Statistical Methods & Applications
https://doi.org/10.1007/s10260-026-00840-5

ORIGINAL PAPER

Two-way capture-recapture methods with emphasis on the bootstrap

Patarawan Sangnawakij[1] · Rattana Lerdsuwansri[1] · Parawan Pijitrattana[1] · Peter Schlattmann[2] · Antonello Maruotti[3,4] · Dankmar Bohning[5]

Received: 21 June 2025 / Accepted: 14 February 2026

Abstract
The Chapman estimator is widely used in dual system estimation for estimating the size of an elusive target population. Two independent sources are required, delivering a two-by-two table of those units identified by both sources, of those identified only by the first, and those only by the second source. Interest is in the frequency of those identified by neither source that remain hidden. While asymptotic variance estimates exist for the Chapman estimator, they often perform poorly with small sample sizes. In addition, the Chapman estimator may be biased for small sample sizes. This study explores the bias of the Chapman and a bias-corrected Chapman estimator by investigating both imputed and non-imputed (simple) bootstrap methods as alternatives for estimating variance and constructing confidence intervals. Through simulation studies, we assess the reliability of these methods by analyzing confidence interval coverage probabilities. Our findings show that the imputed bootstrap consistently delivers better performance, yielding coverage probabilities closer to the nominal level, even under moderate dependence between sources. We demonstrate the practical application of these methods with two case studies: suicide data in Cambodia and heroin use in Thailand.

Keywords Population size estimator · Chapman estimator · Semi-parametric bootstrap · Imputed bootstrap · Uncertainty

1 Introduction

The capture-recapture (CR) method is a powerful statistical technique used to quantify the total population sizes and unobserved units that are missing or difficult to fully observe. It is frequently used in many fields, such as biology, ecology, epidemiology, and social sciences (see McCrea and Morgan 2014; Böhning et al. 2018;

Extended author information available on the last page of the article

Published online: 26 March 2026

Seber and Schofield 2019). For example, scientists in ecology use CR approaches to estimate the population size of animal species or wildlife in a given area (Borchers et al. 2002, 2004; Foley et al. 2025). Social scientists apply CR methods to estimate the population size of illicit drug users, drug offenders, and perpetrators of domestic violence (Böhning et al. 2004; Roberts and Brewer 2006). They are also key for epidemiologists in determining how many people are affected by diseases and estimate the incidence of many diseases and health-related problems, including cancer, stroke, and homeless people (Sukrat et al. 2020; Lerdsuwansri et al. 2022; Domitz et al. 2024). Since the accuracy of population size estimation is an important aspect of the purposes, various methodologies have been developed, depending on count data distributions, inflation in count frequencies, and heterogeneity problems.

The simplest and *most frequently* used CR method uses two trapping sources (or lists) of data to estimate the unobserved count, leading to an estimate of the total population size. We refer to this as *dual system estimation*. Under this method, it is important to note that the duration of the survey is typically conducted over a short period of time, so it is unlikely that the evolution of new cases or the extinction of existing cases will occur during the study. The population of interest must be closed (no changes in the population size during the time period). All members of the population must be independent of each other. Furthermore, all observational occasions must be independent of each other. For details on fundamental and important assumptions in CR studies and their applications, see Böhning et al. (2004); Brittain and Böhning (2009); Mukem et al. (2014); Harris et al. (2016) for example.

Due to the two sources used in dual system estimation, the lists refer to different but overlapping populations and one empty, unknown cell. Estimating the unobserved cell counts, leading to an estimate of the population size, is therefore a matter of interest. The following subsections are detailed and extensive explanations of the traditional types of two-source CR estimation corresponding to this work. They include the Lincoln-Petersen estimator and the Chapman estimator for population size. These are provided to motivate our approaches, which will be described in detail in Sect. 2.

1.1 Settings and notations

Assume that the unknown size N of a population remains unchanged during the study period and also suppose that there are two sources of the identification mechanism. Both sources can be linked, and each covers some part of the same population. According to the settings shown in Table 1, it leads to the following data constellation: n_{11} is the frequency of individuals identified by the two sources, n_{10} is the frequency of those identified by source 1 only but not by source 2, n_{01} is the frequency of those identified by source 2 only but not by source 1, and n_{00} is the frequency of

Table 1 The capture-recapture setting for two-source situation

		Source 2		
		Observed (1)	Not observed (0)	
Source 1	Observed (1)	n_{11}	n_{10}	$n_{1.}$
	Not observed (0)	n_{01}	n_{00}	
		$n_{.1}$		N

individuals identified by neither source. In the marginal of the 2×2 contingency table, $n_{1.}$ is the number observed at source 1, $n_{.1}$ is the number observed at source 2.

If all cell frequencies are given, the population size is simply calculated as $N = n + n_{00}$, where $n = n_{11} + n_{10} + n_{01}$ (Bishop et al. 2007; Baffour et al. 2013). However, the cases known would be *n* and an estimate of the quantity n_{00}, the number of missing cases, is required. So, if n_{00} is accurately estimated using the appropriate CR methodology, so would be *N*.

1.2 Overview of estimators

As a simple method for dual system estimation, an estimate $\hat{N}$ of *N* can be constructed by $\hat{N} = n + \hat{n}_{00}$, where *n* is the observed sample size and $\hat{n}_{00}$ is an estimator to estimate the unknown value n_{00}. Lincoln (1930) and Petersen (1896) derive the estimator for n_{00} using maximum likelihood estimation. It is given as $\hat{n}_{00_{LP}} = n_{10}n_{01}/n_{11}$. The Lincoln-Petersen estimator for the true population size *N* is then formulated by

$$\hat{N}_{LP} = n + \hat{n}_{00_{LP}} = \frac{n_{1.}n_{.1}}{n_{11}}$$

with the estimated variance

$$\widehat{Var}(\hat{N}_{LP}) = \frac{n_{10}n_{01}n_{1.}n_{.1}}{n_{11}^3}.$$

The Lincoln-Petersen estimators given above are often used in applications, as they provide the simple closed-form solutions and are straightforward. However, this method requires the following essential conditions:

1. Both sources must be independent. In other words, the odds ratio (OR) between the sources must be equal to 1. This can be written as $OR = n_{11}n_{00}/n_{10}n_{01} = 1$, leading to $n_{00} = n_{10}n_{01}/n_{11}$.
2. The frequency n_{11} must be greater than 0 to avoid the undefinable $\hat{n}_{00_{LP}}$ of the denominator.

According to condition 2, if there is no overlap between sources ($n_{11} = 0$), $\hat{n}_{00_{LP}}$ cannot be defined. Chapman (1951) proposed the nearly unbiased modification

$$\hat{N}_C = \frac{(n_{1.}+1)(n_{.1}+1)}{n_{11}+1} - 1 = n + \frac{n_{10}n_{01}}{n_{11}+1}, \tag{1}$$

with the implied estimator of the missing cell

$$\hat{n}_{00_C} = \frac{n_{10}n_{01}}{n_{11}+1}.$$

A standard large-sample variance estimator is

$$\widehat{\mathrm{Var}}(\hat{N}_C) = \frac{(n_{1\cdot}+1)(n_{\cdot 1}+1)\, n_{10}\, n_{01}}{(n_{11}+1)^2\,(n_{11}+2)},$$

leading to the usual $(1-\alpha)100\%$ two-sided Wald interval

$$\hat{N}_C \;\pm\; z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{N}_C)}, \tag{2}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)100$th percentile of the standard normal distribution.

Chapman's estimator is exactly unbiased whenever the combined sample size meets or exceeds the population size ($n_{1\cdot}+n_{\cdot 1} \geq N$); otherwise it has a slight downward bias whose magnitude decreases as n_{11} grows. Moreover, the usual variance estimator above is unbiased when $n_{1\cdot}+n_{\cdot 1} > N$ (Seber 1970). Normal-approximation intervals can undercover when n_{11} is small. Alternatives are thus required (Sadinle 2009). Moreover, small sample sizes can bias $\hat{N}_C$, so that:

$$E(\hat{N}_C) = N - N\Delta = N(1-\Delta),$$

where $\Delta = \exp\{-(n_{1\cdot}+1)(n_{\cdot 1}+1)/N\}$. According to this performance, the potential small-sample bias of the Chapman estimator can be constructed. As given in Seber (2002), it is denoted as the bias-corrected estimator and given by

$$\hat{N}_{BC} = \frac{\hat{N}_C}{1-\hat{\Delta}},$$

where $\hat{\Delta} = \exp\{-(n_{1\cdot}+1)(n_{\cdot 1}+1)/\hat{N}_C\}$ is an approximation of Δ.

The Chapman estimator $\hat{N}_C$ corrects for bias in the Lincoln-Petersen estimator $\hat{N}_{LP}$. We have outlined above the conditions under which the Chapman estimator is unbiased. If these conditions are violated, the Chapman estimator can experience a negative bias, in other words provides a lower bound for the true population size. This bias-corrected Chapman estimator $\hat{N}_{BC}$ might be viewed as a small-sample-bias-corrected Chapman estimator, although we keep the name bias-corrected Chapman estimator for simplicity.

1.3 Our ultimate objective

The sole use of the asymptotic variance for statistical inference might be problematic, particularly, in the case of a small observed sample size. Hence, we are interested in exploring alternative ways to provide uncertainty assessment for the dual systems estimator. In doing so, we will focus on estimating the population size using the resampling technique. Since the Chapman estimator has some advantages, it will be used to perform bootstrap estimates and confidence intervals. Several versions of the bootstrap are considered. One is the simplest non-parametric bootstrap, in which data are resampled based on the observed sample. However, as pointed out by Buckland and Garthwaite (1991); Rivest et al. (1995), and Zult et al. (2025), the bootstrap should be

applied with an imputed population size to account for the uncertainty introduced by imputation. The imputed bootstrap and double bootstrap were discussed, but the procedures were described somewhat vaguely. Therefore, they are studied and explained in depth in the next section. A simulation study with different scenarios is then used to investigate the methods. We expect that bootstrapping can address the limitation of existing estimators.

2 Uncertainty assessment of the Chapman estimator using bootstrap

The bootstrap method is a powerful statistical tool to estimate the parameter of interest based on a resampling technique. The main idea of this method is to use sampling with replacement to sample new sets of samples from existing data. The percentiles of the sample statistics of interest then produce the point estimate and the confidence limits (Efron 1988; DiCiccio and Efron 1996). To be more precise, we introduce three types of bootstrap approach for estimating N. These include the imputed bootstrap, the double bootstrap with imputation, and the simple bootstrap without imputation. Since the bootstrap approach can provide the standard error of the statistics, uncertainty estimates can be derived. The following subsections provide details of these bootstraps and their computation procedures.

2.1 Bootstrap I (Bootstrap with imputation)

From the two-source CR setting, it is reasonable to assume that the sampling distribution of the data follows a multinomial distribution with size parameter $\hat{N}$ and a 2×2 probability parameter. However, observing only three of the four cells in the two-way table makes it impossible to define the full probabilities. We then employ the following method to estimate this unknown value.

In this section, the imputed bootstrap (or semiparametric bootstrap method) uses a simple imputation method to estimate the probability parameter of the missing cell. It is modified from the imputation method given in Norris and Kenneth (1996) and the parametric bootstrap approach introduced in Zwane and van der Heijden (2005). Although the imputed bootstrap in the CR setting has been applied to estimate the population size in several works (see Anan et al. 2017; Böhning et al. 2023; Böhning and Friedl 2024), the previous works have not been available for the two-source situation. Therefore, we provide the algorithm for the imputed bootstrap method as follows:

1. According to the settings in Table 1, calculate the event probability vector

$$\hat{p} = \left(\frac{n_{11}}{\hat{N}_C}, \frac{n_{10}}{\hat{N}_C}, \frac{n_{01}}{\hat{N}_C}, \frac{\hat{n}_{00_C}}{\hat{N}_C} \right),$$

where $\hat{N}_C = n + \hat{n}_{00_C}$ is the Chapman estimator for N and $n = n_{11} + n_{10} + n_{01}$.

2. Draw $n_{11}^*, n_{10}^*, n_{01}^*$, and n_{00}^* from a multinomial distribution with size parameter $\hat{N}_C$ and parameter vector $\hat{p}$. We have attached a * to the four frequencies to indicate that they are sampled from the observed/imputed sample.
3. Ignore n_{00}^* and calculate the bootstrap estimates

$$\hat{n}_{00}^* = \frac{n_{10}^* n_{01}^*}{n_{11}^* + 1}$$

and

$$\hat{N}^* = n_{11}^* + n_{10}^* + n_{01}^* + \hat{n}_{00}^*.$$

4. Repeat steps 2 and 3 *B* times and yield *B* Chapman estimates

$$\hat{N}^{*(1)}, \hat{N}^{*(2)}, \cdots, \hat{N}^{*(B)}.$$

5. Calculate the $(1-\alpha)100\%$ percentile bootstrap confidence interval for *N*.

2.2 Bootstrap II (Double bootstrap with imputation)

In Bootstrap I, the estimated, unobserved frequency estimate is kept fixed throughout the process. We think this may ignore the uncertainty involved in estimating the unobserved frequency. In a different approach, we try to incorporate this element in a further step. This approach is referred to as the double bootstrap method. We provide the bootstrap algorithm to compute uncertainty measures for the estimator as follows:

1. From the available data, calculate the event probability vector

$$\tilde{p} = \left(\frac{n_{11}}{n}, \frac{n_{10}}{n}, \frac{n_{01}}{n} \right).$$

2. Draw $n_{11}^*, n_{10}^*, n_{01}^*$ from a multinomial distribution with size parameter *n* and event parameter $\tilde{p}$.
3. Calculate the bootstrap estimate

$$\hat{n}_{00}^* = \frac{n_{10}^* n_{01}^*}{n_{11}^* + 1}.$$

4. Then, calculate

$$\hat{p}^* = \left(\frac{n_{11}}{\hat{N}^*}, \frac{n_{10}}{\hat{N}^*}, \frac{n_{01}}{\hat{N}^*}, \frac{\hat{n}_{00}^*}{\hat{N}^*} \right),$$

where $\hat{N}^* = n_{11} + n_{10} + n_{01} + \hat{n}_{00}^*$.
5. Draw $n_{11}^{**}, n_{10}^{**}, n_{11}^{**}, n_{00}^{**}$ from a multinomial distribution with size parameter $\hat{N}^*$ and probability parameter $\hat{p}^*$.

6. Ignore n_{00}^{**} and compute bootstrap estimates

$$\hat{n}_{00}^{**} = \frac{n_{10}^{**} n_{01}^{**}}{n_{11}^{**} + 1}$$

and

$$\hat{N}^{**} = n_{11}^{**} + n_{10}^{**} + n_{01}^{**} + \hat{n}_{00}^{**}.$$

7. Repeat steps 2 to 6 *B* times to get *B* Chapman estimates

$$\hat{N}^{**(1)}, \hat{N}^{**(2)}, \cdots, \hat{N}^{**(B)}.$$

8. Calculate the $(1-\alpha)100\%$ percentile bootstrap confidence interval for *N* on the basis of estimates given in step 7.

2.3 Bootstrap III (Bootstrap without imputation)

In this subsection, we consider a bootstrap that is solely based on the resampling approach, conditional on the observed data of size *n*. The major distinction between the bootstrap given in this section, called Bootstrap III, and the previous two bootstraps is that *n* is fixed, and there is no imputation involved. Therefore, Bootstrap III is just based on the observed data, which is similar to the simple non-parametric bootstrap (Zwane and van der Heijden 2005). Again, each individual is generated from a multinomial distribution. We outline the process for Bootstrap III as follows:

1. From the available data, calculate the event probability vector

$$\tilde{p} = \left(\frac{n_{11}}{n}, \frac{n_{10}}{n}, \frac{n_{01}}{n} \right).$$

2. Draw $n_{11}^*, n_{10}^*, n_{01}^*$ from a multinomial distribution with size parameter *n* and probability parameter $\tilde{p}$.
3. Compute the bootstrap estimates

$$\hat{n}_{00}^{*} = \frac{n_{10}^{*} n_{01}^{*}}{n_{11}^{*} + 1}$$

and

$$\hat{N}^{*} = n_{11}^{*} + n_{10}^{*} + n_{01}^{*} + \hat{n}_{00}^{*}.$$

4. Repeat steps 2 and 3 *B* times and yield *B* Chapman estimates

$$\hat{N}^{*(1)}, \hat{N}^{*(2)}, \cdots, \hat{N}^{*(B)}.$$

5. Calculate the $(1-\alpha)100\%$ percentile bootstrap confidence interval for N.

The algorithms described in the previous subsections are not limited to the Chapman estimator, which uses $\hat{N}_C$, but can be readily extended to other estimators. In the application and simulation sections, they are applied to calculate the population size using the bias-corrected estimator, which is based on $\hat{N}_{BC}$.

3 Case studies

Two real data examples are used to illustrate the methods introduced in this work. The first example is related to the suicide rate in Cambodia, obtained from Harris et al. (2016). Since there are no adequate Cambodian suicide statistics, the work focusses on estimating the population size of suicides in this country. The reported suicides were collected in 2012 from the two newspapers, including the Raksmey Kampuchea Daily News and the Koh Santepheap Daily News, which were the two largest newspapers in the native language (Khmer language) of Cambodia. The frequency counts of suicides from these two sources are shown in Table 2. Here, a total of 158 suicide deaths are observed. However, there are likely hidden units due to their non-reporting in published news or other media. Therefore, the total number of suicides in the population needs to be estimated.

The second example is related to heroin users in Thailand. Although Thailand is not a heroin-producing country, heroin and narcotic abuse have long been a distressing issue in the country and increased in use among age groups (Centre for Addiction Studies 2023). Question: How many hidden heroin users? To demonstrate, we look at data of heroin users in Pathum Thani province of Thailand to estimate the number of hidden users. The Princess Mother National Institute on Drug Abuse Treatment (PMNIDAT) (see http://www.pmnidat.go.th/thai/), which is organized under the Department of Medical Services, collected data on heroin addicts who attended the treatment facility during Thailand's fiscal year 2023, which started in October 2022 and ended in September 2023. Repeated counts occurred over a given period, leading to a total number of 1447 heroin users observed with their visits. As the context of this work is concerned with the two-source situation and the contribution is built upon the Chapman estimator, we divide the data into two distinct lists: Source 1, representing the first half of the fiscal year (October 2022 to March 2023), and Source 2, which represents the second half of the fiscal year (April to September 2023). The number of heroin users on two occasions is provided in Table 3.

Table 2 Number of suicides in Cambodia obtained from two newspapers in year 2012 ($n = 158$)

		Source 2 (Koh Santepheap News)	
		Observed (1)	Not observed (0)
Source 1 (Raksmey Kampuchea News)	Observed (1)	12	94
	Not observed (0)	52	-

Table 3 Number of heroin users from the health support in Pathum Thani, Thailand, obtained from the fiscal year 2023 ($n = 1447$)

		Source 2 (Second half year)	
		Observed (1)	Not observed (0)
Source 1 (First half year)	Observed (1)	121	747
	Not observed (0)	579	-

Table 4 Estimated number of suicides in Cambodia and heroin users in Pathum Thani, Thailand using Chapman estimator (and bias-corrected Chapman estimator in bracket) with confidence interval (CI) and length of interval from the formula and bootstrap methods

Dataset	Method	$\hat{n}_{00}$	$\hat{N}$	95% CI for N	Interval length
Suicide	Formula	376 (376)	534 (534)	300-769	469
($n = 158$)	Bootstrap I	380 (376)	538 (536)	360-941 (361-935)	581 (574)
	Bootstrap II	376 (381)	534 (539)	358-941 (364-929)	583 (565)
	Bootstrap III	379 (379)	537 (537)	377-926 (374-932)	549 (558)
Heroin	Formula	3545 (3545)	4992 (4992)	4248-5735	1487
($n = 1447$)	Bootstrap I	3542 (3543)	4989 (4990)	4338-5849 (4338-5841)	1511 (1503)
	Bootstrap II	3553 (3552)	5000 (4999)	4341-5868 (4344-5843)	1527 (1499)
	Bootstrap III	3548 (3547)	4995 (4994)	4296-5745 (4364-5795)	1449 (1431)

The approximate number of cases that are never captured from both sources ($\hat{n}_{00} = \hat{N} - n$) and estimated population sizes ($\hat{N}$) from two examples are given in Table 4. Note that the numbers in brackets in Table 4 refer to the bias-corrected Chapman estimator. Due to the large sample size, there is almost no difference between $\hat{N}_C$ and $\hat{N}_{BC}$. The formula method refers to parameter estimation based on $\hat{N}_C$ and the asymptotic normal confidence interval given in (1 and 2), respectively. Since the variance formula for $\hat{N}_{BC}$ is not available in a closed form, the confidence interval is then only based on the bootstrap method. The bootstrap estimate in each approach is averaged from the median of 10,000 bootstrap samples. The statistical analysis is done through the `R` programming language (R Core Team 2024). From the results, the point estimates from the formula method are similar to those of bootstraps. However, interval estimation reveals an effect. For suicide data, where n is small, the confidence intervals of all methods are different. For heroin users data, where n is larger, the confidence intervals based on the formula and Bootstrap III methods behave quite similarly, but are different from Bootstraps I and II. We note that the formula and Bootstrap III methods do not allow imputation of missing cells. Meanwhile, the rest are imputed bootstraps.

The results of all population estimators seem to be realistic, as the lower and upper limits for N are greater than the observed sample size. However, to investigate the performance of these methods, especially the uncertainty of estimation in terms of confidence interval, we provide a simulation study on the basis of several scenarios in the next section.

4 Simulation study

4.1 Simulation scenarios

A simulation study is performed to investigate the performance of the confidence intervals for N. We design the study to cover scenarios in CR studies with two sources. The parameter settings are given as follows. The population sizes N are 10, 25, 50, 100, and 250. The capture probabilities are varied as shown in Table 5, where in case A, we choose populations with increasing p_{00} to study this effect, while in case B populations are chosen with increasing dependency, so that the estimator becomes more conservatively biased. The case B simulation is designed with increasing levels of positive dependency, which is the most likely case in reality. See also Brittain and Böhning (2009). The odds ratio can be calculated as OR $= p_{11}p_{00}/p_{10}p_{01}$. Since the odds ratio between sources plays a crucial factor in the behaviour of the Chapman estimator, we study the cases where two sources are independent (OR = 1) and positive dependent (OR > 1). For each of the populations, the frequencies $\mathrm{f} = (n_{11}, n_{10}, n_{01}, n_{00})$ are generated from a multinomial distribution with parameters (N, p), where $\mathrm{p} = (p_{11}, p_{10}, p_{01}, p_{00})$. Then, the unobserved count n_{00} is ignored as the CR data.

The formula method and three bootstrap approaches are used to estimate N. Each scenario is repeated 10,000 times, and 5000 bootstrap samples are used. All are done using `R` (R Core Team 2024). On average, $\hat{N}$ obtained from the formula method is computed by

Table 5 Capture probabilities of event occurrence in two-source CR studies and odds ratio (OR) for simulations

Case	Population	p_{11}	p_{10}	p_{01}	p_{00}	OR
A	A1	0.320	0.480	0.080	0.120	1
	A2	0.250	0.250	0.250	0.250	1
	A3	0.125	0.125	0.375	0.375	1
	A4	0.050	0.050	0.450	0.450	1
	A5	0.040	0.160	0.160	0.640	1
	A6	0.020	0.080	0.180	0.720	1
B	B1	0.200	0.150	0.350	0.300	1.143
	B2	0.200	0.120	0.380	0.300	1.316
	B3	0.300	0.200	0.250	0.250	1.500
	B4	0.250	0.200	0.200	0.350	2.188
	B5	0.300	0.150	0.200	0.350	3.500
	B6	0.350	0.100	0.200	0.350	6.125

$$\frac{1}{R}\sum_{r=1}^{R}\hat{N}_r,$$

where $\hat{N}_r$ is the estimated population size in the *r*-th replication. Furthermore, $\hat{N}$ based on the bootstrap approach is calculated by

$$\frac{1}{R}\sum_{r=1}^{R}\left[median\left(\hat{N}^{*(1)},\hat{N}^{*(2)},\cdots,\hat{N}^{*(B)}\right)\right]_r,$$

where $\hat{N}^{*(b)}$ are the bootstrap estimates in the *b*-th sample, for $b = 1, 2, \cdots, B$, depending on Bootstrap I, II, or III. The performance of the estimator is evaluated in terms of relative bias, approximated by

$$\frac{1}{N}\left(\frac{1}{R}\sum_{r=1}^{R}\hat{N}_r - N\right).$$

We investigate the performance of the 95% confidence interval by evaluating how often the confidence limits cover the true *N*. Therefore, the estimated coverage probability of the confidence interval from simulation is calculated by

$$\frac{1}{R}\sum_{r=1}^{R}C_r,$$

where C_r is an indicator variable defined as 1 if $CI_{L,r} \leq N \leq CI_{U,r}$ and 0 otherwise. $CI_{L,r}$ and $CI_{U,r}$ are the lower and upper limits of *N* in the *r*-th iteration. In conclusion, the confidence interval with a coverage probability close to the nominal level of 0.95, or the confidence interval that has the largest coverage probability, is more efficient than the other comparator.

4.2 Simulation results

The simulation results show the relative biases of estimated population size and the coverage probabilities of the confidence intervals for *N* using the data generated in *case A* and *case B*. They are presented in Tables 6, 7, 8, 9. The coverage rates are also shown by graphs displayed in Figs. 1 and 2.

We first examine the simulation results under *case A* (see Tables 6 and 7), which assumes the independence between two sources. The main findings are summarized as follows. The estimates of *N* obtained from the three bootstraps are quite close to those of the formula method. They underestimate if $N \leq 50$ and the probability of any individual not appearing in either source (p_{00}) is large, corresponding to populations A5 and A6. However, all approaches provide a small estimation bias when *N* is increased. The Chapman and biased-corrected estimators show similar behaviour in

Table 6 Relative bias of Chapman estimator and biased-corrected Chapman estimator under simulations (Case A: OR = 1)

N	Population	Chapman				Biased-corrected Chapman			
		Formula	Bootstrap I	Bootstrap II	Bootstrap III	Formula	Bootstrap I	Bootstrap II	Bootstrap III
10	A1	−0.0025	−0.0096	−0.0089	−0.0117	0.0217	0.0156	0.0166	0.0134
	A2	−0.0205	−0.0333	−0.0331	−0.0345	0.0501	0.0334	0.0357	0.0381
	A3	−0.1148	−0.1511	−0.1327	−0.1236	0.0774	0.0239	0.0535	0.0692
	A4	−0.2908	−0.3291	−0.3004	−0.2939	−0.0529	−0.1107	−0.0661	−0.0563
	A5	−0.4658	−0.4726	−0.4725	−0.4757	−0.4096	−0.4129	−0.4139	−0.4150
	A6	−0.5734	−0.5799	−0.5798	−0.5815	−0.5324	−0.5345	−0.5349	−0.5352
25	A1	−0.0017	−0.0041	−0.0051	−0.0119	0.0027	−0.0016	−0.0026	−0.0096
	A2	−0.0020	−0.0024	−0.0031	−0.0106	0.0083	0.0072	0.0063	−0.0030
	A3	−0.0133	−0.0216	−0.0212	−0.0189	0.0752	0.0655	0.0652	0.0636
	A4	−0.1396	−0.1701	−0.1577	−0.1419	0.0950	0.0420	0.0559	0.0741
	A5	−0.2369	−0.2708	−0.2662	−0.2450	0.0019	−0.0460	−0.0394	−0.0096
	A6	−0.4301	−0.4717	−0.4571	−0.4345	−0.2151	−0.2726	−0.2529	−0.2245
50	A1	0.0017	0.0003	0.0000	−0.0042	0.0007	0.0004	0.0000	−0.0042
	A2	−0.0007	−0.0007	−0.0010	−0.0055	0.0011	−0.0005	−0.0008	−0.0054
	A3	−0.0011	−0.0009	−0.0010	−0.0043	0.0123	0.0137	0.0135	0.0078
	A4	−0.0357	−0.0488	−0.0457	−0.0380	0.0999	0.0885	0.0910	0.0922
	A5	−0.0894	−0.0971	−0.0974	−0.0931	0.1029	0.0893	0.0885	0.0831
	A6	−0.2628	−0.2893	−0.2872	−0.2649	−0.0171	−0.0479	−0.0454	−0.0179
100	A1	−0.0013	−0.0019	−0.0021	−0.0045	0.0012	−0.0019	−0.0021	−0.0045
	A2	−0.0004	−0.0004	−0.0005	−0.0029	−0.0006	−0.0004	−0.0005	−0.0029
	A3	−0.0010	−0.0009	−0.0009	−0.0039	−0.0005	−0.0007	−0.0007	−0.0038
	A4	0.0100	0.0096	0.0097	0.0092	0.0254	0.0492	0.0491	0.0453
	A5	−0.0104	0.0026	0.0025	−0.0086	0.0615	0.0721	0.0718	0.0516
	A6	−0.0942	−0.0920	−0.0919	−0.0855	0.1027	0.1085	0.1085	0.1097
250	A1	0.0001	−0.0002	−0.0002	−0.0012	0.0012	−0.0002	−0.0002	−0.0012
	A2	0.0002	0.0002	0.0002	−0.0008	0.0007	0.0002	0.0002	−0.0008
	A3	0.0012	0.0012	0.0011	−0.0004	0.0029	0.0012	0.0011	−0.0004
	A4	−0.0010	−0.0009	−0.0008	−0.0017	0.0027	−0.0005	−0.0004	−0.0015
	A5	−0.0007	0.0064	0.0064	0.0005	−0.0016	0.0083	0.0083	0.0018
	A6	0.0006	0.0184	0.0185	0.0130	0.0372	0.0641	0.0643	0.0544

terms of relative bias, except when the population size is small ($N = 10$). As can be seen in Fig. 1, the confidence intervals for N from Bootstrap I and Bootstrap II have similar coverage probabilities. They are also close to the target probability of 0.95 for $N > 50$ and show better coverage than the comparators in all cases of the study. The difference in coverage probability of the confidence intervals from Bootstraps I and II to Bootstrap III and the confidence interval using the variance formula is particularly pronounced if p_{00} becomes large (see populations A4-A6). In conclusion, Bootstraps I and II perform equally well in general cases where two sources are independent. However, Bootstrap II is computationally more expensive as it takes time in the calculation. More specifically, a complete simulation run for one setting of the simulation design takes approximately two minutes for Bootstraps I and III,

Table 7 Coverage probability of the 95% confidence interval for N using Chapman estimator and biased-corrected Chapman estimator under simulations (Case A: OR = 1)

N	Population	Chapman				Biased-corrected Chapman		
		Formula	Bootstrap I	Bootstrap II	Bootstrap III	Bootstrap I	Bootstrap II	Bootstrap III
10	A1	0.6424	0.6552	0.6620	0.6518	0.7232	0.7363	0.5421
	A2	0.7193	0.8036	0.8130	0.7658	0.8636	0.8759	0.7688
	A3	0.6071	0.6986	0.6805	0.6207	0.7700	0.7702	0.6803
	A4	0.3423	0.3840	0.3670	0.3147	0.4467	0.4460	0.3597
	A5	0.3649	0.5654	0.4257	0.3642	0.5670	0.5670	0.5657
	A6	0.2021	0.3658	0.2503	0.2021	0.3677	0.3677	0.3675
25	A1	0.8093	0.8425	0.8434	0.8064	0.8426	0.8435	0.7737
	A2	0.8540	0.9280	0.9302	0.8663	0.9275	0.9302	0.8640
	A3	0.7804	0.9032	0.9046	0.8653	0.9062	0.9073	0.8777
	A4	0.6210	0.7026	0.7006	0.6875	0.7248	0.7246	0.7001
	A5	0.5916	0.7790	0.7577	0.4878	0.9037	0.8986	0.6967
	A6	0.4700	0.5780	0.5580	0.2490	0.7650	0.7490	0.4800
50	A1	0.8858	0.9187	0.9212	0.8599	0.9186	0.9209	0.8577
	A2	0.8921	0.9331	0.9349	0.8748	0.9327	0.9346	0.8733
	A3	0.8469	0.9248	0.9248	0.8928	0.9265	0.9262	0.8941
	A4	0.7450	0.8735	0.8743	0.8540	0.8773	0.8783	0.8682
	A5	0.7280	0.9058	0.9051	0.8126	0.9087	0.9079	0.8388
	A6	0.6061	0.7918	0.7887	0.5813	0.9100	0.9106	0.7619
100	A1	0.9136	0.9331	0.9345	0.8677	0.9331	0.9345	0.8677
	A2	0.9227	0.9422	0.9430	0.8768	0.9422	0.9430	0.8768
	A3	0.8979	0.9430	0.9427	0.9127	0.9430	0.9427	0.9126
	A4	0.8375	0.9256	0.9255	0.9112	0.9275	0.9265	0.9143
	A5	0.8308	0.9327	0.9342	0.9061	0.9351	0.9357	0.9139
	A6	0.7388	0.9036	0.9030	0.8387	0.8899	0.8900	0.8498
250	A1	0.9321	0.9426	0.9430	0.8764	0.9426	0.9430	0.8764
	A2	0.9392	0.9483	0.9492	0.8889	0.9483	0.9492	0.8889
	A3	0.9275	0.9484	0.9480	0.9178	0.9484	0.9480	0.9178
	A4	0.8930	0.9409	0.9397	0.9285	0.9408	0.9398	0.9285
	A5	0.8859	0.9443	0.9442	0.9265	0.9446	0.9442	0.9266
	A6	0.8468	0.9373	0.9372	0.9213	0.9373	0.9374	0.9241

whereas it requires approximately five minutes for Bootstrap II. Bootstrap I is then recommended in this situation.

We then consider the situation in which the odds ratio between two sources is greater than 1 (*case B*). This allows for dependence between sources, which can often occur in applications. In this case, we use the expected value of the Chapman estimator, called the *estimand*, as the true parameter. It is defined by

$$E(\hat{N}) \approx N\left(\frac{p_{10}p_{01}}{p_{11}} + p_{11} + p_{10} + p_{01}\right).$$

Tables 8 and 9 show the relative biases of estimates and 95% coverage probabilities for $E(\hat{N})$ from simulations. All methods provide relative biases close to zero in all

Table 8 Relative bias of Chapman estimator and biased-corrected Chapman estimator under simulations (Case B: OR > 1)

N	Estimand	Population	Chapman				Biased-corrected Chapman			
			Formula	Bootstrap I	Bootstrap II	Bootstrap III	Formula	Bootstrap I	Bootstrap II	Bootstrap III
10	10	B1	−0.0738	−0.0939	−0.0882	−0.0852	0.0555	0.0268	0.0379	0.0456
	9	B2	−0.0978	−0.1180	−0.1102	−0.1081	0.0092	−0.0161	−0.0048	0.0006
	9	B3	−0.0900	−0.0980	−0.0990	−0.1044	−0.0378	−0.0460	−0.0460	−0.0497
	8	B4	−0.2005	−0.2126	−0.2102	−0.2154	−0.1374	−0.1516	−0.1474	−0.1490
	8	B5	−0.2529	−0.2611	−0.2607	−0.2678	−0.2297	−0.2372	−0.2361	−0.2408
	7	B6	−0.2926	−0.2987	−0.2989	−0.3040	−0.2582	−0.2610	−0.2616	−0.2622
25	24	B1	−0.0395	−0.0413	−0.0420	−0.0473	−0.0220	−0.0199	−0.0207	−0.0285
	23	B2	−0.0714	−0.0742	−0.0749	−0.0792	−0.0526	−0.0515	−0.0522	−0.0590
	23	B3	−0.0850	−0.0850	−0.0861	−0.0940	−0.0802	−0.0791	−0.0803	−0.0889
	20	B4	−0.1895	−0.1897	−0.1907	−0.1978	−0.1864	−0.1851	−0.1862	−0.1944
	19	B5	−0.2506	−0.2503	−0.2517	−0.2588	−0.2494	−0.2484	−0.2500	−0.2573
	18	B6	−0.2934	−0.2929	−0.2946	−0.3005	−0.2918	−0.2909	−0.2927	−0.2988
50	48	B1	−0.0365	−0.0369	−0.0373	−0.0422	−0.0371	−0.0372	−0.0374	−0.0427
	46	B2	−0.0736	−0.0746	−0.0749	−0.0796	−0.0712	−0.0717	−0.0720	−0.0770
	46	B3	−0.0854	−0.0853	−0.0856	−0.0903	−0.0835	−0.0834	−0.0837	−0.0884
	40	B4	−0.1907	−0.1906	−0.1909	−0.1953	−0.1918	−0.1916	−0.1919	−0.1965
	38	B5	−0.2507	−0.2503	−0.2507	−0.2551	−0.2496	−0.2491	−0.2495	−0.2540
	35	B6	−0.2917	−0.2907	−0.2915	−0.2956	−0.2923	−0.2913	−0.2921	−0.2962
100	96	B1	−0.0370	−0.0372	−0.0373	−0.0406	−0.0383	−0.0386	−0.0387	−0.0419
	93	B2	−0.0708	−0.0713	−0.0714	−0.0742	−0.0724	−0.0729	−0.0730	−0.0759
	92	B3	−0.0821	−0.0820	−0.0821	−0.0846	−0.0829	−0.0829	−0.0829	−0.0855
	81	B4	−0.1891	−0.1891	−0.1891	−0.1914	−0.1897	−0.1896	−0.1897	−0.1921
	75	B5	−0.2485	−0.2482	−0.2483	−0.2508	−0.2501	−0.2498	−0.2499	−0.2524
	71	B6	−0.2925	−0.2920	−0.2921	−0.2946	−0.2925	−0.2920	−0.2921	−0.2945
250	241	B1	−0.0374	−0.0375	−0.0375	−0.0390	−0.0381	−0.0383	−0.0383	−0.0398
	232	B2	−0.0720	−0.0722	−0.0723	−0.0735	−0.0722	−0.0725	−0.0725	−0.0738
	229	B3	−0.0835	−0.0835	−0.0835	−0.0847	−0.0835	−0.0834	−0.0834	−0.0846
	202	B4	−0.1900	−0.1900	−0.1900	−0.1910	−0.1889	−0.1889	−0.1889	−0.1899
	188	B5	−0.2500	−0.2499	−0.2499	−0.2510	−0.2498	−0.2497	−0.2497	−0.2508
	177	B6	−0.2929	−0.2927	−0.2927	−0.2938	−0.2932	−0.2930	−0.2930	−0.2940

situations. The coverage probabilities of the confidence intervals are then evaluated and shown in Fig. 2. Most importantly, the confidence intervals based on Bootstraps I and II provide coverage probabilities very close to 0.95 when OR < 3.5 (see populations B1-B4). For the higher values of the OR (populations B5 and B6), all interval estimators have low performance, as their coverage probabilities are lower than 0.95. However, the confidence intervals from Bootstraps I and II still outperform those of Bootstrap III and the asymptotic normal confidence interval, in particular for small N. The confidence interval based on the biased-corrected estimator provides coverage rates smaller than that based on the Chapman method when $N = 10$. As N increases, the two become increasingly similar.

Finally, we consider a simulation study on the performance of the estimation methods. The variances of the Chapman estimates from the three bootstrap methods

Table 9 Coverage probability of the 95% confidence interval for estimand $E(N)$ using Chapman estimator and biased-corrected Chapman estimator under simulations (Case B: OR > 1)

N	Estimand	Population	Chapman				Biased-corrected Chapman		
			Formula	Bootstrap I	Bootstrap II	Bootstrap III	Bootstrap I	Bootstrap II	Bootstrap III
10	10	B1	0.6439	0.7328	0.7311	0.6913	0.8179	0.8234	0.7400
	9	B2	0.6697	0.7260	0.7264	0.6720	0.7555	0.7608	0.6019
	9	B3	0.7310	0.8181	0.8246	0.7536	0.7870	0.8137	0.5782
	8	B4	0.6964	0.7509	0.7601	0.6690	0.7282	0.7536	0.5119
	8	B5	0.6233	0.6690	0.6867	0.6113	0.6237	0.6563	0.4607
	7	B6	0.5217	0.5637	0.5927	0.4580	0.6288	0.6288	0.5867
25	24	B1	0.8276	0.9085	0.9101	0.8621	0.9150	0.9174	0.8650
	23	B2	0.8160	0.9083	0.9094	0.8503	0.9150	0.9154	0.8539
	23	B3	0.8418	0.9117	0.9137	0.8185	0.9114	0.9111	0.8009
	20	B4	0.8339	0.8955	0.9012	0.7704	0.8969	0.9024	0.7557
	19	B5	0.7576	0.8306	0.8340	0.6989	0.8205	0.8246	0.6485
	18	B6	0.6575	0.7024	0.7213	0.5600	0.6736	0.7016	0.4617
50	48	B1	0.8840	0.9354	0.9356	0.8801	0.9329	0.9344	0.8819
	46	B2	0.8713	0.9257	0.9267	0.8665	0.9206	0.9207	0.8621
	46	B3	0.8727	0.9101	0.9115	0.8164	0.9161	0.9156	0.8242
	40	B4	0.8775	0.9057	0.9075	0.7856	0.9035	0.9077	0.7861
	38	B5	0.8034	0.8423	0.8473	0.7030	0.8386	0.8431	0.6940
	35	B6	0.6984	0.7324	0.7388	0.5336	0.7368	0.7448	0.5227
100	96	B1	0.9097	0.9329	0.9336	0.8826	0.9393	0.9403	0.8887
	93	B2	0.8993	0.9311	0.9322	0.8773	0.9318	0.9309	0.8796
	92	B3	0.9026	0.9232	0.9236	0.8338	0.9193	0.9218	0.8309
	81	B4	0.8888	0.9069	0.9093	0.8079	0.9067	0.9064	0.8092
	75	B5	0.8411	0.8536	0.8561	0.7064	0.8552	0.8573	0.7067
	71	B6	0.7251	0.7465	0.7506	0.5588	0.7553	0.7595	0.5637
250	241	B1	0.9341	0.9449	0.9454	0.8938	0.9410	0.9421	0.8919
	232	B2	0.9276	0.9377	0.9381	0.8850	0.9386	0.9401	0.8891
	229	B3	0.9210	0.9271	0.9286	0.8376	0.9301	0.9300	0.8386
	202	B4	0.9050	0.9112	0.9118	0.8171	0.9126	0.9118	0.8110
	188	B5	0.8450	0.8507	0.8518	0.7052	0.8579	0.8578	0.7163
	177	B6	0.7521	0.7579	0.7596	0.5665	0.7468	0.7501	0.5600

are provided and compared to the true variance. The results given in Fig. 3 show that with increasing N the variances of all bootstrap methods increase and behave similarly to the true variance. If the OR increases (population B6 has the largest OR), the variances of the bootstrap methods increase much less than the true variance, with strong underestimation for large OR. In particular, they are much lower variance than the true variance when OR ≥ 3.5 (see populations B5 and B6). This could lead to a confidence interval which is too narrow and, consequently, has too low level of coverage probability. To clarify, we show in Table 10 the simulated percentiles of Chapman estimates for the true and bootstrap distributions under the three situations: $N = 250$ for population A1, $N = 250$ for population B6, and $N = 1000$ for population B6. Under the first situation (independence of sources), the medians of the estimates of the bootstrap distributions are close to the true N. This is in contrast to the

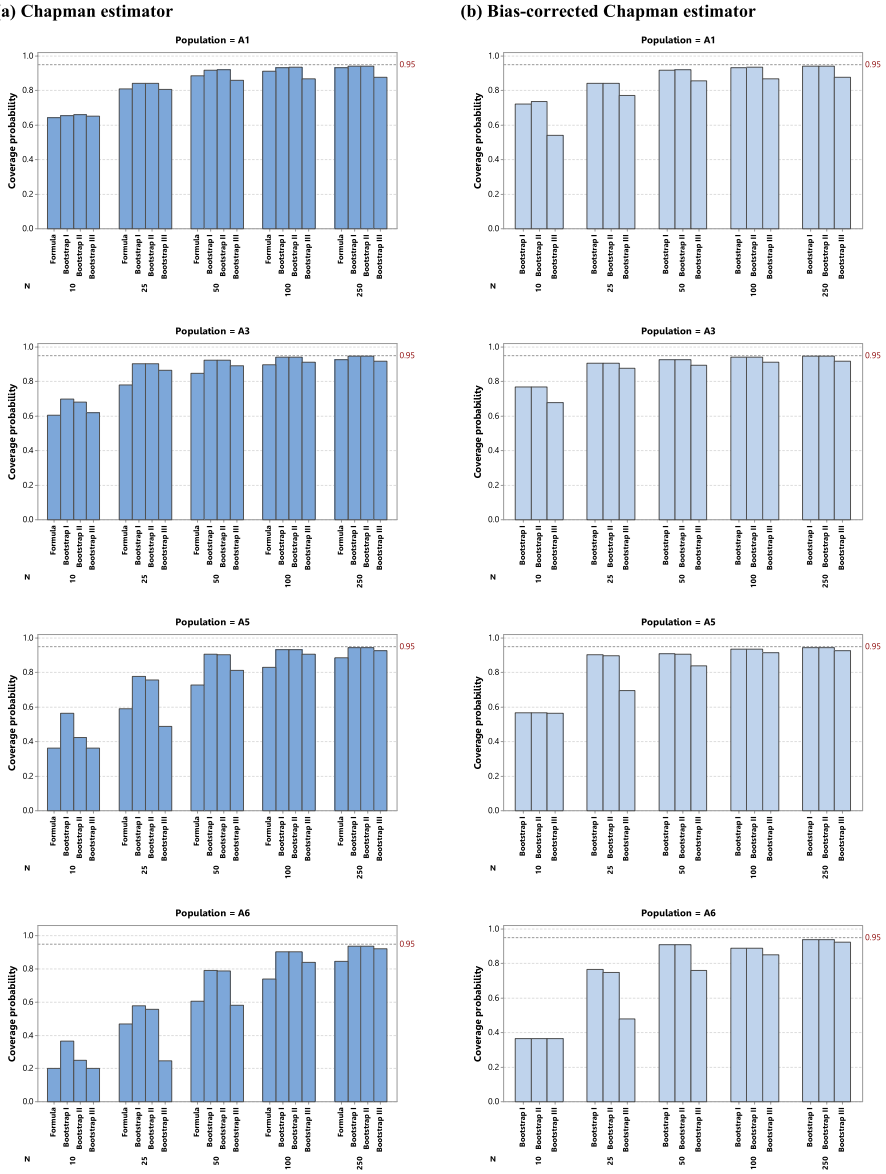

Fig. 1 Coverage probability of the 95% confidence interval for N using two estimators under simulations (Case A: OR = 1)

latter two situations (dependence setting). The medians of the estimates of the bootstrap distributions are still correctly estimating the estimand $E(\hat{N})$. However, the other relevant percentiles of the estimates from the true and bootstrap distributions clearly differ, the more the further in the tails of the distribution. Here, the bootstrap methods have interval lengths narrower than the true distribution. This point corresponds to the simulation study given in Fig. 2, where the bootstrap variance is less

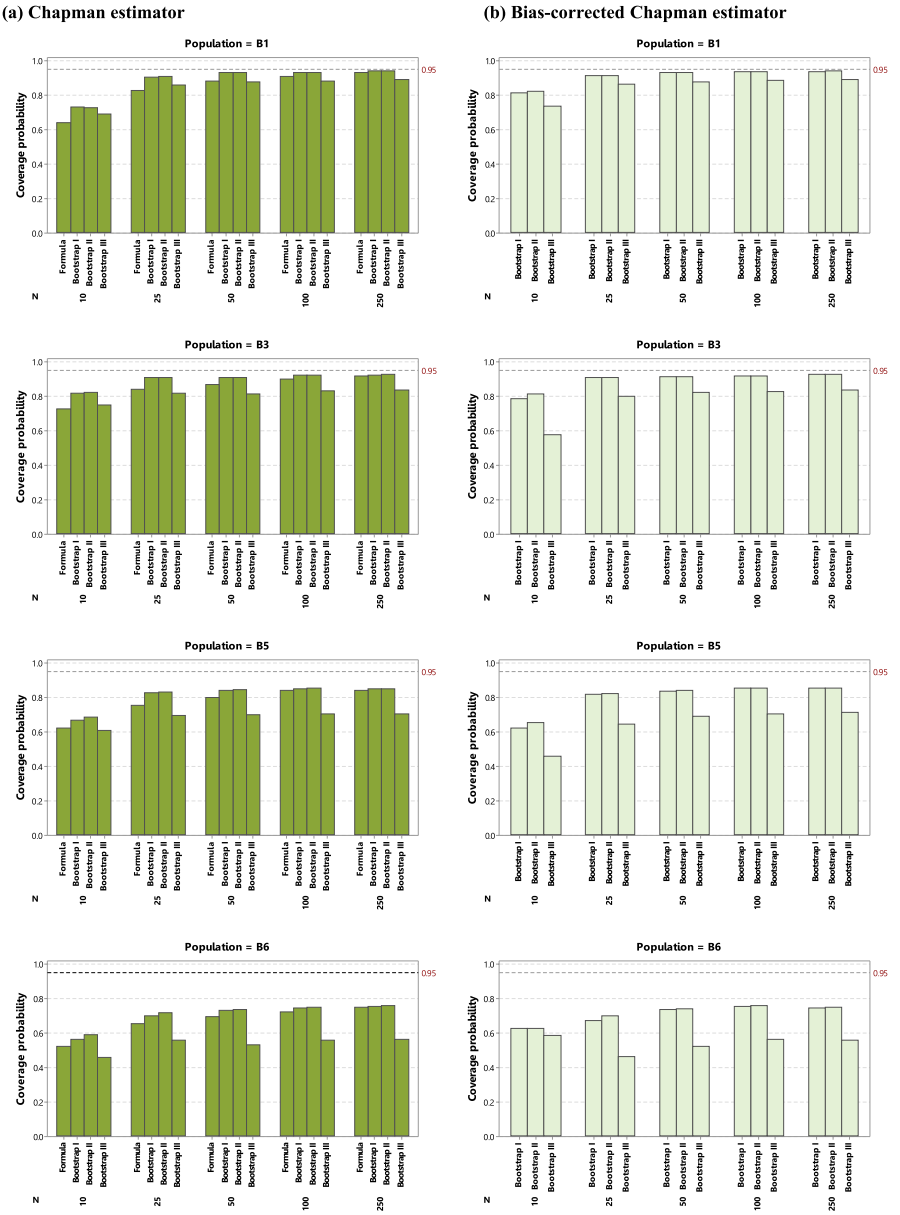

Fig. 2 Coverage probability of the 95% confidence interval for estimand $E(N)$ using two estimator under simulations (Case B: OR > 1)

than the true variance, leading to a small interval length and low coverage probability of confidence interval for the extremely large dependency situation. The reason for this performance lies in the fact that in case of high association of the two sources, the imputed bootstrap will consistently sample from a population with a potentially much smaller population size (the estimand) rather than the true (potentially imputed)

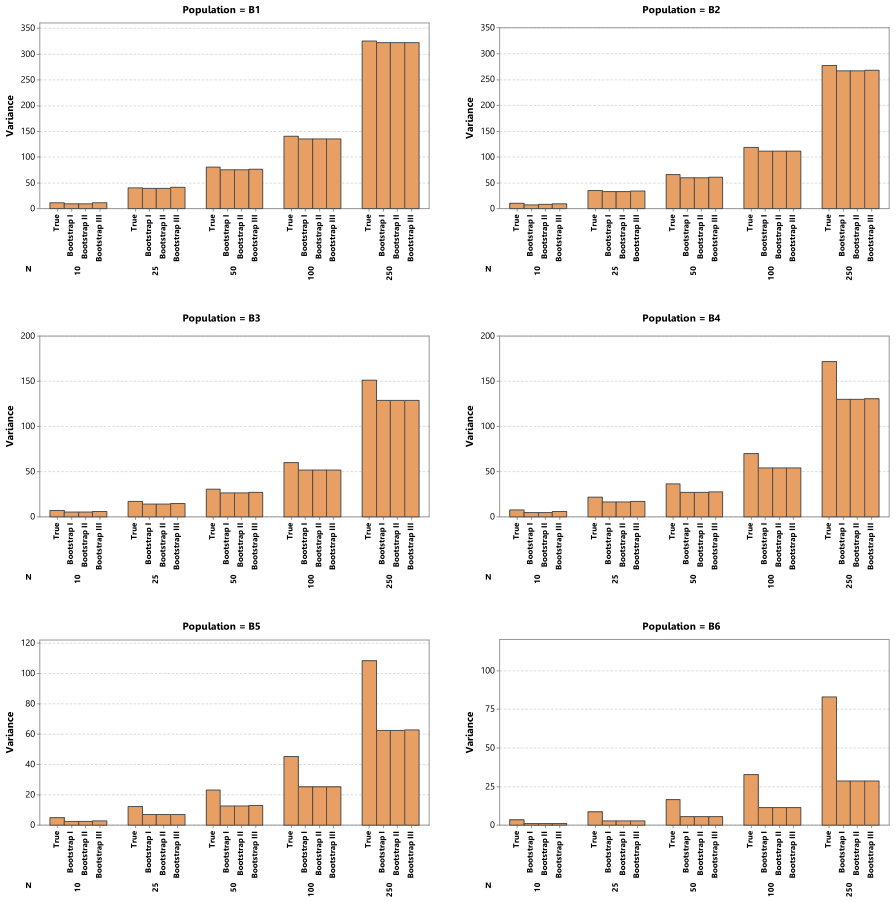

Fig. 3 True variance of the Chapman estimator and bootstrap variances of three methods from simulated data under Case B

N, and hence underestimate the variability of the Chapman estimator. There is no way that this can be corrected, as the true odds ratio would need to be known to do so. Therefore, undercoverage of confidence intervals will occur in dependence situations even if the estimand is used instead of the true population size.

5 Concluding remarks

The Lincoln-Petersen estimator (1896; 1930) is the simplest tool in capture-recapture studies with two sources used to estimate the hidden population size. Due to a limitation of a zero frequency count of individuals identified at both sources (no overlap between sources), Chapman (1951) modifies the estimator of Lincoln-Petersen and shows that it is less biased than the original one, under certain conditions even being unbiased. Hence, it can be viewed as a bias-corrected Lincoln-Petersen estimator. We have outlined the conditions under which the Chapman estimator is unbiased. If

Table 10 Simulated percentiles of Chapman estimates from the true distribution and bootstrap distribution based on Bootstrap I, II, and III

		Bootstrap distribution		
Percentile	True distribution	Bootstrap I	Bootstrap II	Bootstrap III
Situation: $N = 250$ and population A1				
2.5th	231.6152	232.4173	232.3927	236.9602
5th	234.5968	235.0416	235.0408	238.6338
25th	243.1481	243.4503	243.4629	244.6365
50th	249.5747	249.6567	249.6667	249.3904
75th	256.0909	256.2191	256.1631	254.7612
95th	266.7746	266.6519	266.6709	263.8889
97.5th	270.5200	270.2577	270.3365	267.1757
Interval length of 95% CI	38.9048	37.8404	37.9438	30.2155
Situation: $N = 250$ and population B6 (Estimand $E(\hat{N}) = 177$)				
2.5th	159.7861	166.5509	166.5120	170.6292
5th	162.2846	168.1612	168.1526	171.4776
25th	170.6129	173.1921	173.1935	174.2000
50th	176.7826	176.7242	176.6844	176.4382
75th	182.8325	180.2871	180.3046	179.0000
95th	191.7146	185.7734	185.8120	183.4525
97.5th	194.7676	187.6952	187.7344	185.0580
Interval length of 95% CI	34.9815	21.1443	21.2224	14.4288
Situation: $N = 1000$ and population B6 (Estimand $E(\hat{N}) = 707$)				
2.5th	672.0109	686.5784	686.6159	690.1391
5th	677.2965	689.8878	689.9573	696.0588
25th	694.9621	700.1330	700.1367	702.3058
50th	706.9611	707.2899	707.2914	707.0318
75th	719.1297	714.5415	714.5897	712.1243
95th	736.5640	725.1644	725.2258	720.0085
97.5th	742.7284	728.7827	728.7849	722.7434
Interval length of 95% CI	70.7175	42.2043	42.1690	32.6043

these conditions are violated, the Chapman estimator can experience a negative bias, in other words provides a lower bound for the true population size. We have seen in the simulation work that this underestimation bias is mainly relevant for population sizes below 50. In these cases, one might apply the bias-correction as worked in Sect. 1.2. Uncertainty assessment of these estimators requires knowledge of their variances, and estimates have been developed based on asymptotic normality. However, these turn out to be poor for small and moderate sample sizes. Hence, an alternative approach based on bootstrapping has been explored here.

Bootstrapping can be implemented in several ways. In fact, bootstrapping methods have already been considered to estimate uncertainty in population size estimation in recent papers including the simple bootstrap and the imputed bootstrap. Here, we have also suggested a double bootstrap to cope with the uncertainty induced by the imputation. The imputed bootstrap, the double bootstrap with imputation, and the simple bootstrap (non-parametric bootstrap without imputation) have not yet studied

in depth in two-source CR estimation and compared through simulations. In this paper, we provide novel insights on the behaviour of these methods.

A large-scale simulation study is used to evaluate the performance of our bootstraps and the normal approximation-based variance formula approach in terms of the confidence intervals. What works and what does not work? The findings are divided into two situations. Under the independence of sources situation (assumption holds), the formula method and bootstrap without imputation do not perform well, as their coverage probabilities of the confidence intervals are lower than those of the imputed bootstrap and double bootstrap methods. From the results, the performances of the imputed bootstrap and double bootstrap are similar. However, as we have noted, the computational effort for the double bootstrap is larger by a factor of 2.5. We therefore recommend using the imputed bootstrap in practice due to its reduced computational effort compared to the double bootstrap. Under the violation of the independence assumption, all methods do not work well when an extreme dependence between two sources occurs, such as the odds ratios being 3 or more. For small and moderate dependence between sources setting, the imputed bootstrap and double bootstrap perform satisfactorily. They provide a coverage probability larger than the simple bootstrap and the one based on the asymptotic normal approximation. Again, there are only slight differences in the confidence intervals between these two imputed bootstrap versions. In conclusion, the imputed bootstrap seems to be the best choice among all the methods considered here.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Anan O, Böhning D, Maruotti A (2017) Uncertainty estimation in heterogeneous capture-recapture count data. J Stat Comput Simul 87(10):2094–2114

Baffour B, Brown J, Smith P (2013) An investigation of triple system estimators in censuses. Stat J IAOS 29(1):53–68

Bishop Y, Fienberg S, Holland P (2007) Discrete multivariate analysis: theory and practice. Springer, New York

Böhning D, Friedl H (2024) One-inflation and zero-truncation count data modelling revisited with a view on Horvitz–Thompson estimation of population size. Int Stat Rev 92(3):406–430

Böhning D, Suppawattanabodee B, Kusolvisitkul W et al (2004) Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. Eur J Epidemiol 19(12):1075–1083

Böhning D, van der Heijden PGM, Bunge J (2018) Capture-recapture methods for the social and medical sciences. CRC Press, Boca Raton
Böhning D, Lerdsuwansri R, Sangnawakij P (2023) Modeling COVID-19 contact-tracing using the ratio regression capture-recapture approach. Biometrics 79(4):3818–3830
Borchers D, Buckland S, Zucchini W (2002) Estimating animal abundance. Springer, Berlin
Borchers D, Buckland S, Zucchini W (2004) Estimating animal abundance closed populations. Springer, London
Brittain S, Böhning D (2009) Estimators in capture-recapture studies with two sources. Adv Stat Anal 93(1):23–47
Buckland S, Garthwaite P (1991) Quantifying precision of mark-recapture estimates using the bootstrap and related methods. Biometrics 47(1):255–268
Centre for Addiction Studies (2023) Background of academic networks on addiction research in Thailand. https://cads.in.th/cads/content?id=339, Accessed: 2025-02-01
Chapman DG (1951) Some properties of the hypergeometric distribution with applications to zoological censuses. University of California Press, Los Angeles
DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11(3):189–228
Domitz R, Gabriel K, Linhart J (2024) Using the capture-recapture technique to supplement a point-in-time count of homeless adults in Kittitas county. Washington Soc Work Res 48(4):265–274
Efron B (1988) Bootstrap confidence intervals: Good or bad? Psychol Bull 104(2):293
Foley M, Lato K, Fuirst M et al (2025) Spatial and temporal predictability drive foraging movements of coastal birds. Mov Ecol 13(5):1–16
Harris KM, Thandrayen J, Samphoas C et al (2016) Estimating suicide rates in developing nations: a low-cost newspaper capture-recapture approach in Cambodia. Asia Pac J Public Health 28(3):262–270
Lerdsuwansri R, Sangnawakij P, Böhning D et al (2022) Sensitivity of contact-tracing for COVID-19 in Thailand: a capture-recapture application. BMC Infect Dis 22(101):1–10
Lincoln FC (1930) Calculating waterfowl abundance on the basis of banding returns. US Department of Agriculture, Washington
McCrea R, Morgan B (2014) Analysis of capture-recapture data. CRC Press, London
Mukem S, Sriplung H, McNeil E et al (2014) Breast cancer screening among women in Thailand: analyses of population-based household surveys. J Med Assoc Thai 97(11):1106–1118
Norris J, Kenneth H (1996) Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics 52(2):639–649
Petersen CGJ (1896) The yearly immigration of young plaice in the Limfjord from the German sea. Rept Danish Biol Sta 6:1–48
R Core Team (2024) R: a language and environment for statistical computing. Tech. rep., R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
Rivest L, Potvin F, Crepeau H et al (1995) Statistical methods for aerial surveys using the double-count technique to correct visibility bias. Biometrics 51(2):461–470
Roberts J, Brewer D (2006) Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture-recapture method. J R Stat Soc A Stat Soc 169(4):745–756
Sadinle M (2009) Transformed logit confidence intervals for small populations in single capture-recapture estimation. Commun Stat Simul Comput 38(9):1909–1924
Seber G (1970) The effects of trap response on tag recapture estimates. Biometrics 26(1):13–22
Seber G (2002) The estimation of animal abundance and related parameters. Blackburn Press, New Jersey
Seber G, Schofield M (2019) Capture-recapture: parameter estimation for open animal populations. Springer, Boca Raton
Sukrat B, Okascharoen C, Rattanasiri S et al (2020) Estimation of the adolescent pregnancy rate in Thailand 2008–2013: an application of capture-recapture method. BMC Pregnancy Childb 19(20):1–7
Zult D, van der Heijden P, Bakker B (2025) Bias correction in multiple systems estimation. J Off Stat 41(1):495–518
Zwane E, van der Heijden P (2005) Population estimation using the multiple system estimator in the presence of continuous covariates. Stat Model 5(1):39–52

Authors and Affiliations

Patarawan Sangnawakij[1] · Rattana Lerdsuwansri[1] · Parawan Pijitrattana[1] · Peter Schlattmann[2] · Antonello Maruotti[3,4] · Dankmar Bohning[5]

✉ Dankmar Bohning
D.A.Bohning@soton.ac.uk

[1] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand

[2] Department of Medical Statistics, Computer Sciences and Documentation, Jena University Hospital, 07743 Jena, Germany

[3] Department of Public Health and Epidemiology, Khalifa University, 127788 Abu Dhabi, UAE

[4] Department GEPLI, LUMSA University, 00193 Rome, Italy

[5] Mathematical Sciences and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, UK