Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities

by Martín H. Félix-Medina, Pedro E. Monjardin
and Aida N. Aceves-Castro

Release date: December 17, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 349-376
Statistics Canada, Catalogue No. 12-001-X

Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities

Martín H. Félix-Medina, Pedro E. Monjardin and Aida N. Aceves-Castro[1]

Abstract

Félix-Medina and Thompson (2004) proposed a variant of link-tracing sampling to sample hidden and/or hard-to-detect human populations such as drug users and sex workers. In their variant, an initial sample of venues is selected and the people found in the sampled venues are asked to name other members of the population to be included in the sample. Those authors derived maximum likelihood estimators of the population size under the assumption that the probability that a person is named by another in a sampled venue (link-probability) does not depend on the named person (homogeneity assumption). In this work we extend their research to the case of heterogeneous link-probabilities and derive unconditional and conditional maximum likelihood estimators of the population size. We also propose profile likelihood and bootstrap confidence intervals for the size of the population. The results of simulations studies carried out by us show that in presence of heterogeneous link-probabilities the proposed estimators perform reasonably well provided that relatively large sampling fractions, say larger than 0.5, be used, whereas the estimators derived under the homogeneity assumption perform badly. The outcomes also show that the proposed confidence intervals are not very robust to deviations from the assumed models.

Key Words: Bootstrap; Capture-recapture; Chain referral sampling; Maximum likelihood estimator; Profile likelihood confidence interval; Snowball sampling.

1 Introduction

Conventional sampling methods are not appropriate for sampling hidden or hard-to-reach human populations, such as drug users, sexual-workers and homeless people, because of the lack of suitable sampling frames. For this reason, several specific sampling methods for this type of population have been proposed. See Magnani, Sabin, Saidel and Heckathorn (2005) and Kalton (2009) for reviews of some of them. According to Heckathorn (2002) two types of sampling methods for hidden populations are the most commonly used in actual studies. One is location sampling, also known as time-and-space sampling, aggregation point sampling or intercept point sampling. The other is snowball sampling, also known as link-tracing sampling (LTS) or chain referral sampling.

In location sampling a frame of primary units is constructed. The primary units are combinations of places and time segments where the elements of the population tend to gather. The frame is not assumed to cover the whole population. A probability sample of primary units is selected and from each sampled unit a sort of systematic sample of elements is drawn. Although design-based estimators of different parameters can be constructed, the main drawback of location sampling is that inferences are valid only for the part of the population covered by the frame. For reviews of this method see MacKellar, Valleroy, Karon, Lemp and Janssen (1996), Munhib, Lin, Stueve, Miller, Ford, Johnson and Smith (2001), McKenzie and Mistianen (2009), Semaan (2010) and Karon and Wejnert (2012). Location sampling has been used in the Young Men's Survey to estimate HIV seroprevalence in young men who have sex with

1. Martín H. Félix-Medina, Pedro E. Monjardin and Aida N. Aceves-Castro, Facultad de Ciencias F sico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México. E-mail: mhfelix@uas.edu.mx.

men. (See McKellar et al. 1996.) In this study the primary units were venues attended by young men such as dance clubs, bars and street locations.

In LTS an initial sample of members of the population is selected and the sample size is increased by asking the sampled people to name or to refer other members of the population to be included in the sample. The named people who are not in the initial sample might be asked to refer other persons, and the process might continue in this way until a specified stopping rule is satisfied. For reviews of several variants of LTS see Spreen (1992), Thompson and Frank (2000) and Johnston and Sabin (2010). LTS was used in the Colorado Springs study on heterosexual transmission of HIV/AIDS. (See Potterat, Woodhouse, Rothenberg, Muth, Darrow, Muth and Reynolds 1993; Rothemberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl 1995 and Potterat, Woodhouse, Muth, Rothenberg, Darrow, Klovdahl and Muth 2004.) In this research an initial non probabilistic sample of people presumably at high risk of acquiring and transmitting HIV was obtained and they were asked for a complete enumeration of their personal contacts who were also included in the sample.

Frank and Snijders (1994) proposed a variant of LTS that allows the sampler to estimate the population size. In their variant they assume an initial Bernoulli sample, that is, that every element of the population has the same probability of being included in the sample and that the inclusions are independent. In addition, they assume that the probability that person i in the initial sample refers person j in the population, which we will call link-probability, is a constant, and that the referrals are independent. We will name the first of the additional premises the assumption of homogeneity of the link-probabilities. Based on these hypotheses these authors derive several estimators of the population size. They indicate that their method yielded reasonable estimates of the number of heroin users in Groningen. However, Dávid and Snijders (2002) reported an underestimate of the number of homeless in Budapest using this method. They indicate that the underestimation might be caused by deviations from the assumption of an initial Bernoulli sample.

The problem of satisfying in actual applications of LTS the assumption of an initial Bernoulli sample of members of the population motivated Félix-Medina and Thompson (2004) to develop a variant of LTS in which the initial sample is selected by a probabilistic design. To do this they assume, as in location sampling, that the sampler can construct a sampling frame of sites or venues where the members of the population tend to gather, such as bars, parks and blocks. The frame is not assumed to cover the whole population, but only a portion of it. Then, a simple random sample without replacement (SRSWOR) of sites is selected and the members of the population who belong to the sampled sites are identified. Finally, as in ordinary LTS, the people in the initial sample are asked to name other members of the population.

These authors propose maximum likelihood estimators (MLEs) of the population size derived under a probability model that describes the numbers of people found in the sampled sites and a model that regards that the link-probabilities between the elements of the population and the sampled sites are homogeneous, that is, that they depend on the sampled sites, but not on the potentially named people. Later, Félix-Medina and Monjardin (2006) consider this same variant of LTS and propose estimators of the population size derived also under the assumption of homogeneity, but using a Bayesian-assisted approach, that is, the functional forms of the estimators are obtained using the Bayesian approach, but inferences are made under the frequentist approach.

Although the variant of LTS proposed by Félix-Medina and Thompson (2004) has not been used in any actual study, we would expect that estimators of the population size derived under the assumption of

homogeneity will present problems of underestimation if this hypothesis is not satisfied as occurs in capture-recapture studies. We think this because these estimators resemble those used in that field.

In this paper, we extend the work by Félix-Medina and Thompson (2004) to the case in which the link-probabilities depend on the named people, that is, we assume heterogeneous link-probabilities. The structure of the paper is as follows. In Section 2 we introduce the LTS variant proposed by Félix-Medina and Thompson (2004). In Section 3 we present a model for the link-probabilities that takes into account their heterogeneity and derive unconditional and conditional MLEs of the population size. In Section 4 we construct profile likelihood and bootstrap confidence intervals for the population size. In Section 5 we present a procedure for determining the size of the initial sample in order to achieve a specified value of the relative error of the estimation. In Section 6 we describe the results of two simulation studies, and finally, in Section 7 we present some conclusions and suggestions for future research.

2 Sampling design and notation

Since in this work we consider the variant of LTS proposed by Félix-Medina and Thompson (2004), we will briefly describe it. Thus, let U be a finite population of an unknown number τ of people. We assume that a portion U_1 of U is covered by a sampling frame of N sites $A_1, \ldots, A_N$, where the members of the population can be found with high probability. We suppose that we have a criterion that allows us to assign a person in U_1 to only one site in the frame. Notice that we are not assuming that a person could not be found in different sites, but that, as in ordinary cluster sampling, we are able to assign him or her to only one site, for instance, the site where he or she spends most of his or her time. Thus, we can consider the sites in the frame as clusters of people. Let M_i denote the number of members of the population that belong to the site $A_i, i = 1, \ldots, N$. From the previous assumption it follows that the number of people in U_1 is $\tau_1 = \sum_1^N M_i$ and the number of people in the portion $U_2 = U - U_1$ of U that is not covered by the frame is $\tau_2 = \tau - \tau_1$.

The sampling design is as follows. A SRSWOR S_A of n sites $A_1, \ldots, A_n$ is selected from the frame and the M_i members of the population who belong to the sampled site A_i are identified, $i = 1, \ldots, n$. Let S_0 be the set of people in the initial sample. Observe that the size of S_0 is $M = \sum_1^n M_i$. The people in each sampled site are asked to name other members of the population. We will say that a person and a site are linked if any of the people who belong to that site names him or her. Finally, let S_1 and S_2 be the sets of people in $U_1 - S_0$ and U_2, respectively, who are linked to some site or sites in S_A.

3 Maximum likelihood estimators of τ_1, τ_2 and τ

3.1 Probability models

To construct MLEs of the τ's we need to specify models for the observed variables. Thus, as in Félix-Medina and Thompson (2004), we will suppose that the numbers $M_1, \ldots, M_N$ of people who belong to

the sites $A_1, \ldots, A_N$ are independent Poisson random variables with mean λ_1. Therefore, the joint conditional distribution of $(M_1, \ldots, M_n, \tau_1 - M)$ given that $\sum_1^N M_i = \tau_1$ is multinomial with probability mass function (pmf):

$$f(m_1, \ldots, m_n, \tau_1 - m) = \frac{\tau_1!}{\prod_1^n m_i!(\tau_1 - m)!}\left(\frac{1}{N}\right)^m \left(1 - \frac{n}{N}\right)^{\tau_1 - m}. \tag{3.1}$$

To model the links between the members of the population and the sampled sites we will define the following random variables: $X_{ij}^{(k)} = 1$ if person j in $U_k - A_i$ is linked to site A_i and $X_{ij}^{(k)} = 0$ if $j \in A_i$ or that person is not linked to A_i, $j = 1, \ldots, \tau_k$, $i = 1, \ldots, n$. We will suppose that given the sample S_A of sites the $X_{ij}^{(k)}$'s are independent Bernoulli random variables with means $p_{ij}^{(k)}$'s, where the link-probability $p_{ij}^{(k)}$ satisfies the following Rasch model:

$$p_{ij}^{(k)} = \Pr\left(X_{ij}^{(k)} = 1 \middle| \beta_j^{(k)}, S_A\right) = \frac{\exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}{1 + \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}, \; j \in U_k - A_i; \; i = 1, \ldots, n. \tag{3.2}$$

It is worth noting that this model was considered by Coull and Agresti (1999) in the context of capture-recapture sampling. In this model $\alpha_i^{(k)}$ is a fixed (not random) effect that represents the potential that the cluster A_i has of forming links with the people in $U_k - A_i$, and $\beta_j^{(k)}$ is a random effect that represents the propensity of the person $j \in U_k$ to be linked to a cluster. We will suppose that $\beta_j^{(k)}$ is normally distributed with mean 0 and unknown variance σ_k^2 and that these variables are independent. The parameter σ_k^2 determines the degree of heterogeneity of the $p_{ij}^{(k)}$'s : great values of σ_k^2 imply high degree of heterogeneity.

Before we end this subsection, we will make some comments about the assumed models. First, the multinomial distribution of the observed M_i's (which is the one used in the likelihood function) implies that people are distributed independently and with equal probability on the sites of the sampling frame. This assumption is difficult to satisfy in actual situations; however, as will be shown later, the likelihood function depends on the observed M_i's basically through their sum M and since NM/n is a design-based estimator of τ_1, that is, it is a distribution free estimator, it follows that the MLE of τ_1 will be also robust to deviations from the multinomial distribution of the M_i's. Nevertheless, deviations from this model will affect the performance of variance estimators and confidence intervals derived under this assumption. Second, the Rasch model given by (3.2) implies the following: (i) the link-probability $p_{ij}^{(k)}$ depends only on two effects: the sociability of the people in cluster A_i and that of person $j \in U_k - A_i$; (ii) the two effects are additive, and (iii) for any site A_i in the frame and any person $j \in U - A_i$, $p_{ij}^{(k)} > 0$. Model (3.2) is a particular case of a generalized linear mixed model. (See Agresti 2002, Section 2.1, for a brief review of this type of model.) Therefore, we could incorporate the network structures of the people in cluster A_i and person $j \in U_k - A_i$ to model the link-probability $p_{ij}^{(k)}$ by extending model (3.2) to one that includes covariates associated with person j, with cluster A_i, and their interaction terms. However, if we used a more general model than (3.2), we would make the problem of inference much more difficult than that we face in this work. Thus, in spite of the relative simplicity of

model (3.2), we expect that it still captures the heterogeneity of the link-probabilities and allow us to make inferences about the τ's at least at the correct order of magnitude.

3.2 Likelihood function

The easiest way of constructing the likelihood function is to factorize it into different components. One of them is associated with the probability of selecting the initial sample S_0, which is given by the multinomial distribution (3.1), that is,

$$L_{\text{MULT}}(\tau_1) \propto \frac{\tau_1!}{(\tau_1 - m)!}(1 - n/N)^{\tau_1 - m}.$$

Two other components are associated with the conditional probabilities of the configurations of links between the people in $U_k - S_0, k = 1, 2$, and the clusters $A_i \in S_A$, given S_A. To derive these factors we need to compute the probabilities of some events. Let $\mathbf{X}_j^{(k)} = \left(X_{1j}^{(k)}, \ldots, X_{nj}^{(k)}\right)$ be the $n-$dimensional vector of link-indicator variables $X_{ij}^{(k)}$ associated with the j^{th} person in $U_k - S_0$. Notice that $\mathbf{X}_j^{(k)}$ indicates which clusters $A_i \in S_A$ are linked to that person. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a vector whose i^{th} element is 0 or $1, i = 1, \ldots, n$. Because of the assumptions we made about the distributions of the variables $X_{ij}^{(k)}$'s, we have that the conditional probability, given $\beta_j^{(k)}$, that $\mathbf{X}_j^{(k)}$ equals $\mathbf{x}$, that is, the probability that the j^{th} person is linked to only those clusters $A_i \in S_A$ such that the i^{th} element x_i of $\mathbf{x}$ equals 1, is

$$\Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x} \,\middle|\, \beta_j^{(k)}, S_A\right) = \prod_{i=1}^{n}\left[p_{ij}^{(k)}\right]^{x_i}\left[1 - p_{ij}^{(k)}\right]^{1-x_i} = \prod_{i=1}^{n}\frac{\exp\left[x_i\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}.$$

Therefore, the probability that the vector of link-indicator variables associated with a randomly selected person in $U_k - S_0$ equals $\mathbf{x}$ is

$$\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k) = \int\prod_{i=1}^{n}\frac{\exp\left[x_i\left(\alpha_i^{(k)} + \sigma_k z\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \sigma_k z\right)}\phi(z)\,dz,$$

where $\boldsymbol{\alpha}_k = \left(\alpha_1^{(k)}, \ldots, \alpha_n^{(k)}\right)$ and $\phi(\cdot)$ denotes the probability density function of the standard normal distribution $[\mathrm{N}(0,1)]$.

As in Coull and Agresti (1999), instead of using $\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ in the likelihood function we will use its Gaussian quadrature approximation $\tilde{\pi}_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ given by

$$\tilde{\pi}_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k) = \sum_{t=1}^{q}\prod_{i=1}^{n}\frac{\exp\left[x_i\left(\alpha_i^{(k)} + \sigma_k z_t\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \sigma_k z_t\right)}\nu_t, \tag{3.3}$$

where q is a fixed constant and $\{z_t\}$ and $\{\nu_t\}$ are obtained from tables.

We are now in conditions of computing the two above mentioned factors of the likelihood function. Let $\Omega = \{(x_1, \ldots, x_n) : x_i = 0,1; i = 1, \ldots, n\}$, the set of all n – dimensional vectors such that each one of their elements is 0 or 1. For $\mathbf{x} = (x_1, \ldots, x_n) \in \Omega$, let $R_{\mathbf{x}}^{(k)}$ be the random variable that indicates the number of distinct people in $U_k - S_0$ whose vectors of link-indicator variables are equal to $\mathbf{x}$. Finally, let R_k be the random variable that indicates the number of distinct people in $U_k - S_0$ that are linked to at least one cluster $A_i \in S_A$. Notice that $R_k = \sum_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} R_{\mathbf{x}}^{(k)}$, where $\mathbf{0}$ denotes the n – dimensional vector of zeros.

Because of the assumptions we made about the distributions of the variables $X_{ij}^{(k)}$'s, we have that the conditional joint probability distribution of the variables $\{R_{\mathbf{x}}^{(1)}\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ and $\tau_1 - m - R_1$, given that $\{M_i = m_i\}_{i=1}^n$, is a multinomial distribution with parameter of size $\tau_1 - m$ and probabilities $\{\pi_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ and $\pi_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)$, and that of the variables $\{R_{\mathbf{x}}^{(2)}\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ and $\tau_2 - R_2$ is a multinomial distribution with parameter of size τ_2 and probabilities $\{\pi_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\}_{\mathbf{x} \in \Omega - \{\mathbf{0}\}}$ and $\pi_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)$.

Therefore, the factors associated with the probabilities of the configurations of links between the people in $U_k - S_0, k = 1, 2,$ and the clusters $A_i \in S_A$ are

$$L_1(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) \propto \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!} \prod_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{r_{\mathbf{x}}^{(1)}} \left[\tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{\tau_1 - m - r_1}$$

and

$$L_2(\tau_2, \boldsymbol{\alpha}_2, \sigma_2) \propto \frac{\tau_2!}{(\tau_2 - r_2)!} \prod_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{r_{\mathbf{x}}^{(2)}} \left[\tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{\tau_2 - r_2}.$$

The last component of the likelihood function is associated with the conditional probability, given S_A, of the configuration of links between the people in S_0 and the clusters $A_i \in S_A$. To derive this factor firstly observe that by the definition of the indicator variables $X_{ij}^{(k)}$'s, the i^{th} element of the vector of link-indicator variables associated with a person in $A_i \in S_A$ is equal to zero. Thus, let $\Omega_{-i} = \{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) : x_j = 0,1,\ j \neq i,\ j = 1, \ldots, n\}$, that is, the set of all $(n-1)$ – dimensional vectors obtained from the vectors in Ω by omitting their i^{th} coordinate. For $\mathbf{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \Omega_{-i}$ let $R_{\mathbf{x}}^{(A_i)}$ be the random variable that indicates the number of distinct people in $A_i \in S_A$ such that their vectors of link-indicator variables, when the i^{th} coordinate is omitted, equal $\mathbf{x}$. Finally, let $R^{(A_i)}$ be the random variable that indicates the number of distinct people in $A_i \in S_A$ that are linked to at least one site $A_j \in S_A, j \neq i$. Notice that $R^{(A_i)} = \sum_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}} R_{\mathbf{x}}^{(A_i)}$, where $\mathbf{0}$ denotes the $(n-1)$ – dimensional vector of zeros. Then, as in the previous cases, the conditional joint probability distribution of the variables $\{R_{\mathbf{x}}^{(A_i)}\}_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}}$ and $m_i - R^{(A_i)}$, given that $\{M_i = m_i\}_{i=1}^n$, is a multinomial distribution with parameter of size m_i and probabilities $\{\pi_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)\}_{\mathbf{x} \in \Omega_{-i} - \{\mathbf{0}\}}$ and $\pi_{\mathbf{0}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)$, where $\boldsymbol{\alpha}_1^{(-i)} = (\alpha_1^{(1)}, \ldots, \alpha_{i-1}^{(1)}, \alpha_{i+1}^{(1)}, \ldots, \alpha_n^{(1)})$ and

$$\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)=\int\prod_{j\neq i}^{n}\frac{\exp\left[x_j\left(\alpha_j^{(1)}+\sigma_1 z\right)\right]}{1+\exp\left(\alpha_j^{(1)}+\sigma_1 z\right)}\phi(z)\,dz.$$

Therefore, the probability of the configuration of links between the people in S_0 and the clusters $A_j \in S_A$ is given by the product of the previous multinomial probabilities (one for each $A_i \in S_A$), and consequently the factor of the likelihood associated with that probability is

$$L_0\left(\boldsymbol{\alpha}_1,\sigma_1\right)\propto\prod_{i=1}^{n}\prod_{\mathbf{x}\in\Omega_{-i}-\{\mathbf{0}\}}\left[\tilde{\pi}_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)\right]^{r_{\mathbf{x}}^{(A_i)}}\left[\tilde{\pi}_{\mathbf{0}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)\right]^{m_i-r^{(A_i)}},$$

where

$$\tilde{\pi}_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)=\sum_{t=1}^{q}\prod_{j\neq i}^{n}\frac{\exp\left[x_j\left(\alpha_j^{(1)}+\sigma_1 z_t\right)\right]}{1+\exp\left(\alpha_j^{(1)}+\sigma_1 z_t\right)}\nu_t, \tag{3.4}$$

is the Gaussian quadrature approximation to the probability $\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)$.

From the previous results we have that the likelihood function is given by

$$L\left(\tau_1,\tau_2,\alpha_1,\alpha_2,\sigma_1,\sigma_2\right)=L_{(1)}\left(\tau_1,\boldsymbol{\alpha}_1,\sigma_1\right)L_{(2)}\left(\tau_2,\boldsymbol{\alpha}_2,\sigma_2\right),$$

where

$$L_{(1)}\left(\tau_1,\boldsymbol{\alpha}_1,\sigma_1\right)=L_{\text{MULT}}\left(\tau_1\right)L_1\left(\tau_1,\boldsymbol{\alpha}_1,\sigma_1\right)L_0\left(\boldsymbol{\alpha}_1,\sigma_1\right)$$

and

$$L_{(2)}\left(\tau_2,\boldsymbol{\alpha}_2,\sigma_2\right)=L_2\left(\tau_2,\boldsymbol{\alpha}_2,\sigma_2\right).$$

In the comments at the end of Subsection 3.1 was indicated that the likelihood function depends on the M_i's basically through their sum M. This can be seen by noting that only the factor L_0 depends directly through the M_i's. The factors L_{MULT} and L_1 depend on the M_i's through M, whereas the factor $L_{(2)}$ does not depend on the M_i's.

3.3 Unconditional maximum likelihood estimators

Numerical maximization of the likelihood function $L\left(\tau_1,\tau_2,\alpha_1,\alpha_2,\sigma_1,\sigma_2\right)$ with respect to the parameters yields the ordinary or unconditional maximum likelihood estimators (UMLEs) $\hat{\tau}_k^{(U)},\hat{\alpha}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ of τ_k,α_k and $\sigma_k, k=1,2$. Consequently the UMLE of $\tau=\tau_1+\tau_2$ is $\hat{\tau}^{(U)}=\hat{\tau}_1^{(U)}+\hat{\tau}_2^{(U)}$. Closed forms for the UMLEs do not exist; however, using the asymptotic approximation $\partial\ln\left(\tau_k!\right)/\partial\tau_k\approx\ln\left(\tau_k\right)$, we get the following approximations to $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_2^{(U)}$:

$$\hat{\tau}_1^{(U)}=\frac{M+R_1}{1-(1-n/N)\,\tilde{\pi}_{\mathbf{0}}^{(1)}\left(\hat{\boldsymbol{\alpha}}_1^{(U)},\hat{\sigma}_1^{(U)}\right)}\quad\text{and}\quad\hat{\tau}_2^{(U)}=\frac{R_2}{1-\tilde{\pi}_{\mathbf{0}}^{(2)}\left(\hat{\boldsymbol{\alpha}}_2^{(U)},\hat{\sigma}_2^{(U)}\right)}. \tag{3.5}$$

Notice that these expressions are not closed forms since $\hat{\alpha}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ depend on $\hat{\tau}_k^{(U)}$, $k = 1, 2$. Nevertheless, these expressions are useful to get formulae for the asymptotic variances of $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_2^{(U)}$.

3.4 Conditional maximum likelihood estimators

Another way to get MLEs of τ_k, α_k and σ_k is by using Sanathanan's (1972) approach, which yields conditional maximum likelihood estimators (CMLEs). These estimators are numerically simpler to compute than UMLEs. In addition, if covariates were used in the model for the link-probability $p_{ij}^{(k)}$, this approach could still be used to get estimators of τ_k, α_k and σ_k, whereas the unconditional likelihood approach could not since the values of the covariates associated with the non sampled elements would be unknown.

The idea in Sanathanan's approach is to factorize the pmf of the multinomial distributions of the frequencies $R_{\mathbf{x}}^{(k)}$ of the different configurations of links as follows:

$$
\begin{aligned}
L_1(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) &\propto f\left(\left\{r_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}, \tau_1 - m - r_1 \middle| \{m_i\}, \tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) \\
&= f\left(\left\{r_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \middle| \{m_i\}, \tau_1, r_1, \boldsymbol{\alpha}_1, \sigma_1\right) f\left(r_1 \middle| \{m_i\}, \tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) \\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\frac{\tilde{\pi}_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)}{1 - \tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)}\right]^{r_{\mathbf{x}}^{(1)}} \times \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!}\left[1 - \tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{r_1}\left[\tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{\tau_1 - m - r_1} \\
&= L_{11}(\boldsymbol{\alpha}_1, \sigma_1) L_{12}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1)
\end{aligned}
$$

and

$$
\begin{aligned}
L_2(\tau_2, \boldsymbol{\alpha}_2, \sigma_2) &\propto f\left(\left\{r_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}, \tau_2 - r_2 \middle| \{m_i\}, \tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) \\
&= f\left(\left\{r_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \middle| \{m_i\}, \tau_2, r_2, \boldsymbol{\alpha}_2, \sigma_2\right) f\left(r_2 \middle| \{m_i\}, \tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) \\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\frac{\tilde{\pi}_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)}{1 - \tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)}\right]^{r_{\mathbf{x}}^{(2)}} \times \frac{\tau_2!}{(\tau_2 - r_2)!}\left[1 - \tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{r_2}\left[\tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{\tau_2 - r_2} \\
&= L_{21}(\boldsymbol{\alpha}_2, \sigma_2) L_{22}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2).
\end{aligned}
$$

Observe that in each case the first factor $L_{k1}(\boldsymbol{\alpha}_k, \sigma_k)$ is proportional to the conditional joint pmf of the $\left\{R_{\mathbf{x}}^{(k)}\right\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, given that $\{M_i = m_i\}_1^n$ and $R_k = r_k$, which is the multinomial distribution with parameter of size r_k and probabilities $\left\{\tilde{\pi}_{\mathbf{x}}^{(k)} / \left[1 - \tilde{\pi}_{\mathbf{0}}^{(k)}\right]\right\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, and that this distribution does not depend on τ_k. Notice also that the second factors $L_{12}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1)$ and $L_{22}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2)$ are proportional to the conditional pmfs of R_1 and R_2, given that $\{M_i = m_i\}_1^n$, which are the distributions $\mathrm{Bin}\left(\tau_1 - m, 1 - \tilde{\pi}_{\mathbf{0}}^{(1)}\right)$ and $\mathrm{Bin}\left(\tau_2, 1 - \tilde{\pi}_{\mathbf{0}}^{(2)}\right)$, respectively, where $\mathrm{Bin}(\tau, \theta)$ denotes the Binomial distribution with parameter of size τ and probability θ.

The CMLEs $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$ of $\boldsymbol{\alpha}_k$ and $\sigma_k, k = 1, 2$ are obtained by maximizing numerically

$$L_{11}(\boldsymbol{\alpha}_1, \sigma_1) L_0(\boldsymbol{\alpha}_1, \sigma_1) \quad \text{and} \quad L_{21}(\boldsymbol{\alpha}_2, \sigma_2) \tag{3.6}$$

with respect to $(\boldsymbol{\alpha}_1, \sigma_1)$ and $(\boldsymbol{\alpha}_2, \sigma_2)$, respectively. Note that the factors in (3.6) do not depend on $\tau_k, k = 1, 2.$

Finally, by plugging the estimates $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$ into the factors of the likelihood function that depend on $\tau_k, k = 1, 2,$ and maximizing these factors, that is, maximizing $L_{12}\left(\tau_1, \hat{\boldsymbol{\alpha}}_1^{(C)}, \hat{\sigma}_1^{(C)}\right) L_{\text{MULT}}(\tau_1)$ and $L_{22}\left(\tau_2, \hat{\boldsymbol{\alpha}}_2^{(C)}, \hat{\sigma}_2^{(C)}\right)$, with respect to τ_1 and τ_2, respectively, we get that the CMLEs $\hat{\tau}_1^{(C)}$ and $\hat{\tau}_2^{(C)}$ of τ_1 and τ_2 are given by (3.5) but replacing $\hat{\boldsymbol{\alpha}}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ by $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}, k = 1, 2.$ Observe that these expressions for $\hat{\tau}_1^{(C)}$ and $\hat{\tau}_2^{(C)}$ are closed forms. The CMLE of τ is $\hat{\tau}^{(C)} = \hat{\tau}_1^{(C)} + \hat{\tau}_2^{(C)}$.

4 Confidence intervals

We will consider two types of confidence intervals (CIs) for the population sizes: profile likelihood and bootstrap CIs.

4.1 Profile likelihood confidence intervals

Several authors such as Cormack (1992), Evans, Kim and O'Brien (1996), Coull and Agresti (1999) and Gimenes, Choquet, Lamor, Scofield, Fletcher, Lebreton and Pradel (2005) have indicated that, in the context of capture-recapture sampling, profile likelihood confidence intervals (PLCIs) perform better than traditional Wald CIs when the sample size is not large. Some factors that affect the performance of Wald CIs are biases in the estimators of the population size, biases in the estimators of the variances and asymmetries in the distributions of the estimators of the population size. Besides, a Wald CI for the population size might present the drawback that its lower bound might be less than the number of captured elements. Notice that, with the exception of the first listed factor, none of the others affect the performance of PLCIs. Furthermore, Evans et al. (1996), based on Ratkowsky (1988), indicate that the nonlinear nature of the capture-recapture estimators is approximated by likelihood-based CIs better than by Wald CIs.

Since the proposed estimators resemble those used in capture-recapture sampling and based on the previous comments, we should expect that also in our case PLCIs performance better than Wald CIs. It is worth noting that if we wanted to use Wald CIs, we would need to compute estimators of the variances of the proposed estimators. One alternative is to construct estimators of their asymptotic variances by using Sanathanan's (1972) results; however, for n large, say 20 or greater, obtaining these type of estimator is computationally very expensive because for each estimator is required the construction of a $(n + 1) \times (n + 1)$ symmetric matrix whose elements are sums of 2^n terms.

To get PLCIs for τ_1, τ_2 and τ we will follow Coull and Agresti's (1999) approach. Thus, for fixed values τ_1, τ_2 and τ of the population sizes, let r_{10}, r_{20} and r_{00} be non-negative real numbers such that $\tau_1 = m + r_1 + r_{10}, \tau_2 = r_2 + r_{20}$ and $\tau = m + r_1 + r_2 + r_{00}$, where m, r_1 and r_2 are the observed values of the random variables M, R_1 and R_2. Then $100(1 - \alpha)\%$ PLCIs for τ_1, τ_2 and τ are defined as the

following sets: $\{\tau_1 = m + r_1 + r_{10} : -2\ln[\Lambda_1(r_{10})] \le \chi^2_{1,1-\alpha}\}$, $\{\tau_2 = r_2 + r_{20} : -2\ln[\Lambda_2(r_{20})] \le \chi^2_{1,1-\alpha}\}$ and $\{\tau = m + r_1 + r_2 + r_{00} : -2\ln[\Lambda(r_{00})] \le \chi^2_{1,1-\alpha}\}$, respectively, where

$$\Lambda_1(r_{10}) = \max_{\alpha_1,\sigma_1} L_{(1)}(m + r_1 + r_{10}, \alpha_1, \sigma_1) / L_{(1)}(\hat{\tau}_1, \hat{\alpha}_1, \hat{\sigma}_1),$$

$$\Lambda_2(r_{20}) = \max_{\alpha_2,\sigma_2} L_{(2)}(r_2 + r_{20}, \alpha_2, \sigma_2) / L_{(2)}(\hat{\tau}_2, \hat{\alpha}_2, \hat{\sigma}_2)$$

and

$$\Lambda(r_{00}) = \max_{r_{10},\alpha_1,\alpha_2,\sigma_1,\sigma_2} L(m + r_1 + r_{10}, r_2 + r_{00} - r_{10}, \alpha_1, \alpha_2, \sigma_1, \sigma_2) / L(\hat{\tau}_1, \hat{\tau}_2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}_1, \hat{\sigma}_2),$$

$\hat{\tau}_k, \hat{\alpha}_k$ and $\hat{\sigma}_k$ are either the UMLEs or CMLEs of τ_k, α_k and $\sigma_k, k = 1, 2$, and $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ quantile of the chi-square distribution with 1 degree of freedom.

Although PLCIs for τ_2 are not affected by a possible extra-Poisson variation of the M_i's [or strictly speaking extra-multinomial dispersion of $(M_1, \ldots, M_n)$] because they are obtained from the likelihood function $L_{(2)}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2)$ which does not depend on these variables, we do not expect that the PLCIs for τ_1 and τ be robust to extra-Poisson variation of the M_i's; therefore we will consider adjusted PLCIs for τ_1 and τ that take into account this extra variation. Following the suggestion of Gimenes et al. (2005), the adjusted PLCIs are constructed as the previous ones but replacing the value $\chi^2_{1,1-\alpha}$ by the value $(s^2_M / \bar{m}) F_{1,n-1,1-\alpha}$, where $\bar{m} = m/n$ and $s^2_M = \sum_1^n (m_i - \bar{m})^2 / (n-1)$ are the sample mean and variance of the m_i's, and $F_{1,n-1,1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ quantile of the F distribution with 1 and $n-1$ degrees of freedom. Observe that $s^2_M / \bar{m}$ is obtained by dividing by $n-1$ the value of the Pearson chi-square test statistic to test the hypothesis that the conditional distribution of the observed M_i's, given that $\sum_1^n M_i = m$, is multinomial with parameter of size m and vector of probabilities $(1/n, \ldots, 1/n)$. The adjusted PLCIs should be used if the null hypothesis of the conditional multinomial distribution of the M_i's were rejected at the $100\alpha\%$ level of significance, that is, if $s^2_M / \bar{m} > \chi^2_{n-1,1-\alpha} / (n-1)$, otherwise the unadjusted PLCIs should be used.

It is worth noting that the calculation of PLCIs is a computationally expensive task; therefore, efficient numerical algorithms need to be used, such as the one proposed by Venzon and Moolgavkar (1988).

4.2 Bootstrap confidence intervals

We will present a variant of bootstrap to construct CIs for the population sizes τ_1, τ_2 and τ based on either the UMLEs or the CMLEs. The proposed variant is obtained by combining the bootstrap version for finite populations proposed by Booth, Butler and Hall (1994) and the parametric bootstrap variant (see Davison and Hinkley 1997, Chapter 2). This version of bootstrap is an extension of the one used by Félix-Medina and Monjardin (2006) in the case of homogeneous link-probabilities.

Since our proposed version of bootstrap is a parametric variant, we need to have estimates of all the parameters associated with the assumed models. Until now, the only parameters that have not yet been

estimated are the random effects $\beta_j^{(k)}$'s. We will now derive a predictor of $\beta_j^{(k)}$. Thus, given the subset $U_k - S_0$ or $A_i \in S_A$ that contains the element j, the conditional joint pdf of $\mathbf{X}_j^{(k)}$ and $\beta_j^{(k)}$ is

$$\begin{aligned} f\left(\mathbf{x}_j^{(k)}, \beta_j^{(k)} \middle| j \in U_k - S_0, S_A\right) &= \Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x}_j^{(k)} \middle| \beta_j^{(k)}, j \in U_k - S_0, S_A\right) f\left(\beta_j^{(k)}\right) \\ &\propto \prod_{i=1}^{n} \left[p_{ij}^{(k)}\right]^{x_{ij}^{(k)}} \left[1 - p_{ij}^{(k)}\right]^{1-x_{ij}^{(k)}} \exp\left[-\left(\beta_j^{(k)}\right)^2 \Big/ 2\sigma_k^2\right] \\ &\quad \text{if } j \in U_k - S_0,\ k = 1, 2, \end{aligned}$$

or

$$\begin{aligned} f\left(\mathbf{x}_j^{(1)}, \beta_j^{(1)} \middle| j \in A_{i'} \in S_A, S_A\right) &\propto \prod_{i \neq i'}^{n} \left[p_{ij}^{(1)}\right]^{x_{ij}^{(1)}} \left[1 - p_{ij}^{(1)}\right]^{1-x_{ij}^{(1)}} \exp\left[-\left(\beta_j^{(1)}\right)^2 \Big/ 2\sigma_1^2\right] \\ &\quad \text{if } j \in A_{i'} \in S_A,\ i' = 1, \ldots, n. \end{aligned}$$

We will use as a prediction or estimate of $\beta_j^{(k)}$ the value $\hat{\beta}_j^{(k)}$ that maximizes the conditional joint pdf of $\mathbf{X}_j^{(k)}$ and $\beta_j^{(k)}$ with the parameters α_k and σ_k set at either their UMLEs or their CMLEs. This procedure yields that $\hat{\beta}_j^{(k)}$ is given as the solution to the following equation:

$$\sum_{i=1}^{n} x_{ij}^{(k)} - \sum_{i=1}^{n} \frac{\exp\left[\hat{\alpha}_i^{(k)} + \beta_j^{(k)}\right]}{1 + \exp\left[\hat{\alpha}_i^{(k)} + \beta_j^{(k)}\right]} - \frac{1}{\hat{\sigma}_k^2} \beta_j^{(k)} = 0 \quad \text{if } j \in U_k - S_0,\ k = 1, 2,$$

or

$$\sum_{i \neq i'}^{n} x_{ij}^{(1)} - \sum_{i \neq i'}^{n} \frac{\exp\left[\hat{\alpha}_i^{(1)} + \beta_j^{(1)}\right]}{1 + \exp\left[\hat{\alpha}_i^{(1)} + \beta_j^{(1)}\right]} - \frac{1}{\hat{\sigma}_1^2} \beta_j^{(1)} = 0 \quad \text{if } j \in A_{i'} \in S_A,\ i' = 1, \ldots, n,$$

where $\hat{\alpha}_i^{(k)}$ and $\hat{\sigma}_k$ denote either the UMLEs or the CMLEs of $\alpha_i^{(k)}$ and $\sigma_k, i = 1, \ldots, n; k = 1, 2$. Note that this equation implies that the predictor $\hat{\beta}_j^{(k)}$ of $\beta_j^{(k)}$ depends on the number of clusters that are linked to the element j, but not on the particular clusters to which that element is linked. Thus, if two persons j and j' in $U_k - S_0$ are linked to the same number of clusters in S_A, the predictors $\hat{\beta}_j^{(k)}$ and $\hat{\beta}_{j'}^{(k)}$ are equal one another. The same happens for two persons in $A_i \in S_A$.

Hereinafter, we will denote by $[\hat{\tau}_k]$, the nearest integer to $\hat{\tau}_k$, where $\hat{\tau}_k$ denotes either the UMLE or the CMLE of $\tau_k, k = 1, 2$. The steps of the proposed bootstrap procedure are the following. (i) Construct a population vector $\mathbf{m}_{\text{Boot}}$ of N values of m_i's by repeating N/n times, assuming that N/n is an integer, the observed sample of n cluster sizes $\mathbf{m}_s = \{m_1, \ldots, m_n\}$. If N/n is not an integer, that is, if $N = an + b$, where a and $b, b < n$, are positive integers, then repeat a times $\mathbf{m}_s$ and add to this set a SRSWOR of b values of m_i's selected from $\mathbf{m}_s$. (ii) For each $k = 1, 2$, construct a population vector $\hat{\alpha}_{\text{Boot}}^{(k)}$ of dimension N whose elements are the estimates $\hat{\alpha}_i^{(k)}$'s of the $\alpha_i^{(k)}$'s associated with the clusters whose sizes m_i's are in $\mathbf{m}_{\text{Boot}}$. (iii) Construct a population vector $\hat{\beta}_{\text{Boot}}^{(0)}$ whose elements are the estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the people who belong to the clusters whose sizes m_i's are in $\mathbf{m}_{\text{Boot}}$. Observe that the dimension of this vector is not necessarily $[\hat{\tau}_1]$, but it equals the sum of the m_i's in $\mathbf{m}_{\text{Boot}}$. (iv) Construct a population vector $\hat{\beta}_{\text{Boot}}^{(1)}$ of dimension $[\hat{\tau}_1]$ whose first m elements are the

estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the people in S_0; the remaining $[\hat{\tau}_1] - m$ elements are the r_1 estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the people in S_1 and the $[\hat{\tau}_1] - m - r_1$ estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the non sampled people in U_1. These $[\hat{\tau}_1] - m$ elements $\hat{\beta}_j^{(1)}$'s are randomly placed after the first m elements $\hat{\beta}_j^{(1)}$ of $\hat{\boldsymbol{\beta}}_{\text{Boot}}^{(1)}$. (v) Construct a population vector $\hat{\boldsymbol{\beta}}_{\text{Boot}}^{(2)}$ of dimension $[\hat{\tau}_2]$ whose first r_2 elements are the estimates $\hat{\beta}_j^{(2)}$'s of the $\beta_j^{(2)}$'s associated with the people in S_2 and the remaining $[\hat{\tau}_2] - r_2$ elements are the estimates $\hat{\beta}_j^{(2)}$'s of the $\beta_j^{(2)}$'s associated with the non sampled people in U_2. (vi) Select a SRSWOR of n values m_i from $\mathbf{m}_{\text{Boot}}$. Let $S_A^{\text{Boot}} = \{i_1, \ldots, i_n\}$ be the set of indices of the m_i's in the sample. In addition, let $A_i^{\text{Boot}} = \left(\sum_{t=1}^{i-1} m_t, \sum_{t=1}^{i} m_t\right) \cap \mathbb{Z}$ be the set of indices j associated with the elements in the cluster whose index is $i \in S_A^{\text{Boot}}$, where m_t is the t^{th} element of $\mathbf{m}_{\text{Boot}}$ and $\mathbb{Z}$ is the set of the integer numbers. Finally, let $S_0^{\text{Boot}} = \bigcup_{i \in S_A^{\text{Boot}}} A_i^{\text{Boot}}$. (vii) For each $i \in S_A^{\text{Boot}}$ and $j \in \{1, \ldots, [\hat{\tau}_2]\}$ generate a value $x_{ij}^{(2)}$ by sampling from the Bernoulli distribution with mean $\hat{p}_{ij}^{(2)}$ given by (3.2), but replacing $\alpha_i^{(2)}$ and $\beta_j^{(2)}$ by their estimates $\hat{\alpha}_i^{(2)}$ and $\hat{\beta}_j^{(2)}$. Similarly, for each $i \in S_A^{\text{Boot}}$ and $j \in \{1, \ldots, [\hat{\tau}_1]\} - A_i^{\text{Boot}}$ generate a value $x_{ij}^{(1)}$ by sampling from the Bernoulli distribution with mean $\hat{p}_{ij}^{(1)}$, where the value of $\hat{\beta}_j^{(1)}$ that is used to compute $\hat{p}_{ij}^{(1)}$ is obtained from $\hat{\boldsymbol{\beta}}_{\text{Boot}}^{(0)}$ if $j \in S_0^{\text{Boot}}$, and from $\hat{\boldsymbol{\beta}}_{\text{Boot}}^{(1)}$ otherwise. (viii) Compute the estimates of τ_1, τ_2 and τ using the same procedure as that used to compute the original estimates $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. (ix) Repeat the steps (vi)–(viii) a large enough number B of times. Let $\hat{\tau}_{1,b}^{\text{Boot}}, \hat{\tau}_{2,b}^{\text{Boot}}$ and $\hat{\tau}_b^{\text{Boot}}$ be the estimates obtained in the b^{th} bootstrap sample, $b = 1, \ldots, B$.

The final step of our proposed bootstrap variant consists in constructing the CIs for the population sizes. There exist several alternatives to do this. One is to construct them without assuming any distributions for the estimators $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. As examples of this alternative are the basic and the percentile method. (See Davison and Hinkley 1997, Chapter 5, for descriptions of these methods.) In the basic method a $100(1-\alpha)\%$ CI for τ is $\left[2\hat{\tau} - \hat{\tau}_{1-\alpha/2}^{\text{Boot}}, 2\hat{\tau} - \hat{\tau}_{\alpha/2}^{\text{Boot}}\right]$, and in the percentile method the CI is $\left[\hat{\tau}_{\alpha/2}^{\text{Boot}}, \hat{\tau}_{1-\alpha/2}^{\text{Boot}}\right]$, where $\hat{\tau}_{\alpha/2}^{\text{Boot}}$ and $\hat{\tau}_{1-\alpha/2}^{\text{Boot}}$ are the lower and upper $\alpha/2$ points of the empirical distribution obtained from $\hat{\tau}_b^{\text{Boot}}, b = 1, \ldots, B$. Although this type of alternative has good properties of robustness, it requires a large number B of bootstrap samples, say $B = 1{,}000$, and this might be a serious problem if $\hat{\tau}$ is costly to compute.

Another alternative to construct CIs is to assume a distribution for $\hat{\tau}$ and use the bootstrap sample to estimate the parameters of that distribution. In this case the number B of required bootstrap samples is not so large, say $50 \leq B \leq 200$ is generally enough. Examples of this alternative are the assumption that $\hat{\tau}$ is normally distributed and the one that $\hat{\tau} - \nu$ is lognormally distributed, where ν is the number of sampled elements. In the first case a $100(1-\alpha)\%$ CI for τ is the well known Wald CI given by $\hat{\tau} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})}$, whereas in the second case the CI is $[\nu + (\hat{\tau} - \nu)/c,\ \nu + (\hat{\tau} - \nu) \times c]$, where $c = \exp\left\{z_{\alpha/2}\sqrt{\ln\left[1 + \hat{V}(\hat{\tau})/(\hat{\tau} - \nu)^2\right]}\right\}$, $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution and $\hat{V}(\hat{\tau})$ is an estimate of the variance of $\hat{\tau}$. (See Williams, Nichols and Conroy 2002, Section 14.2, for a description of this type of CI.) It is worth noting that in the lognormal based CIs for τ_1, τ_2 and τ the values of ν are $m + r_1, r_2$ and $m + r_1 + r_2$, respectively.

An estimator $\hat{V}(\hat{\tau})$ of the variance of $\hat{\tau}$ could be computed using the sample variance of the bootstrap sample $\hat{\tau}_b^{\text{Boot}}, b = 1,\ldots,B$. However, this estimator is not robust to extreme values of $\hat{\tau}_b^{\text{Boot}}$, which are likely to occur with the proposed estimators when the sampling rates are not large enough. Therefore, to use a robust estimator of $V(\hat{\tau})$ is a better strategy. One possibility is to use Huber's proposal 2 to jointly estimate the parameters of location and scale from the bootstrap sample. (See Staudte and Sheather 1990, Section 4.5, for a description of this method.) In particular, the estimate of the parameter of scale is an estimate of the standard deviation $\sqrt{\hat{V}(\hat{\tau})}$ of $\hat{\tau}$.

5 Sample size determination

We will present a procedure to determine the initial sample size n. This procedure is based on stringent assumptions, but, as was indicated by one of the reviewers of the paper, it could nevertheless be very useful for researchers who want to apply this sampling design.

The first step is to compute the asymptotic variances of the proposed estimators. Although the variances depend on several unknown parameters, we can simplify them by assuming that the effects $\alpha_i^{(k)}$ of the sampled sites are homogeneous, that is, $\alpha_i^{(k)} = \alpha^{(k)}, i = 1,\ldots,n; k = 1,2$. Under this premise, the probabilities $\pi_{\mathbf{x}}^{(k)}$ and $\pi_{\mathbf{x}}^{(A_i)}$ that the vectors of link-indicator variables associated with randomly selected persons from $U_k - S_0$ and S_0, respectively, equal $\mathbf{x}$ depend only on the number of 1's that appear in the vector $\mathbf{x}$. Thus, their Gaussian quadrature approximations, given by (3.3) and (3.4), are simplified to

$$\tilde{\pi}_x^{(k)}\left(\boldsymbol{\theta}^{(k)}\right) = \tilde{\pi}_x^{(k)}\left(\alpha^{(k)},\sigma_k\right) = \sum_{t=1}^{q}\frac{\exp\left[x\left(\alpha^{(k)}+\sigma_k z_t\right)\right]}{\left[1+\exp\left(\alpha^{(k)}+\sigma_k z_t\right)\right]^n}\nu_t,\ x = 0,1,\ldots,n,$$

and

$$\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(k)}\right) = \tilde{\pi}_x^{(A)}\left(\alpha^{(1)},\sigma_1\right) = \sum_{t=1}^{q}\frac{\exp\left[x\left(\alpha^{(1)}+\sigma_1 z_t\right)\right]}{\left[1+\exp\left(\alpha^{(1)}+\sigma_1 z_t\right)\right]^n}\nu_t,\ x = 0,1,\ldots,n-1,$$

where $\boldsymbol{\theta}^{(k)} = \left(\theta_1^{(k)},\theta_2^{(k)}\right) = \left(\alpha^{(k)},\sigma_k\right)$.

Following Sanathanan's (1972) procedure we get that the asymptotic variances of the proposed estimators are given by

$$V(\hat{\tau}_k) = \tau_k\big/\left(D_k - \mathbf{B}_k'\mathbf{A}_k^{-1}\mathbf{B}_k\right),\ k = 1,2,\ \text{and}\ V(\hat{\tau}) = V(\hat{\tau}_1) + V(\hat{\tau}_2),$$

where $\mathbf{A}_k = \left[a_{ij}^{(k)}\right]$ is a 2×2 matrix whose $a_{ij}^{(k)}$ element is

$$a_{ij}^{(k)} = \begin{cases} \left(1-\dfrac{n}{N}\right)\displaystyle\sum_{x=0}^{n}\binom{n}{x}\frac{1}{\tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}\left[\frac{\partial\tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial\theta_i^{(1)}}\right]\left[\frac{\partial\tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial\theta_j^{(1)}}\right] \\ +\dfrac{n}{N}\displaystyle\sum_{x=0}^{n-1}\binom{n-1}{x}\frac{1}{\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}\left[\frac{\partial\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial\theta_i^{(1)}}\right]\left[\frac{\partial\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial\theta_j^{(1)}}\right] & \text{if } k=1 \\ \displaystyle\sum_{x=0}^{n}\binom{n}{x}\frac{1}{\tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}\left[\frac{\partial\tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}{\partial\theta_i^{(2)}}\right]\left[\frac{\partial\tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}{\partial\theta_j^{(2)}}\right] & \text{if } k=2, \end{cases}$$

$\mathbf{B}_k$ is a bi-dimensional vector whose elements are

$$b_i^{(k)} = -\left[\partial\tilde{\pi}_0^{(k)}\left(\boldsymbol{\theta}^{(k)}\right)/\partial\theta_i^{(k)}\right]/\tilde{\pi}_0^{(k)}(\boldsymbol{\theta}^{(k)}),\ i = 1,2$$

and D_k is a real number given by

$$D_k = \begin{cases} \left[1-(1-n/N)\,\tilde{\pi}_0^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)\right]/\left[(1-n/N)\,\tilde{\pi}_0^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)\right] & \text{if } k = 1 \\ \left[1-\tilde{\pi}_0^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)\right]/\tilde{\pi}_0^{(2)}\left(\boldsymbol{\theta}^{(2)}\right) & \text{if } k = 2. \end{cases}$$

It is worth noting that in the derivation of the asymptotic variances we have made the assumptions that $\tau_k \to \infty, k = 1,2, m_i \to \infty, i = 1,\ldots,N,$ and N and n are fixed numbers.

To obtain numerical values of the variances we need to specify values for $\tau_k, \alpha^{(k)}, \sigma_k$ and n. One way to do this is to assign values to τ_k, σ_k and to the proportion $\tilde{\pi}_{1+}^{(k)}$ of people in U_k who are linked at least to a particular site A_i, which is common to all the sites and it is easier to specify than $\alpha^{(k)}$. Then, for a given n, the value of $\alpha^{(k)}$ is the solution to the equation

$$\tilde{\pi}_{1+}^{(k)} - \sum_{x=1}^{n}\binom{n-1}{x-1}\tilde{\pi}_x^{(k)}\left(\alpha^{(k)},\sigma_k\right) = 0.$$

Once $\alpha^{(k)}, k = 1,2,$ is obtained for a given n, we can compute the numerical values of the variances $V(\hat{\tau}_k)$ and $V(\hat{\tau})$; the square roots of the relative variances $\sqrt{V(\hat{\tau}_k)/\tau_k^2}$ and $\sqrt{V(\hat{\tau})/\tau^2}$, and the sampling fractions $f_1 = 1-(1-n/N)\,\tilde{\pi}_0^{(1)}$, $f_2 = 1-\tilde{\pi}_0^{(2)}$ and $f = (f_1\times\tau_1 + f_2\times\tau_2)/\tau$. If the values of the square roots of the relative variances were not satisfactory, we could try different values of n until we get satisfactory values.

We have programmed this procedure in the R software programming language and it is available to the interested readers by requesting to the authors. To illustrate the procedure, let us suppose that we have a sampling frame of $N = 150$ sites and we assign the values $\tau_1 = 1{,}200$, $\tau_2 = 400, \sigma_1 = \sigma_2 = 1$, $\tilde{\pi}_{1+}^{(1)} = 0.05$ and $\tilde{\pi}_{1+}^{(2)} = 0.04$, then for $n = 15$ we get that $V(\hat{\tau}_1) = 4{,}780.8$, $V(\hat{\tau}_2) = 11{,}525.3$, $V(\hat{\tau}) = 16{,}306.1$, $\sqrt{V(\hat{\tau}_1)}/\tau_1 = 0.06$, $\sqrt{V(\hat{\tau}_2)}/\tau_2 = 0.27$, $\sqrt{V(\hat{\tau})}/\tau = 0.08$, $f_1 = 0.50$, $f_2 = 0.38$ and $f = 0.47$.

6 Monte Carlo studies

6.1 Populations constructed using artificial data

We constructed four artificial populations; a description of each one is presented in Table 6.1. Notice that in Populations I, III and IV the $N = 150$ values of the m_i's were obtained by sampling from a Poisson distribution, whereas in Population II by sampling from a zero truncated negative binomial distribution. In addition, in Populations I and II, the link-probabilities $p_{ij}^{(k)}$ were generated by the Rasch model (3.2). In Population III they were generated by that model but the random effects $\beta_j^{(k)}$ were obtained by sampling from a scaled Student's T distribution with six degrees of freedom and unit-variance

instead of by sampling from the standard normal distribution. Finally, in Population IV, the $p_{ij}^{(k)}$'s were generated by the following latent class model proposed by Pledger (2000) in the context of capture-recapture studies: $p_{ij}^{(k)} = \exp\left[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}\right] \Big/ \left\{1 + \exp\left[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}\right]\right\}$, $i = 1, \ldots, n$; $j = 1, 2$, and $k = 1, 2$. In this model the people in U_k is divided into two latent classes $(j = 1, 2)$ according to their propensities to be linked to the sampled clusters. The probability that a randomly person in U_k is in class j is $p_j^{(k)}$ and the $p_{ij}^{(k)}$'s are the same for all the people in the class j.

The simulation experiment was carried out by repeatedly selecting r samples from each population by using the sampling design described in Section 2 with initial sample size $n = 15$. Thus, each time that the value m_i was included in an initial sample, the value $x_{ij}^{(k)}$ was obtained by sampling from the Bernoulli distribution with mean $p_{ij}^{(k)}$. Because of the values assigned to n and to the parameters that appear in the expression of $p_{ij}^{(k)}$, the resulting sampling rates were $f_1 \approx 0.5$ and $f_2 \approx 0.4$. It is worth noting that the characteristics of the populations and samples considered in this study were not motivated by the ones of an actual study since this sampling design has not been applied yet. Thus, the populations and samples were constructed only with the purpose of analyzing the performance of the proposed point and interval estimators.

Table 6.1
Parameters of the simulated populations

Population I	Population II	Population III	Population IV
$N = 150$	$N = 150$	$N = 150$	$N = 150$
$M_i \sim$ Poisson	$M_i \sim$ Zero trunc. neg. binomial	$M_i \sim$ Poisson	$M_i \sim$ Poisson
$E(M_i) = 8$	$E(M_i) = 8$	$E(M_i) = 8$	$E(M_i) = 8$
$V(M_i) = 8$	$V(M_i) = 24$	$V(M_i) = 8$	$V(M_i) = 8$
$\tau_1 = 1{,}209$	$\tau_1 = 1{,}208$	$\tau_1 = 1{,}209$	$\tau_1 = 1{,}209$
$\tau_2 = 400$	$\tau_2 = 400$	$\tau_2 = 400$	$\tau_2 = 400$
$\tau = 1{,}609$	$\tau = 1{,}608$	$\tau = 1{,}609$	$\tau = 1{,}609$
$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0.001}$	$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0.001}$	$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0.001}$	$\alpha_i^{(k)} = \dfrac{-12}{M_i^{1/2} + 0.001}$
$c_1 = -5.45$	$c_1 = -5.45$	$c_1 = -5.45$	$\mu^{(1)} = -1.1$; $\mu^{(2)} = -1.2$
$c_2 = -5.85$	$c_2 = -5.85$	$c_2 = -5.85$	$\beta_1^{(k)} = 1.5$; $\beta_2^{(k)} = 0$
$\beta_j^{(k)} \sim N(0,1)$	$\beta_j^{(k)} \sim N(0,1)$	$\beta_j^{(k)} \sim \sqrt{2/3}T_6$	$(\alpha\beta)_{ij}^{(k)} \sim N\left(0, 1.25^2\right)$
			$p_1^{(k)} = 0.3 = 1 - p_2^{(k)}$

From each sample the following estimators of τ_1, τ_2 and τ were considered: the proposed UMLEs $\hat{\tau}_1^{(U)}, \hat{\tau}_2^{(U)}$ and $\hat{\tau}^{(U)}$; the proposed CMLEs $\hat{\tau}_1^{(C)}, \hat{\tau}_2^{(C)}$ and $\hat{\tau}^{(C)}$; the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ proposed by Félix-Medina and Thompson (2004) and derived under the assumption of homogeneous link-probabilities, and the Bayesian-assisted estimators $\breve{\tau}_1, \breve{\tau}_2$ and $\breve{\tau}$ proposed by Félix-Medina and Monjardin (2006), derived also under the homogeneity assumption and using the following initial distributions for τ_1, τ_2 and $\alpha_i^{(k)} = \ln\left[p_i^{(k)} \big/ \left(1 - p_i^{(k)}\right)\right]$, where $p_i^{(k)}$ is given by (3.2), but setting $\beta_j^{(k)} = 0$: $\xi(\tau_1|\lambda_1) \propto (N\lambda_1)^{\tau_1} \big/ \tau_1!$

and $\xi(\lambda_1) \propto \lambda_1^{a_1-1}\exp(-b_1\lambda_1)$; $\xi(\tau_2|\lambda_2) \propto \lambda_2^{\tau_2}/\tau_2!$ and $\xi(\lambda_2) \propto \lambda_2^{a_2-1}\exp(-b_2\lambda_2)$, and $\xi\left(\alpha_i^{(k)}|\theta_k\right) \propto \exp\left[-\left(\alpha_i^{(k)}-\theta_k\right)^2/2\sigma_k^2\right]$ and $\xi(\theta_k) \propto \exp\left[-(\theta_k-\mu_k)^2/2\gamma_k^2\right]$, where $a_1 = 1.0$, $b_1 = 0.1, a_2 = 6.0, b_2 = 0.01, \mu_k = -3.5$ and $\sigma_k^2 = \gamma_k^2 = 9.0$. These values assigned to the parameters of the initial distributions made them practically non-informative. The Gaussian quadrature approximations (3.3) and (3.4) to the probabilities $\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ and $\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{-i}, \sigma_1\right)$ were computed using $q = 40$ terms.

The performance of an estimator $\hat{\tau}$ of τ, say, was evaluated by means of its relative bias $(\mathrm{r-bias})$, the square root of its relative mean square error $\left(\sqrt{\mathrm{r-mse}}\right)$, and the medians of its relative estimation errors (mdre) and its absolute relative estimation errors (mdare) defined by $\mathrm{r-bias} = \sum_1^r(\hat{\tau}_i-\tau)/(r\tau)$, $\sqrt{\mathrm{r-mse}} = \sqrt{\sum_1^r(\hat{\tau}_i-\tau)^2/(r\tau^2)}$, $\mathrm{mdre} = \mathrm{median}\{(\hat{\tau}_i-\tau)/\tau\}$ and $\mathrm{mdrae} = \mathrm{median}\{|(\hat{\tau}_i-\tau)/\tau|\}$, where $\hat{\tau}_i$ was the value of $\hat{\tau}$ obtained in the i^{th} sample, which in the case of the point estimators was 10,000.

We also considered the following 95% CIs for the τ's : the proposed PLCIs and adjusted for extra-Poisson variation PLCIs; the proposed bootstrap CIs based on $B = 100$ bootstrap samples and constructed assuming a lognormal distribution for $\hat{\tau} - \nu$ and estimating $\sqrt{\hat{V}(\hat{\tau})}$ by Huber's proposal 2 estimator of scale with tuning value $d = 1.5$; the design-based Wald CIs obtained from the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ and proposed by Félix-Medina and Thompson (2004), and the design-based Wald CIs obtained from the Bayesian estimators $\bar{\tau}_1, \bar{\tau}_2$ and $\bar{\tau}$ and proposed by Félix-Medina and Monjardin (2006). It is worth noting that the PLCIs and adjusted PLCI were computed using the Venzon and Moolgavkar's (1988) method or an algorithm based on the definition of a PLCI when the first method failed to find the endpoints of the intervals.

The performance of a CI was evaluated by its coverage probability (cp), the mean of its relative lengths (mrl) and the median of its relative lengths (mdrl) defined as the proportion of samples in which the parameter was contained in the interval and the the mean and the median of the lengths of the intervals divided by the value of the parameter, respectively. Since carrying out a simulation study on the CIs using a large number of replicated samples is a very time consuming task, we evaluated the performance of the PLCIs for τ_1 and τ_2 using $r = 1{,}000$ samples; that of the PLCIs for τ using $r = 500$ samples and that of the bootstrap CIs using $r = 250$ samples. The performance of the CIs based on the estimators derived under the homogeneity assumption was evaluated by using $r = 10{,}000$ samples. The numerical study was carried out using the R software programming language [R Development Core Team (2013)].

The results of the study on the estimators of the population sizes are shown in Table 6.2 and in Figures 6.1 and 6.2. The main outcomes are the following. The distributions of the estimators UMLE $\hat{\tau}_1^{(U)}$ and CMLE $\hat{\tau}_1^{(C)}$ were more or less symmetrical about τ_1; thus, the two measures of bias $(\mathrm{r-bias}$ and mdre) showed similar values, as well as the two measures of variability $(\sqrt{\mathrm{r-mse}}$ and mdare). Both of these estimators performed acceptably well, except in Population III where the estimator $\hat{\tau}_1^{(U)}$ presented moderate problems of bias and $\hat{\tau}_1^{(C)}$ showed something more serious problems of bias. The distributions of the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ were skewed to the right with very long tails. This caused that the values of their $\mathrm{r-bias}$ and $\mathrm{r-mse}$ tended to be large. However, in terms of the medians of their relative errors

(mdre), these estimators presented moderate problems of bias in Populations I and III and serious problems in Population IV. In terms of the medians of their absolute relative errors (mdare) these estimators showed moderate problems of instability in the first three populations and serious problems in the fourth population. The distributions of the estimators $\hat{\tau}^{(U)}$ and $\hat{\tau}^{(C)}$ were similar to those of the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$; thus, the quantities $\mathrm{r-bias}$ and $\sqrt{\mathrm{r-mse}}$ were more sensitive to large values than the quantities mdre and mdare. Both of these estimators performed acceptably well in Populations I, II and IV; although in this last population the values of their $\sqrt{\mathrm{r-mse}}$ were large because of the reasons previously indicated. In Population III both estimators presented problems of bias.

Although the deviation from the assumed Poisson distribution of the M_i's increased the variability of all the proposed estimators, the increments were not large so that we consider that they have some robust properties against deviations from this assumption. The proposed estimators of τ_1 and τ were in addition robust to the deviation from the assumed Rasch model for the $p_{ij}^{(1)}$'s (although the values of the $\sqrt{\mathrm{r-mse}}$ of the estimators of τ were large, those of the median of their absolute relative errors were not). The deviation from the assumed normal distribution of the effects $\beta_j^{(k)}$ caused that all the proposed estimators presented problems of overestimation. Neither of the two types of proposed estimators UMLEs and CMLEs performed uniformly better than the other, but the UMLEs performed in a greater number of cases slightly better than the CMLEs.

Table 6.2
Relative biases, square roots of relative mean square errors and medians of relative errors and absolute relative errors of the estimators of the population sizes

Population		I				II				III				IV			
Sampling rates		f_1 0.51		f_2 0.40		f_1 0.50		f_2 0.40		f_1 0.51		f_2 0.40		f_1 0.51		f_2 0.40	
Estimator		r bias	$\sqrt{\text{r mse}}$	mdre	mdare	r bias	$\sqrt{\text{r mse}}$	mdre	mdare	r bias	$\sqrt{\text{r mse}}$	mdre	mdare	r bias	$\sqrt{\text{r mse}}$	mdre	mdare
Uncond. heter.	$\hat{\tau}_1^{(U)}$	-0.01	0.06	-0.01	0.04	-0.00	0.08	-0.00	0.05	0.10	0.11	0.10	0.10	$-0.04_6^{0.20}$	0.08	-0.03	0.05
	$\hat{\tau}_2^{(U)}$	-0.06	0.26	-0.11	0.17	$0.06^{0.02}$	0.35	-0.01	0.16	0.16	0.43	0.07	0.18	0.04_2^{15}	2.2	-0.19	0.25
MLEs	$\hat{\tau}^{(U)}$	-0.02	0.08	-0.03	0.05	$0.01^{0.02}$	0.10	0.01	0.06	0.11	0.15	0.10	0.10	-0.02_7^{15}	0.55	-0.6	0.08
Cond. heter.	$\hat{\tau}_1^{(C)}$	-0.00	0.07	-0.01	0.05	0.01	0.07	0.00	0.05	0.18	0.19	0.17	0.17	$-0.05^{1.6}$	0.09	-0.05	0.07
	$\hat{\tau}_2^{(C)}$	-0.04	0.26	-0.09	0.17	0.09	0.38	0.01	0.16	0.18	0.46	0.10	0.18	0.12_2^{21}	2.4	-0.14	0.23
MLEs	$\hat{\tau}^{(C)}$	-0.1	0.08	-0.02	0.05	0.03	0.11	0.01	0.06	0.18	0.22	0.16	0.16	-0.00_2^{23}	0.61	-0.06	0.08
Homo-geneous MLEs	$\tilde{\tau}_1$	-0.28	0.28	-0.28	0.28	-0.31	0.31	-0.31	0.31	-0.30	0.30	-0.30	0.30	-0.18	0.19	-0.18	0.18
	$\tilde{\tau}_2$	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.30	0.32	-0.32	0.32
	$\tilde{\tau}$	-0.31	0.31	-0.31	0.31	-0.33	0.33	-0.33	0.33	-0.32	0.33	-0.32	0.32	-0.21	0.22	-0.21	0.21
Homo-geneous BEs	$\bar{\tau}_1$	-0.28	0.28	-0.28	0.28	-0.31	0.31	-0.31	0.31	-0.30	0.30	-0.30	0.30	-0.18	0.19	-0.18	0.18
	$\bar{\tau}_2$	-0.39	0.39	-0.39	0.39	-0.39	0.40	-0.39	0.39	-0.39	0.40	-0.39	0.39	-0.27	0.30	-0.29	0.29
	$\bar{\tau}$	-0.31	0.31	-0.31	0.31	-0.33	0.33	-0.33	0.33	-0.32	0.32	-0.32	0.32	-0.20	0.21	-0.20	0.20

Notes Results based on 10^4 samples. A superscript number indicates the percentage of samples in which the estimator was not computed because of numerical convergence problems and a subscript figure indicates the number of values of the estimator that exceeded 10^5 and that were not used to compute its $\mathrm{r-bias}$ and $\mathrm{r-mse}$. Upper bounds for the Monte Carlo errors of the estimates of the $\mathrm{r-bias}$ and the $\sqrt{\mathrm{r-mse}}$ of the estimators of the τ's were the following: τ_1 : 0.001 and 0.001; τ_2 : 0.004 and 0.011 in Pops. I–III, and 0.027 and 0.39 in Pop. IV; τ : 0.001 and 0.002 in Pops. I–III, and 0.007 and 0.095 in Pop. IV.

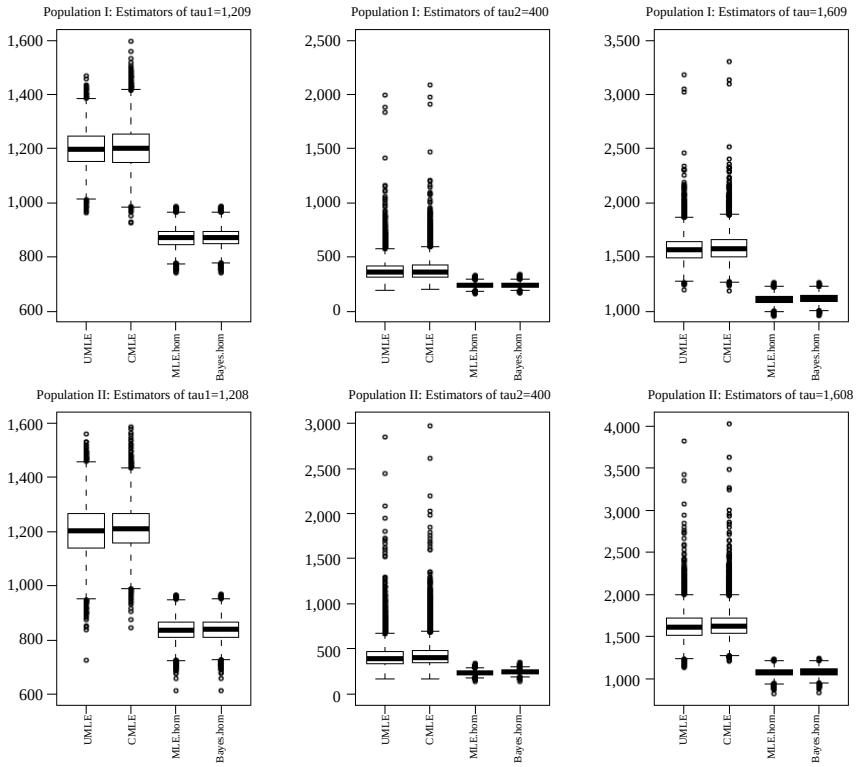

Figure 6.1 Boxplots for the values of the estimators of τ_1, τ_2 and τ in Populations I and II.

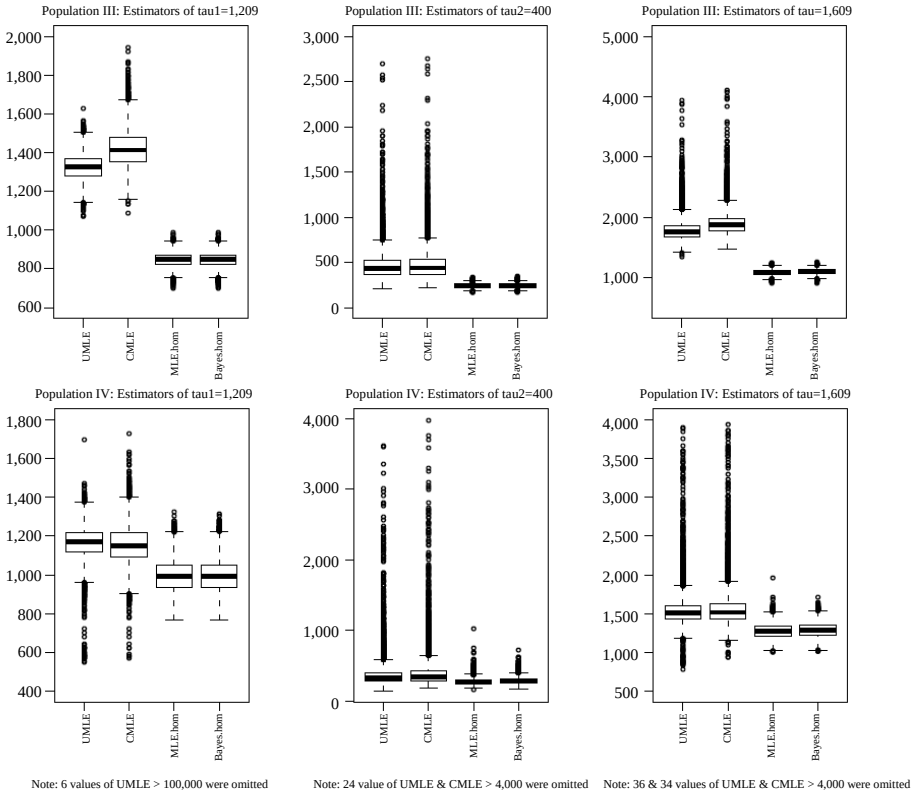

Figure 6.2 Boxplots for the values of the estimators of τ_1, τ_2 and τ in Populations III and IV.

With regard to the estimators derived under the assumption of homogeneous $p_{ij}^{(k)}$, both the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ and the Bayesian assisted estimators $\breve{\tau}_1, \breve{\tau}_2$ and $\breve{\tau}$ showed very similar behavior which was characterized by serious problems of bias that deteriorated their performance.

Notice that the percentages of samples in which the proposed estimators were not computed because of numerical convergence problems, as well as the number of samples in which the values of the estimators exceeded 10^5 and that not were used in calculating the reported results of $\mathrm{r-bias}$ and $\sqrt{\mathrm{r-mse}}$ because they would have been seriously affected, were not large, except in Population IV. As was indicated by a reviewer, computing the $\mathrm{r-bias}$ and $\mathrm{r-mse}$ of an estimator using only its available values lower than 10^5 favors the proposed estimators. We agree with that observation and for this reason we also reported measures of the performance of the estimators based on the medians of the relative errors and absolute relative errors which are robust to large values of the estimators. Thus, if we supposed that any time that an estimator was not computed its value had been very large and we computed the values of the measures of the performance of the estimators that are based on the medians using the complete set of observations the results would not have been different from those reported in Table 6.2, and our conclusions based on these measures would not have changed.

The results of the simulation study on the 95% CIs are shown in Table 6.3. The main outcomes are the following. All the PLCIs and adjusted PLCIs for τ_1 : the ones based on the UMLE $\hat{\tau}_1^{(U)}$ and those based on the CMLE $\hat{\tau}_1^{(C)}$, showed good values of the cp in Population I. The adjusted PLCIs presented also good values of the cp in Population II, but not the unadjusted PLCIs whose values of the cp were relatively low. In Population III the values of the cp of all the PLCIs and adjusted PLCIs for τ_1 were low, whereas in Population IV the values were only slightly low. A good characteristic of these CIs was that they showed pretty acceptable values of their mrl and mdrl in each of the situations that were considered. The PLCI for τ_2 based on $\hat{\tau}_2^{(U)}$ and the one based on $\hat{\tau}_2^{(C)}$ presented acceptable values of the cp in all the populations, except in Population IV, where the values were something low. However, in all the cases the mrl and mdrl of these CIs were so large that they were not useful for making reasonable inferences. Both PLCI for τ : the one based on the UMLE $\hat{\tau}^{(U)}$ and that based on the CMLE $\hat{\tau}^{(C)}$, performed acceptably well in Populations I and II, although the means of their relative lengths were large in Population II because this measure is not robust to great values of the lengths of the intervals. In the other populations these CIs showed problems of low coverage and/or large relative lengths; thus their performance was not good. Both types of adjusted PLCIs performed well only in Population I, in the other populations they presented large values of their relative lengths. Neither of the two types of CIs: the ones based on the UMLEs and those based on CMLEs performed uniformly better than the other, but those based on the UMLEs performed in a greater number of cases slightly better than those based on the CMLEs.

With respect to the bootstrap CIs, we have that each of the two types of CIs for τ_1 : the one based on $\hat{\tau}_1^{(U)}$ and that based on $\hat{\tau}_1^{(C)}$ performed well in Populations I, II and IV, although in this last population the values of their cp were slightly low. In Population III the values of their cp were very low because of the biases of the point estimators of τ_1. The two types of bootstrap CIs for τ_2 performed badly in all the populations because the values of their relative lengths were large. Finally, the two types of CIs for τ performed in general well. Notice that the values of their mrl tended to be large because this measure is not robust to great values of the lengths, whereas the values of their mdrl were acceptable. Neither of the

two types of bootstrap CIs performed uniformly better than the other, but the CIs based on the UMLEs performed in most cases better than those based on the CMLEs.

Table 6.3
Coverage probabilities and means and medians of relative lengths of the 95% confidence intervals for the population sizes

Population	I			II			III			IV		
Sampling rates	f_1		f_2	f_1		f_2	f_1		f_2	f_1		f_2
	0.51		0.40	0.50		0.40	0.51		0.40	0.51		0.40
Conf. interval	cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl
PLCI- $\hat{\tau}_1^{(U)}$	0.95	0.22	0.22	0.85	0.22	0.22	0.61	0.24	0.24	$0.89^{1.6}$	0.23	0.24
Adj-PLCI- $\hat{\tau}_1^{(U)}$	0.94	0.24	0.24	0.98	0.42	0.41	0.69	0.27	0.26	$0.90^{1.6}$	0.25	0.26
PLCI- $\hat{\tau}_2^{(U)}$	0.94	1.4	0.98	0.95	2.8_2	1.3	0.95	2.6	1.6	0.77^{19}	7.1_6	1.4
PLCI- $\hat{\tau}^{(U)}$	0.95	0.66	0.59	0.97	0.91	0.65	1.0	1.0	0.79	0.86^{21}	2.1_1	0.65
Adj-PLCI- $\hat{\tau}^{(U)}$	0.92	0.75	0.62	$1.0^{7.0}$	5.8_{33}	2.1	$1.0^{0.20}$	1.7_1	0.87	0.90^{22}	2.7_4	0.78
Bootstr-CI- $\hat{\tau}_1^{(U)}$	0.94	0.23	0.23	0.94	0.33	0.31	0.59	0.25	0.25	$0.89^{0.40}$	0.24	0.25
Bootstr-CI- $\hat{\tau}_2^{(U)}$	0.87	1.4	0.86	0.97	3.9	1.6	0.97	4.7	2.1	0.83^{13}	4.6_3	0.90
Bootstr-CI- $\hat{\tau}^{(U)}$	0.93	0.38	0.28	0.99	0.97	0.49	0.98	1.1	0.52	0.89^{13}	1.2_3	0.31
PLCI- $\hat{\tau}_1^{(C)}$	0.96	0.24	0.23	0.91	0.25	0.24	0.76	0.31	0.29	$0.90^{2.1}$	0.25	0.25
Adj-PLCI- $\hat{\tau}_1^{(C)}$	0.95	0.26	0.25	0.99	0.44	0.42	0.81	0.33	0.32	$0.90^{2.1}$	0.27	0.27
PLCI- $\hat{\tau}_2^{(C)}$	0.94	1.4	0.98	0.95	2.7_2	1.3	0.95	2.5	1.6	0.85^{25}	7.5_4	1.7
PLCI- $\hat{\tau}^{(C)}$	0.95	0.64	0.54	$0.94^{1.6}$	1.2	0.62	$0.86^{2.0}$	3.1_1	0.74	0.93^{30}	2.9_1	0.84
Adj-PLCI- $\hat{\tau}^{(C)}$	0.96	0.82	0.59	$1.0^{7.6}$	7.2_{33}	2.6	$0.90^{3.8}$	3.6_6	1.2	0.94^{32}	2.9_5	0.90
Bootstr-CI- $\hat{\tau}_1^{(C)}$	0.96	0.29	0.29	0.98	0.31	0.31	0.48	0.40	0.39	$0.89^{1.6}$	0.31	0.30
Bootstr-CI- $\hat{\tau}_2^{(C)}$	0.90	1.5	1.0	0.98	4.7	1.7	0.97	5.5	2.4	0.94^{24}	3.4_6	1.3
Bootstr-CI- $\hat{\tau}^{(C)}$	0.95	0.44	0.34	1.0	1.1	0.49	0.96	1.3	0.62	0.94^{25}	0.88_6	0.43
Wald-CI- $\tilde{\tau}_1$	0.00	0.09	0.09	0.00	0.08	0.08	0.00	0.08	0.08	0.08	0.13	0.13
Wald-CI- $\tilde{\tau}_2$	0.00	0.17	0.16	0.00	0.17	0.16	0.00	0.16	0.16	0.18	0.34	0.31
Wald-CI- $\tilde{\tau}$	0.00	0.08	0.08	0.00	0.07	0.07	0.00	0.07	0.07	0.03	0.13	0.13
Wald-CI- $\bar{\tau}_1$	0.00	0.09	0.09	0.00	0.09	0.09	0.00	0.08	0.08	0.11	0.15	0.15
Wald-CI- $\bar{\tau}_2$	0.00	0.17	0.17	0.00	0.17	0.17	0.00	0.17	0.17	0.25	0.36	0.33
Wald-CI- $\bar{\tau}$	0.00	0.08	0.08	0.00	0.08	0.08	0.00	0.08	0.07	0.04	0.15	0.14

Notes A superscript number indicates the percentage of samples in which the confidence interval (CI) was not computed because of numerical convergence problems and a subscript figure indicates the number of values of the relative length of the CI that exceeded 10^5 and that were not used to compute its mrl. An upper bound (UB) for the Monte Carlo errors (MCEs) of the cps was 0.03. UBs for the MCEs of the estimates of the mrl were the following: PLCIs and adj. PLCIs for τ_1: 0.003; PLCIs for τ_2: 0.78; PLCIs and Adj. PLCIs for τ: 0.07 in Pop. I and 0.56 in Pops. II – IV; Bootstrap CIs for τ_1: 0.005; Bootstrap CIs for τ_2: 0.1 in Pop. I and 1.5 in Pops II – IV; Bootstrap CIs for τ: 0.02 in Pop. I and 0.37 in Pops. II – IV.

With regard to the CIs based on the point estimators derived under the homogeneity assumption, all of them showed null values of the cp, except in Population IV where the values were different from zero, but still too low. The bad performance of these CIs in terms of the cp was a result of the large biases of the point estimators. Thus, despite of the very small values of the $\mathrm{r}-\mathrm{ml}$ of these intervals the very low values of their cp did not allow making reasonable inferences.

Observe that in the first three populations the percentages of samples in which the proposed CIs were not computed because of numerical convergence problems as well as the number of samples in which the values of their relative lengths exceeded 10^5 and that not were used in calculating the reported results of the mrl because they would have been seriously affected, were not large (less than 4%), except in the case of the adjusted PLCIs based on $\hat{\tau}^{(U)}$ and on $\hat{\tau}^{(C)}$ which in Population II were not computed in about 7% of the samples and the means of their relative lengths were computed without using 33 values greater than 10^5. However, in Population IV some percentages were close to 20% and others close to 30%. The large values of these percentages were, in part, consequence of the relatively large values of the percentages of samples in which the corresponding point estimators were not computed. It is clear that computing the measures of performance of a CI using only the cases in which the interval was obtained or computing the rml using only the samples in which the relative lengths were lower than 10^5 favors the proposed CIs. However, notice that practically in all cases in which those percentages were large, say larger than 5%, both the mrl and the mdrl of the intervals were large enough that those intervals were not useful for making inferences. Therefore, if the performance of these CIs was not good under this favorable assessment, it will not be good under a fairer evaluation. The exceptions to this pattern were the PLCI based on $\hat{\tau}_1^{(U)}$, and the two types of bootstrap CIs for τ which showed acceptable performance and large percentage of samples in which were not computed. So, the results of these CIs should be taken with reserve.

6.2 Population constructed using data from the Colorado Springs study on HIV/AIDS transmission

In this simulation study we constructed a population using data from the Colorado Springs study on heterosexual transmission of HIV/AIDS. As was indicated in the introduction to this paper, this epidemiological research was focused on a population of people who lived in the Colorado Springs metropolitan area from $1982-1992$ and who were at high risk of acquiring and transmitting HIV. That population included drug users, sex workers and their personal contacts, defined as those persons with whom they had close social, sexual or drug-associated relations. In that study, 595 initial responders were selected in a non-random fashion through a sexually transmitted disease clinic, a drug clinic, self-referral and street outreach. The responders were asked for a complete enumeration of their personal contacts and a total of 7,379 contacts who were not in the set of the initial responders were named and included in the study. In our simulation study the set U_1 was defined as the set of the 595 initial responders and, as in Félix-Medina and Monjardin (2010), they were grouped into $N = 105$ clusters of sizes m_i's generated by sampling from a zero-truncated negative binomial distribution with parameter of size 2.5 and

probability $2/3$. The sample mean and variance of the 105 values m_i's were 5.67 and 15.03, respectively. It is worth noting that most of the people who were assigned to the same cluster came from the same original source of recruitment. A person was defined to be linked to a cluster if he or she was a personal contact of at least one element in that cluster. Since, approximately 95% of the 7,379 contacts of the initial responders were linked to only one cluster, and this could affect the performance of the proposed estimators, in our study we defined the set U_2 as the subset of the 7,379 contacts formed by the 415 persons who were linked to at least two clusters plus the 379 sex workers who were linked to only one cluster. Thus, $\tau_1 = 595, \tau_2 = 794$ and $\tau = 1{,}389$. It is worth noting that this population is the same as the one called "reduced population" by Félix-Medina and Monjardin (2010).

We set the sizes of the initial samples selected from the population to $n = 25$. This value of n yielded the sampling rates: $f_1 = 0.46$ and $f_2 = 0.37$. The simulation experiment was carried out as the previous one, except that each time that the value m_i was contained in an initial sample, all the people linked to cluster i were included in the sample. We used the same number of replications r and the same number B of bootstrap samples as those used in the previous study. In addition, the values of the parameters of the initial distributions that were used to construct the Bayesian-assisted estimators $\breve{\tau}_k$ and the value of q used to compute the Gaussian quadrature formulas (3.3) and (3.4) were the same as those used in the previous study.

The results of the simulation study are shown in Table 6.4. We can see that among the proposed estimators of the population sizes, only the estimators of τ_1 did not present problems of bias nor problems of instability. The estimators of τ_2 and τ exhibited serious problems of bias, particularly the estimators of τ_2, which affected their performance. As a result of the performance of the point estimators, only the adjusted PLCIs and bootstrap CIs for τ_1 performed acceptably well, although the values of the cp of the bootstrap CIs were slightly low. The unadjusted PLCIs for τ_1 showed low values of the cp because of the deviation from the assumed Poisson distribution of the M_i's. The PLCIs and bootstrap CIs for τ_2 and τ presented very large values of the mrl and mdrl that these intervals were not useful. Observe that the percentages of samples in which the proposed point and interval estimators of τ_1 were not computed because of numerical convergence problems were small (less than 1.2%). Therefore, they were virtually not favored by the evaluation procedure. In the case of the proposed point and intervals estimators of τ_2 and τ those percentages were large. However, if their performance was not good under this favorable assessment, it will not be good under a fairer evaluation.

With regard to the point estimators derived under the homogeneity assumption, we have that the MLEs $\tilde{\tau}_1$ and $\tilde{\tau}_2$ showed problems of bias which affected their performance; however, the estimator $\tilde{\tau}$ did not show problems of bias and its performance was acceptable. The small bias exhibited by this estimator might be explained by the fact that the negative bias of $\tilde{\tau}_1$ was canceled out by the positive bias of $\tilde{\tau}_2$. The Bayesian-assisted estimators performed similarly to the previous ones, although in this case the estimator $\breve{\tau}_2$ of τ_2 showed only mild problems of bias. The Wald CIs based on the MLEs and on the Bayesian-assisted estimators showed low values of the cp. However, since the values of the $\mathrm{r-ml}$ of these intervals were acceptable, the intervals for τ_2 and τ might provide some information about these parameters.

Table 6.4
Simulation results obtained for estimators and confidence intervals in a population constructed using data from the Colorado Springs study

		Point estimators				Confidence intervals			
Estimator		r − bias	$\sqrt{\text{r}-\text{mse}}$	mdre	mdare	Conf. interval	cp	mdre	mdare
Uncond. heter. MLEs	$\hat{\tau}_1^{(U)}$	$-0.00_1^{0.03}$	0.10	-0.01	0.07	PLCI- $\hat{\tau}_1^{(U)}$	0.75	0.24	0.24
						Adj-PLCI- $\hat{\tau}_1^{(U)}$	0.95	0.41	0.41
						PLCI- $\hat{\tau}_2^{(U)}$	$0.39^{8.3}$	10_{18}	3.7
	$\hat{\tau}_2^{(U)}$	$1.7_{16}^{3.5}$	4.5	0.79	0.79	PLCI- $\hat{\tau}^{(U)}$	$0.83^{8.6}$	5.7_6	2.1
						Adj-PLCI- $\hat{\tau}^{(U)}$	0.99^{21}	11_{47}	7.5
	$\hat{\tau}^{(U)}$	$0.95_{17}^{3.5}$	2.6	0.46	0.46	Bootstr-CI- $\hat{\tau}_1^{(U)}$	0.91	0.37	0.37
						Bootstr-CI- $\hat{\tau}_2^{(U)}$	$0.86^{3.2}$	11_{29}	3.6
						Bootstr-CI- $\hat{\tau}^{(U)}$	$0.88^{3.2}$	6.3_{28}	2.0
Cond. heter. MLEs	$\hat{\tau}_1^{(C)}$	$0.01^{0.57}$	0.12	0.01	0.08	PLCI- $\hat{\tau}_1^{(C)}$	$0.81^{1.2}$	0.30	0.27
						Adj-PLCI- $\hat{\tau}_1^{(C)}$	$0.97^{1.2}$	0.45	0.44
						PLCI- $\hat{\tau}_2^{(C)}$	0.39^{10}	9.6_{17}	3.6
	$\hat{\tau}_2^{(C)}$	$1.7_{10}^{4.7}$	4.5	0.80	0.80	PLCI- $\hat{\tau}^{(C)}$	0.89^{20}	6.2_9	2.6
						Adj-PLCI- $\hat{\tau}^{(C)}$	1.0^{27}	14_{32}	9.3
	$\hat{\tau}^{(C)}$	$0.96_{10}^{5.2}$	2.6	0.46	0.46	Bootstr-CI- $\hat{\tau}_1^{(C)}$	$0.86^{1.2}$	0.35	0.35
						Bootstr-CI- $\hat{\tau}_2^{(C)}$	$0.90^{8.0}$	9.7_{35}	3.9
						Bootstr-CI- $\hat{\tau}^{(C)}$	$0.91^{9.2}$	5.9_{34}	2.2
Homo-geneous MLEs	$\tilde{\tau}_1$	-0.22	0.23	-0.22	0.22	Wald-CI- $\tilde{\tau}_1$	0.06	0.16	0.16
	$\tilde{\tau}_2$	0.21	0.34	0.16	0.18	Wald-CI- $\tilde{\tau}_2$	0.71	0.60	0.53
	$\tilde{\tau}$	0.02	0.17	-0.00	0.10	Wald-CI- $\tilde{\tau}$	0.73	0.35	0.31
Homo-geneous BEs	$\bar{\tau}_1$	-0.22	0.23	-0.22	0.22	Wald-CI- $\bar{\tau}_1$	0.13	0.22	0.22
	$\bar{\tau}_2$	0.12	0.22	0.10	0.13	Wald-CI- $\bar{\tau}_2$	0.72	0.45	0.43
	$\bar{\tau}$	-0.02	0.13	-0.04	0.09	Wald-CI- $\bar{\tau}$	0.70	0.27	0.26

Notes A superscript number indicates the percentage of samples in which the estimator or confidence interval (CI) was not computed because of numerical convergence problems. A subscript figure indicates the number of values of the estimator or the relative length of a CI that exceeded 10^5 and that were not used to compute the $\text{r}-\text{bias}$ and $\sqrt{\text{r}-\text{mse}}$ of the estimator or the mrl of the CI. Upper bounds (UBs) for the Monte Carlo errors (MCEs) of the estimates of $\text{r}-\text{bias}$ and $\sqrt{\text{r}-\text{mse}}$ of the estimators of the τ's were the following: τ_1 : 0.001 and 0.001; τ_2 : 0.05 and 0.31, and τ : 0.03 and 0.18. An UB for the MCEs of the estimates of cp was 0.02. UBs for the MCEs of the estimates of the mrl were the following: PLCIs and adj. PLCIs for τ_1 : 0.003; PLCIs for τ_2 and τ : 0.59; adj PLCIs for τ : 0.93; Bootstrap CIs for τ_1, τ_2 and τ : 0.005, 1.5 and 0.88, respectively.

7 Conclusions and suggestions for future research

The results of the simulation studies carried out in this research indicate that the two proposed estimators of τ_1 : $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_1^{(C)}$, perform reasonably well in different situations. This evidences their robustness to several types of deviations from the assumed model. (Although $\hat{\tau}_1^{(C)}$ seems to be sensitive to deviations from the assumed normal distribution of the $\beta_j^{(k)}$'s.) On the other hand, the two proposed

estimators of τ_2 : $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$, present problems of bias and especially problems of instability if the sampling fraction in U_2 is not large enough, say it is not larger than 50%. In addition, small sampling fractions along with deviations from the assumed model for the link-probabilities increase the risk of numerical convergence problems. The two proposed estimators of τ : $\hat{\tau}^{(U)}$ and $\hat{\tau}^{(C)}$, perform similarly to the estimators $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_1^{(C)}$ if τ_1 is much greater than τ_2 (as in the case of the artificial populations), perform similarly to the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ if τ_2 is much greater than τ_1 (as in the case of the Colorado Springs population), and perform as a combination of the performance of the estimators of τ_1 and τ_2 if the values of these parameters are not very different from each other. Finally, the estimators derived under the assumption of homogeneous link-probabilities present serious problems of bias if this assumption is not satisfied.

It is worth noting that our conclusion about the proposed estimators of τ_2 is based on the results of several small simulation studies that we carried out using sampling fractions greater than those used in the Monte Carlo studies reported in this paper. In one study carried out with the artificial populations, we increased the values of the link-probabilities $p_{ij}^{(k)}$ so that their average values were $\bar{p}_{ij}^{(1)} \approx 0.088$ and $\bar{p}_{ij}^{(2)} \approx 0.071$ and kept the sizes of the initial samples at $n = 15$. These changes yielded the sampling fractions $f_1 \approx 0.65$ and $f_2 \approx 0.55$. In another study also with the artificial populations, we reduced the values of the $p_{ij}^{(k)}$ so that $\bar{p}_{ij}^{(1)} = \bar{p}_{ij}^{(2)} \approx 0.016$ and increased the sizes of the initial samples to $n = 78$ in Populations I – III and to $n = 67$ in Population IV. These changes yielded the sampling fractions $f_1 \approx 0.78$ and $f_2 \approx 0.55$. In both studies the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ performed acceptably well. (The results are not shown.) These outcomes indicate that these estimators also seem to have properties of robustness to deviations from the assumed models provided that large sampling fractions be used.

However, in a study with the Colorado Springs population using initial samples of sizes $n = 42$ which yielded sampling fractions $f_1 \approx 0.64$ and $f_2 \approx 0.56$, the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ presented serious problems of bias ($\text{r}-\text{bias} \approx 1.0$ and $\text{mdre} \approx 0.85$) which affected the values of their $\text{r}-\text{mse}$ $\left(\sqrt{\text{r}-\text{mse}} \approx 1.0\right)$ and mdare $(\text{mdare} \approx 0.85)$. Why these estimators did not perform well even with large sampling fractions? We think that the bad performance of these estimators is consequence of the very small average value of the $p_{ij}^{(2)}$'s $\left(\bar{p}_{ij}^{(2)} \approx 0.018\right)$ and the way the Monte Carlo studies were carried out. To clarify this statement, note the following. When $\bar{p}_{ij}^{(2)}$ is very small, say less than 0.02, the expected number of elements in U_2 that are linked to at least one site in the frame is much less than τ_2. For instance, in Population I when $\bar{p}_{ij}^{(2)} = 0.015$, this expected number was about $300 < 400 = \tau_2$. Therefore, if the sampling fraction f_2 is large enough, the estimates $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ will be close to τ_2 and will be much greater than the expected number of elements linked to at least one site in the frame. Thus, if we supposed that the Colorado Springs population was generated by a random process, then the 794 contacts linked to at least one site in the frame, and which we used as the value of τ_2, would be a much smaller value than the actual size of U_2. Consequently, the performance of $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ as estimators of the assumed value 794 of τ_2 would be very bad. This explanation was suggested and confirmed by the results of a small simulation study in which we considered Population I (in which every one of the

assumptions is satisfied), but instead of carrying the study as is described in Subsection 5.1, we generated the complete set of values $x_{ij}^{(2)}$ of $X_{ij}^{(2)}$ by sampling from the Bernoulli distributions with means $p_{ij}^{(2)}, i = 1,\ldots,N; j = 1,\ldots,400$ and we kept them fixed. Then, we defined the value of τ_2 as the number of elements of U_2 linked to at least one site in the frame. We considered two cases: large values of $p_{ij}^{(2)}$ $\left(\bar{p}_{ij}^{(2)} = 0.071\right)$ and small values $\left(\bar{p}_{ij}^{(2)} = 0.015\right)$. In the first case $\tau_2 = 388$, whereas in the second one $\tau_2 = 300$. To have comparable results the sizes of the initial samples were set to $n = 15$ in the first case and to $n = 78$ in the second case, so that in both cases the number of sampled elements were about 220. The results of the numerical study showed that in the case of large values of $p_{ij}^{(2)}$ the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ performed well (because $388 \approx 400$), whereas in the case of small values of $p_{ij}^{(2)}$ these estimators performed badly (because $300 << 400$). We think that the results obtained in the last case are illustrative and explain the ones obtained in the Colorado Springs population.

With respect to the two types of proposed CIs: profile likelihood and bootstrap CIs, we can conclude that they need larger sample sizes than the point estimators to perform reasonably well. They are more sensitive to deviations from the assumed models than the point estimators. In addition, if small sampling fractions are used and deviations from the assumed model for the $p_{ij}^{(k)}$ are present, the occurrence of numerical convergence problems will be greater than in the case of point estimators.

From the previous observations we can conclude that in actual applications of this sampling methodology, a good strategy is to construct a sampling frame that covers the largest possible portion of the target population. This way, τ_1 would be close to τ, and the estimates of τ_1 could be used as estimates of τ. The advantage of this strategy is that the estimators $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_1^{(C)}$ perform better than the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ because the first ones incorporate the information about the cluster sizes m_i. Furthermore, this strategy makes possible to use the design-based estimator $N\sum_1^n m_i/n$ as an estimator of τ. The other factor that must be taken into account to have good estimates is to use large sampling fractions, say larger than 0.5. This suggestion is in agreement with the result reported by Xi, Watson and Yip (2008), who in the context of capture-recapture studies indicate that in presence of heterogeneous capture probabilities, a population size between 300 and 500, and a number of sampling occasions between 10 and 20, the minimum sampling fraction to have reliable estimates is at least 60%. Since the estimation of τ_2 is basically the same problem as that of estimating the population size in a capture-recapture study, we think that this conclusion also applies to our situation. In this line, we have developed a method to determine the size of the initial sample in order to have desired values of $\sqrt{V(\hat{\tau}_k)/\tau_k^2}, k = 1, 2$ and $\sqrt{V(\hat{\tau})/\tau^2}$. Although this procedure is based on stringent assumptions such as the homogeneity of the effects $\alpha_i^{(k)}$'s associated with the sites and the necessity of large values of the m_i's, the results seem to be satisfactory. For instance, the situation illustrated at the end of Section 5 correspond to that of the artificial populations considered in the Monte Carlo study and we can see that the results obtained by our procedure are very close to those reported for Populations I and II (Table 6.2), where the estimators of τ_2 and τ performed acceptably.

Finally, despite the drawbacks of the proposed point and interval estimators, they are a better alternative for making inferences about the population size than those based on the assumption of homogeneous link-probabilities. Obviously, our proposal need to be improved. The two major problems that need to be considered in future research are the instability of the estimators of τ_2 when the sampling fraction is not large enough and the not satisfactory performance of the confidence intervals. A possible solution to both problems is to use the Bayesian approach to construct estimators that incorporate information prior to sampling that the researcher has about the parameters. The point and interval estimators obtained by this approach might be more stable than those proposed in this paper because of the additional information used to construct them. Other possible solution to the problem of lack of robustness of the confidence intervals is to replace the assumption of the Poisson distribution of the M_i's by a more flexible distribution such as the negative binomial, and the assumption of the normal distribution of the effects $\beta_j^{(k)}$ by one of the distributions ordinarily used to increase the robustness of the estimators such as a mixture of normal distributions or a Student's T distribution.

Acknowledgements

This research was supported by grant PROFAPI 2008/054 of the *Universidad Autónoma de Sinaloa* and grant APOY-COMPL-2008/89777 of the *Consejo Nacional de Ciencia y Tecnología*, Mexico. We thank John Potterat and Steve Muth for allowing us to use the data from the Colorado Springs study. We also thank the reviewers for their helpful suggestions and comments which improved this work.

References

Agresti, A. (2002). *Categorical Data Analysis, Second edition*. New York: John Wiley & Sons, Inc.

Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association,* 89, 1282-1289.

Cormack, R.M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics,* 48, 567-576.

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

Dávid, B., and Snijders, T.A.B. (2002). Estimating the size of the homeless population in Budapest, Hungary. *Quality & Quantity,* 36, 291-303.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. New York: Cambridge University Press.

Evans, M.A., Kim, H.-M. and O'Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological and Environmental Statistics*, 1, 131-140.

Félix-Medina, M.H., and Monjardin, P.E. (2006). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A bayesian-assisted approach. *Survey Methodology*, 32, 2, 187-195.

Félix-Medina, M.H., and Monjardin, P.E. (2010). Combining link-tracing sampling and cluster sampling to estimate totals and means of hidden human populations. *Journal of Official Statistics,* 26, 603-631.

Félix-Medina, M.H., and Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics,* 20, 19-38.

Frank, O., and Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics,* 10, 53-67.

Gimenes, O., Choquet, R., Lamor, R., Scofield, P., Fletcher, D., Lebreton, J.-D. and Pradel, R. (2005). Efficient profile-likelihood confidence intervals for capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics,* 10, 1-13.

Heckathorn, D.D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems,* 49, 11-34.

Johnston, L.G., and Sabin, K. (2010). Sampling hard-to-reach populations with respondent driven sampling. *Methodological Innovations Online,* 5, 2, 38-48.

Kalton, G. (2009). Methods for oversampling rare populations in social surveys. *Survey Methodology,* 35, 2, 125-141.

Karon, J.M., and Wejnert, C. (2012). Statistical methods for the analysis of time-location sampling data. *Journal of Urban Health,* 89, 565-586.

MacKellar, D., Valleroy, L., Karon, J., Lemp, G. and Janssen, R. (1996). The young men's survey: Methods for estimating HIV sero-prevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111, supplement 1, 138-144.

Magnani, R., Sabin, K., Saidel, T. and Heckathorn, D. (2005). Review of sampling hard-to-reach populations for HIV surveillance. *AIDS*, 19, S67-S72.

McKenzie, D.J., and Mistiaen, J. (2009). Surveying migrant households: a comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society, Series A*, 172, 339-360.

Munhib, F.B., Lin, L.S., Stueve, A., Miller, R.L., Ford, W.L., Johnson, W.D. and Smith, P. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports,* 116, supplement 1, 216-222.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56, 434-442.

Potterat, J.J., Woodhouse, D.E., Muth, S.Q., Rothenberg, R.B., Darrow, W.W., Klovdahl, A.S. and Muth, J.B. (2004). Network dynamism: History and lessons of the Colorado Springs study. In *Network Epidemiology: A Handbook for Survey Design and Data Collection*, (Ed., M. Morris), New York: Oxford University Press, 87-114.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. and Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratkowsky, D.A. (1988). *Handbook of Nonlinear Regression Models*. New York: Marcel Dekker.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks, Drug Abuse, and HIV Transmission*, (Eds., R.H. Needle, S.G. Genser and R.T. II Trotter) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

Semaan, S. (2010). Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodological Innovations Online*, 5, 2, 60-75.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.

Staudte, R.G., and Sheather, S.J. (1990). *Robust Estimation and Testing*. New York: John Wiley & Sons, Inc.

Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 1, 87-98.

Venzon, D.J., and Moolgavkar, S.H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37, 87-94.

Williams, B.K., Nichols, J.D. and Conroy, M.J. (2002). *Analysis and Management of Animal Populations*. San Diego, California: Academic Press.

Xi, L., Watson, R. and Yip, P.S.F. (2008). The minimum capture proportion for reliable estimation in capture-recapture models. *Biometrics*, 64, 242-249.