first trapping sample a number of animals are captured; the animals are uniquely tagged or marked and released into the population. Then at each subsequent trapping sample we record and attach a unique tag to every unmarked, record the capture of any animal that has been previously tagged, and return all animals to the population. At the end of the experiment the complete capture history for each animal is known. Such experiments are also called mark-recapture, tag-recapture and multiple-record system.

According to Seber [10], the first use of the capture-recapture technique can be traced back to Laplace, who used it to estimate the population size of France in 1786. The earliest applications to ecology include Petersen's and Dahl's work on fish populations and Lincoln's use of band returns to estimate waterfowl in 1930. More sophisticated statistical theory and inference procedures have been proposed since Darroch's paper [17], in which the mathematical framework of this topic was founded. Seber [10–12] and Schwarz and Seber [13] provided comprehensive reviews on the methodology and applications.

The models are generally classified as either closed or open. As stated in Section 1, the size of a closed population is assumed to be a constant throughout the study period. The closure

assumption is usually valid for data collected in a relatively short time during a non-breeding season. In an open model, births and deaths are allowed and thus the size varies with time. Open models are usually used for modelling the data from long-term investigations of animals or migratory birds. The capture-recapture technique has also been adopted to estimate survival rates, birth rates and migration rates.

We restrict this paper to closed models because the population size for most epidemiological studies can be assumed to be approximately a constant in a fixed time period. We also assume that individuals do not lose their marks and all marks are recorded and matched correctly. Marking or tagging is mainly used to distinguish individuals, and thus the recapture information can be applied to evaluate the degree of undercount. When recaptures in the subsequent samples are few, we know intuitively for independent samples that the size is much higher than the number of distinct captures. On the other hand, if the recapture rate is high, then we are likely to have caught most of the animals.

We remark that overlap information in some studies can be obtained without marking. For example, when the main purpose in regular bird counts is to count the number of species, marking or tagging is not needed because species identification for each sighting would suffice for providing species overlap and that information can be properly used to estimate the number of undiscovered species.

2.2. *Human populations*

The capture-recapture technique originally developed for animal studies has been applied to human populations under the term 'multiple-record system' [6, 8, 9, 18–21]. The special two-sample cases are often referred to as the 'dual-system' or 'dual-record system'. For ascertainment data, if each list is regarded as a trapping sample and identification numbers and/or names are used as tags or marks, then this framework is similar to a capture-recapture set-up for wildlife estimation. Thus 'capture in a sample' corresponds to 'being recorded or identified in a list', and 'capture probability' becomes 'ascertainment probability'.

The earliest references to the application of the capture-recapture techniques to health science included the pioneering paper by Sekar and Deming [18] for two samples, Wittes and Sidel [19] for three samples, Wittes [20] for four samples, Wittes *et al.* [5] and Fienberg [6] for five samples. Epidemiologists recently have shown renewed and growing interest in the use of the capture-recapture models [16, 21]. Hook and Regal [22] also suggested the use of capture-recapture models even for apparently exhaustive surveys. Hook and Regal [23], IWGDMF [8, 9] and Chao [24] provided overviews of the applications of the capture-recapture models specifically to epidemiological data. However, some critical comments and practical concerns about the use of capture-recapture models have been expressed by several authors [14, 25–28]. We will address their main concerns in Section 6.

Three main differences between wildlife and human applications are noted:

- (i) There are usually more trapping samples in wildlife studies, whereas in most epidemiological surveys only two to four lists are available.
- (ii) There is a natural time ordering in animal experiments, but generally no such order exists in epidemiological lists, or the order may vary with individuals.
- (iii) In animal studies, identical trapping methods are usually used in all trapping samples. Hence animals' behavioural response to capture is often present and is modelled in analysis. In human populations, different types of ascertainment sources are utilized to

Table I. Individual ascertainment data for three lists.

Individual	List 1	List 2	List 3
1	X_{11}	X_{12}	X_{13}
2	X_{21}	X_{22}	X_{23}
...
M	X_{M1}	X_{M2}	X_{M3}
$M + 1$	0	0	0
...
N	0	0	0

search all individuals. The behavioural response due to the sampling scheme is not commonly considered in models.

Researchers in wildlife and human populations have developed models and methodologies along separate lines. Three of these approaches will be introduced after the formulation of the data structure and the concept of dependence among samples.

3. DATA STRUCTURE AND DEPENDENCE

3.1. Data structure

We first introduce some notation. Assume that the true population size is N which is our parameter of interest. The individuals can be conceptually indexed by $1, 2, \dots, N$ and all individuals act independently. Assume that there are t samples (lists, records or sources) and they are indexed by $1, 2, \dots, t$. Presence and absence in any source are denoted by 1 and 0, respectively. For a three-list case as given in Table I, we can use three numbers (each is either 0 or 1) to denote the record of each individual. For example, individual 1 was identified in list 1 only. Then it is associated with the record (100) in the data; individual 2 was identified in all three lists, it is recorded as the record (111). Each individual in a three-list case is associated with one of the following seven possible 'capture histories' or 'ascertainment records': (001); (010); (011); (100); (101); (110), and (111). Suppose there are M identified individuals and $N - M$ uncounted. Without losing generality, assume that these M identified individuals are indexed by $1, 2, \dots, M$. If we augment all the identified records by $N - M$ individuals with history (000) as in Table I, then the ascertainment data for all individuals can be conveniently expressed by an $N \times t$ matrix $X = (X_{ij})$. Here $X_{ij} = 1$ if the i th individual is listed in the j th sample, 0 otherwise. A record (000) means that an individual is not identified in any of the three samples.

The ascertainment data for all identified individuals can be aggregated as a categorical data format as shown in Table II for the HAV data. That is, the frequencies of the same record are grouped. Let Z_{s_1, s_2, \dots, s_t} be the number of individuals with record s_1, s_2, \dots, 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_{001} , Z_{010} , Z_{011} , Z_{100} , Z_{101} , Z_{110} and Z_{111} , where Z_{001} is the number of individuals listed in sample 3 only, Z_{011} is the number of individuals listed in samples 2 and 3 but not in sample 1. A similar interpretation pertains

Table II. Aggregated data on hepatitis A virus.

Hepatitis A list			Data
P	Q	E	
0	0	0	$Z_{000} = ??$
0	0	1	$Z_{001} = 63$
0	1	0	$Z_{010} = 55$
0	1	1	$Z_{011} = 18$
1	0	0	$Z_{100} = 69$
1	0	1	$Z_{101} = 17$
1	1	0	$Z_{110} = 21$
1	1	1	$Z_{111} = 28$

to other capture histories. The missing cell $Z_{000} = N - M$ denotes the uncounted. 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 $Z_{++1} = Z_{001} + Z_{011} + Z_{101} + Z_{111}$, and $Z_{+++} = N$. Let n_j , $j = 1, 2, \dots, t$ be the number of individuals listed in sample j . For $t = 3$, we have $n_1 = Z_{1++}$, $n_2 = Z_{+1+}$, $n_3 = Z_{++1}$.

For the HAV data in Table II, there were 63 people listed in the E-list only, 55 people listed in the Q-list only, and 18 people listed in both lists Q and E but not in the P-list. Similarly, we can interpret the other records. The purpose here is to estimate the number of total individuals (that is, N) who were infected in the outbreak. It is thus equivalent to predicting the number of missed (that is, $Z_{000} = N - M$) by all three sources.

In a typical approach in epidemiology, cases in various lists are merged and any duplicate cases are eliminated. That is, the capture histories in Table I and the categories in Table II are ignored in the analysis and only the final merged list is obtained. This typical approach assumes complete ascertainment and does not correct or adjust for under-ascertainment. However, there were non-negligible uncounted cases in many epidemiological surveillance studies. For example, before the screen serological check for all students of that college, epidemiologists suspected that the observed number of cases (271) in Table II considerably undercounted the true number of infected and an evaluation of the degree of undercount was needed [1, 29].

3.2. Dependence among samples

A crucial assumption in the traditional statistical approach is that the samples are independent. Since individuals can be cross-classified according to their presence or absence in each list, the dependence for any two samples is usually interpreted from a 2×2 categorical data analysis in human applications. In animal studies, traditional ‘equal-catchability assumption’ is even more restrictive, that is, in each fixed sample all animals including marked and unmarked have the same capture probability. (Equal catchability assumption implies independence among samples but the reverse is not true; see Section 4.3.) Non-independence or unequal catchabilities may be caused by the following two sources:

- (i) Local dependence (also called list dependence or local list dependence) within each individual; conditional on any individual, the inclusion in one source has a direct causal

effect on his/her inclusion in other sources. That is, the response of a selected individual to one source depends on his/her response to the other sources. For example, the probability of going to a hospital for treatment for any individual depends on his/her result on the serum test of the HAV. The ascertainment of the serum sample and that of the hospital sample then becomes dependent. We remark that 'local independence' has been a fundamental assumption in many statistical methodologies [30].

- (ii) Heterogeneity between individuals; even if the two lists are independent within individuals, the ascertainment of the two sources may become dependent if the capture probabilities are heterogeneous among individuals. This phenomenon is similar to Simpson's paradox in categorical data analysis. That is, aggregating two independent 2×2 tables might result in a dependent table. Hook and Regal [31] presented an interesting epidemiological example.

These two types of dependencies are usually confounded and cannot be easily disentangled in a data analysis. Lack of independence leads to a bias (called 'correlation bias' in census undercount estimation [32]) for the usual estimator which assumes independence. We use a two-sample animal experiment to explain the direction of the bias. Assume that a first sample of n_1 animals is captured. Therefore, the marked rate in the population is n_1/N . A second sample of n_2 animals is subsequently drawn and there are m_2 (that is, Z_{11} in our notation for grouped data) previously marked. The capture rate for the marked (recapture rate, overlap rate) in the second sample can be estimated by m_2/n_2 . If the two samples are independent, then the recapture rate should be approximately equal to the marked rate in the population. Therefore we have $m_2/n_2 = n_1/N$, which yields an estimate of the population size under independence: $\hat{N}_P = n_1 n_2 / m_2$ (the well-known Petersen estimator or dual-system estimator). However, if the two samples are positively correlated, then those individuals captured in the first sample are more easily captured in the second sample. The recapture rate in the second sample tends to be larger than the marked rate in the population. That is, we would expect that $m_2/n_2 > n_1/N$, which yields $N > n_1 n_2 / m_2$. As a result, Petersen's estimator underestimates the true size if both samples are positively dependent. Conversely, it overestimates for negatively dependent samples. A similar argument is also valid for a general number of samples. That is, a higher (lower) overlap rate is observed for positively (negatively) dependent samples, which implies fewer (more) estimated missing cases. Therefore, a negative (positive) bias exists for any estimator which assumes independence.

When only two lists are available, three cells are observable: people identified in list 1 only; people identified in list 2 only, and people listed in both. However, there are four parameters: N , two mean capture probabilities and a dependence measure. The data are insufficient for estimating dependence unless additional covariates are available. All existing methods unavoidably encounter this problem and adopt the independence assumption. This independence assumption has become the main weak point in the use of the capture-recapture method for two lists.

A variety of models incorporating dependence among samples have been proposed in the literature. We will review three classes of models: ecological models; log-linear models, and the sample coverage approach. The latter two approaches can be used to provide estimates for some ecological models, but they are considered separately because of their different ways of dealing with dependence.

Table III. Two types of capture probabilities for ecological models.

Model	Multiplicative model in log-linear form	Logistic model
\mathbf{M}_{tbh}	$\log(P_{ij}) = \alpha_i + \beta_j + \gamma Y_{ij}$	$\text{logit}(P_{ij}) = \alpha_i + \beta_j + \gamma Y_{ij}$
\mathbf{M}_{bh}	$\log(P_{ij}) = \alpha_i + \gamma Y_{ij}$	$\text{logit}(P_{ij}) = \alpha_i + \gamma Y_{ij}$
\mathbf{M}_{tb}	$\log(P_{ij}) = \beta_j + \gamma Y_{ij}$	$\text{logit}(P_{ij}) = \beta_j + \gamma Y_{ij}$
\mathbf{M}_{th}	$\log(P_{ij}) = \alpha_i + \beta_j$	$\text{logit}(P_{ij}) = \alpha_i + \beta_j$ (Rasch model)
\mathbf{M}_h	$\log(P_{ij}) = \alpha_i$	$\text{logit}(P_{ij}) = \alpha_i$
\mathbf{M}_b	$\log(P_{ij}) = \alpha + \gamma Y_{ij}$ ($\alpha_i \equiv \alpha$)	$\text{logit}(P_{ij}) = \alpha + \gamma Y_{ij}$ ($\alpha_i \equiv \alpha$)
\mathbf{M}_t	$\log(P_{ij}) = \beta_j$	$\text{logit}(P_{ij}) = \beta_j$

4. MODELS AND ESTIMATORS

4.1. Ecological models

This approach specifies various forms of capture probabilities based on empirical investigations of animal ecology. Although most authors in this field did not aim to model dependence between samples, dependence is induced when some special types of capture probabilities are formulated. Two types of probabilities have been proposed: multiplicative and logistic.

The multiplicative class of models was first proposed by Pollock [33] and was fully discussed in the two monographs by Otis *et al.* [34] and White *et al.* [35]. Three sources of variation in capture probability are considered: time-varying, behavioural response, and heterogeneity. The corresponding models are denoted by model \mathbf{M}_t , \mathbf{M}_b and \mathbf{M}_h , respectively. Various combinations of these three types of unequal capture probabilities (that is, models \mathbf{M}_{tb} , \mathbf{M}_{th} , \mathbf{M}_{bh} and \mathbf{M}_{tbh}) are also considered. These models specify the conditional probability of capturing the i th animal in the j th sample given the capture history of samples $1, 2, \dots, j - 1$. Denote this conditional probability by P_{ij} for notational simplicity. A multiplicative form of model \mathbf{M}_{tbh} is

$$P_{ij} = \begin{cases} p_i e_j & \text{until first capture} \\ \phi p_i e_j & \text{for any recapture} \end{cases}$$

where $0 < p_i e_j, \phi p_i e_j < 1$. Here the parameters $\{e_1, e_2, \dots, e_t\}$, $\{p_1, p_2, \dots, p_N\}$ and ϕ are used to model the time effects, individual heterogeneity and the behavioural response to capture, respectively. Reparameterize $\alpha_i = \log(p_i)$, $\beta_j = \log(e_j)$, $\gamma = \log(\phi)$, and define $Y_{ij} = I$ [the i th animal has been captured before the j th sample] where $I(A)$ is an indicator function of the event A , that is, $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. The time-dependent variable Y_{ij} is used to denote the prior capture history of individual i for sample j . Then the multiplicative type of probability can be conveniently expressed as the following log-linear form:

$$\log(P_{ij}) = \alpha_i + \beta_j + \gamma Y_{ij} \tag{1}$$

All submodels can be easily formulated as shown in Table III.

Logistic types of models have also been proposed [36–38] in the literature and the form of a logistic model \mathbf{M}_{tbh} is

$$\text{logit}(P_{ij}) \equiv \log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \alpha_i + \beta_j + \gamma Y_{ij} \tag{2}$$

which is equivalent to $P_{ij} = \exp(\alpha_i + \beta_j + \gamma Y_{ij}) / [1 + \exp(\alpha_i + \beta_j + \gamma Y_{ij})]$. All submodels of the logistic form are also shown in Table III. The two types of models and submodels can thus be integrated in a unified expression and they only differ in the link function. The logistic model \mathbf{M}_{th} , that is, $\gamma = 0$ in equation (2), is the well-known Rasch model [39], which plays an important role in educational statistics and in the analysis of survey data.

To reduce the number of parameters and to remove the non-identification caused by the numerous parameters $\{\alpha_1, \alpha_2, \dots, \alpha_N\}$, one of the following is usually assumed: (i) they are a random sample from a parametric distribution characterized by a few parameters, for example, beta or normal distributions (random-effects model) [40, 41]; (ii) they can be partitioned as two or more homogeneous groups (latent class model or mixture model) [42]; (iii) They are determined by the first two moments totally (usually, mean and coefficient of variation) [43].

For both types of ecological models, the samples are independent only for model \mathbf{M}_t . Local list dependence is present for models \mathbf{M}_b and \mathbf{M}_{tb} ; heterogeneity arises for model \mathbf{M}_h ; both types of dependencies exist for models \mathbf{M}_{bh} and \mathbf{M}_{tbh} . As noted in the reviews [10–13], estimators for various models may be found in the literature. These estimators rely on a wide range of methodologies.

An analysis of the closed model can be performed using a comprehensive computer program, CAPTURE [44]. The program is readily available from Gary White's website (<http://www.cnr.colostate.edu/~gwhite/software.html>). See Section 5 for another program calculating some recent estimates and their standard errors.

As indicated in Section 2.2, there is a natural time ordering in ecological experiments. Models with behavioural response (that is, models \mathbf{M}_b , \mathbf{M}_{tb} , \mathbf{M}_{tb} and \mathbf{M}_{tbh}) allow the capture of any individual depending on its own 'previous' capture histories and thus ordering is implicitly involved in these four models. Meanwhile, almost all estimation procedures derived under these models depend on the ordering of the lists. Since there is usually no sequential time order in ascertainment lists or sources, any model involving behavioural response or any estimator depending on the list order has limited use in epidemiology. Therefore, only models \mathbf{M}_t , \mathbf{M}_h and \mathbf{M}_{th} are potentially useful for our applications.

For model \mathbf{M}_t , the multiplicative model and the logistic model are equivalent because either model is only a reparameterization of the other. The MLE (conditional on the total number of identified) of population size under model \mathbf{M}_t is identical to that obtained under the independent log-linear model [10, 17, 45]. Therefore, model \mathbf{M}_t will be included in the log-linear model approach in Section 4.2.

For model \mathbf{M}_h , both types of models are also equivalent. Model \mathbf{M}_h assumes that each individual has its own unique probability that remains constant over samples. Thus it is usually applied to the situations where similar types of trapping methods are taken. Model \mathbf{M}_{th} extends model \mathbf{M}_h by allowing for various sampling efforts or time effects. A widely used estimator for model \mathbf{M}_h is the jackknife estimator proposed by Burnham and Overton [46]. The jackknife estimator is a linear function of the capture frequencies $\{f_1, f_2, \dots, f_i\}$ where f_k denotes the number of animals captured exactly k times in the t samples. The jackknife estimator is invariant to the order of the lists. As indicated by Otis *et al.* in their monograph (reference [34], p. 34), the bias of the jackknife is within a tolerable range if the number of trapping samples is greater than five.

The logistic model \mathbf{M}_{th} (that is, the Rasch model) is equivalent to a quasi-symmetric log-linear model with some moment constraints [32, 47]. Hence it will be discussed in Section 4.2. For multiplicative models \mathbf{M}_{th} and \mathbf{M}_h , Chao *et al.* [48] and Lee and Chao [43] proposed

some estimators focusing on animal data, but those estimators are suggested for use when the number of samples is sufficiently large (say at least five, as in the jack-knife method). Therefore, except for the Rasch model, heterogeneous ecological models are recommended only when at least five lists are available. An example with five samples is given in Section 5.4 for illustration.

4.2. Log-linear models

The log-linear models have been proposed [6, 40, 45, 47, 49, 50] to handle dependence among samples. This approach is well discussed in the two review papers by IWGDMF [8, 9], thus we only provide a brief description here. In this 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 log-linear models are fitted to the observed cells. How well a model fits the data is assessed using the deviance statistic and a model is usually selected based on the Akaike information criterion. The chosen model is then projected onto the unobserved cell by assuming that there is no t -sample interaction. The two types of dependencies can be modelled by including some specific interactions or common interaction in the models.

We use the three-list data for illustration. The log-linear approach models the logarithm of the expected value of each observable category, that is, the most general model is

$$\begin{aligned} \log E(Z_{ijk}) = & u + u_1 I(i=1) + u_2 I(j=1) + u_3 I(k=1) + u_{12} I(i=j=1) + u_{13} I(i=k=1) \\ & + u_{23} I(j=k=1) + u_{123} I(i=j=k=1) \end{aligned} \quad (3)$$

That is, $\log E(Z_{001}) = u + u_3$, $\log E(Z_{010}) = u + u_2$, $\log E(Z_{100}) = u + u_1$, $\log E(Z_{110}) = u + u_1 + u_2 + u_{12}$, $\log E(Z_{101}) = u + u_1 + u_3 + u_{13}$, $\log E(Z_{011}) = u + u_2 + u_3 + u_{23}$, and $\log E(Z_{111}) = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$. This is a reparameterization of the eight expected values.

For three-list data, we have seven observed categories, whereas there are eight parameters in equation (3). Therefore, a natural assumption is that there is no three-list interaction term, that is, $u_{123} = 0$. Intuitively, this means the complete 2×2 table formed with respect to lists 2 and 3 for individuals in list 1 and the incomplete 2×2 table for individuals not in list 1 have the same odds ratio. The sample odds ratio for the former table is $Z_{111}Z_{100}/(Z_{110}Z_{101})$ whereas the odds ratio for the latter table is $Z_{011}Z_{000}/(Z_{010}Z_{001})$. The assumption of $u_{123} = 0$ allows the following extrapolation formula:

$$\hat{Z}_{000} = \hat{Z}_{001}\hat{Z}_{010}\hat{Z}_{100}\hat{Z}_{111}/(\hat{Z}_{110}\hat{Z}_{011}\hat{Z}_{101}) \quad (4)$$

which expresses the estimated missing cases as a function of the fitted values of other categories [6, 45]. The fitted values of the observable cells are determined by the chosen model.

The independent model includes only main effects as given by

$$\log E(Z_{ijk}) = u + u_1 I(i=1) + u_2 I(j=1) + u_3 I(k=1)$$

The resulting estimator under this model using (4) is equivalent to the MLE for model \mathbf{M}_1 [45]. The interaction terms are used to model dependence. If local list dependence arises in samples 1 and 2, then the interaction term u_{12} is included, and the model is denoted as model

(12, 3) or 12/3 as used in categorical data analysis. If local dependence also appears in samples 1 and 3, then the two interactions u_{12} and u_{13} are needed. The model is denoted as model (12, 13) or 12/13 and similarly for models 13/2, 23/1, 13/23 and others.

The log-linear model can also be motivated by the Rasch model and its generalizations which incorporate heterogeneity among individuals. As shown in Table III and Section 4.1, the Rasch model assumes $\text{logit}(P_{ij}) = \alpha_i + \beta_j$. Only dependence due to heterogeneity arises in this model and there is no local list dependence. A generalized Rasch model allows the heterogeneity effects $\{\alpha_1, \alpha_2, \dots, \alpha_N\}$ to be different for two or more separate groups of samples. For example, in a three-list case, it assumes

$$\text{logit}(P_{ij}) = \begin{cases} \alpha_i + \beta_j, & j = 1, 2 \\ \alpha_i^* + \beta_j, & j = 3 \end{cases} \quad (5)$$

It has been verified [32, 47, 51] that the Rasch (generalized Rasch) model is equivalent to a quasi-symmetric (partial quasi-symmetric) model with some moment constraints. Except for the constraints, a quasi-symmetric model for the three-list case with no second-order interaction, that is, $u_{123} = 0$, is equivalent to the model with first-order interactions identical; this is denoted by (12 = 13 = 23) or simply H1 (first-order heterogeneity [8, 9]). Only one degree of freedom is used to model heterogeneity. A partial quasi-symmetric model in equation (5) with $u_{123} = 0$ is equivalent to the model with $u_{13} = u_{23}$. This model is denoted as (13 = 23, 12). Similarly, we have models (12 = 13, 23) and (12 = 23, 13) corresponding to other two partial quasi-symmetric models. Therefore, the dependence due to heterogeneity can be modelled by either a quasi-symmetric or a partial quasi-symmetric model. We remark that when both types of dependencies occur, they are inevitably confounded in the interaction or common interaction terms and cannot be separated.

The model in (3) can be similarly formulated when there are more than three lists. The basic assumption for four lists is the third-order interaction vanishes (that is, $u_{1234} = 0$); that is, the three-list interaction for individuals in list 1 is the same as that for individuals not in list 1. Local list dependence can be modelled by including the first-order interaction term ($u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34}$) and/or the second-order interaction ($u_{123}, u_{134}, u_{124}, u_{234}$). The Rasch model is equivalent to a model with first-order heterogeneity H1 (that is, 12 = 13 = 14 = 23 = 24 = 34) and second-order heterogeneity H2 (that is, 123 = 124 = 134 = 234). Thus two parameters are used to model heterogeneity in the Rasch model for four lists. If additional local dependencies also occur between lists 1 and 2, lists 1 and 3, and lists 2 and 4, then we add three more parameters u_{12} , u_{13} and u_{24} to the model and the resulting model is denoted as (12/13/24, H1, H2). Refer to the sample program output in Section 5.3 for other types of models. For the general case of t lists, the Rasch model with no highest order interaction is equivalent to a model with interactions of equal order assumed identical; a partial quasi-symmetric Rasch model is equivalent to assuming some of the interactions are equal; see Lloyd [51] for details.

Other useful models, including normal random-effects models, latent class models [40, 42, 50], non-parametric models [52] and Bayesian approaches [47], have also been proposed to reflect heterogeneity in the Rasch model, thus widening the scope and application areas of the log-linear model approach.

4.3. Sample coverage approach

As discussed in Section 2.1, overlap information plays an important role in estimating the number of missing cases. The main purposes of the sample coverage approach proposed by Chao and Tsay [53] and Tsay and Chao [54] are to provide a measure to quantify the overlap information and also to propose parameters to quantify source dependence.

Dependence is modelled by parameters called the ‘coefficient of covariation’ (CCV). To better understand the CCV parameters, we only consider heterogeneity. Define P_{ij} as the conditional probability of identifying individual i in the j th list. The CCV of samples j and k for a fixed-effect approach is defined as

$$\gamma_{jk} = \frac{1}{N} \sum_{i=1}^N \frac{(P_{ij} - \mu_j)(P_{ik} - \mu_k)}{\mu_j \mu_k} = \frac{1}{N} \frac{\sum_{i=1}^N P_{ij} P_{ik}}{\mu_j \mu_k} - 1 \quad (6a)$$

where $\mu_j = \sum_{i=1}^N P_{ij}/N = E(n_j)/N$ denotes the average probability for being listed in the j th sample. The magnitude of γ_{jk} measures the degree of dependence between samples j and k . The two heterogeneous samples are independent if and only if $\gamma_{jk} = 0$, that is, $N^{-1} \sum_{i=1}^N P_{ij} P_{ik} = \mu_j \mu_k$, which means that the covariance between the two sets of probabilities, $\{P_{ij}; i = 1, 2, \dots, N\}$ and $\{P_{ik}; i = 1, 2, \dots, N\}$, is zero. They are positively (negatively) dependent if $\gamma_{jk} > 0$ ($\gamma_{jk} < 0$), which is equivalent to $N^{-1} \sum_{i=1}^N P_{ij} P_{ik} > \mu_j \mu_k$ ($N^{-1} \sum_{i=1}^N P_{ij} P_{ik} < \mu_j \mu_k$), that is, the average probability of jointly being listed in the two samples is greater (less) than that in the independent case. It follows from (6a) that the CCV is zero if the capture probabilities for one sample are constant (that is, a random sample); in this case, no correlation bias arises even if the other sample is highly heterogeneous provided that no local dependence exists. We can prevent local dependence by taking the second sample random because equal catchability implies probabilities are the same regardless of marking status. For a random-effect model, we assume that $\{(P_{i1}, P_{i2}, \dots, P_{it}), i = 1, 2, \dots, N\}$ are a random sample from a t -dimensional distribution $F_{P_1, P_2, \dots, P_t}(p_1, p_2, \dots, p_t)$. The CCV for samples j and k then becomes

$$\gamma_{jk} = \frac{E[(P_j - \mu_j)(P_k - \mu_k)]}{\mu_j \mu_k} = \frac{\text{cov}(P_j, P_k)}{\mu_j \mu_k} = \frac{E(P_j P_k)}{\mu_j \mu_k} - 1 \quad (6b)$$

where $\mu_j = E(P_j)$ denotes the average capture probability for the j th sample. Researchers in fishery sciences have suggested that correlation bias due to heterogeneity could be reduced if two different sampling schemes were used (for example, trapping and then resighting, or netting and then angling). This was justified by Seber (reference [10], p. 86); it also could be seen from formula (6b) because there is almost no covariance between the distributions for two distinct samplings.

The CCV for more than two samples can be similarly defined and interpreted. For example, the CCV for samples k_1, k_2, \dots, k_m in a random-effect model is defined as

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

The CCV for the general cases with two types of dependencies has been developed [53], but it will not be addressed here. We only remark that all CCVs in the general cases measure the overall effect of the two types of dependencies.

For two lists, the usual independence assumption is equivalent to setting the two-sample CCV at 0 ($\gamma_{12} = 0$). It is not possible to model dependence between two lists as we discussed

in Section 3.2. For the three-list cases, there are seven observable categories (as in the HAV data in Table II) and eight parameters: $N; \mu_1; \mu_2; \mu_3; \gamma_{12}; \gamma_{13}; \gamma_{23}; \gamma_{123}$. One constraint is still needed, yet it is possible to model dependence. Consequently, at least three samples are required to reasonably estimate any dependence parameters.

The concept of sample coverage was originally proposed by Turing and Good [55]. This concept has played an important role in the classical species estimation for heterogeneous communities [56] and has been modified for multiple-sample cases [53], in which the sample coverage is used as a measure of overlap fraction. The basic idea is that the sample coverage can be well estimated even in the presence of two types of dependencies. Thus an estimate of population size can be derived via the relationship between the population size and the sample coverage. Chao and Tsay [53] dealt mainly with the three-sample case. Extension to cases with more than three samples was provided by Tsay and Chao [54]. Below we will separately summarize the estimation procedures for the three-list case and the general case.

If an additional case were selected from the third list, then a proper overlapping measure would be the conditional probability of finding this case that had already been identified in the combined list of the other two sources (that is, finding a case i for which $X_{i1} + X_{i2} > 0$). The overlap fraction can be quantified as $\sum_{i=1}^N P_{i3} I(X_{i1} + X_{i2} > 0) / \sum_{i=1}^N P_{i3}$. Considering that this additional individual could be selected from any of the three lists, we define the sample coverage as the average of the three possible overlap fractions as follows:

$$C = \frac{1}{3} \left[\frac{\sum_{i=1}^N P_{i3} I(X_{i1} + X_{i2} > 0)}{\sum_{i=1}^N P_{i3}} + \frac{\sum_{i=1}^N P_{i2} I(X_{i1} + X_{i3} > 0)}{\sum_{i=1}^N P_{i2}} + \frac{\sum_{i=1}^N P_{i1} I(X_{i2} + X_{i3} > 0)}{\sum_{i=1}^N P_{i1}} \right]$$

An estimator of the sample coverage is [53]

$$\hat{C} = 1 - \frac{1}{3} \left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3} \right) = \frac{1}{3} \left[\left(1 - \frac{Z_{100}}{n_1} \right) + \left(1 - \frac{Z_{010}}{n_2} \right) + \left(1 - \frac{Z_{001}}{n_3} \right) \right] \quad (7)$$

which is the average (over three lists) of the fraction of cases found more than once. Note that Z_{100} , Z_{010} and Z_{001} are the numbers of individuals listed only in one sample. Hence this estimator is the complement of the fraction of singletons. Obviously, singletons cannot contain any overlapping information. Define

$$D = \frac{1}{3} [(M - Z_{100}) + (M - Z_{010}) + (M - Z_{001})] = M - \frac{1}{3} (Z_{100} + Z_{010} + Z_{001}) \quad (8)$$

Here $(Z_{100} + Z_{010} + Z_{001})/3$ represents the average of the non-overlapped cases and recall that M denotes the total number of identified cases. Thus D can be interpreted as the average of the overlapped cases. The sample coverage estimation procedures for the three-list case are summarized in the following:

1. When the three sources are independent, a simple population size estimator is derived as

$$\hat{N}_0 = D / \hat{C} \quad (9)$$

The above estimator is obtained by noting that C is reduced to D/N under independence. It can also be intuitively thought of as ratio of overlapped cases to overlap fraction.

2. When dependence exists and the overlap information is large enough (how large it should be will be discussed further below), we take into account the dependence by adjusting the simple estimator given in (9) based on a function of two-sample CCVs. The adjustment expansion formula [54] is

$$N = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}} [(Z_{1+0} + Z_{+10})\gamma_{12} + (Z_{10+} + Z_{+01})\gamma_{13} + (Z_{01+} + Z_{0+1})\gamma_{23}] + R \quad (10)$$

where R is the remainder term in the above expansion and $R/N \rightarrow \psi = \mu_1\mu_2[\gamma_{12}(\gamma_{13} + \gamma_{23}) - \gamma_{123}] + \mu_1\mu_3[\gamma_{13}(\gamma_{12} + \gamma_{23}) - \gamma_{123}] + \mu_2\mu_3[\gamma_{23}(\gamma_{12} + \gamma_{13}) - \gamma_{123}]$ in probability when N becomes large. Our constraint is set by letting $\psi = 0$. The constraint is satisfied under a multiplicative model \mathbf{M}_{th} where the heterogeneity effects follow a gamma type of distribution [53]. Since gamma distribution can cover a wide range of heterogeneity patterns, this is our main motivation for setting this constraint. In equation (10), if R is ignored and CCVs are substituted by the following functions of N :

$$\gamma_{13} = N \frac{Z_{1+1}}{n_1 n_3} - 1, \quad \gamma_{23} = N \frac{Z_{+11}}{n_2 n_3} - 1, \quad \gamma_{12} = N \frac{Z_{11+}}{n_1 n_2} - 1 \quad (11)$$

then we end up with an estimating equation of N . The following solution of the resulting estimating equation is the estimator:

$$\hat{N} = \left[\frac{Z_{+11} + Z_{1+1} + Z_{11+}}{3\hat{C}} \right] / \left\{ 1 - \frac{1}{3\hat{C}} \left[\frac{(Z_{1+0} + Z_{+10})Z_{11+}}{n_1 n_2} + \frac{(Z_{10+} + Z_{+01})Z_{1+1}}{n_1 n_3} + \frac{(Z_{0+1} + Z_{01+})Z_{+11}}{n_2 n_3} \right] \right\} \quad (12)$$

3. For relatively low sample coverage data, we feel the data do not contain sufficient information to accurately estimate the population size. In this case, the following ‘one-step’ estimator \hat{N}_1 is suggested (the estimator is called ‘one-step’ because it is obtained by one iterative step from the adjustment formula (10) with γ_{ij} ’s being replaced by (11)):

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}} [(Z_{1+0} + Z_{+10})\hat{\gamma}_{12} + (Z_{10+} + Z_{+01})\hat{\gamma}_{13} + (Z_{01+} + Z_{0+1})\hat{\gamma}_{23}] \quad (13)$$

where CCV estimates are

$$\hat{\gamma}_{13} = \hat{N}' \frac{Z_{1+1}}{n_1 n_3} - 1, \quad \hat{\gamma}_{23} = \hat{N}' \frac{Z_{+11}}{n_2 n_3} - 1, \quad \hat{\gamma}_{12} = \hat{N}' \frac{Z_{11+}}{n_1 n_2} - 1 \quad (13a)$$

and

$$\hat{N}' = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}} \left[(Z_{1+0} + Z_{+10}) \left(\frac{D}{\hat{C}} \frac{Z_{11+}}{n_1 n_2} - 1 \right) + (Z_{10+} + Z_{+01}) \left(\frac{D}{\hat{C}} \frac{Z_{1+1}}{n_1 n_3} - 1 \right) + (Z_{01+} + Z_{0+1}) \left(\frac{D}{\hat{C}} \frac{Z_{+11}}{n_2 n_3} - 1 \right) \right]$$

This one-step estimator can be regarded as a lower (upper) bound for positively (negatively) dependent samples. Hook and Regal [7] noted that most data sets used in epidemiological applications tend to have a net positive dependence. Thus the one-step estimator is often used as a lower bound as will be shown in Section 5.

A bootstrap resampling method [57] is proposed to obtain estimated standard errors for the above three estimators and to construct confidence intervals using a log-transformation [58]. A relatively low overlap fraction means that there are relatively many singletons. In this case, the undercount cannot be measured accurately due to insufficient overlap. Consequently, a large standard error is usually associated with the estimator in equation (12). How large should the overlap information be? Previous simulation studies [29] have suggested that the estimated sample coverage should be at least 55 per cent. A more practical data-dependent guideline can be determined from the estimated bootstrap SE associated with the estimator given in (12). If the estimated bootstrap standard error becomes unacceptable (say, it exceeds one-third of the population size estimate), then only the lower or upper bound in (13) is recommended.

We now outline the result for the general t -sample case. The sample coverage for the general case can be defined in a similar manner and the estimator is

$$\hat{C} = 1 - \frac{1}{t} \sum_{k=1}^t \frac{S_k}{n_k} \tag{14}$$

where $S_n = Z_{k_1 k_2 \dots k_t} I[k_n = 1, k_j = 0, j \neq n]$ denotes the number of individuals that are listed in sample n only (that is, singletons). For $t=4$, we have $S_1 = Z_{1000}$, $S_2 = Z_{0100}$, $S_3 = Z_{0010}$, and $S_4 = Z_{0001}$ and $\hat{C} = 1 - (Z_{1000}/n_1 + Z_{0100}/n_2 + Z_{0010}/n_3 + Z_{0001}/n_4)$, which is an extension of (7). In the independent case, a valid estimator is D/\hat{C} , where

$$D = \frac{1}{t} \sum_{k=1}^t \left\{ \sum_{i=1}^N I \left[\sum_{j \neq k} X_{ij} > 0 \right] \right\} = M - \frac{1}{t} \sum_{k=1}^t S_k \tag{15}$$

Define $H(i, j) = Z_{k_1 k_2 \dots k_t} I[k_i = 1, k_j = +, k_n = 0, n \neq i, n \neq j]$, and $A(i, j) = H(i, j) + H(j, i)$. For example, $A(1, 2) = Z_{1+00} + Z_{+100}$, $A(2, 3) = Z_{01+0} + Z_{0+10}$. When any type of dependence exists, a generalized formula of (10) becomes [54]

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

where R^* denotes the remainder term. As in the three-list case, R^*/N tends to zero under a multiplicative model \mathbf{M}_{th} where the heterogeneity effects follow a gamma type of distribution when N is large enough. Let $B(i, j) = Z_{k_1 k_2 \dots k_t} I[k_i = k_j = 1, k_n = +, n \neq i, n \neq j]$. For example, $B(1, 2) = Z_{11++}$ and $B(2, 3) = Z_{+11+}$. Using the relationship that $\gamma_{ij} = NB(i, j)/(n_i n_j) - 1$ and substituting it into the above equation, an estimator for the t -sample case can be shown to be

$$\hat{N} = \left[\frac{D}{\hat{C}} - \frac{1}{t\hat{C}} \sum_{i < j} \sum A(i, j) \right] \left\{ 1 - \frac{1}{t\hat{C}} \sum_{i < j} \sum \frac{A(i, j)B(i, j)}{n_i n_j} \right\}^{-1} \tag{16}$$

Table IV. Features of the program CARE.

	CARE-1	CARE-2
Application	Epidemiological data (without a natural time ordering)	Animal data (usually with a natural time ordering)
Environment/language	S-plus	C
Data format	Aggregated categorical data	(With covariates) individual capture history (Without covariates) both types of data
Model	Log-linear models Sample coverage approach	Ecological multiplicative and logistic models
Number of samples (t)	$t = 2$ to 6	$t \geq 2$ for homogeneous models $t \geq 5$ is suggested for heterogeneous models
Estimators	All estimators are independent of the ordering of the lists	Some estimators depend on the ordering of the lists

Note that when $t = 3$, (16) reduces to (12) because $tD - \sum \sum A(i, j) = Z_{+11} + Z_{1+1} + Z_{+11}$. The one-step estimator is given by

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{t\hat{C}} \sum_{i < j} \sum A(i, j) \hat{\gamma}_{ij} \quad (17)$$

where $\hat{\gamma}_{ij}$ is similarly defined as those in (13a). Standard error estimate and confidence interval can be analogously constructed using a bootstrap method.

5. PROGRAM 'CARE' WITH EXAMPLES

The program CAPTURE, specifically developed for analysing closed ecological data, has not been updated since 1991 [44]. It is felt that an additional program might be needed because some new estimators have been proposed. We have developed a program CARE (for capture-recapture) containing two parts: CARE-1 and CARE-2. The size of the program is about 400 kB. CARE-1 is an S-plus [59] program for analysing epidemiological data; CARE-2, written in C language, calculates various estimates for multiplicative ecological models. The features for these two subprograms are presented in Table IV. The program CARE is available on the first author's website at <http://www.stat.nthu.edu.tw/~chao/>.

Since our focus here is on epidemiological applications, we only demonstrate in this section the use of CARE-1, but some results from CARE-2 will be given in Section 5.4 for a data with five lists. The reader is referred to the previously-mentioned website for the use of CARE-2. The four data sets mentioned in Section 1 are used for illustration. The HAV data are displayed in Table II (three lists) and the other three sets on spina bifida (three lists), diabetes (four lists) and birth defects (five lists) are shown in Table V.

Table V. Data on spina bifida, diabetes and congenital anomaly.

Spina bifida list			Data
B	D	M	
0	0	0	$Z_{000} = ??$
0	1	0	$Z_{010} = 49$
0	1	1	$Z_{011} = 4$
1	0	0	$Z_{100} = 247$
1	0	1	$Z_{101} = 112$
1	1	0	$Z_{110} = 142$
1	1	1	$Z_{111} = 12$

Diabetes list				Data
1	2	3	4	
0	0	0	0	$Z_{0000} = ??$
0	0	1	0	$Z_{0010} = 182$
0	0	1	1	$Z_{0011} = 8$
0	1	0	0	$Z_{0100} = 74$
0	1	0	1	$Z_{0101} = 7$
0	1	1	0	$Z_{0110} = 20$
0	1	1	1	$Z_{0111} = 14$
1	0	0	0	$Z_{1000} = 709$
1	0	0	1	$Z_{1001} = 12$
1	0	1	0	$Z_{1010} = 650$
1	0	1	1	$Z_{1011} = 46$
1	1	0	0	$Z_{1100} = 104$
1	1	0	1	$Z_{1101} = 18$
1	1	1	0	$Z_{1110} = 157$
1	1	1	1	$Z_{1111} = 58$

Congenital anomaly list					Data
1	2	3	4	5	
0	0	0	0	0	$Z_{00000} = ??$
0	0	0	0	1	$Z_{00001} = 83$
0	0	0	1	0	$Z_{00010} = 97$
0	0	0	1	1	$Z_{00011} = 30$
0	0	1	0	0	$Z_{00100} = 4$
0	0	1	0	1	$Z_{00101} = 3$
0	0	1	1	0	$Z_{00110} = 2$
0	0	1	1	1	$Z_{00111} = 0$
0	1	0	0	0	$Z_{01000} = 37$
0	1	0	0	1	$Z_{01001} = 34$
0	1	0	1	0	$Z_{01010} = 37$
0	1	0	1	1	$Z_{01011} = 23$
0	1	1	0	0	$Z_{01100} = 1$
0	1	1	0	1	$Z_{01101} = 0$
0	1	1	1	0	$Z_{01110} = 3$

Table V. (Continued)

Congenital anomaly list					Data
1	2	3	4	5	
0	1	1	1	1	$Z_{01111} = 0$
1	0	0	0	0	$Z_{10000} = 27$
1	0	0	0	1	$Z_{10001} = 36$
1	0	0	1	0	$Z_{10010} = 22$
1	0	0	1	1	$Z_{10011} = 5$
1	0	1	0	0	$Z_{10100} = 4$
1	0	1	0	1	$Z_{10101} = 5$
1	0	1	1	0	$Z_{10110} = 1$
1	0	1	1	1	$Z_{10111} = 3$
1	1	0	0	0	$Z_{11000} = 19$
1	1	0	0	1	$Z_{11001} = 18$
1	1	0	1	0	$Z_{11010} = 25$
1	1	0	1	1	$Z_{11011} = 8$
1	1	1	0	0	$Z_{11100} = 1$
1	1	1	0	1	$Z_{11101} = 2$
1	1	1	1	0	$Z_{11110} = 5$
1	1	1	1	1	$Z_{11111} = 2$

5.1. Hepatitis A virus data (three-sample, low sample coverage)

The analysis procedures for the HAV data given in Table II are the following (the program CARE-1 must be executed in an S-plus environment [59]; what the user needs to input is shown in bold face throughout this section):

1. Insert the CARE disk in a floppy disk drive, say in drive a. Invoke S-plus and type **source("a:/care-1.txt")** after a prompt sign, then press the <Enter> key. The following display is shown:

```
CARE-1: for applications to epidemiological data.
This program is used to estimate population size
based on incomplete sources by capture-recapture methods.
The models considered include the log-linear models and the
sample coverage approach. Output includes population size
estimate and its associated standard error as well as a 95%
confidence interval (cil,ciu).
```

The necessary change in the S-PLUS environment:

```
* Under the S-PLUS toolbox, please select Options under Main
Menu -> General Settings -> Computations -> Max Recursion,
then change the default value 256 to 1024.
```

Before using this program, please check the following assumptions:

```
* Interpretation or definition for the characteristic of the
target population should be consistent for all data sources.
```

- * Closure assumption: the size of the population is approximately a constant during the study period.
- * Ascertainable assumption: each case must be ascertainable for all sources, although the probability of ascertainment is allowed to be heterogeneous.
- * For all sources, identification marks are correctly recorded and matched.

Please select:

- 1: three-source case
- 2: four-source case
- 3: five-source case
- 4: six-source case
- 5: exit

Selection: 1

2. The next step is data entry. CARE-1 can only handle categorical data. Since the HAV data consist of three lists, select 1 (three-source case) as above. Then press the (Enter) key and do the following step-by-step data entry:

Your selection is 1 (three-source)

Please key in Z001: 63

Please key in Z010: 55

Please key in Z011: 18

Please key in Z100: 69

Please key in Z101: 17

Please key in Z110: 21

Please key in Z111: 28

3. When data entry is finished, press (Enter). After a while, the output is shown as further below. For three-list data, the output includes: (i) estimates based on any pair of samples; this part includes the standard Petersen estimator and the nearly unbiased estimator (the Chapman estimator [10]). Although these two estimates are valid only under the restrictive independence assumption, they are practically useful as a preliminary analysis [22, 47]; (ii) estimates based on various log-linear models; and (iii) estimates obtained from the sample coverage approach. For programming convenience, the lists are consecutively labelled as list 1, 2 and 3. The correspondence to the user's label of lists should be clear.

OUTPUT:

Number of identified cases in each list:

n1	n2	n3
135	122	126

(1) ESTIMATES BASED ON ANY PAIR OF SAMPLES:

	Petersen	Chapman	se	cil	ciu
pair(1,2)	336	334	29	289	403

pair(1,3)	378	374	36	319	461
pair(2,3)	334	331	30	285	404

Note 1: Refer to Seber (1982, pages 59 and 60) for the Petersen estimator and the Chapman estimator as well as s.e. formula.

Note 2: A log-transformation is used to obtain the confidence interval so that the lower limit is always greater than the number of ascertained. Refer to Chao (1987, *Biometrics*, 43, 783-791) for the construction of the confidence interval.

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:

	dev.	df	est	se	cil	ciu
independent	24.36	3	388	23	352	442
13/2	24.25	2	393	28	350	461
23/1	21.33	2	413	31	364	488
12/3	21.14	2	416	32	365	494
12/23	13.20	1	527	80	412	735
12/13	19.42	1	464	60	377	622
23/13	19.90	1	452	54	373	592
symmetry	2.05	4	1314	520	685	2899
quasi-sy	0.96	2	1313	520	685	2899
part-qs1	0.03	1	1309	519	682	2891
part-qs2	0.86	1	1306	517	681	2882
part-qs3	0.55	1	1325	528	688	2934
saturated	0.00	0	1313	522	683	2904

DEFINITIONS for the log-linear models:

dev.: deviance statistic for testing goodness of fit.

df: degree of freedom.

est: estimate.

se: asymptotic standard error.

cil: 95% confidence interval lower limit (using a log transformation).

ciu: 95% confidence interval upper limit (using a log transformation).

For the 3-list case, all models are special cases of the following:

$$\log E(Z_{ijk}) = u + u_1 I(i=1) + u_2 I(j=1) + u_3 I(k=1) + u_{12} I(i=j=1) + u_{13} I(i=k=1) + u_{23} I(j=k=1)$$

independent: (independent model) $u_{12} = u_{13} = u_{23} = 0$.

13/2: (model with one interaction) $u_{12} = u_{23} = 0$.

23/1: (model with one interaction) $u_{12} = u_{13} = 0$.

12/3: (model with one interaction) $u_{13} = u_{23} = 0$.

12/23: (model with two interactions) $u_{13} = 0$.

12/13: (model with two interactions) $u_{23} = 0$.

13/23: (model with two interactions) $u_{12} = 0$.

symmetry:(symmetry model) $u_1 = u_2 = u_3$, $u_{12} = u_{13} = u_{23}$.
 quasi-sy:(quasi-symmetry model) $u_{12} = u_{13} = u_{23}$.
 part-qs1:(partial-quasi-symmetry model) $u_{12} = u_{23}$.
 part-qs2:(partial-quasi-symmetry model) $u_{12} = u_{13}$.
 part-qs3:(partial-quasi-symmetry model) $u_{23} = u_{13}$.
 saturated:(saturated model) no restriction.

(3) SAMPLE COVERAGE APPROACH:

	M	D	\hat{C}	est	se	cil	ciu
Nhat-0	271	208.667	0.513	407	28	363	472
Nhat	271	208.667	0.513	971	925	369	5290
Nhat-1	271	208.667	0.513	508	40	442	600

parameter estimates:

	u_1	u_2	u_3	r_{12}	r_{13}	r_{23}	r_{123}
Nhat-0	0.33	0.30	0.31	0.21	0.08	0.22	0.73
Nhat	0.14	0.13	0.13	1.89	1.57	1.91	6.35
Nhat-1	0.27	0.24	0.25	0.51	0.34	0.52	1.11

DEFINITIONS for the sample coverage approach:

M: number of individuals ascertained in at least one list.

D: the average of the number of individuals listed in the combination of two lists omitting the third.

\hat{C} : sample coverage estimate, see (7), or Equation (14) of Chao and Tsay (1998).

est: population size estimate.

se: estimated standard error of the population size estimation based on 1000 bootstrap replications. Note this s.e. might vary with trials.

cil: 95% confidence interval lower limit (using a log-transformation).

ciu: 95% confidence interval upper limit (using a log-transformation).

Nhat-0: population size estimate for independent samples; see (9), or Equation (15) of Chao and Tsay (1998).

Nhat: Population size estimate for sufficiently high sample coverage cases; see (12), or Equation (20) of Chao and Tsay (1998).

Nhat-1: One-step population size estimate for low sample coverage cases; see (13), or Equation (2.21) of Chao et al. (1996). This estimator is suggested for use when the estimated s.e. of Nhat is relatively large.

u_1, u_2, u_3 : estimated mean probabilities depending on the estimate of N.

$r_{12}, r_{13}, r_{23}, r_{123}$: estimated coefficient of covariation (CCV) depending on the estimate of N.

For the HAV data, the Petersen and Chapman estimates are in the range of 330 to 380. As discussed in Section 3.2, the Petersen estimator based on two samples is biased downwards (upwards) if these two samples are positively (negatively) dependent. However, the narrow range of these estimates would not indicate the possible direction of dependence at this stage.

The estimated dependence parameters are provided in the output of the sample coverage approach.

The second part of the output includes the results for all possible log-linear models fitted to these data. The outputs show the corresponding deviances and estimates of the total number of infected. The notation and definitions for various models are introduced in Section 4.2 as well as in the output. The independent model produces an estimate of 388, which is close to the results for any two samples. Except for the saturated model, all the log-linear models, which consider local independence only and do not take into account heterogeneity (that is, models PE/Q, QE/P, PQ/E, PQ/QE, PQ/PE and QE/PE), do not fit the data well. All other models, which take heterogeneity only into account (quasi-symmetric and partial quasi-symmetric models) fit well. Those adequate models produce approximately the same estimates of 1300 with an approximate estimated SE of 520. This relatively large estimated SE shows that the data are actually insufficient to model heterogeneous models.

The third part of the output contains the sample coverage approach. The estimators \hat{N}_0 , \hat{N} and \hat{N}_1 derived in equations (9), (12) and (13) correspond to Nhat-0, Nhat and Nhat-1, respectively, in the output. Other statistics can be easily identified in the output because of similar consistent notation. The sample coverage based on (7) is estimated to be

$$\hat{C} = 1 - \frac{1}{3} \left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3} \right) = 1 - \frac{1}{3} \left(\frac{69}{135} + \frac{55}{122} + \frac{63}{126} \right) = 51.3 \text{ per cent}$$

where 69, 55 and 63 are the numbers of singletons. The average of the overlapped cases is equal to $D = 271 - (69 + 55 + 63)/3 = 208.67$. If we ignore the possible dependence between samples, an estimate based on (9) for the HAV data is $\hat{N}_0 = D/\hat{C} = 208.67/0.513 = 407$, which is slightly higher than the estimate of 388 based on the independent log-linear model. The estimator given in (12) is $\hat{N} = 971$, but a large estimated bootstrap SE (925) renders the estimate useless. The estimated SE was calculated by using a bootstrap method based on 1000 replications. We feel these data with a sample coverage estimate of 51 per cent do not contain enough information to correct for undercount. The proposed one-step estimator in equation (13) is $\hat{N}_1 = 508$ with an estimated SE of 40 using 1000 bootstrap replications. The same bootstrap replications produce a 95 per cent confidence interval of (442, 600). We remark that the estimated SE might vary from trial to trial because replications vary in the bootstrap procedures.

It follows from (11) that the CCV measures depend on the value of N . In the output, the CCV estimates based on the three estimates of N show that any two or three samples are positively dependent. As a result, the estimate $\hat{N}_1 = 508$ can only serve as a lower bound. Also, the estimates (Petersen and Chapman's estimates) assuming independence based on two samples should have a negative bias. However, we cannot distinguish which type of dependence (local dependence or heterogeneity) is the main cause of the bias.

In December 1995, the National Quarantine Service of Taiwan conducted a screen serum test for the HAV antibody for all students of the college at which the outbreak of the HAV occurred [1]. After suitable adjustments, they have concluded that the final figure of the number infected was about 545. Thus this example presents a very valuable data set with the advantage of a known true parameter. Our estimator \hat{N}_1 does provide a satisfactory lower bound. This example shows the need for undercount correction and also the usefulness of the capture-recapture method in estimating the number of missing cases.

5.2. *Spina bifida data (three-sample)*

The data set on spina bifida [2, 3] was reproduced in Table V. Three lists were collected: birth certificates (B-list); death certificates (D-list), and medical rehabilitation files (M-list). There were 513, 207 and 188 cases, respectively, on B-, D- and M-lists and in total 626 ascertained cases. After the data entry, part of the output shows:

OUTPUT:

Number of identified cases in each list:

n1	n2	n3
513	207	188

(1) ESTIMATES BASED ON ANY PAIR OF SAMPLES:

	Petersen	Chapman	se	cil	ciu
pair(1,2)	690	689	23	651	743
pair(1,3)	778	776	35	718	857
pair(2,3)	2432	2311	498	1554	3556

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:

	dev.	df	est	se	cil	ciu
independent	58.35	3	764	21	728	812
13/2	58.09	2	756	25	715	816
23/1	3.86	2	731	17	702	771
12/3	46.85	2	831	37	770	919
12/23	0.00	1	758	26	716	820
12/13	37.50	1	1361	396	899	2602
23/13	0.67	1	711	18	683	754
symmetry	370.66	4	658	13	640	696
quasi-sy	29.01	2	649	10	636	678
part-qs1	29.00	1	649	10	636	679
part-qs2	3.67	1	762	86	670	1051
part-qs3	15.79	1	659	14	641	700
saturated	0.00	0	763	87	670	1053

(3) SAMPLE COVERAGE APPROACH:

	M	D	C^{\sim}	est	se	cil	ciu
Nhat-0	626	507.333	0.654	775	22	738	826
Nhat	626	507.333	0.654	752	36	699	844
Nhat-1	626	507.333	0.654	767	33	716	848

parameter estimates:

	u1	u2	u3	r12	r13	r23	r123
Nhat-0	0.66	0.27	0.24	0.12	0.00	-0.68	-0.08
Nhat	0.68	0.28	0.25	0.09	-0.03	-0.69	-0.03
Nhat-1	0.67	0.27	0.25	0.11	-0.01	-0.68	-0.06

Note that the Petersen and Chapman estimates based on the D- and M-lists are substantially higher than the other two estimates. This implies that a possible negative dependence exists between these two lists. Using three samples, Regal and Hook [3] showed that there is a strong

negative dependence between the D- and M-lists, moderate positive dependence between the B- and D-lists and weak dependence for the B- and M-lists. Regal and Hook indicated that an adequate model is DM/BD, which gives a population size estimate of 758 with a 95 per cent confidence interval of (707, 809), but they also commented that this confidence interval might be artificially narrow. A pairwise DM/BD/BM model yields an estimate of 763 with a 95 per cent confidence interval of (590, 936). Their confidence bounds are slightly different from those in the output because a log-transformation is adopted in our approach.

The sample coverage estimate for this data set is 65.4 per cent, and $D = 507.33$, so an estimate without taking into account the possible dependence is $\hat{N}_0 = 775$. The proposed estimate $\hat{N} = 752$ (SE = 36) and the corresponding 95 per cent confidence interval is (699, 844) using 1000 bootstrap replications. These data are sufficient to produce a reliable population size estimate. Both the log-linear and our approaches produce very close estimates. Our interval is slightly wider than the one obtained by the DM/BD model but much narrower than that obtained by the pairwise DM/BD/BM model. Based on (11) and an estimated population size of 752, the output shows that $\hat{\gamma}_{DM} = -0.69$, $\hat{\gamma}_{BD} = 0.09$ and $\hat{\gamma}_{BM} = -0.03$. These estimated CCV values support the finding of Regal and Hook about the dependencies between two samples. Here the CCV estimates represent the mixed effects of two types of dependencies.

5.3. Diabetes data (four-sample)

The data given in Table V on diabetes were collected and discussed by Bruno *et al.* [4] and IWGDMF [8, 9]. 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 purchases of reagent strips and insulin syringes (list 4, 173 cases). A total of 2069 cases were identified. When the number of samples is at least four, there are many available log-linear models. We provide six selections in the program. The basic models include all the estimates based on any pair of samples, some selected log-linear models and the sample coverage approach. The other five selections include different types of log-linear models. This program does not provide a model selection procedure and the users have to select their own model. In the following, we show the procedure for analysing the diabetes data using basic models as well as models which include five and six two-factor interactions with/without heterogeneity:

```

Please select:
1: three-source case
2: four-source case
3: five-source case
4: six-source case
5: exit
Selection: 2
your selection is 2 (four-source)
Please key in Z0001: 10
Please key in Z0010: 182
Please key in Z0011: 8
Please key in Z0100: 74
Please key in Z0101: 7

```

Please key in Z0110: 20
 Please key in Z0111: 14
 Please key in Z1000: 709
 Please key in Z1001: 12
 Please key in Z1010: 650
 Please key in Z1011: 46
 Please key in Z1100: 104
 Please key in Z1101: 18
 Please key in Z1110: 157
 Please key in Z1111: 58

Please select:

- 1: basic models (including pair-sample models, an independent model, symmetric model, quasi-symmetric model, models with one 3-factor interaction and the sample coverage estimates)
- 2: models with one 2-factor interaction w/o H1, H2
 (H1: the first order heterogeneity, i.e., all 2-factor interactions are identical)
 (H2: the second order heterogeneity, i.e., all 3-factor interactions are identical)
- 3: models with two 2-factor interactions w/o H1, H2
- 4: models with three 2-factor interactions w/o H1, H2
- 5: models with four 2-factor interactions w/o H1, H2
- 6: models with five & six 2-factor interactions w/o H1, H2
- 7: exit

Selection: 1

OUTPUT:

Number of identified cases in each list:

n1	n2	n3	n4
1754	452	1135	173

(1) ESTIMATES BAS

ED ON ANY PAIR OF SAMPLES:

	Petersen	Chapman	se	cil	ciu
pair(1,2)	2353	2351	58	2250	2478
pair(1,3)	2185	2185	22	2146	2233
pair(1,4)	2264	2261	88	2117	2468
pair(2,3)	2060	2057	77	1922	2224
pair(2,4)	806	803	47	725	913
pair(3,4)	1558	1555	67	1445	1712

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:

	dev.	df	est	se	cil	ciu
independent	217.48	10	2251	19	2217	2292
123/4	165.76	6	2185	40	2130	2291

124/3	92.23	6	2247	21	2211	2293
134/2	154.38	6	2386	45	2309	2487
234/1	55.24	6	2283	22	2244	2331
H1	105.63	9	2669	83	2528	2854
symmetry	3156.50	11	2197	50	2130	2338
quasi-sy	93.95	8	2239	68	2148	2431
saturated	0.00	0	5367	2771	2856	15883

(3) SAMPLE COVERAGE APPROACH:

	M	D	C [^]	est	se	cil	ciu
Nhat-0	2069	1825.25	0.803	2272	26	2226	2330
Nhat	2069	1825.25	0.803	2609	81	2472	2792
Nhat-1	2069	1825.25	0.803	2458	50	2372	2568

parameter estimates:

	u1	u2	u3	r12	r13	r14	r23	r24	r34
Nhat-0	0.77	0.20	0.50	0.08	-0.03	0.04	0.00	0.10	1.82
Nhat	0.67	0.17	0.44	0.07	0.11	0.19	0.15	0.27	2.24
Nhat-1	0.71	0.18	0.46	0.07	0.04	0.12	0.09	0.19	2.05

If you want to continue to fit other models
please select:

- 1: basic models (including pair-sample models, an independent model, symmetric model, quasi-symmetric model, models with one 3-factor interaction and the sample coverage estimates)
- 2: models with one 2-factor interaction w/o H1, H2
(H1: the first order heterogeneity, i.e., all 2-factor interactions are identical)
(H2: the second order heterogeneity, i.e., all 3-factor interactions are identical)
- 3: models with two 2-factor interactions w/o H1, H2
- 4: models with three 2-factor interactions w/o H1, H2
- 5: models with four 2-factor interactions w/o H1, H2
- 6: models with five & six 2-factor interactions w/o H1, H2
- 7: exit

Selection: 6

OUTPUT:

Estimates based on log-linear models:

	dev.	df	est	se	cil	ciu
12/13/14/23/24	52.03	5	2648	122	2455	2939
12/13/14/23/34	135.33	5	2630	119	2440	2916
12/13/14/24/34	16.74	5	2637	116	2452	2912
12/13/23/24/34	7.62	5	2771	146	2538	3120

	12/14/23/24/34	53.10	5	2271	23	2230	2323
	13/14/23/24/34	13.72	5	2562	75	2435	2733
H1	12/13/14/23/24	7.05	4	2790	153	2547	3155
H1	12/13/14/23/34	7.05	4	2790	153	2547	3155
H1	12/13/14/24/34	7.05	4	2790	153	2547	3155
H1	12/13/23/24/34	7.05	4	2790	153	2547	3155
H1	12/14/23/24/34	7.05	4	2790	153	2547	3155
H1	13/14/23/24/34	7.05	4	2790	153	2547	3155
H1 H2	12/13/14/23/24	0.92	3	4501	1319	2968	8647
H1 H2	12/13/14/23/34	0.92	3	4501	1319	2968	8647
H1 H2	12/13/14/24/34	0.92	3	4501	1319	2968	8647
H1 H2	12/13/23/24/34	0.92	3	4501	1319	2968	8647
H1 H2	12/14/23/24/34	0.92	3	4501	1319	2968	8647
H1 H2	13/14/23/24/34	0.92	3	4501	1319	2968	8647
	12/13/14/23/24/34	7.05	4	2790	153	2547	3155

Bruno *et al.* [4] found that the log-linear model 12/13/23/24/34 fits the data well and presented an estimate of 2771 with a 95 per cent confidence interval of (2492, 3051). Their estimate is shown in the above output. The use of a log-transformation shifts their interval rightward and results in little increase in interval length. When they further stratified for the pattern of treatment (dietary control, hypoglycaemia agents and insulin), an estimate of 2586 was obtained with a 95 per cent interval of (2341, 2830). The two review papers by IWGDMF [8, 9] analysed the data by including heterogeneity terms to several proper log-linear models and selected the final model by the Akaike information criterion. They obtained an estimate of 2834 based on the stratified data.

The sample coverage for these data is estimated to be 80.3 per cent. Since the coverage estimate is sufficiently high, a precise estimate is expected. The proposed estimate is $\hat{N} = 2609$ with an estimated SE of 81 using 1000 bootstrap replications. Positive dependencies exist in any two samples as shown in the output. The corresponding 95 per cent 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 the unstratified result.

5.4. Infants' congenital anomaly data (five-sample)

For this data set, Wittes *et al.* [5] obtained an estimate of 638 (SE 15) for the total number of cases under an independent assumption. This result can be seen from the second part of our output (see below) for the log-linear model approach. Fienberg [6] further modelled the possible dependencies between samples and fitted a log-linear model (12/14/25/34) to the data. He found that all the four interactions were highly significant and obtained an estimate of 634 (SE 18). Fienberg's estimate is very close to the result under independence because there are both positive and negative dependence-effects in the model; see the CCV estimates below. Part of the output from the program CARE-1 after data entry is shown below.

```
Number of identified in each list:
      n1      n2      n3      n4      n5
    183     215     36     263     252
(1) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
```

	Petersen	Chapman	se	cil	ciu
pair(1,2)	492	490	32	437	565
pair(1,3)	286	283	31	240	368
pair(1,4)	678	674	53	587	796
pair(1,5)	584	581	40	515	674
pair(2,3)	553	532	99	392	797
pair(2,4)	549	547	30	498	617
pair(2,5)	623	620	41	552	714
pair(3,4)	592	574	96	437	830
pair(3,5)	605	584	103	438	860
pair(4,5)	933	927	78	796	1106

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:

	dev.	df	est	se	cil	ciu
independent	93.45	25	638	15	613	673
symmetry	396.96	26	756	424	556	3064
quasi-sy	87.69	22	806	521	560	3647

(3) SAMPLE COVERAGE APPROACH:

	M	D	C [~]	est	se	cil	ciu
Nhat0	537	487.4	0.774	630	15	604	664
Nhat	537	487.4	0.774	659	35	607	750
Nhat-1	537	487.4	0.774	649	27	608	714

parameter estimates:

	u1	u2	u3	u4	u5
Nhat-0	0.29	0.34	0.06	0.42	0.40
Nhat	0.28	0.33	0.05	0.40	0.38
Nhat-1	0.28	0.33	0.06	0.41	0.39

	r12	r13	r14	r15	r23	r24	r25	r34	r35
Nhat-0	0.28	1.20	-0.07	0.08	0.14	0.15	0.01	0.06	0.04
Nhat	0.34	1.30	-0.03	0.13	0.19	0.20	0.06	0.11	0.09
Nhat-1	0.32	1.27	-0.04	0.11	0.17	0.18	0.04	0.10	0.07

	r45
Nhat-0	-0.33
Nhat	-0.29
Nhat-1	-0.30

First notice that two of the ten pairwise estimates lie below the other values, especially the estimate using list 1 and list 3. This shows strong evidence of positive dependence between these two lists. Also, negative dependence arises between samples 4 and 5 as the estimate using this pair of samples is much higher. The other estimates are in the range of 550 to 680.

For the sample coverage approach, the overlap fraction is estimated by the sample coverage estimate, which is 77.4 per cent. Our proposed estimator based on equation (16) is 659 (SE

35) with a 95 per cent confidence interval of (607, 750). Our estimator that incorporates both types of dependencies agrees well with the previous findings, but the variation is larger due to estimating more dependence parameters. The CCV estimates show that there are relatively large positive values (γ_{12} and γ_{13}) and negative value (γ_{45}). The dependence between samples 1 and 3 is significantly higher than the others, but it was not included in Fienberg's model.

For this five-sample data set, we can illustrate the use of heterogeneous ecological models, although those models are typically applied to situations with identical trapping methods. As discussed in Section 4.1, only the following models are potentially useful: multiplicative model \mathbf{M}_{th} ; logistic model \mathbf{M}_{th} (that is, the Rasch model); model \mathbf{M}_h , and model \mathbf{M}_t (for the latter two models, the multiplicative and logistic types of models are equivalent). Model \mathbf{M}_t is equivalent to the independent log-linear model and this is the model considered by Wittes *et al.* [5]. For the Rasch model, it is equivalent to a quasi-symmetric model with some constraints. From the second part of the output, the estimate under a quasi-symmetric model is 806 (SE 521). However, the estimated SE is large so that the model is unlikely to be useful. For model \mathbf{M}_h , the first-order and second-order jackknife estimators (Burnham and Overton [46]) are, respectively, 735 (SE 19) and 800 (SE 27). The interpolated jackknife combining the first- and second-order jackknife is 772 (SE 49). The two estimators proposed by Lee and Chao [43] are 770 (SE 32) and 641 (SE 21) under model \mathbf{M}_h , and 789 (SE 36) and 654 (SE 25) under model \mathbf{M}_{th} . Although these two heterogeneous models do not consider the possible local dependence, it is interesting to notice that one of the estimates under each model is close to the proposed estimate that considers two types of dependencies. All the estimates and SEs for models \mathbf{M}_{th} and \mathbf{M}_h were obtained using our program CARE-2.

6. REMARKS AND DISCUSSION

Three classes of capture-recapture models have been reviewed in this tutorial: ecological models; log-linear models, and the sample coverage approach. Most ecological models allowing for heterogeneous capture probabilities are recommended only when there are at least five trapping samples, whereas the other two approaches are mainly useful for two to five samples. We have focused on the latter two models for epidemiological applications and demonstrated the use of the program CARE developed by the authors.

Hook and Regal [60, 61] presented 17 recommendations for the use of the capture-recapture method in epidemiology. There are several basic assumptions that should be fulfilled or checked to validate the application of the method. In addition to the closure assumption, a basic assumption is an explicit definition or interpretation of the 'target population'. Gutteridge and Collin, in a prevalence study of physical disability [62], reported that two sources might have different interpretations of disability and its severity. Thus the 'target population' for two sources might become inconsistent. Another basic assumption is that all identification 'marks' should be correctly recorded and matched. Although in most epidemiological studies this assumption obviously can be fulfilled, in reality it might be an impediment in developing countries, as indicated by Black and McLarty [63].

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. This implies that any individual must have a positive probability to be ascertained by any source and unascertainment is purely a 'random zero' (missing due to small chance), not a 'structural zero' (missing due to

impossibility). If some cases are systematically missed by one or more sources, then there is no overlap information and those 'uncatchable' individuals cannot be included in our target population and should be treated separately. 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 get source intersection, consequently capture-recapture cannot be used to estimate the total number of cases in the combined whole area. The estimating target is actually the size of those jointly 'ascertainable' or 'catchable' individuals by *all* sources. Therefore, two complementary lists collected from disjoint areas cannot be utilized as two separate lists and they should be combined into a joint list.

We also note that another limitation of the capture-recapture methods is that sufficiently high overlapping information is required to produce reliable population size estimates and to model dependence among samples. Otherwise, Coull and Agresti [40] demonstrated that the likelihood functions under some random-effect models for sparse information might become flat and the resulting estimates based on equivalent log-linear models are likely to become unstable. Chao *et al.* [29] also showed in a sample coverage approach that a large variation might be associated with the resulting estimator due to insufficient overlap. In such cases, we have proposed a plausible lower bound given in equations (13) and (17) for positively dependent samples. We feel that a precise bound is of more practical use than an imprecise point estimate.

As indicated by the IWGDMF [8, 9], the log-linear model approach has many advantages as follows: (a) all models are under a unified framework; (b) model selection can be easily implemented and carried out in a flexible fashion, based on data and guided by prior information; (c) tests are available for comparing models; (d) dependence can be incorporated by adding proper interactions; and (e) all inference is within the mainstream of statistical data analysis. An untestable assumption involved is that the highest-order interaction does not exist. The IWGDMF commented that the existence of heterogeneity in three-list data might result in the lack of a reliable estimate. Another concern is that two equally fitted models might produce quite different estimates [reference 40, p. 299]. As the number of lists increases, the number of adequate models increases rapidly and thus model selection causes further problems [23].

The sample coverage approach provides an alternative approach, which makes use of overlap information to incorporate source-dependence in the estimation. The advantages for this approach include the following: (a) overlap information can be quantified by the estimated sample coverage; (b) dependence among samples can be quantified by estimated CCVs, and thus can be detected by data; (c) no model selection or model comparison is needed; and (d) no further difficulty arises when the number of lists increases. The program CARE-1 can deal with data up to six lists and can be easily extended to handle data with more than six lists. Nevertheless, there is an untestable and quite complicated assumption. Recall that this approach is mainly derived from an expansion (equation (10)) and its generalization; the assumption is that the limiting value of the remainder term of that expansion tends to zero. This assumption is satisfied under the gamma type of heterogeneity, but the robustness of the resulting estimator to the departure from the gamma distribution needs further investigation [29, 53].

Some major problems with the use of capture-recapture models have been raised by previous authors [14, 25–28] and have been encountered in our consulting services with researchers in health science. Their problems (listed below) represent the major concerns among epidemiologists about the method. To clarify the use of this methodology and to enhance the

understanding of its application, we focus on the following principal concerns and include some discussion from a statistical point of view.

6.1. Is it necessary to have at least a random sample?

If a random sample could be obtained, then two samples would suffice for estimating population size. A random sample implies that all individuals in the population have the same probability to be ascertained in that list. As we discussed in Section 4.3, the usual Petersen estimator is valid if the second sample is random. No correlation bias arises even when the first non-random sample is highly selective or extremely heterogeneous. This can be understood intuitively. For example, if we only tag large fish in the first sample, but fish of any size are captured in an equally likely manner in the recapture sample, then local dependence clearly does not arise because marked and unmarked have identical chances of being caught. The marked rate in the sample is approximately the rate in the population, which justifies the use of the Petersen estimator (see Section 3.2). Moreover, when only heterogeneity is present, correlation bias vanishes if either sample is random. A random tagging in the first sample means that the tag rates are nearly the same in any groups of different sizes. Even if we only catch large fish in the recapture sample, the fraction of tagged individuals in the sample is still approximately equal to the fraction tagged in the whole population. These findings could also be theoretically justified by the definition of CCV in the sample coverage approach (Section 4.3). One of the motivations for developing models incorporating dependence is that a random sample is almost unfeasible for animal studies. Animals cannot be drawn in a randomized manner. An advantage of using more sophisticated dependent models to estimate population size is that a random sample is not necessary.

6.2. Is it better to use an identical ascertainment method for all lists?

As we have discussed in Section 2.2, identical trapping methods are usually used in animal studies. The advantages from a statistical point of view are threefold: (a) dependence patterns between samples are similar, so one or two parameters are sufficient to model dependence no matter how many samples are taken, so simpler models can be adopted for modelling heterogeneous populations; (b) as more samples are conducted, overlap information is generally increased without inducing additional dependence measures; and (c) finally, systematic missing patterns are unlikely to occur and all animals are 'jointly catchable'. However, there are disadvantages: (a) possible local dependence due to a behavioural response to identical trapping experience might be induced; and (b) dependence due to heterogeneity arises because of strong correlation between two sets of similar capture probabilities. To reduce correlation bias, different trapping methods have been proposed by researchers, especially in fishery science [reference 10, p. 86]. Thus, identical ascertainment methods are not necessary, but we need to model more dependence measures and must be cautious about the possible missing patterns or structural zeros if different surveys are applied.

6.3. The traditional assumption of independence

As we have explained for the two-list cases in Section 3.2, this is indeed a problem unless one is willing to assume some value for the two-sample interaction. When there are three or more sources, the log-linear models and the sample coverage approach provide viable

ways to model dependence. Previous studies [8, 9, 29, 53] have shown in many cases that the performance of the two approaches is encouraging.

6.4. *Almost zero probability of being captured by any source*

As we have remarked before, those 'uncatchable' cases cannot be included in our target 'population' and should be treated separately. We can only estimate the size of a subpopulation that contains only catchable individuals.

6.5. *Heterogeneity among individuals ('variable catchability')*

This problem has been discussed extensively in animal population studies. Stratified analysis has been suggested in the literature. For example, data for males and females are treated separately, or the data can be stratified by the type of treatment, such as in the case of data on diabetes. However, even in a stratum, residual heterogeneity may still exist. As indicated before, heterogeneity among individuals induces possible source dependence. Therefore, the heterogeneity problem can be partially solved by proper adjustment for dependence. We emphasize that heterogeneity does not always induce dependence and two heterogeneous samples may be independent (see Section 4.3).

6.6. *How many sources are needed?*

Hay [64], Chang *et al.* [65] and Ismail *et al.* [66] had some interesting observations and suggestions regarding the number of sources used in capture-recapture studies. When identical trapping methods are used as in ecological studies, more individuals would be caught, and overlap information generally would increase without inducing more dependence measures. As a result, a more precise estimator may be produced. In the log-linear model approach, the higher order interaction is likely to be less significant. Thus the basic assumption of there being no highest order interaction is more reasonable when there are more lists. However, in health science, different ascertainment methods are used and thus more dependence parameters are involved as the number of lists is increased. Moreover, the probability of producing structural zero is increased as there are more cells in the data. Increasing the number of lists often costs more and requires additional effort, but it does not necessarily yield better results, especially when different ascertainment methods are applied. We would thus recommend that only three or four samples be used unless similar types of identification surveys are conducted.

In summary, capture-recapture models provide a potentially useful method to estimate population size in epidemiological studies but there are assumptions and limitations to this approach. The four data sets discussed in this paper have provided examples to show the usefulness of the capture-recapture analysis in assessing the extent of incomplete ascertainment. Efforts are needed to study the relative merits of the existing models and to provide practical guidelines. More collaboration is certainly needed between epidemiologists and statisticians to pursue additional methodological and conceptual research work.

ACKNOWLEDGEMENTS

This paper is an expanded revision of our earlier manuscript on analysing the hepatitis A virus data in Taiwan. We thank all reviewers for providing helpful comments and suggestions on the earlier versions and for pointing out some recent publications (references [28, 47, 60] and [61]). We also thank one

reviewer and the editors (Professors Machin and D'Agostino) for having suggested expansion and resubmission as a tutorial paper. This research was supported by the National Science Council of Taiwan.

REFERENCES

1. Chao DY, Shau WY, Lu CWK, Chen KT, Chu CL, Shu HM, Horng CB. A large outbreak of hepatitis A in a college school in Taiwan: associated with contaminated food and water dissemination. *Epidemiology Bulletin*, Department of Health, Executive Yuan, Taiwan Government, 1997.
2. Hook EB, Albright SG, Cross PK. Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York State. *American Journal of Epidemiology* 1980; **112**:750–758.
3. Regal RR, Hook EB. The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* 1991; **10**:717–721.
4. Bruno GB, Biggeri A, LaPorte RE, McCarty D, Merletti F, Pagono G. Application of capture-recapture to count diabetes. *Diabetes Care* 1994; **17**:548–556.
5. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of cases ascertainment when using multiple information sources. *Journal of Chronic Diseases* 1974; **27**:25–36.
6. Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* 1972; **59**:591–603.
7. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *American Journal of Epidemiology* 2000; **152**:771–779.
8. International Working Group for Disease Monitoring and Forecasting (IWGDMF). Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology* 1995; **142**:1047–1058.
9. International Working Group for Disease Monitoring and Forecasting (IWGDMF). Capture-recapture and multiple-record systems estimation II: application in human diseases. *American Journal of Epidemiology* 1995; **142**:1059–1068.
10. Seber GAF. *The Estimation of Animal Abundance*, 2nd edn. Griffin: London, 1982.
11. Seber GAF. A review of estimating animal abundance. *Biometrics* 1986; **42**:267–292.
12. Seber GAF. A review of estimating animal abundance II. *International Statistical Review* 1992; **60**:129–166.
13. Schwarz CJ, Seber GAF. A review of estimating animal abundance III. *Statistical Science* 1999; **14**:427–456.
14. Schouten LJ, Straatman H, Kiemeny LALM, Gimbrere CHF, Verbeek ALM. The capture-recapture method for estimation of cancer registry completeness; a useful tool? *International Journal of Epidemiology* 1994; **23**:1111–1116.
15. Hook EB, Regal RR. Validity of Bernoulli census, log-linear, and truncated binomial models for correction for underestimates in prevalence studies. *American Journal of Epidemiology* 1982; **116**:168–176.
16. LaPorte RE, McCarty DJ, Tull ES, Tajima N. Counting birds, bees and NCDs. *Lancet* 1992; **339**:494.
17. Darroch JN. The Multiple-Recapture Census I. Estimation of a closed population. *Biometrika* 1958; **45**:343–359.
18. Sekar C, Deming WE. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* 1949; **44**:101–115.
19. Wittes JT, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases* 1968; **21**:287–301.
20. Wittes JT. Applications of a multinomial capture-recapture method to epidemiological data. *Journal of the American Statistical Association* 1974; **69**:93–97.
21. McCarty DJ, Tull ES, Moy CS, Kwok CK, LaPorte RE. Ascertained corrected rates: Applications of capture-recapture methods. *International Journal of Epidemiology* 1993; **22**:559–565.
22. Hook EB, Regal RR. The value of capture-recapture methods even for apparently exhaustive surveys: the need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *American Journal of Epidemiology* 1992; **135**:1060–1067.
23. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitation. *Epidemiological Reviews* 1995; **17**:243–264.
24. Chao A. Capture-recapture. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998.
25. Kiemeny LALM, Schouten LJ, Straatman H. Ascertainment corrected rates (Letter to Editor). *International Journal of Epidemiology* 1994; **23**:203–204.
26. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* 1994; **23**:1322–1323.

27. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitation of methods based on multiple data sources. *International Journal of Epidemiology* 1996; **25**: 474–477.
28. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *Journal of Clinical Epidemiology* 1999; **52**:909–914.
29. Chao A, Tsay PK, Shau WY, Chao DY. Population size estimation for capture-recapture models with applications to epidemiological data. *Proceedings of Biometrics Section, American Statistical Association* 1996; 108–117.
30. Lazarsfeld PF, Henry NW. *Latent Structure Analysis*. Houghton Mifflin: Boston, 1968.
31. Hook EB, Regal RR. Effects of variation in probability of ascertainment by sources ('variable catchability') upon 'capture-recapture' estimates of prevalence. *American Journal of Epidemiology* 1993; **137**:1148–1166.
32. Darroch JN, Fienberg SE, Glonek GFV, Junker BW. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* 1993; **88**:1137–1148.
33. Pollock KH. Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife population: past, present, and future. *Journal of the American Statistical Association* 1991; **86**:225–238.
34. Otis DL, Burnham KP, White GC, Anderson DR. Statistical inference from capture data on closed animal populations. *Wildlife Monographs* 1978; **62**:1–135.
35. White GC, Anderson DR, Burnham KP, Otis DL. *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos National Lab, LA-8787-NERP: Los Alamos, New Mexico, USA, 1982.
36. Huggins RM. On the statistical analysis of capture experiments. *Biometrika* 1989; **76**:133–140.
37. Alho JM. Logistic regression in capture-recapture models. *Biometrics* 1990; **46**:623–635.
38. Huggins RM. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* 1991; **47**:725–732.
39. Rasch G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Neyman J (ed.). University of California Press: 1961; 321–333.
40. Coull BA, Agresti A. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* 1999; **55**:294–301.
41. Lloyd CJ, Yip P. A unification of inference for capture-recapture studies through martingale functions. In *Estimating Equations*, Godambe VP (ed.). Clarendon Press: Oxford, 1991; 65–88.
42. Pledger S. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* 2000; **56**:434–442.
43. Lee SM, Chao A. Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* 1994; **50**:88–97.
44. Rexstad E, Burnham KP. *User's Guide for Interactive Program CAPTURE*. Colorado Cooperative Fish and Wildlife Research Unit: Fort Collins, 1991.
45. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press: Cambridge, Mass., 1975.
46. Burnham KP, Overton WS. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 1978; **65**: 625–633.
47. Fienberg SE, Johnson MS, Junker BW. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of Royal Statistical Society, Series A* 1999; **162**:383–405.
48. Chao A, Lee SM, Jeng SL. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics* 1992; **48**:201–216.
49. Cormack RM. Loglinear models for capture-recapture. *Biometrics* 1989; **45**:395–413.
50. Agresti A. Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* 1994; **50**:494–500.
51. Lloyd CJ. *Statistical Analysis of Categorical Data*. Wiley: New York, 1999.
52. Norris JL, Pollock KH. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* 1996; **52**:639–649.
53. Chao A, Tsay PK. A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association* 1998; **93**: 283–293.
54. Tsay PK, Chao A. Population size estimation for capture-recapture models with applications to epidemiological data. *Journal of Applied Statistics* 2001; **28**:25–36.
55. Good IJ. The population frequencies of species and the estimation of population parameters. *Biometrika* 1953; **40**:237–264.
56. Bunge J, Fitzpatrick M. Estimating the number of species: recent developments. *Journal of the American Statistical Association* 1993; **88**:364–373.
57. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall: New York, 1993.
58. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; **43**:783–791.

59. MathSoft. *S-PLUS User's Manual, Version 4.0*. MathSoft, Inc.: Seattle, WA, 1997.
60. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology* 1999; **52**:917–926.
61. Hook EB, Regal RR. On the need for a 16th and 17th recommendation for capture-recapture analysis. *Journal of Clinical Epidemiology* 2000; **53**:1275–1277.
62. Gutteridge W, Collin C. Capture-recapture technique: quick and cheap (Letter). *British Medical Journal* 1994; **308**:531.
63. Black JFP, McLarty DG. Capture-recapture technique: difficult to use in developing countries (Letter). *British Medical Journal* 1994; **308**:531.
64. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; **46**:515–520.
65. Chang YF, LaPorte RE, Aaron DJ, Songer TJ. The importance of source selection and pilot study in the capture-recapture application. *Journal of Clinical Epidemiology* 1999; **52**:927–928.
66. Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture-recapture? *International Journal of Epidemiology* 2000; **29**:536–541.