AStA Adv Stat Anal (2009) 93: 23–47
DOI 10.1007/s10182-008-0085-y

ORIGINAL PAPER

Estimators in capture–recapture studies with two sources

Sarah Brittain · Dankmar Böhning

Received: 1 March 2008 / Accepted: 12 September 2008 / Published online: 21 October 2008

Abstract This paper investigates the applications of capture–recapture methods to human populations. Capture–recapture methods are commonly used in estimating the size of wildlife populations but can also be used in epidemiology and social sciences, for estimating prevalence of a particular disease or the size of the homeless population in a certain area. Here we focus on estimating the prevalence of infectious diseases. Several estimators of population size are considered: the Lincoln–Petersen estimator and its modified version, the Chapman estimator, Chao's lower bound estimator, the Zelterman's estimator, McKendrick's moment estimator and the maximum likelihood estimator. In order to evaluate these estimators, they are applied to real, three-source, capture-recapture data. By conditioning on each of the sources of three source data, we have been able to compare the estimators with the true value that they are estimating. The Chapman and Chao estimators were compared in terms of their relative bias. A variance formula derived through conditioning is suggested for Chao's estimator, and normal 95% confidence intervals are calculated for this and the Chapman estimator. We then compare the coverage of the respective confidence intervals. Furthermore, a simulation study is included to compare Chao's and Chapman's estimator. Results indicate that Chao's estimator is less biased than Chapman's estimator unless both sources are independent. Chao's estimator has also the smaller mean squared error. Finally, the implications and limitations of the above methods are discussed, with suggestions for further development.

Keywords Estimators of Chapman · Chao · Lincoln–Petersen · McKendrick · Zelterman · Covariate adjustment · Variance estimators

We are grateful to the Medical Research Council for supporting this work.

S. Brittain · D. Böhning (✉)
Quantitative Biology and Applied Statistics, School of Biological Sciences, University of Reading, Reading RG6 6BX, UK
e-mail: d.a.w.bohning@reading.ac.uk

1 Introduction

Capture–recapture methods are well established for estimating wildlife population sizes (see Seber 2002; Borchers et al. 2002). Typically, we cannot take a complete census of an entire animal population, so capture–recapture methods are used to formulate estimates of population size. In a wildlife setting, surveys of the population are carried out, where animals are captured, marked, re-released and allowed to mix with the population. On the second survey, the number of animals captured is again counted and marked, noting the number of animals which have already been marked on the first sampling occasion. This capture–recapture method continues for k surveys, and we can use the numbers of captured and recaptured animals obtained from all the surveys to estimate the total population size or the number of animals which were not caught in any of the k surveys. More recently, these methods have been applied to human populations in epidemiology (see Verstraeten et al. 2001; Gallay et al. 2000; Chao et al. 2001; Carrao et al. 2000) and in the social sciences for estimating number of drug users or the size of homeless populations (Smit et al. 2002; Roberts and Brewer 2006). In epidemiology, we can determine an estimate of disease occurrence by using a number of different sources (lists). We can treat each source as a survey occasion, and if an individual appears within a source, this is analogous to an animal being caught in a trap. If the same individual then appears in another source, this is analogous to an animal being caught in two surveys. Typically we have two or three incomplete lists (not including the entire population that is under study), available to us, usually in the form of hospital lists, treatment center registries or pharmacy records. Each list will identify some of the cases with the disease under study, but some individuals will remain undetected in any list (see also Hook and Regal 1995; IWGDMF 1995). The number of cases that does not appear in either list is unknown, and therefore the quantity that we wish to estimate. The methods used in estimating wildlife population can be applied in these public health situations if we can match individuals between lists. Animals in capture–recapture experiments are uniquely marked, and so we need to uniquely match individuals in separate lists. Matching is usually performed using information such as date of birth and initials of a patient. Obviously, the more matching criteria that is used, the more likely we are to achieve perfect matching between sources. If the matching between lists is not perfect, the validity of the estimates obtained may be affected. The aim of this contribution is to compare several different estimators of population size, using real data, to see how well they are performing, in terms of bias and variance.

2 Methods

We will begin by looking at the properties of some estimators of population size that are available to us. Here we will only be concerned with the situation where we have two sources, so we can compare newer methods with the classic Lincoln–Petersen and Chapman estimators, which can only be applied to two source data. Following this, comparisons will be made between the estimators when they are applied to real data. The data come from 19 three-source capture–recapture studies in epidemiology,

compiled by Van Hest et al. (2007), from current and recent papers on disease occurrence. Details of the disease under study and the sources that were used in each study can be found in the Appendix. For each study, we have three identifying sources. We have the number of patients identified in just one source, the number identified in two of the three sources and the number of cases observed in all three of the sources, giving us a total of six pieces of information for each data set.

Often, however, two-source applications occur in public health and epidemiology where the Lincoln–Petersen estimator can be applied. For this situation, we would like to investigate the Chao estimator as an alternative. The Lincoln–Petersen estimate is a well-known and therefore, commonly used estimator of population size. We wish to see if Chao's lower bound estimator is an improvement on this classic estimator. Also available to us is the Zelterman estimator, which is said to be an upper bound for population size.

To evaluate the estimators, we look at the three-source situations in the Appendix, where we condition on one of the three sources. We can then check the validity of our estimates by comparing them with the truth and with each other. Several assumptions must be made about our data when applying the estimates:

1. Independence between sources
2. A closed population
3. Independence between individuals.

However, in human populations, it is highly unlikely to have complete independence between sources. The nature of the sources used means there is sometimes high dependence between sources. For example, study 9 examines bacterial meningitis. The data sources for this study include the notifiable disease surveillance system, a voluntary hospital laboratory based surveillance system and the hospital discharge code registry. We would expect that hospitals have strict notification procedures, and so if a patient has been discharged from hospital after being diagnosed with meningitis, we would expect that individual to also appear in the notifiable disease surveillance system list. We can see from Table 8 that the log odds ratio between these two sources is in fact relatively high.

The second of our assumptions is also not easily satisfied. If the disease under study is fatal and an individual has been observed by one of the sources say, and subsequently dies, it would be no longer possible for that individual to be identified by a list which only identified live cases. In this case the closed population assumption would be violated, as we would have an exit from the population. We would expect independence between individuals unless we were studying a disease that could be passed to immediate family members. In this case, affected individuals in the same family may follow the same treatment route and therefore be identified by the same sources. We can check the independence between two sources by calculating the odds ratio. We can use this to see what effect any departures from independence have on the validity of our estimators in terms of relative bias. Hence if we take the top four lines of Table 1 (condition on source 1), we can calculate an estimate for the missing cases from sources 2 and 3 and then check this estimate against f_{100}. We can then condition on each of the three sources for the 19 data sets we have available, giving us 57 population estimates. We will assume a closed population and independence

Table 1 The three-source situation

Source 1	Source 2	Source 3	Frequency
1	1	1	f_{111}
1	1	0	f_{110}
1	0	1	f_{101}
1	0	0	f_{100}
0	1	1	f_{011}
0	1	0	f_{010}
0	0	1	f_{001}
0	0	0	f_{000} (unknown)

Table 2 The two-source situation

		Source 1		
		1	0	
Source 2	1	f_{11}	f_{01}	n_2
	0	f_{10}	f_{00}	
		n_1		

between individuals throughout, and that cases have been perfectly matched in each of the datasets.

3 An overview of estimators

3.1 Lincoln–Petersen estimator

The Lincoln–Petersen estimator is based on the odds ratio and can be used in the two source situation. It is assumed that identifying sources are independent and that cases are equally likely to be identified within each source. The two source situation leads to Table 2, with a case taking value 1 if present in a source and 0 otherwise.

If the two sources in the table were independent, the odds ratio would be close to unity. Without conditioning, f_{00}, the number of individuals that did not appear in either source, is unknown, and the quantity we wish to estimate. Under independence we have:

$$\frac{f_{11} f_{00}}{f_{10} f_{01}} \approx 1. \tag{1}$$

We can use this assumption of independence to give us an estimate for f_{00}:

$$\hat{f}_{00} = \frac{f_{10} f_{01}}{f_{11}}. \tag{2}$$

If $m_2 =$ number observed in both sources and $n_i =$ the number observed in source i, then the Lincoln–Petersen estimate is given by

$$\begin{aligned} \hat{N}_{\mathrm{LP}} &= f_{11} + f_{10} + f_{01} + \hat{f}_{00} = f_{11} + f_{10} + f_{01} + \frac{f_{10} f_{01}}{f_{11}} \\ &= m_2 + (n_1 - m_2) + (n_2 - m_2) + \frac{(n_1 - m_2)(n_2 - m_2)}{m_2} = \frac{n_1 n_2}{m_2}. \end{aligned} \tag{3}$$

If there is no overlap between sources ($m_2 = 0$), we cannot compute the Lincoln–Petersen estimator for population size. Another form of this estimator is the Chapman estimator, which is given by

$$\hat{f}_{00} = \frac{f_{10} f_{01}}{f_{11} + 1}. \tag{4}$$

The Chapman estimator of total population size is given by

$$\hat{N}_{\mathrm{CPM}} = \frac{(n_1 + 1)(n_2 + 1)}{m_2 + 1} - 1. \tag{5}$$

This estimate is less affected by zeros and is said to be less biased than the Lincoln–Petersen estimator. For these reasons, we will use the Chapman's modified form of the Lincoln–Petersen estimate to compare with the other estimators.

3.2 Chao's lower bound

Anne Chao (1987, 1989) proposed an alternative estimator of population size which relaxes the assumption that identifying sources are independent. For the two-source situation, we begin by looking at the mixed binomial with size parameter 2:

$$E(f_j) = N \int_0^1 \binom{2}{j} p^j (1-p)^{2-j} f(p)\, dp, \quad j = 0, 1, 2, \tag{6}$$

where $f_j =$ the number present in j sources, i.e. $f_1 = f_{10} + f_{01}$ and $f_2 = f_{11}$. We apply the Cauchy–Schwarz inequality to two random variables X and Y (see Kolmogorov and Formin 1970, for example):

$$\left[E(XY)\right]^2 \le E\left(X^2\right) E\left(Y^2\right). \tag{7}$$

Now if we choose $X = p$ and $Y = (1 - p)$ in (7) we have:

$$\left(\int_0^1 p(1-p) f(p)\, dp\right)^2 \le \int_0^1 (1-p)^2 f(p)\, dp \int_0^1 p^2 f(p)\, dp,$$

$$\left(\frac{1}{2} E(f_1)\right)^2 \le E(f_0) E(f_2),$$

from where

$$E(f_0) \ge \left\{ \frac{[E(f_1)]^2}{4E(f_2)} \right\}$$

follows. Replacing expected frequencies with observed frequencies provides the Chao lower bound estimate of f_{00}:

$$\hat{f}_0 = \frac{f_1^2}{4 f_2}. \tag{8}$$

If we take $n = f_{11} + f_{01} + f_{10}$, Chao's estimate of total population size is given by

$$\hat{N}_{\mathrm{C}} = n + \frac{f_1^2}{4f_2}. \tag{9}$$

3.3 Zelterman's upper bound

The Horvitz–Thompson estimator of population size is given by

$$\hat{N} = \frac{n}{1 - p_0}. \tag{10}$$

Zelterman (1988) proposed to estimate p_0 using only ones and twos from the zero-truncated count distribution which should not only perform well in terms of bias if the underlying count model is a Poisson, but also if contaminations occur such as expressed in a Poisson mixture model. The estimator of Zelterman (1988) is

$$\hat{N}_{\mathrm{Z}} = \frac{n}{1 - \exp(-2f_2/f_1)}. \tag{11}$$

In the two-source situation, we can apply the same method used by Zelterman to the mixed binomial likelihood, size parameter $m = 2$, with zeros truncated, as the cases that appear in neither of the two sources are unobserved in the capture–recapture setting. The binomial probabilities, with zeros truncated are given by the following:

$$P(X = 1) = \frac{2p(1-p)}{p^2 + 2p(1-p)} = \frac{2(1-p)/p}{1 + 2(1-p)/p},$$
$$P(X = 2) = \frac{p^2}{p^2 + 2p(1-p)} = \frac{1}{1 + 2(1-p)/p}.$$

The likelihood is then given by

$$\begin{aligned} L(p) &= P(X = 1)^{f_1} P(X = 2)^{f_2} \\ &= \left(\frac{2(1-p)/p}{1 + 2(1-p)/p}\right)^{f_1} \left(\frac{1}{1 + 2(1-p)/p}\right)^{f_2}. \end{aligned} \tag{12}$$

Now we can re-parameterize this likelihood using $\theta = \frac{1-p}{p} \rightarrow p = \frac{1}{1+\theta}$. This provides a binomial likelihood

$$L(\theta) = \left(\frac{2\theta}{1 + 2\theta}\right)^{f_1} \left(\frac{1}{1 + 2\theta}\right)^{f_2}$$

with associated log-likelihood

$$l(\theta) = f_1 \log(2\theta) + f_2 \log(1) - f \log(2\theta + 1),$$

where $f = f_1 + f_2$. Now, we can find the maximum likelihood estimate of θ by taking the derivative $\frac{\partial l}{\partial \theta} = \frac{f_1}{\theta} - \frac{2f}{2\theta+1}$, and equating it to zero leads to

$$\hat{\theta} = \frac{f_1}{2f_2}. \tag{13}$$

Using the unique transformation θ into p, we achieve an estimate for p, which in turn can be used to construct an estimate of p_0. This is detailed in the following using $\hat{p} = \frac{1}{1+\hat{\theta}}$:

$$\hat{p}_0 = (1-\hat{p})^2 = \left(1 - \frac{1}{1+\hat{\theta}}\right)^2 = \left(\frac{\hat{\theta}}{1+\hat{\theta}}\right)^2 = \left(\frac{f_1}{f_1+2f_2}\right)^2. \tag{14}$$

Putting this into the Horvitz–Thompson estimator will then give us the Zelterman estimator of population size in the binomial setting, namely

$$\hat{N}_{\mathrm{Z}} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-(\frac{f_1}{f_1+2f_2})^2}. \tag{15}$$

4 Some characterizations

Now we note two relationships between the estimators given in Sect. 3, which helps us to understand how they will perform with real data.

4.1 The identity of the Chao and the Zelterman estimators

Under the binomial model, with the number of capture occasions equal to 2, the estimators for the population sizes as proposed by Chao and Zelterman are identical. The formula for the estimated population size using the Chao estimate of f_0 is given by

$$\hat{N}_{\mathrm{C}} = f_1 + f_2 + \hat{f}_0 = f_1 + f_2 + \frac{f_1^2}{4f_2}. \tag{16}$$

According to (15), the Zelterman estimator is

$$\hat{N}_{\mathrm{Z}} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-(\frac{f_1}{f_1+2f_2})^2}, \tag{17}$$

and rearranging leads to

$$\begin{aligned}\hat{N}_{\mathrm{Z}} &= \frac{f_1+f_2}{1-\frac{f_1^2}{(f_1+2f_2)^2}} = \frac{(f_1+f_2)(f_1+2f_2)^2}{4f_1f_2+4f_2^2} = \frac{(f_1+f_2)(f_1+2f_2)^2}{4f_2(f_1+f_2)} \\ &= \frac{(f_1+2f_2)^2}{4f_2} = \frac{f_1^2+4f_1f_2+4f_2^2}{4f_2} = f_1+f_2+\frac{f_1^2}{4f_2} = \hat{N}_{\mathrm{C}},\end{aligned}$$

showing that the Zelterman estimator is identical to the Chao estimator for our situation.

4.2 Relationship between the Lincoln–Peterson estimator and the Chao estimator

If there is symmetry in the two by two table, then we can see that the Lincoln–Petersen estimator of f_0 is equal to the Chao estimator of f_0, and hence their estimates of the entire population are equal. Conformation of this is shown below. We have the following equivalence:

$$\hat{f}_{0,\text{L-P}} = \hat{f}_{0,C} \quad \Longleftrightarrow \quad \frac{f_{01} f_{10}}{f_2} = \frac{f_1^2}{4 f_2},$$

where $f_1 = f_{10} + f_{01}$. Assume that the following symmetry condition holds:

$$f_{01} = f_{10} = \frac{f_1}{2}.$$

Note that this assumption of symmetry is identical to the assumption that both sources have identical marginal probabilities of identifying a unit. Then, using the definition of the Lincoln–Petersen estimator we have:

$$\hat{f}_{0,\text{L-P}} = \frac{(\frac{f_1}{2})^2}{f_2},$$

which simplifies to $\frac{f_1^2}{4 f_2} = \hat{f}_{0,\text{C}}$. If the expected frequencies for the number of cases included in only one of the sources, were the same for both sources, then we would expect the Lincoln–Petersen and the Chao estimates to give us similar results.

4.3 Two alternative estimators

Two more estimators of population size are examined in this section, with similar findings to those in Sect. 4.1. Both estimators are developed under the assumption that the observed data follow a binomial distribution. Firstly the maximum likelihood estimator for p, which can then used in the Horvitz–Thompson estimator, $\hat{N} = \frac{n}{1-\hat{p}_0}$, to obtain the maximum likelihood estimate of population size. Similarly, we can use the McKendricks moment estimator of p to estimate population size in the capture–recapture setting.

Maximum Likelihood Estimator We can derive a general form for the maximum likelihood estimator of p_0 using the binomial distribution with trial parameter m and probability p, with zeros truncated. We have the following general zero-truncated likelihood:

$$\frac{\binom{m}{j} p^j (1-p)^{m-j}}{1-(1-p)^m}, \quad j = 1, \ldots, m.$$

Now, for $j = 1, 2$, the log-likelihood becomes

$$\begin{aligned} l(p) = {} & f_1 \big\{ \log(p) + (m-1)\log(1-p) \big\} \\ & + f_2 \big\{ 2\log(p) + (m-2)\log(1-p) \big\} - f \log\big(1-(1-p)^m\big) \end{aligned}$$

$$
\begin{aligned}
&= (f_1 + 2f_2)\log(p) + (fm - f_1 - 2f_2)\log(1-p) \\
&\quad - f\log\bigl(1-(1-p)^m\bigr) \\
&= (f_1 + 2f_2)\log(p) + (fm - f_1 - 2f_2)\log(1-p) \\
&\quad - f\log\bigl(1-(1-p)^m\bigr).
\end{aligned}
$$

Now, taking the derivative of this log-likelihood, we have:

$$
\begin{aligned}
\frac{\partial l}{\partial p} &= \frac{f_1 + 2f_2}{p} - \frac{fm - f_1 - 2f_2}{1-p} - \frac{fm(1-p)^{m-1}}{1-(1-p)^m} \\
&= (f_1 + 2f_2)\bigl[(1-p) - (1-p)^{m+1}\bigr] \\
&\quad - (fmp - f_1 p - 2f_2 p)\bigl[1-(1-p)^m\bigr] - fmp(1-p)^m \\
&= (f_1 + 2f_2)(1-p) - (f_1 + 2f_2)(1-p)^{m+1} \\
&\quad - fmp + f_1 p + 2f_2 p - (f_1 + 2f_2)p(1-p)^m \\
&= f_1 + 2f_2 - fmp - (f_1 + 2f_2)(1-p)^m\bigl[(1-p) + p\bigr].
\end{aligned}
$$

Now we equate the derivative of the log-likelihood to zero to get an expression for the maximum likelihood estimate $\hat{p}$ of p:

$$
\begin{aligned}
fm\hat{p} &= (f_1 + 2f_2)\bigl[1-(1-\hat{p})^m\bigr], \\
\hat{p} &= \frac{(f_1 + 2f_2)}{(f_1 + f_2)m}\bigl[1-(1-\hat{p})^m\bigr],
\end{aligned}
\tag{18}
$$

and, for $m = 2$,

$$
\begin{aligned}
\hat{p} &= \frac{(f_1 + 2f_2)}{(f_1 + f_2)2}\bigl[1-(1-\hat{p})^2\bigr] \\
&= \frac{(f_1 + 2f_2)}{2f}\bigl[2\hat{p} - \hat{p}^2\bigr],
\end{aligned}
$$

and a nonzero solution of this equation is provided by

$$
\hat{p}_{\mathrm{MLE}} = \frac{2f_2}{(f_1 + 2f_2)}. \tag{19}
$$

Using this estimate $\hat{p}$ of p, we can get an estimate for N using the Horvitz–Thompson estimator:

$$
\hat{N} = \frac{n}{1-(1-\hat{p})^m}
$$

and, for $m = 2$,

$$
\hat{N}_{\mathrm{MLE}} = \frac{n}{2\hat{p} - \hat{p}^2} = \frac{n}{\frac{4f_2}{f_1+2f_2} - \frac{4f_2^2}{(f_1+2f_2)^2}}
$$

$$= \frac{n}{\frac{4f_2(f_1+2f_2)-4f_2^2}{(f_1+2f_2)^2}} = \frac{n(f_1+2f_2)^2}{4f_2(f_1+f_2)}$$

$$= \frac{(f_1+2f_2)^2}{4f_2} = \frac{f_1^2}{4f_2} + f_1 + f_2 = \hat{N}_{\mathrm{C}},$$

showing that the maximum likelihood estimate agrees with Chao's and Zelterman's estimator. More generally, when m is greater than two, we can see from (18) that there is no closed form expression for the mle $\hat{p}$ of p. To obtain an estimate for p in this situation an iterative procedure could be used, or the Taylor series expansion of $(1-\hat{p})^m$ could be applied to find an approximate solution.

McKendricks Moment Estimator We know that when X_i is binomial with size parameter m and success parameter p, then $E(X_i) = mp$ and

$$\mathrm{Var}(X_i) = E\big(X_i^2\big) - \big[E(X_i)\big]^2 = mp(1-p), \quad \text{or}$$

$$E\big(X_i^2\big) = \mathrm{Var}(X_i) + \big[E(X_i)\big]^2 = mp(1-p) + (mp)^2.$$

We also know that the mean of $\bar{X} = \frac{1}{N}\sum X_i$ is given by $E(\bar{X}) = mp$. Using the *method of moments*, we equate the expected value of the $\bar{X}$ with the sample mean $\bar{x}$ to obtain an expression for the sum of the x_i in terms of the binomial parameters:

$$mp = \frac{1}{N}\sum x_i. \tag{20}$$

Since the expectation and summation can be interchanged and using (20) we have:

$$E\left(\sum X_i^2\right) = \sum E\big(X_i^2\big) = Nmp\big[(1-p)+mp\big]. \tag{21}$$

Averaging the expression (21) gives:

$$E\left(\frac{1}{N}\sum X_i^2\right) = mp(1-p) + (mp)^2,$$

which we equate to the second sample moment $\frac{1}{N}\sum x_i^2$ giving the second moment equation

$$\sum x_i^2 = Nmp\big[(1-p)+mp\big] = \sum x_i\big[(1-p)+mp\big],$$

where we have used the first moment equation (20). Rearranging this, we can obtain an expression for $\hat{p}_{\mathrm{MO}}$

$$\frac{\sum x_i^2}{\sum x_i} = 1 + \hat{p}_{\mathrm{MO}}(m-1), \quad \text{or}$$

$$\hat{p}_{\mathrm{MO}} = \frac{\frac{\sum x_i^2}{\sum x_i} - 1}{m-1} = \frac{\sum x_i^2 - \sum x_i}{\sum x_i(m-1)}.$$

Now, we know the sum of the x_i to be $f_1 + 2f_2$, and the sum of the x_i^2 to be $f_1 + 4f_2$, since f_1 is the frequency of those x_i that take the value 1, and f_2 is the frequency of those x_i that take the value 2. Substituting these values into the above gives:

$$\hat{p}_{\mathrm{MO}} = \frac{f_1 + 4f_2 - f_1 - 2f_2}{(f_1 + 2f_2)(m-1)} = \frac{2f_2}{(f_1 + 2f_2)(m-1)}.$$

As a fundamental result we see that this moment estimator of p_0 is equal to the maximum likelihood estimate of p_0, when we have only two sources.

We can use this estimate to obtain an estimate for N by substituting into (20):

$$\sum x_i = Nm\frac{\sum x_i^2 - \sum x_i}{\sum x_i(m-1)}, \quad \text{or}$$

$$\hat{N}_{\mathrm{MO}} = \frac{(m-1)(\sum x_i)^2}{m(\sum x_i^2 - \sum x_i)} = \frac{(m-1)(f_1 + 2f_2)^2}{m(f_1 + 4f_2 - f_1 - 2f_2)} = \frac{(m-1)(f_1 + 2f_2)^2}{m2f_2}.$$

This is the moment estimator for the total population. In the two source case, $m = 2$. Substituting this value into the above, we have that

$$\hat{N}_{\mathrm{MO}} = \frac{(f_1 + 2f_2)^2}{4f_2} = \frac{f_1^2 + 4f_1f_2 + 4f_2^2}{4f_2} = \frac{f_1^2}{4f_2} + f_1 + f_2 = \hat{N}_{\mathrm{C}},$$

which again coincides with the Chao estimator.

To summarize we have looked at the following approaches and estimators:

1. Chao's estimator
2. Zelterman's estimator
3. Maximum likelihood estimator for a zero-truncated binomial with size parameter 2
4. The associated moment estimator (McKendrick's estimator)

and shown that all four agree in our situation of a capture–recapture setting with two sources. The only substantially different estimator remaining is the Lincoln–Petersen or Chapman estimator. Hence, in our further analysis we will concentrate on comparing *only* Chao's and Chapman's estimator.

5 Analysis

In order to see how these estimators were behaving with real data, they were applied to the 19 data sets available to us. The data were collected from recent and current papers on disease occurrence. As we are interested in the two-source situation and we have three-source data sets, we can condition on each of the three sources from each data sets so we have 57 values with which to compare the different estimators. The resulting estimates can be found in the Appendix.

To compare the estimators we will look at several measures. One of these is the relative bias, which is calculated as the difference between true and the estimated value of f_{00} and then scaled by the estimate of the number of unobserved values ($\hat{f}_{00}$).

We will also look at a variance formula for the estimators of Chao and Zelterman in this two-source binomial situation when their estimates coincide. We can then compare this with the variance of the Chapman estimator by computing the associated confidence intervals, to see if these confidence intervals do in fact include the true value of N.

5.1 Relative bias

In order to compare the different estimates for the number of missing cases, we can calculate the relative bias. Here, we take this to be the following:

$$\text{Relative bias} = \frac{f_{00} - \hat{f}_{00}}{\hat{f}_{00}}. \tag{22}$$

A good estimate of population size will have an associated relative bias close to zero. Calculating this relative bias for the Chao and the Chapman estimators for the 57 estimates that we have (three for each dataset), we can see that there are only seven occurrences when the Chapman estimator performs 'better' than the Chao estimator, i.e. it has a relative bias closer to zero. This is important for capture–recapture studies, as the Chapman estimator is still the most commonly used. However, here we see that Chao's alternative estimator is less biased than the Chapman estimator and is just as simple to calculate.

Figure 1 shows the relative bias plotted against the log odds ratio, and we can see a smoothness in the relationship between the log odds ratio and the relative bias of

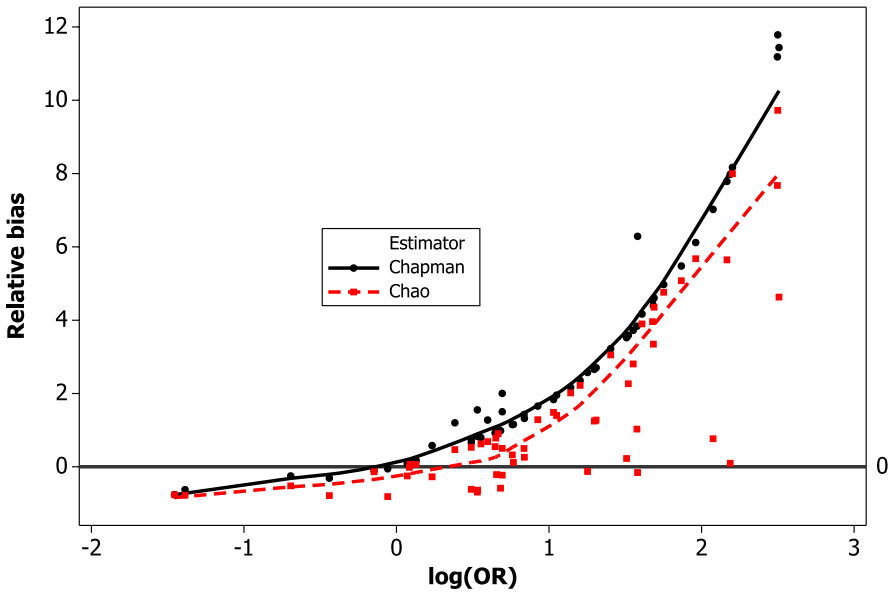

Fig. 1 Relative bias against log odds ratio for the Chapman and Chao estimators

the Chapman estimator. Investigating this relationship further, we have:

$$\log(OR) = \log\left(\frac{f_{11}f_{00}}{f_{10}f_{01}}\right), \tag{23}$$

$$\text{Chapman Relative bias} = \frac{f_{00} - \hat{f}_{00}}{\hat{f}_{00}}. \tag{24}$$

Now,

$$\begin{aligned}
\hat{f}_{00} &= \frac{f_{10}f_{01}}{f_{11}+1} \\
\text{Substituting into (24)} &= \frac{f_{00}}{\hat{f}_{00}} - 1 \\
&= \frac{f_{00}(f_{11}+1)}{f_{10}f_{01}} - 1 \\
&= \frac{f_{11}f_{00}}{f_{10}f_{01}} + \frac{f_{00}}{f_{10}f_{01}} - 1 \\
&= OR\left(1 + \frac{1}{f_{11}}\right) - 1 \quad \text{so that}
\end{aligned}$$

$$\log(\text{Chapman relative bias} + 1) = \log(OR) + \log\left(1 + \frac{1}{f_{11}}\right). \tag{25}$$

We can see that the relative bias for the Chapman estimator is a function of the odds ratio, hence the smoothness when the relative bias is plotted against the $\log(OR)$. We can see that not all points lie exactly on the line that has been fitted. When f_{11} is large, the second term on the right-hand side of (25) is close to zero. This fact, along with a positive odds ratio, makes it appear as though these points are anomalous results. Six of these points come from studies 3a and 3b where the number of cases identified by all three sources was four and two respectively. The other occurrence is study 5, where f_{111} is equal to two. However, it is only when we condition on source one that this point departs from the fitted line in Fig. 1, as this is the only occasion when conditioning on study 5 produces a positive log odds ratio.

As we have seen in Sect. 4.2, we can expect that the Chapman and Chao estimators are close when the identifying probabilities of the two sources are similar. This comes out in the analysis if we plot the relative bias against the difference of the estimated capture probabilities of source 1 and source 2. Figure 2 shows the associated scatterplot with the LOWESS-smoother included. Clearly, the benefit of using the Chao estimator diminishes if this difference is close to zero.

We have defined the relative bias as $\frac{f_{00}-\hat{f}_{00}}{\hat{f}_{00}}$. Alternatively, the relative bias could have been defined as the Pearson-type residual $\frac{f_{00}-\hat{f}_{00}}{\sqrt{\hat{f}_{00}}}$. However, the results and graphs were not substantially different so that they are not presented here for the sake of brevity.

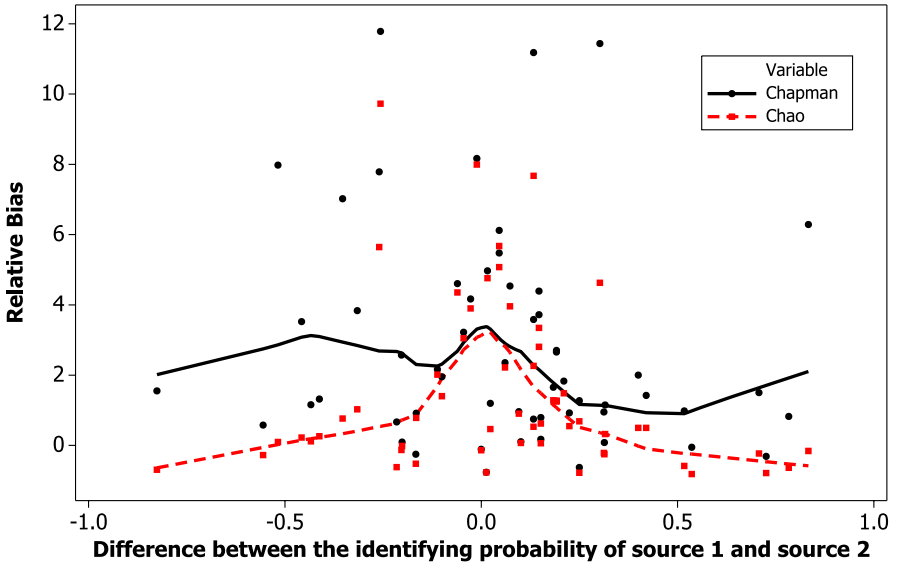

Fig. 2 Relative bias against the difference of the capture probabilities of the two sources for the Chapman and Chao estimators

5.2 Variance estimation

We now consider variance estimation for our estimates. An approximately unbiased estimate of the variance of $\hat{N}_{\text{CPM}}$ was derived by Seber (1970) to be as follows:

$$\text{Var}(\hat{N}) = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}, \tag{26}$$

where $n_1 = f_{11} + f_{10}$, $n_2 = f_{11} + f_{01}$, $m_2 = f_{11}$ We can calculate this variance for our 57 estimates of N and also construct normal based confidence intervals.

Similarly, this can be done for the Chao estimate. A relatively simple variance estimate can be found by applying a general technique suggested in Böhning (2008) (see also Van der Heijden et al. 2003) to the Zelterman estimator (which is identical here to Chao's estimator) and leads to

$$\widehat{\text{Var}}(\hat{N}_Z) = f_2\hat{\theta}^2(\hat{\theta}+1)^2 = \frac{f_1^2}{4f_2}\left(\frac{f_1}{2f_2}+1\right)^2. \tag{27}$$

More details can be found in Brittain and Böhning (2007).[1]

We now consider the two variance estimators for our data sets. Figure 3 shows that as the f_1/f_2 ratio increases, the variance for both the Chapman and Chao estimators increases. In Fig. 3, the relative variance was plotted (variance divided by n), and the natural log of this quantity was taken to scale the sometimes large estimates of

[1] Unpublished, available on request.

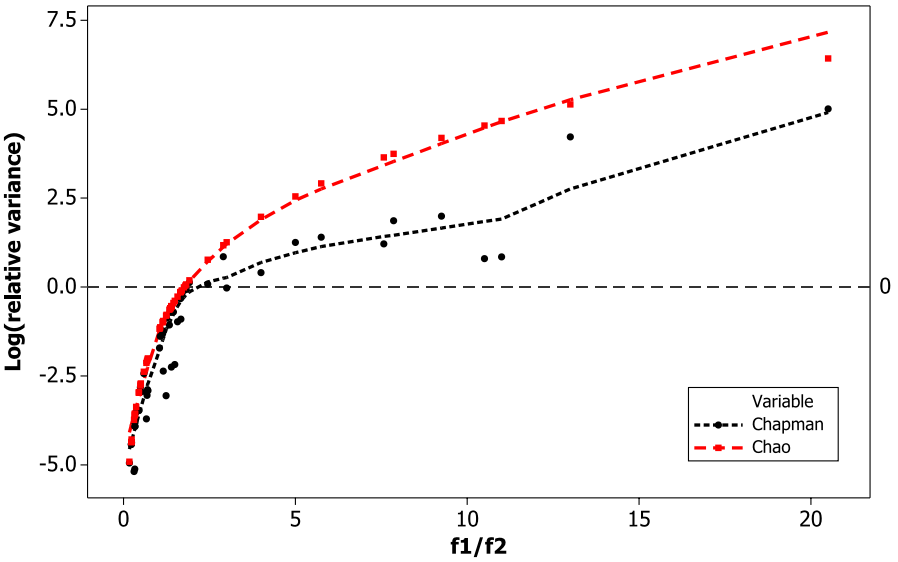

Fig. 3 Log(relative variance) against the f_1/f_2 ratio for all 57 estimates

the variance. When the f_1/f_2 ratio is close to zero or smaller, the variances for the different estimators are fairly similar. However, as the ratio increases, the variance for the Chao estimator is larger than the Chapman variance. The variance of Chao estimator is a smooth function of the f_1/f_2 ratio, which we can also see from Fig. 3. Although the Chapman variance is less than the variance of Chao's estimator in all of the datasets, this does not make it necessarily a better estimate.

We cannot compare the size of the variances directly as they are variances of different quantities. To make any meaningful comparison, we need to calculate confidence intervals to see how often these contain the true value of N for each of the estimators. Obviously, this will also depend on how good the original estimates were, and we have already seen in Sect. 5.1 that the relative bias for Chao's estimator was lower than that of Chapman's estimator in most cases.

The normal based 95% confidence intervals were calculated with the variance formulae for the Chapman and Chao estimators given in (26) and (27) respectively. The results can be found in the Appendix. We can then compare these intervals to see whether or not they contain the true value of N for each of the 57 confidence intervals. Table 3 is a summary of how often each of the confidence intervals for the two estimators contained the truth. ($1 =$ Covered N, $0 =$ Did not cover N.) We can see that the confidence interval for the Chao estimator, using the variance derived above, includes the true value of N 26 times out of the 57, whereas the Chapman confidence interval only includes the truth 19 times. We can also see from the table that there are only two occasions out of 19 (11%) when the Chapman confidence interval covers the truth, and the Chao confidence interval does not. From 26 occasions in which Chao's estimator covered the true N, there were 9 occasions (35%) in which the Chapman

Table 3 Summary table of number of times 95% confidence intervals cover the true value of N (1 = C.I covers N, 0 = True value not covered by 95% C.I)

		Chapman		Total
		1	0	
Chao	1	17	9	26
	0	2	29	31
	Total	19	38	57

Table 4 Design of the simulation study with capture probabilities p_{00}, p_{10}, p_{01}, p_{11} and odds ratio $OR = \frac{p_{00}p_{11}}{p_{10}p_{01}}$

Population number	p_{00}	p_{10}	p_{01}	p_{11}	OR
1	0.25	0.25	0.25	0.25	1.00000
2	0.30	0.20	0.25	0.25	1.50000
3	0.30	0.15	0.30	0.25	1.66667
4	0.35	0.20	0.20	0.25	2.18750
5	0.35	0.15	0.20	0.30	3.50000
6	0.35	0.10	0.20	0.35	6.12500

estimator did not. We can also see that when only the confidence interval for one of the estimators covers the truth, it is Chao's estimator nine times out of 11.

This would suggest that the Chao estimator is a more appropriate estimator of population size, particularly when the independence assumption is in doubt.

6 Simulation study

To provide further insights into the behaviour of the two estimators a simulation study was designed. The population size was chosen $N = 100$ and capture–recapture probabilities were chosen according to Table 4. For example, in the second population the probability p_{11} that an individual is identified by both sources is 0.25, the probability p_{10} that it is identified by the first source and not by the second is 0.20, the probability p_{01} that it is identified by the second source but not the first source is 0.25, and the probability p_{00} that it remains unidentified at all is 0.30. For each of the 6 populations, 1000 samples were generated, each of these leading to frequencies f_{00}, f_{10}, f_{01} and f_{11} fulfilling that their sum is 100. The fact that the frequency f_{00} is known was ignored and was estimated instead using Chapman's and Chao's estimator. Since this was done 1000 times, their bias, variance and mean squared error could be calculated.

Note that the population differ in various aspects. Most importantly, they carry different dependence structures between the two sources. For example, in the first population the sources are independent. This can be seen by looking at the odds ratio defined as $OR = \frac{p_{00}p_{11}}{p_{10}p_{01}}$, which takes the value 1 in this case. For all other populations, the odds ratio is calculated in the last column of Table 4. Clearly, the populations are ordered with respect to their dependence structure as measured by the odds ratio. We expect that the odds ratio is a key factor in the behaviour of the estimators. In

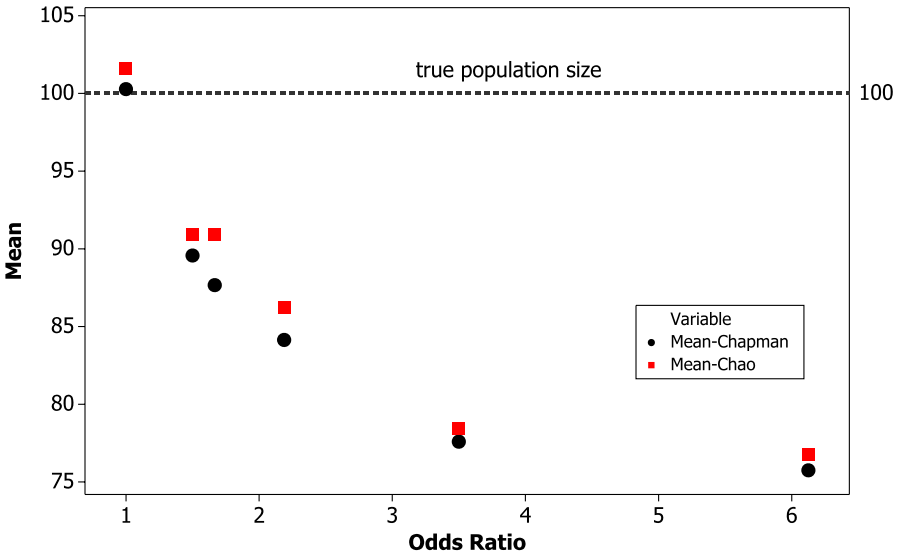

Fig. 4 Mean of Chapman and Chao estimator against the odds ratio for the six populations of Table 4

fact, it is known that with increasing dependencies between sources the Lincoln–Petersen/Chapman estimator will underestimate the population size. The reason for this behaviour in empirical data has been given in (25), which relates the relative bias of the Chapman estimator to the odds ratio. Another argument on population level is as follows. Consider the odds ratio estimate $\widehat{OR} = \frac{f_{00} f_{11}}{f_{10} f_{01}}$. The Lincoln–Petersen estimate is found by setting the $\widehat{OR} = 1$ and solving the occurring equation for f_{00}. If there is positive dependency, the estimate $\hat{f}_{00} = \widehat{OR} \frac{f_{10} f_{01}}{f_{11}}$ should be used. Note that

$$\widehat{OR} \frac{f_{10} f_{01}}{f_{11}} \geq \frac{f_{10} f_{01}}{f_{11}}, \tag{28}$$

where the expression on the right-hand side of (28) is conventionally used in estimation. Hence, the underestimation.

Figure 4 shows that both estimators indeed underestimate the true population size, the more the higher the odds ratio. However, Chao's estimator shows consistently less bias unless we are in the case of independence where we know that Chapman's estimator is unbiased.

As known from the analysis in the previous chapters, Chao's estimators has a larger variance than the one of Chapman. This is also visible in Fig. 5. Hence, it might be valuable to look at the trade-off measure of mean squared error (MSE), which is a sum of squared bias and variance. The resulting Fig. 6 indicates that Chao's estimator has superior MSE in all cases unless we are in the independence case.

We have seen in Sect. 4.2 that Chao's estimator will be close to the Chapman estimator if the identifying sources have similar marginal probabilities. In other words, we expect that the estimators of Chao and Chapman agree if $p_{11} + p_{10} = p_{11} + p_{01}$.

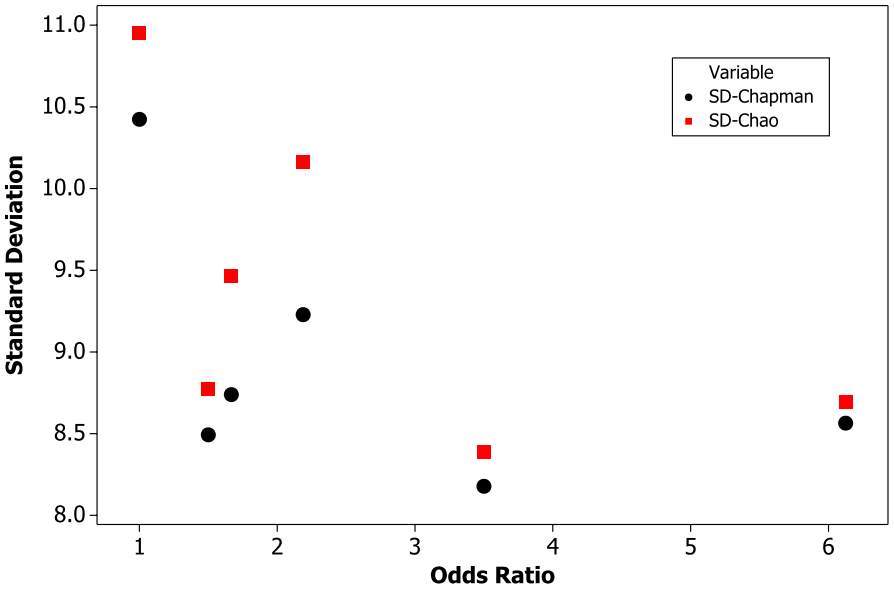

Fig. 5 Standard deviation of Chapman and Chao estimator against the odds ratio for the six populations of Table 4

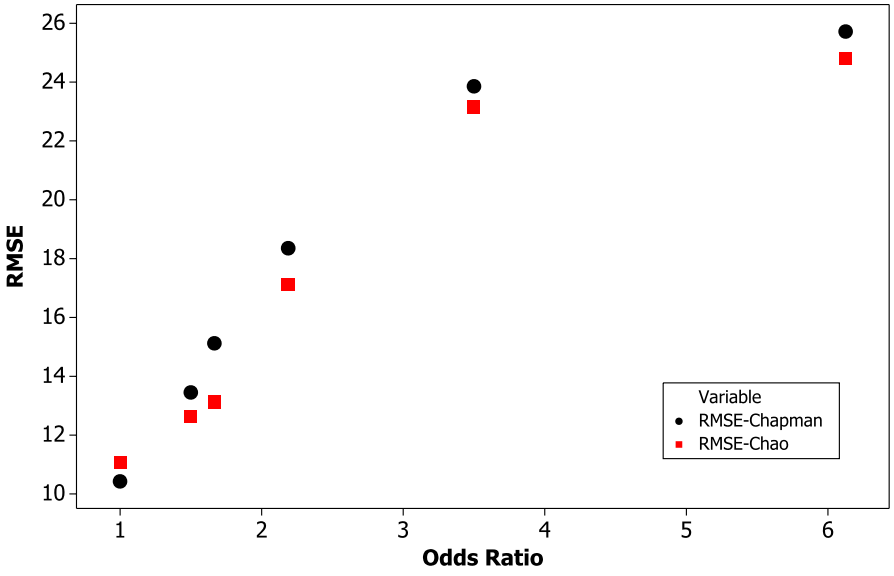

Fig. 6 Root mean squared error of Chapman and Chao estimator against the odds ratio for the six populations of Table 4

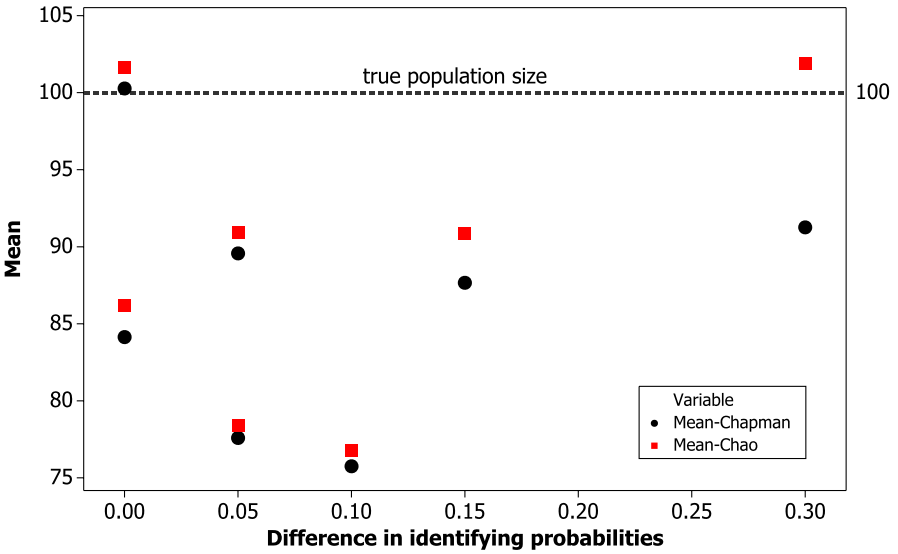

Fig. 7 Mean of Chapman and Chao estimator against the odds ratio for the six populations of Table 4 and a further one with more extreme difference in the identifying probabilities

Table 5 Design of the simulation study with capture probabilities p_{00}, p_{10}, p_{01}, p_{11} and difference between identifying probabilities $p_{01} - p_{10}$

Population number	p_{00}	p_{10}	p_{01}	p_{11}	$p_{01} - p_{10}$
1	0.25	0.25	0.25	0.25	0.00
2	0.30	0.20	0.25	0.25	0.05
3	0.30	0.15	0.30	0.25	0.15
4	0.35	0.20	0.20	0.25	0.00
5	0.35	0.15	0.20	0.30	0.05
6	0.35	0.10	0.20	0.35	0.10
7	0.25	0.10	0.40	0.25	0.30

Table 5 displays the populations with respect to their difference in the marginal, identifying probabilities $(p_{11} + p_{01}) - (p_{11} + p_{10}) = p_{01} - p_{10}$. As can be seen, these differences range from 0 to 0.15 for the populations 1 to 6. To study a more extreme case, we have added population 7 with a difference of 0.3 and an odds ratio of 1.56. The results of the simulation study (including population 7 now) with respect of bias are provided in Fig. 7. With the new population included, it becomes clear that Chao's estimator will be most beneficial if there is dependency between sources as well as both sources differ in their identification probabilities.

7 Discussion

We have seen several estimators being applied to epidemiological data. We have been able to see how well they are performing, as we have been able to compare each of our estimates with the truth and each other. Overall, we have seen that Chao's estimator has a lower relative bias, a larger variance, but with better confidence interval coverage than the more commonly used Chapman estimator.

However, Van Hest et al. (2007) suggest that this technique for evaluating estimators has its limitations, particularly when dealing with infectious disease incidence data. We have seen that the dependencies between two of the sources varies greatly between data sets, and most of the two sources do have a positive dependence. With this data, we should expect to see positive interdependence between the three sources (Van Hest et al. 2007). Extracting 2×2 tables in this manner ignores these possible dependencies, thus the estimates that we obtained for the total population size are not using all possible information available.

Other limitations of these methods include the violation of assumptions. Already discussed is the independence between sources, but we also assume that each individual has equal ascertainment probability within sources. This assumption is more likely to be violated in human populations than animal populations as we would expect human behaviour to be less homogeneous than that of animals. Whether a member of the population is identified by a particular source may depend on several covariates such as severity of disease, the location of that individual in the study area, or gender. This suggests that regression modelling could be a valuable tool in achieving a more accurate estimate of the population, especially when individuals have different ascertainment probabilities.

The assumption that we are dealing with a closed population is also likely to be violated in capture–recapture experiments. Particularly when we are considering fatal diseases, where death will mean an exit from the population. There are other methods available for dealing with open populations (Seber 2002), and a comparison of these methods would test the validity of this assumption in our situation.

Using real data sources as well as simulated data, we have been able to demonstrate that Chao's lower bound does provide a more reliable estimate of population size, especially when the odds ratio is high between sources and the difference between the identifying probabilities of the two sources are large, by looking at the relative bias for the real data sources and by looking at bias and mean squared error for the simulated data. This important result means that we have an estimator that is as simple to calculate as the more commonly used Chapman estimator and is a better predictor of population size. The ease of calculation of Chao's estimator is particularly beneficial as some studies are carried out by epidemiologists with limited mathematical ability. Here we have presented evidence for wider use of Chao's estimator as a more reliable alternative to the Chapman estimator.

We have also provided a simple formula to calculate the variance of Chao's estimator (27). Although the confidence interval for this estimate is wider, it is more likely to include the true value of N than the confidence interval for the Chapman estimate.

In conclusion, Chao's estimator has been shown to be more viable than Chapman's estimator, and efforts should be made to introduce it more widely to epidemiologists

and biologists, many of whom preferentially use the Chapman estimator. An evaluation of the use of covariates in estimating population size would also be a topic for further investigation. Whether or not the extra effort of data collection gave us a significantly more accurate estimate of population size, compared with Chao's estimator would need to be assessed.

Acknowledgements The first author wishes to thank the Medical Research Council for providing sponsorship for her M.Sc. Both authors would like to thank the Editor of AStA *Advances in Statistical Analysis* for the opportunity to publish this work. They are also very grateful to the Associate Editor for many very helpful comments.

Appendix

In the following we provide more details on the studies which formed the empirical base for all comparisons.

Table 6 Overview of studies: Three-source capture–recapture studies of infectious diseases (Van Hest et al. 2007)

Study	Disease	Objective	Data sources
1	Legionnaires' disease	To estimate the level of underreporting of Legionnaires' disease and to evaluate the feasibility of a laboratory based reporting system in France in 1995	1. National Notification System 2. Reference laboratory 3. Hospital laboratory survey
2	AIDS/HIV	To estimate the completeness of the prison AIDS register in Spain in 2000	1. Prison register of AIDS patients 2. Prison register of tuberculosis patients 3. Prison register of hospital admissions
3a, 3b	Pertussis	To estimate under notification of whooping cough in the north west of England, 1994–1996	1. Notification data from the office for national statistics (ONS) 2. Hospital admission data 3. Public health laboratory reports
4	Salmonella infection	To assess the number of foodborne Salmonella outbreaks in France, 1995	1. Mandatory public health notification 2. Mandatory veterinary notification 3. National Salmonella reference centre
5	Pertussis	To improve estimates of pertussis deaths in England and identify reasons for under ascertainment, 1994–1999	1. Hospital episode statistics 2. Enhanced laboratory pertussis surveillance 3. ONS mortality data

Table 6 (*Continued*)

Study	Disease	Objective	Data sources
6	Meningococcal Meningitis	To evaluate the exhaustiveness of three information sources on meningococcal disease in Tenerife, Spain, 1999–2001	1Mandatory notifiable disease surveillance system 2. Laboratory survey 3. Hospital discharge codes
8	Tuberculosis	Assessment of completeness of the tuberculosis systems in the Piedmont region of Italy, 2001	1. Physician notification system 2. Reference laboratory 3. Hospital admission statistics
9	Meningitis, bacterial	to estimate the incidence of bacterial meningitis and to assess the quality of the surveillance systems in the Lazio region of Italy, 1995–1996	1. Notifiable disease surveillance system 2. Voluntary hospital laboratory-based-surveillance system 3. Hospital discharge code registry
10	Malaria	To estimate the completeness of notification of malaria by physicians and laboratories in The Netherlands in 1996	1. Passive national notification register 2. Active laboratory survey 3. National hospital admission registration
11	Legionnaires' disease	To evaluate improvements made to the mandatory notification system for Legionnaires' disease in France in 1998	1. National Notification System 2. Reference laboratory 3. Hospital laboratory survey
12	Hepatitis A	Estimation of individuals infected with hepatitis A during an outbreak in Taiwan, 1995	1. Laboratory serological test records 2. Hospital records 3. Epidemiological questionnaires
13a & 13b	Tuberculosis	Description of systematic examination and case-verification, record-linkage, capture–recapture analysis and assessment of the completeness of three tuberculosis registers in The Netherlands, 1998	1. Physician notification system 2. Reference laboratory 3. Hospital admission statistics
14	Tuberculosis	Description of case-verification, record-linkage, capture–recapture analysis and assessment of completeness of three tuberculosis registers in England, 1999–2002	1. Notification 2. Laboratory reports 3. Hospital discharge codes
15	Legionnaires' disease	Assessment of Legionnaires' disease incidence and completeness of notification in The Netherlands, 2000–2001	1. Passive national notification register 2. Active laboratory survey 3. National hospital admission registration
16a & 16b	Meningococcal disease	Assessment of completeness of three data sources for meningococcal disease after correction for false-positive diagnoses in The Netherlands, 1993–1999	1. Notification register 2. Hospital Episode Statistics 3. Reference Laboratory for bacterial meningitis records

Table 7 Population size estimates: Conditioning on Source 1

Study	f_{100}	f_{110}	f_{101}	f_{111}	Estimates of f_{100}		log-OR	$\frac{f_1}{f_2}$	Relative bias	
					Chap.	Chao			Chap.	Chao
1	7	6	10	14	4.00	4.57	0.49	1.14	0.75	0.53
2	26	17	29	33	14.50	16.03	0.55	1.39	0.79	0.62
3a	24	19	4	4	15.20	33.06	0.23	5.75	0.58	−0.27
3b	17	20	1	2	6.67	55.13	0.53	10.50	1.55	−0.69
4	45	10	39	20	18.57	30.01	0.84	2.45	1.42	0.50
5	12	2	6	2	4.00	8.00	0.69	4.00	2.00	0.50
6	5	4	7	30	0.90	1.01	1.68	0.37	4.54	3.96
7	1	14	15	49	4.20	4.29	−1.46	0.59	−0.76	−0.77
8	125	183	96	153	114.08	127.19	0.08	1.82	0.10	−0.02
9	5	7	6	76	0.55	0.56	2.20	0.17	8.17	7.99
10	54	37	94	123	28.05	34.88	0.65	1.07	0.93	0.55
11	132	77	52	95	41.71	43.79	1.14	1.36	2.16	2.01
12	69	21	17	28	12.31	12.89	1.69	1.36	4.61	4.35
13a	78	30	510	388	39.33	187.89	0.68	1.39	0.98	−0.58
13b	40	30	548	388	42.26	215.26	−0.06	1.49	−0.05	−0.81
14	7777	6503	3789	6075	4055.28	4359.06	0.65	1.69	0.92	0.78
15	56	31	131	155	26.03	42.33	0.76	1.05	1.15	0.32
16a	189	189	314	2234	26.55	28.31	1.96	0.23	6.12	5.68

Table 8 Population size estimates: Conditioning on Source 2

Study	f_{100}	f_{110}	f_{101}	f_{111}	Estimates of f_{100}		log-OR	$\frac{f_1}{f_2}$	Relative bias	
					Chap.	Chao			Chap.	Chao
1	73	6	100	14	40.00	200.64	0.53	7.57	0.83	−0.64
2	17	17	29	33	14.50	16.03	0.13	1.39	0.17	0.06
3a	285	19	33	4	125.40	169.00	0.60	13.00	1.27	0.69
3b	308	20	21	2	140.00	210.13	0.38	20.50	1.20	0.47
4	24	10	19	20	9.05	10.51	0.93	1.45	1.65	1.28
5	1	2	4	2	2.67	4.50	−1.39	3.00	−0.63	−0.78
6	2	4	3	30	0.39	0.41	1.61	0.23	4.17	3.90
7	1	14	1	49	0.28	1.15	1.25	0.31	2.57	−0.13
8	64	183	6	153	7.13	58.37	2.19	1.24	7.98	0.10
9	52	7	46	76	4.18	9.24	2.51	0.70	11.43	4.63
10	41	37	127	123	37.90	54.67	0.07	1.33	0.08	−0.25
11	93	77	105	95	84.22	87.17	0.09	1.92	0.10	0.07
12	55	21	18	28	13.03	13.58	1.40	1.39	3.22	3.05
13a	93	30	99	388	7.63	10.72	2.50	0.33	11.18	7.67
13b	35	30	99	388	7.63	10.72	1.52	0.33	3.58	2.26
14	2478	6503	512	6075	547.98	2025.11	1.51	1.15	3.52	0.22
15	30	31	45	155	8.94	9.32	1.20	0.49	2.35	2.22
16a	253	189	808	2234	68.33	111.24	1.31	0.45	2.70	1.27
16b	250	189	808	2234	68.33	111.24	1.30	0.45	2.66	1.25

Table 9 Population size estimates: Conditioning on Source 3

Study	f_{100}	f_{110}	f_{101}	f_{111}	Estimates of f_{100}		log-OR	$\frac{f_1}{f_2}$	Relative bias	
					Chap.	Chao			Chap.	Chao
1	46	10	100	14	66.67	216.07	−0.44	7.86	−0.31	−0.79
2	22	29	29	33	24.74	25.48	−0.15	1.76	−0.11	−0.14
3a	66	4	33	4	26.40	85.56	0.69	9.25	1.50	−0.23
3b	51	1	21	2	7.00	60.50	1.58	11.00	6.29	−0.16
4	451	39	19	20	35.29	42.05	2.50	2.90	11.78	9.73
5	6	6	4	2	8.00	12.50	−0.69	5.00	−0.25	−0.52
6	2	7	3	30	0.68	0.83	1.05	0.33	1.95	1.40
7	0.5	15	1	49	0.30	1.31	0.49	0.33	0.67	−0.62
8	30	96	6	153	3.74	17.00	2.08	0.67	7.02	0.76
9	7	6	46	76	3.58	8.89	0.66	0.68	0.95	−0.21
10	189	94	127	123	96.27	99.27	0.67	1.80	0.96	0.90
11	161	52	105	95	56.88	64.87	1.03	1.65	1.83	1.48
12	63	17	18	28	10.55	10.94	1.75	1.25	4.97	4.76
13a	301	510	99	388	129.79	238.97	0.84	1.57	1.32	0.26
13b	301	548	99	388	139.47	269.72	0.77	1.67	1.16	0.12
14	1544	3789	512	6075	319.28	761.26	1.58	0.71	3.84	1.03
15	332	131	45	155	37.79	49.96	2.17	1.14	7.79	5.65
16a	612	314	808	2234	113.52	140.88	1.68	0.50	4.39	3.34
16b	536	314	808	2234	113.52	140.88	1.55	0.50	3.72	2.80

References

Böhning, D.: A simple variance formula for population size by conditioning. Stat. Methodol. **5**, 410–423 (2008)

Borchers, D.L., Buckland, S.T., Zucchini, W.: Estimating Animal Abundance. Springer, Berlin (2002)

Brittain, S., Böhning, D.: Supplementary material for 'Estimators in capture–recapture studies with two sources' (2007, unpublished)

Carrao, G., Bagnardi, V., Vittadini, G., Favilli, S.: Capture–recapture methods to size alcohol related problems in a population. J. Epidemiol. Community Health **54**, 603–610 (2000)

Chao, A.: Estimating the population size for capture–recapture data with unequal catchability. Biometrics **43**, 783–791 (1987)

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

Chao, A., Tsay, P.K., Lin, S.H., Shau, W.Y., Chao, D.Y.: Tutorial in Biostatistics: The applications of capture–recapture models to epidemiological data. Stat. Med. **20**, 3123–3157 (2001)

Gallay, A., Vaillant, V., Bouvet, P., Grimont, P., Desenclos, J.C.: How many foodborne outbreaks of salmonella infection occurred in France in 1995? Application of the capture–recapture method to three surveillance systems. Am. J. Epidemiol. **152**, 171–177 (2000)

Hook, E.B., Regal, R.R.: Capture–recapture methods in epidemiology: methods and limitations. Epidemiol. Rev. **17**, 243–264 (1995)

International Working Group for Disease Monitoring and Forecasting: Capture–recapture and multiple-record systems estimation I: history and theoretical development. Am. J. Epidemiol. **142**, 1047–1058 (1995)

Kolmogorov, A.N., Formin, S.V.: Metric spaces. In: Silverman, R.A. (ed.) Introductory Real Analysis, Prentice-Hall, Englewood Cliffs (1970)

Roberts, J.M., Brewer, D.D.: Estimating the Prevalence of male clients of prostitute women in Vancouver with a simple capture–recapture method. J. R. Stat. Soc. (Ser. A) **169**, 745–756 (2006)

Seber, G.A.F.: The effects of trap response on tag recapture estimates. Biometrics **26**, 13–22 (1970)

Seber, G.A.F.: Estimation of Animal Abundance and Related Parameters. Blackburn, Caldwell (2002)
Smit, F., Reinking, D., Reijerse, M.: Estimating the number of people eligible for health service use. Eval. Program. Plan. **25**, 101–105 (2002)
Van der Heijden, P., Bustami, R., Cruyff, M.J., Engbersan, G., van Houwelingen, H.C.: Point and interval estimation of population size using the truncated Poisson regression model. Stat. Model. Int. J. **3**, 305–322 (2003)
Van Hest, N.H.A., Grant, A.D., Smit, F., Story, A., Richardus, J.H.: Estimating infectious diseases incidence: validity of capture–recapture analysis and truncated models for incomplete count data. Epidemiol. Infect. **136**, 14–22 (2007)
Verstraeten, T., Baughman, A.L., Cadwell, B., Zanardi, L., Haber, P., Chen, R.T., The Vaccine Adverse Event Reporting System Team: Enhancing vaccine safety surveillance: a capture–recapture analysis of itussusception after rotavirus vaccination. Am. J. Epidemiol. **154**, 1006–1012 (2001)
Zelterman, D.: Robust estimation in truncated discrete distributions with applications to capture–recapture experiments. J. Stat. Plan. Inference **18**, 225–237 (1988)