Combining Cluster Sampling and Link-Tracing Sampling to Estimate Totals and Means of Hidden Populations in Presence of Heterogeneous Probabilities of Links

Martín Humberto Félix-Medina

We propose Horvitz-Thompson-like and Hájek-like estimators of the total and mean of a response variable associated with the elements of a hard-to-reach population, such as drug users and sex workers. A portion of the population is assumed to be covered by a frame of venues where the members of the population tend to gather. An initial cluster sample of elements is selected from the frame, where the clusters are the venues, and the elements in the sample are asked to name their contacts who belong to the population. The sample size is increased by including in the sample the named elements who are not in the initial sample. The proposed estimators do not use design-based inclusion probabilities, but model-based inclusion probabilities which are derived from a Rasch model and are estimated by maximum likelihood estimators. The inclusion probabilities are assumed to be heterogeneous, that is, they depend on the sampled people. Variance estimates are obtained by bootstrap and are used to construct confidence intervals. The performance of the proposed estimators and confidence intervals is evaluated by two numerical studies, one of them based on real data, and the results show that their performance is acceptable.

Key words: Capture-recapture; Hájek estimator; Horvitz-Thompson estimator; maximum likelihood estimator; snowball sampling.

1. Introduction

The problem of selecting samples from hidden or hard-to-detect populations, such as drug users, sex workers and homeless people, that allow reasonably valid statistical inferences is challenging because of the following six factors (1) lack of appropriate sampling frames for those populations; (2) rareness of those populations; (3) elusiveness of their members to be sampled; (4) difficulty in identifying their elements due to a stigmatized or illegal behavior; (5) difficulty in locating their members, and (6) difficulty in persuading their elements to participate in the study, among others. See Tourangean (2014) for a detailed discussion about these and other issues. Because of these factors, conventional sampling methods are not appropriate for this type of population, and consequently several specially tailored sampling methods that take into account the particular characteristics of those populations have been proposed. Among these methods we can mention multiplicity sampling, venue-based sampling, link-tracing sampling and capture-recapture sampling. (For descriptions of these methods, see Spreen 1992; Magnani et al. 2005; Kalton 2009;}

1 Facultad de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Av. de las Américas y Universitarios, Culiacán, Sinaloa 80013, México. Email: mhfelix@uas.edu.mx

Acknowledgments: This research was supported by Grant PROFAPI-2011/057 of Universidad Autónoma de Sinaloa.
Marpsat and Razafindratsima 2010; UNAIDS/WHO 2010; Lee et al. 2014; Spreen and Bogaerts 2015; Heckathorn and Cameron 2017). In addition, special estimation methods based on conventional samples, such as the scale up method (Killworth et al. 1998a, 1998b; Bernard et al. 2010; McCormick et al. 2010; Maltiel et al. 2015) or a combination of some of the previously mentioned sampling designs, such as the multiplier method (UNAIDS/WHO 2010; Johnston et al. 2013; Men and Gustafson 2017) and the one step network based method (Dombrowski et al. 2012; Khan et al. 2018) have been proposed.

It is worth noting that most of the recently published research papers on sampling from hidden populations focus on estimating the population size. (See Cheng et al. 2020 for a review and analysis, from an asymptotic approach, of most of these methods.) Thus, the scale up, multiplier and one step network methods were developed with this goal in mind. The interest in developing methods for estimating the size of a hidden population is mainly because information about this parameter allows the design of appropriate plans to address the problems associated with this type of population. However, information about other population parameters, such as average monthly spending on drugs and average age at which consumption begins in a population of drug addicts, and average weekly number of clients and average weekly income in a population of sex workers, is also important. This is because this type of information increases our knowledge about the population, and in addition, this knowledge could be used to improve the plans for its care that are based only on its population size.

On the other hand, among the sampling designs for hidden populations that allow estimating parameters different from or in addition to the population size, we have venue-based sampling and link-tracing sampling. Venue-based sampling (MacKellar et al. 1996) is a probability sampling method specifically developed to estimate the means of variables of interest, and particularly proportions. The method consists in carrying out an ethnographic study to construct a sampling frame of venues where the members of the population tend to gather. Venues are not only sites, such as bars, parks and street locations, but could be combinations of sites, days of the week and time segments. For instance, a venue could be a specific bar from 4:00 p.m. to 12:00 a.m. on Fridays and Saturdays, whereas another venue could be the same bar, over the same time segment, but from Monday to Thursday. Furthermore, some venues could be events such as gay parades. A probability sample of venues is selected, and from each chosen venue a sort of systematic sample of members of the population who are present at the venue is selected. For each sampled element the values of the variables of interest associated with that person are recorded, and in addition, information about his or her attendance to the venues in the sampling frame is obtained so that his or her inclusion probability can be estimated, and consequently, suitable unbiased weighted estimators of the means can be computed. It is evident that the estimates obtained by using this sampling design are valid only for the portion of the population that attends the venues in the frame. Therefore, the extension of the results to the entire population requires the assumption that with respect to the distributions of the variables of interest there are no differences between the members who visit the places in the frame and those who do not.

Link-tracing sampling (LTS) is an umbrella term that encompasses a set of sampling designs in which an initial sample of elements from the target population is selected and every sampled person is asked to name his or her contacts (defined according to a certain
criterion) who are also members of the target population. The elements in the initial sample form wave zero, and the named elements who are not in the initial sample form wave one. People in wave one might also be asked to name their contacts. The named elements who have not been previously sampled form wave two. The sampling procedure might continue in this way until a stopping rule is satisfied. For instance, a specified number of waves or a specified sample size. Several variants of LTS have been proposed. For example, Klovdahl (1989) developed a variant, known as random walk. Heckathorn (1997, 2002) proposed a variant of LTS, called respondent driven sampling (RDS) to estimate proportions of some subpopulations of the population of interest. It is worth noting that in subsequent works, such as Volz and Heckathorn (2008), Handcock et al. (2014) and Crawford et al. (2018) estimators of several population parameters have been proposed to be used along with this sampling design. Thompson and Frank (2000), Chow and Thompson (2003) and St. Clair and O’Connell (2012) have also proposed estimators of population parameters to be used with a pretty general variant of LTS. Finally, a variant of LTS related to the one considered in this work is the one proposed by Frank and Snijders (1994). In this variant, the initial sample is assumed to be selected by Bernoulli sampling, (Särndal et al. 1992, subsec. 3.2) that is, every element of the population has the same probability of being included in the initial sample and the inclusions are independent. Furthermore, those authors assumed that the probability that a specific element of the population be named as a contact by a particular person in the initial sample, which in this document is called link probability, is a constant, that is, it does not depend on the named person nor on the person who names. This supposition is known as the homogeneity assumption.

In this article, we consider the problem of estimating the total and the mean of a variable of interest, such as the weekly drug expense of a drug user, the number of weekly clients of a sex worker, and an indicator (positive = 1/negative = 0) of a person’s drug use, from a sample selected by the LTS variant proposed by Félix-Medina and Thompson (2004). This sampling variant was devised to avoid the assumption of an initial Bernoulli sample required by the sampling variant proposed by Frank and Snijders (1994), which is difficult to satisfy in real-world applications. To achieve this goal, those authors proposed that a sampling frame of venues where the members of the population tend to gather be constructed, as in venue-based sampling. Those authors made the assumptions that the frame does not cover the whole population and that an element in the portion covered by the frame is assigned to only one venue. The last assumption can be achieved by means of a specified criterion, for instance, the venue where the person spends most of his or her time. Then, a simple random sample without replacement (SRSWOR) of venues is selected and every one of the members of the population who belongs to each sampled venue is sampled and interviewed. Notice that the assumption that each element in the frame is assigned to only one venue implies that if a person were found in a sampled place where he or she does not meet the criterion to be assigned to, that person should not be included in the initial sample. Next, from each sampled venue its elements are asked to name their contacts who are also members of the population, either they belong or not to the portion covered by the frame. Félix-Medina and Thompson (2004) proposed maximum likelihood estimators (MLEs) of the population size which were derived under the assumption that the probability that a person is linked to a sampled venue, that is, that he or she is a contact of an element in that venue, depends on the venue, but not on the named
person. This means that the estimators were derived under the assumption of homogeneous link probabilities. This assumption might be unrealistic in most actual populations because people with many social relationships may be more likely to be linked to a given venue than those with few social relationships. In order to avoid the homogeneity assumption, Félix-Medina et al. (2015) derived MLEs of the population size under the assumption that the link probabilities also depend on the named persons, that is, that the probabilities are heterogeneous. In their work, the authors showed by means of a Monte Carlo study, that if the assumption of homogeneous link probabilities is not satisfied, the estimators derived under that assumption are negatively biased. This result agrees with those reported in capture-recapture studies (see e.g., Burnham and Overton 1978; Hwang and Huggins 2005) and it should be expected as the estimators that have been proposed to be used with this LTS variant are similar to those used in capture-recapture studies. Since the LTS variant proposed by Félix-Medina and Thompson (2004) has not been used in any study with a real population, there are no results about the effect of the heterogeneity of the link probabilities on the estimators derived under the homogeneity assumption based on real populations. However, Dávid and Snijders (2002) used the variant of Frank and Snijders (1994) to estimate the number of homeless in Hungary and obtained an underestimation. They attributed this result to the failure of the initial Bernoulli sample assumption. However, it is possible that the estimate was also affected by not satisfying the assumption of homogeneity of the link probabilities on which the estimator they used is based.

In this article, we used the model proposed by Félix-Medina et al. (2015) to construct model-based Horvitz-Thompson-like estimators (HTLEs) and model-based Hájek-like estimators (HKLEs) of the total and the mean of a variable of interest. It should be noted that Félix-Medina and Monjardin (2010) also considered the problem addressed in this article, but they proposed estimators of the total and the mean derived under the assumption of homogeneous link probabilities. Thus, our work is an extension of theirs. The structure of this article is as follows. In Section 2, we introduce the LTS variant proposed by Félix-Medina and Thompson (2004), as well as the notation to be used throughout this article. In Section 3, we present the models and the MLEs of the population sizes proposed by Félix-Medina et al. (2015). In Section 4, we develop the strategy to construct the proposed model-based HTLEs and HKLEs of the total and the mean. In Section 5, we describe the construction of confidence intervals for the total and mean based on the proposed HTLEs and HKLEs of these parameters and on estimates of the standard deviations of these estimators obtained by a variant of bootstrap proposed in this article. In Section 6, we present the results of two numerical studies carried out to observe the performance of the proposed estimators and confidence intervals and to compare their performance with that of the proposed by Félix-Medina and Monjardin (2010). In Section 7, we state some conclusions and suggestions for future research. Finally, in the Appendix (Section 8) we described the technical aspects of the proposed bootstrap procedure.

2. Sampling Design and Notation

In this article, we consider the variant of LTS proposed by Félix-Medina and Thompson (2004) which we will describe next. Let $U$ be a finite population of an unknown number $\tau$
of people. A portion $U_1$ of $U$ is assumed to be covered by a sampling frame of $N$ venues $A_1, \ldots, A_N$, where the members of the population can be found with high probability. As in ordinary cluster sampling, each person in $U_1$ is assumed that can be assigned, by means of a specified criterion, to only one venue in the frame, for instance, the venue where he or she spends most of his or her time. Let $m_i$ denote the number of members of the population that belong, that is, that are assigned to the venue $A_i, i = 1, \ldots, N$. 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 first step of the sampling design is to select a SRSWOR $S_A$ of $n$ venues $A_1, \ldots, A_n$ from the frame. The $m_i$ members of the population who belong to the sampled venue $A_i$ are identified and their associated $y$-values of a variable of interest $y$ are recorded, $i = 1, \ldots, n$. Notice that the assumption that a person in $U_1$ is assigned to only one venue by a specified criterion implies that if a person were found in a sampled venue where he or she does not meet the criterion to be assigned, that person should not be included in the initial sample. Let $S_0$ be the set of people in the initial sample. Notice that the size of $S_0$ is $m = \sum_1^n m_i$. The second step is to ask the people in each sampled venue to name other members of the population. We will say that a person and a venue are linked if any of the people who belong to that venue names him or her. Let $x_{ij}^{(k)} = 1$ if person $j \in U_k - A_i$ is linked to venue $A_i \in S_A$ and $x_{ij}^{(k)} = 0$ if $j \in A_i$ or $j$ is not linked to $A_i, i = 1, \ldots, n; k = 1, 2$. For each named person, the following information is recorded: the value of the variable of interest $y$ associated with him or her, the sampled venues that are linked to him or her, and the subset of $U$: $U_1 - S_0$, a specific $A_i \in S_A$ or $U_2$, that contains him or her. Let $S_1$ be the set of people in $U_1 - S_0$ who are linked to at least one venue in $S_A$, and let $S_2$ be the set of people in $U_2$ who are linked to at least one venue in $S_A$. We will denote by $r_k$ the size of $S_k$, $k = 1, 2$. Finally, let $S_1^* = S_0 \cup S_1$ and $S_2^* = S_2$ be the sets of the sampled people from $U_1$ and $U_2$, respectively. Notice that the respective sizes of these sets are $m + r_1$ and $r_2$.

3. Maximum Likelihood Estimators of the Population Sizes

Félix-Medina et al. (2015) proposed MLEs of the population sizes $\tau_1, \tau_2$ and $\tau$, which were derived from the following assumptions. The values $m_1, \ldots, m_N$ are considered as realizations of the random variables $M_1, \ldots, M_N$, which are supposed to be independent and identically distributed Poisson random variables with mean $\lambda_1$. This implies that the joint conditional distribution of the vector of variables $M = (M_1, \ldots, M_n, \tau_1 - M)$, where $M = \sum_1^n M_i$ given that $\sum_1^n M_i = \tau_1$, is multinomial with parameter of size $\tau_1$ and vector of probabilities $(1/N, \ldots, 1/N, 1 - n/N)$. The assumption that the $M_i$s are independent and identically distributed Poisson random variables is not as restrictive as it seems at first glance. This assumption contributes to the likelihood function with the previously indicated multinomial distribution through the term \[\frac{\tau_1!}{(\tau_1 - m)!}(1 - n/N)^{\tau_1},\] which depends on the $M_i$s by means of the value $m$ of $M = \sum_1^n M_i$. Since $NM/n$ is a design-based estimator of $\tau_1$ (based only on the information contained in the initial sample), it follows that the Poisson assumption does not weaken the robustness of the maximum likelihood estimators. However, it does affect the variability of the estimators. For this reason, variance estimators and confidence intervals must be constructed without taking into account this assumption.
The values $x_{ij}^{(k)}$'s are assumed to be realizations of the random variables $X_{ij}^{(k)}$, which given the sample $S_A$ of venues, are supposed to be independent Bernoulli random variables with means $p_{ij}^{(k)}$, where the means or link probabilities $p_{ij}^{(k)}$'s are given by the following Rasch model:

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

As is indicated in Félix-Medina et al. (2015), this model was considered by Coull and Agresti (1999) in the context of multiple capture-recapture sampling. The parameter $\alpha(ji)$ is a fixed (not random) effect that represents the potential that the venue $A_i$ has of forming links with people in $U_k - A_i$, and $\beta_{(k)}$ is a random effect that represents the propensity of the person $j \in U_k$ to be linked to a sampled venue. Those authors suppose that $\beta_{(k)}$ is normally distributed with mean 0 and unknown variance $\sigma_k^2$, and that these variables are independent. The parameter $\sigma_k^2$ determines the degree of heterogeneity of the $p_{ij}^{(k)}$: great values of $\sigma_k^2$ imply high degrees of heterogeneity.

Henceforth, all probability statements will be conditioned on the sample $S_A$ of venues unless otherwise is specified. Let $X_j^{(k)} = (X_{ij}^{(k)}, \ldots, X_{nj}^{(k)})$ be the $n$-dimensional vector of link indicator variables $X_{ij}^{(k)}$'s associated with the $j$-th person in $U_k - S_0$, and let $\Omega = \{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_i = 0 \text{ or } x_i = 1; i = 1, \ldots, n \}$. Then, the probability that $X_j^{(k)}$ equals $x = (x_1, \ldots, x_n) \in \Omega$, that is, the probability that the $j$-th person in $U_k - S_0$ is linked to only the venues $A_i \in S_A$ such that the $i$-th element $x_i$ of $x$ equals 1, is

$$\Pr \left( X_j^{(k)} = x | \alpha_k, \beta_{(k)} \right) = \prod_{i=1}^n \left[ p_{ij}^{(k)} \right]^{x_i} \left[ 1 - p_{ij}^{(k)} \right]^{1-x_i} = \prod_{i=1}^n \frac{\exp [x_i (\alpha(ji) + \beta_{(k)})]}{1 + \exp (\alpha(ji) + \beta_{(k)})},$$

where $\alpha_k = (\alpha(1), \ldots, \alpha(m))$. Therefore, the probability that the vector of link indicator variables associated with a randomly selected person in $U_k - S_0$ equals $x$ is

$$\pi_{(k)x} = \pi_{(k|x)}(\alpha_k, \sigma_k) = \int \prod_{i=1}^n \frac{\exp [x_i (\alpha(ji) + \sigma_k z_i)]}{1 + \exp (\alpha(ji) + \sigma_k z_i)} \phi(z)dz,$$

where $\phi(\cdot)$ denotes the probability density function of the standard normal distribution $[N(0, 1)]$.

Félix-Medina et al. (2015), following Coull and Agresti (1999), used the Gaussian quadrature method to obtain the following approximation to $\pi_{(k)x}$:

$$\hat{\pi}_{(k)x} = \hat{\pi}_{(k|x)}(\alpha_k, \sigma_k) = \sum_{t=1}^q \prod_{i=1}^n \frac{\exp [x_i (\alpha(ji) + \sigma_k z_{it})]}{1 + \exp (\alpha(ji) + \sigma_k z_{it})} v_t,$$ (2)

where $q$ is a fixed constant and $\{ z_t \}$ and $\{ v_t \}$ are obtained from tables (see Table 25.5 in Abramowitz and Stegun 1964) or statistical software (see R library statmod developed by Giner and Smyth 2016).

Similarly, for person $j$ in $A_i \subseteq S_A$, let $X_j^{(A_i)} = (X_{ij}^{(A_i)}, \ldots, X_{i+1j}^{(A_i)}, X_{i+1j+1}^{(A_i)}, \ldots, X_{nj}^{(A_i)})$ be the $(n - 1)$-dimensional vector of link indicator variables $X_{ij}^{(A_i)}$'s associated with that person, and let $\Omega_{-i} = \{ x = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in [\mathbb{R}^{n-1}] : x_j = 0 \text{ or } x_j = 1; j = 1, \ldots, n; j \neq i \}$. Then, the probability that $X_j^{(A_i)}$ equals $x = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \Omega_{-i}$,
that is, the probability that the \(j\)-th person in \(A_i\) is linked to only the venues \(A_{i'} \in S_A, i' \neq i\), such that the \(i'\)-th element \(x_{i'}\) of \(x\) equals 1, is

\[
\Pr \left( X_j^{(A_i)} = x | \alpha_1, \beta_{(1)n} \right) = \prod_{i' \neq 1} \left[ p_{i'}^{(1)} \right]^{x_{i'}} \left[ 1 - p_{i'}^{(1)} \right]^{1-x_{i'}} = \prod_{i' \neq 1} \exp \left[ x_{i'} (\alpha_{(1)i'} + \beta_{(1)j}) \right] \left/ \left[ 1 + \exp (\alpha_{(1)i'} + \beta_{(1)j}) \right) \right.
\]

and the Gaussian quadrature approximation to the probability \(\pi_{(A_i)x} = \pi_{(A_i)x}(\alpha_1, \sigma_1)\) that the vector of link indicator variables associated with a randomly selected person from the sampled venue \(A_i\) equals the \((n-1)\)-dimensional vector \(x = (x_1, \ldots, x_i, \ldots, x_n)\) is

\[
\hat{\pi}_{(A_i)x} = \hat{\pi}_{(A_i)x}(\alpha_1, \sigma_1) = \sum_{i=1}^n \prod_{i' \neq 1} \exp \left[ x_{i'} (\alpha_{(1)i'} + \sigma_1 z_{i'}) \right] \left/ \left[ 1 + \exp (\alpha_{(1)i'} + \sigma_1 z_{i'}) \right) \right.
\]

Under the previous assumptions, Félix-Medina et al. (2015) constructed the likelihood function of \(\tau_k, \alpha_k\) and \(\sigma_k, k = 1, 2\), which is proportional to a product of several multinomial distributions. One multinomial distribution is the conditional distribution of the vector of variables \(M_x = (M_1, \ldots, M_n, \tau_1 - M)\), given that \(\sum_{i=1}^n M_i = \tau_1\), and that was indicated at the beginning of this section. Other two multinomial distributions are obtained by considering each one of the \(2^n\) vectors \(x \in \Omega\) as the label of a cell of a contingency table. Then, the vectors of cell-frequencies \(R_x^{(1)} \in \Omega\) and \(R_x^{(2)} \in \Omega\), where \(R_x^{(1)}\) and \(R_x^{(2)}\) denote the numbers of people in \(U_1 - S_0\) and in \(U_2\) whose vectors \(X^{(1)}\) and \(X^{(2)}\) of link indicator variables are equal to \(x \in \Omega\), have joint multinomial distributions with parameters of size \(\tau_1 - m\) and \(\tau_2\), and vectors of probabilities \((\pi_{(1)x})_x \in \Omega\) and \((\pi_{(2)x})_x \in \Omega\), respectively. The last \(n\) multinomial distributions, one for each \(A_i \in S_A\), are obtained as the previous two. Thus, for each \(i = 1, \ldots, n\), it follows that the vector of variables \(R_x^{(A_i)} \in \Omega\), where \(R_x^{(A_i)}\) denotes the number of people in \(A_i\) whose vectors of link indicator variables \(X^{(A_i)}\) are equal to the vector \(x \in \Omega\), has a multinomial distribution with parameter of size \(m_i\) and vector of probabilities \((\pi_{(A_i)x})_x \in \Omega\).

Those authors proposed maximum likelihood estimators of \(\tau_k, \alpha_k\) and \(\sigma_k, k = 1, 2\), whose values are obtained by numerically maximizing the likelihood function expressed in terms of the Gaussian quadrature approximations \(\hat{\pi}_{(k)x}\) and \(\hat{\pi}_{(A_i)x}\) to the probabilities \(\pi_{(k)x}\) and \(\pi_{(A_i)x}\). They called these estimators unconditional maximum likelihood estimators (UMLE) and denoted them as \(\hat{\tau}_1^{(U)}\), \(\hat{\alpha}_1^{(U)}\) and \(\hat{\sigma}_1^{(U)}\), \(k = 1, 2\). Although these estimators do not have closed forms, the authors provided the following asymptotic approximations for \(\hat{\tau}_1^{(U)}\) and \(\hat{\tau}_2^{(U)}\):

\[
\hat{\tau}_1^{(U)} = \frac{M + R_1}{1 - (1 - n/N)\pi_{(1)0} (\hat{\alpha}_1^{(U)}, \hat{\sigma}_1^{(U)})} \quad \text{and} \quad \hat{\tau}_2^{(U)} = \frac{R_2}{1 - \pi_{(2)0} (\hat{\sigma}_2^{(U)}, \hat{\sigma}_2^{(U)})},
\]

where \(R_1\) and \(R_2\) denote the numbers of distinct people in \(U_1 - S_0\) and \(U_2\), respectively, that are linked to at least one venue in \(S_A\). Notice that these are not close forms because \(\hat{\alpha}_k^{(U)}\) and \(\hat{\sigma}_k^{(U)}\) depend on \(\hat{\tau}_k^{(U)}\). Once \(\hat{\tau}_1^{(U)}\) and \(\hat{\tau}_2^{(U)}\) are obtained, the UMLE of \(\tau\) is \(\hat{\tau}^{(U)} = \hat{\tau}_1^{(U)} + \hat{\tau}_2^{(U)}\).

Also, Félix-Medina et al. (2015), following Coull and Agresti (1999), used Sanathanan’s (1972) approach to derive conditional MLEs \(\hat{\alpha}_k^{(C)}\) and \(\hat{\sigma}_k^{(C)}\) of \(\alpha_k\) and \(\sigma_k\), given \(R_k, k = 1, 2\). The conditional likelihood function of these parameters is obtained by considering the conditional distribution of the vector \((R_x^{(k)})_x \in \Omega\) given that \(R_k = r_k\),
which is multinomial with parameter of size $r_k$ and vector of probabilities $(\pi_{(k|x)}^{(c)})_{x \in \Omega - \{0\}}$, where $\pi_{(k|x)}^{(c)} = \pi_{(k|x)}/[1 - \pi_{(k|x)}]$, $k = 1, 2$. The product of these two distributions and the $n$ multinomial distributions of the vectors $(R_x^{(A)})_{x \in \Omega - \{0\}}$, $i = 1, \ldots, n$, forms the conditional likelihood function. The values of the estimators $\hat{\alpha}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$ are obtained by maximizing numerically this function expressed in terms of the Gaussian quadrature approximations to the probabilities $\pi_{(k|x)}$ and $\pi_{(A|x)}$. The values of the conditional estimators $\hat{C}_1$ and $\hat{C}_2$ of $\tau_1$ and $\tau_2$ are obtained by using the fact that the respective distributions of $R_1$ and $R_2$ are binomial with parameters of size $\tau_1 - m$ and $\tau_2$ and probabilities $1 - \pi_{(k|x)}$, $k = 1, 2$. The product of these two distributions and the conditional multinomial distribution of $M_k$ forms the likelihood function of $\tau_1$ and $\tau_2$. The maximization of this function, after replacing the values of $\alpha_k$ and $\sigma_k$ by their estimates $\hat{\alpha}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$, yields the conditional MLEs $\hat{C}_1$ and $\hat{C}_2$, which are given by Equation (4), but replacing $\hat{\alpha}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ by $\hat{\alpha}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$. Note that in this case the expressions (4) are closed forms. The conditional MLE of $\tau$ is $\hat{\tau}^{(C)} = \hat{C}_1^{(C)} + \hat{C}_2^{(C)}$.

It is worth noting that for each $k = 1, 2$, the assumption that the link probabilities follow the Rasch model (1) could be tested by a Pearson chi-square statistic. One possibility is to use the previously indicated conditional multinomial distribution of $(R_x^{(k)})_{x \in \Omega - \{0\}}$ given that $R_k = r_k$. In this case, the Pearson chi-square statistic to test the hypothesis that the link probabilities associated with the elements in $U_k - S_0$ follow the Rasch model (1) is

$$X_k^2 = r_k \sum_{x \in \Omega - \{0\}} \left[ \frac{R_x^{(k)}}{r_k} - \frac{\hat{\pi}_x^{(c)}}{\pi_{(k|x)}} \right]^2 / \pi_{(k|x)}, k = 1, 2,$$

where $\hat{\pi}_x^{(c)} = \pi_x(\hat{\alpha}_k^{(a)}, \hat{\sigma}_k^{(a)})/[1 - \pi_x(\hat{\alpha}_k^{(a)}, \hat{\sigma}_k^{(a)})]$, and $a = U$ or $a = C$ depending on whether UMLEs or CMLEs are used. Since the contingency tables on which these statistics are based are sparse, their distributions are not well approximated by chi-square distributions. One alternative to compute the $p$-values of these tests is by bootstrap. See Reiser (2019) for this and other alternatives. Specific assumptions about the Rasch model (1) such as the non-interaction between the link indicator variables that form the vector $X^{(k)}$ could be tested using Rasch models that take into account these interactions. See Fienberg et al. (1999) for more general Rasch models than the one used in this article.

4. Estimators of the Total and Mean

In this section, we will focus on the problem of estimating the total and the mean of the values of a variable of interest $Y$. Let $y_j^{(k)}$ be the value of $y$ associated with the $j$-th element of $U_k$, $j = 1, \ldots, \tau_k$, $k = 1, 2$. In this work we will suppose that the $y$-values are fixed numbers and not random variables. Note that this assumption is the one made in traditional sampling, in which inferences are made from the design-based approach. (See Thompson 2012, 2) Then $Y_k = \sum_{j \in U_k} y_j^{(k)}$ and $Y_k / \tau_k$ represent the total and the mean of the portion $U_k$, $k = 1, 2$, of the population. Similarly, $Y = Y_1 + Y_2$ and $Y / \tau = Y / \tau_2$ represent the total and the mean of the whole population $U$.

We cannot compute the design-based inclusion probabilities of the sampled elements because we do not know the venues in the frame that are linked to each sampled person. However, we can compute conditional model-based inclusion probabilities given the venues $A_i \in S_A$. These probabilities are given by
\[ \pi_{(1)j}(\alpha_1, \sigma_1, \beta_{(1)j}) = 1 - (1 - n/N) \prod_{i=1}^{n} (1 - p_{ij}^{(1)}) \text{ if } j \in U_1, \text{ and} \]

\[ \pi_{(2)j}(\alpha_2, \sigma_2, \beta_{(2)j}) = 1 - \prod_{i=1}^{n} (1 - p_{ij}^{(2)}) \text{ if } j \in U_2. \]

We do not know the probabilities \( \pi_{(k)j}(\alpha_k, \sigma_k, \beta_{(k)j}) \) because they depend on unknown parameters. However, we could estimate them by estimating those parameters and replacing the parameters in Equations (5) and (6) by their estimates. The computation of both UMLES and CMLES of \( \alpha_k \) and \( \sigma_k \) was described in the previous section. See also Félix-Medina et al. (2015). Next, we will derive a predictor of the random effect \( \beta_{(k)j} \).

Thus, if the element \( j \in U_k - S_j, k = 1, 2 \), then the conditional joint probability density function of the vector \( X_{j}^{(k)} \) of link indicator variables associated with that element and the random effect \( \beta_{(k)j} \) is

\[ f(x_j^{(k)}, \beta_{(k)j}|\alpha_k, \sigma_k) = \text{Pr}(X_j^{(k)} = x_j^{(k)}|\beta_{(k)j}, \alpha_k)f(\beta_{(k)j}|\sigma_k) \]

\[ \propto \prod_{i=1}^{n} [p_{ij}^{(k)}]^{x_{ij}} [1 - p_{ij}^{(k)}]^{1-x_{ij}} \exp \left[ -\frac{(\beta_{(k)j})^2}{2\sigma_k^2} \right], \]

whereas, if the element \( j \in A_{i'}, i' = 1, \ldots, n \), then

\[ f(x_j^{(A_{i'})}, \beta_{(1)j}|\alpha_1, \sigma_1) \propto \prod_{i=1}^{n} [p_{ij}^{(1)}]^{y_{ij}} [1 - p_{ij}^{(1)}]^{1-y_{ij}} \exp \left[ -\frac{(\beta_{(1)j})^2}{2\sigma_1^2} \right]. \]

We will propose as a predictor of \( \beta_{(k)j} \) the conditional expected value of \( \beta_{(k)j} \) given \( X_j^{(k)} = x_j^{(k)} \), evaluated either at the UMLES \( \hat{\alpha}_k^{(U)} \) and \( \hat{\sigma}_k^{(U)} \) or at the CMLES \( \hat{\alpha}_k^{(C)} \) and \( \hat{\sigma}_k^{(C)} \), that is

\[ z_{(a)}^{(a)} = E(\beta_{(k)j}|x_j^{(k)}, \hat{\alpha}_k^{(a)}, \hat{\sigma}_k^{(a)}) = \frac{\int \beta_{(k)j} f(x_j^{(k)}, \beta_{(k)j}|\hat{\alpha}_k^{(a)}, \hat{\sigma}_k^{(a)}) d\beta_{(k)j}}{\int f(x_j^{(k)}, \beta_{(k)j}|\hat{\alpha}_k^{(a)}, \hat{\sigma}_k^{(a)}) d\beta_{(k)j}}, \quad a = U, C. \]

We will approximate \( \hat{z}_{(a)}^{(a)} \) by using the Gaussian quadrature method, that is by

\[ \hat{z}_{(a)}^{(a)} = \sum_{i=1}^{q} z_i \prod_{i=1}^{n} \left\{ \exp \left[ x_i(\hat{\alpha}_{(k)}^{(a)} + \hat{\sigma}_{(k)}^{(a)}z_i) \right] / \left[ 1 + \exp (\hat{\alpha}_{(k)}^{(a)} + \hat{\sigma}_{(k)}^{(a)}z_i) \right] \right\} v_i \]

\[ \sum_{i=1}^{q} \prod_{i=1}^{n} \left\{ \exp \left[ x_i(\hat{\alpha}_{(k)}^{(a)} + \hat{\sigma}_{(k)}^{(a)}z_i) \right] / \left[ 1 + \exp (\hat{\alpha}_{(k)}^{(a)} + \hat{\sigma}_{(k)}^{(a)}z_i) \right] \right\} v_i \]

\[ = \frac{\hat{\alpha}_k^{(a)} \sum_{i=1}^{q} z_i \sum_{i=1}^{n} x_i / \prod_{i=1}^{n} \left[ 1 + \exp (\hat{\alpha}_{(k)}^{(a)} + \hat{\sigma}_{(k)}^{(a)}z_i) \right] \right\} v_i }{\sum_{i=1}^{q} \left\{ \exp (\hat{\sigma}_{(k)}^{(a)}z_i \sum_{i=1}^{n} x_i / \prod_{i=1}^{n} \left[ 1 + \exp (\hat{\alpha}_{(k)}^{(a)} + \hat{\sigma}_{(k)}^{(a)}z_i) \right] \right\} v_i }, \quad a = U, C, \]
if \( j \in U_k - S_0 \), \( k = 1, 2 \), and by

\[
\hat{\beta}_{(1)y}^{(a)} = \frac{\hat{\sigma}_{1}^{(a)} \sum_{t=1}^{q} z_{i} \left\{ \exp \left( \hat{\sigma}_{1}^{(a)} z_{i} \sum_{j \neq t}^{n} x_{j} / \Pi_{i \neq j}^{n} \left[ 1 + \exp (\hat{\sigma}_{1}^{(a)} z_{i}) \right] \right) \right\} v_{t} }{\sum_{t=1}^{q} \left\{ \exp \left( \hat{\sigma}_{1}^{(a)} z_{i} \sum_{j \neq t}^{n} x_{j} / \Pi_{i \neq j}^{n} \left[ 1 + \exp (\hat{\sigma}_{1}^{(a)} z_{i}) \right] \right) \right\} v_{t}}, a = U, C,
\]

if \( j \in A_{i} \subseteq S_{A}, i' = 1, \ldots, n \).

The previous expressions imply that \( \hat{\beta}_{(1)y}^{(a)} \) depends on the \( x_{j} \)s through their sum, that is, on the number of venues that are linked to the element \( j \), but not on the particular venues 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 venues in \( S_{A} \), the predictors \( \hat{\beta}_{(k)j}^{(a)} \) and \( \hat{\beta}_{(k)j'}^{(a)} \) are equal one another. The same happens for two persons in \( A_{k} \subseteq S_{A} \).

Thus, model-based Horvitz-Thompson-like estimators (HTLEs) of the totals \( Y_{k}, k = 1, 2 \), and \( Y \) are

\[
\hat{Y}_{HT,k}^{(a)} = \sum_{j \in S_{k}} y_{j}^{(k)} / \hat{\pi}_{(k)j}^{(a)} \left( \hat{\sigma}_{1}^{(a)}, \hat{\sigma}_{1}^{(a)}, \hat{\beta}_{(k)j}^{(a)} \right), k = 1, 2, \text{ and } \hat{Y}_{HT}^{(a)} = \hat{Y}_{HT,1}^{(a)} + \hat{Y}_{HT,2}^{(a)}, a = U, C.
\]  

(8)

Similarly, model-based HTLEs of the means \( \hat{Y}_{k} \) and \( \hat{Y} \) are

\[
\hat{Y}_{HT,k}^{(a)} = \hat{Y}_{HT,k}^{(a)} / \hat{\pi}_{(k)k}^{(a)} , k = 1, 2, \text{ and } \hat{Y}_{HT}^{(a)} = \hat{Y}_{HT,k}^{(a)} / \hat{\pi}_{(k)k}^{(a)} , a = U, C.
\]

Notice that if we set \( y_{j}^{(k)} = 1 \), for \( j = 1, \ldots, \tau_{k} \), then \( Y_{k} = \tau_{k}, k = 1, 2, \) and \( Y = \tau \). Therefore, HTLEs of \( \tau_{k} \) and \( \tau \) are \( \hat{\tau}_{HT,k}^{(a)} = \hat{Y}_{HT,k}^{(a)}, k = 1, 2, \) and \( \hat{\tau}_{HT}^{(a)} = \hat{Y}_{HT,k}^{(a)}, a = U, C, \) where \( \hat{Y}_{HT,k}^{(a)} \) and \( \hat{Y}_{HT}^{(a)} \) are given by (8) with \( y_{j}^{(k)} = 1 \).

We could also define Hájek-like estimators (HKLEs) of the population totals and means. Thus, HKLEs of the means \( \hat{Y}_{k} \) and \( \hat{Y} \) are

\[
\hat{Y}_{HK,k}^{(a)} = \hat{Y}_{HK,k}^{(a)} / \hat{\pi}_{(k)k}^{(a)} , k = 1, 2, \text{ and } \hat{Y}_{HK}^{(a)} = \hat{Y}_{HK,k}^{(a)} / \hat{\pi}_{(k)k}^{(a)} , a = U, C,
\]

and HKLEs of the totals \( Y_{k} \) and \( Y \) are

\[
\hat{Y}_{HK,k}^{(a)} = \hat{Y}_{HK,k}^{(a)} \hat{\pi}_{(k)k}^{(a)} , k = 1, 2, \text{ and } \hat{Y}_{HK}^{(a)} = \hat{Y}_{HK,k}^{(a)} \hat{\pi}_{(k)k}^{(a)} , a = U, C.
\]

5. **Bootstrap Variance Estimators and Confidence Intervals**

We propose the use of bootstrap to construct estimators of the variances of the proposed estimators of the totals and means, as well as confidence intervals (CIs) for those population parameters. The proposed bootstrap variant is obtained by combining the bootstrap version for finite populations proposed by Booth et al. (1994) and the parametric bootstrap variant (see Davison and Hinkley, 1997, chap. 2). This version of bootstrap is an extension of the one used by Félix-Medina et al. (2015) to construct CIs based on their proposed MLEs of the population sizes. We will describe the technical aspects of the proposed bootstrap variant in Appendix and for now it is enough to indicate that after applying that procedure to an estimator \( \hat{\theta} \) of a population parameter \( \theta \), we will get a bootstrap sample of \( B \) values \( \theta_{1}, \ldots, \theta_{B} \).
Félix-Medina: Estimating Totals Using Link Tracing Sampling

To construct the CIs for the population totals and means we could use any of the different bootstrap alternatives that have been proposed. For instance, if we did not want to assume any probability distribution for an estimator, we could use the basic or the percentile method. (See Davison and Hinkley 1997, chap. 5) Although this type of alternative has good properties of robustness, it requires a large number $B$ of bootstrap samples, say $B \geq 1000$, and this might be a serious problem if the estimator requires much time to be computed. On the other hand, if we were willing to assume a distribution probability for an estimator, we could use the $B$ values $\hat{\theta}_1, \ldots, \hat{\theta}_B$ to estimate the variance of $\hat{\theta}$ and construct the CI using the estimated variance and the assumed distribution of $\hat{\theta}$. In this case, the number $B$ of required bootstrap samples is not so large, say $50 \leq B \leq 200$ is generally enough. We will follow this approach, using some ideas taken from Félix-Medina et al. (2015).

Thus, as in that article, we will estimate the variance of an estimator $\hat{\theta}$ of the population parameter $\theta$, by using Huber's proposal 2 to jointly estimate the parameters of location and scale from the bootstrap sample of $\hat{\theta}_b$. (See Staudte and Sheather 1990, subsec. 4.5). In particular, the estimate of the parameter of scale is an estimate of the standard deviation $\sqrt{\hat{V}(\hat{\theta})}$ of $\hat{\theta}$. The idea behind the use of this estimator is that it yields an estimate of the standard deviation that is robust to very large values $\hat{\theta}$ which are likely to occur.

To construct the CIs we will use the following approach. (1) If the parameter is $\tau_k$, $k = 1, 2, 3$ or $\tau$, then, as in Félix-Medina et al. (2015), we will assume that $\hat{\tau}_k^{(a)} - \tau_k$ is lognormally distributed, where $\hat{\tau}_k^{(a)}$, $a = U, C$, is an estimator of $\tau_k$ and $\tau_k$ is the number of sampled elements from $U_k$. Thus, a CI for $\tau_k$ is $(\hat{\tau}_k^{(a)} - \tau_k)/c_k$, $\tau_k = (\hat{\tau}_k^{(a)} - \tau_k) \times c_k$, where $c_k = \exp \left\{ \frac{z_{\alpha/2}}{\ln \left[ 1 + \hat{V}(\hat{\tau}_k^{(a)})/(\hat{\tau}_k^{(a)} - \tau_k)^2 \right]} \right\}$, $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution and $\hat{V}(\hat{\tau}_k^{(a)})$ is an estimate of the variance of $\hat{\tau}_k^{(a)}$. A CI for $\tau$ is built analogously. The values of $\nu_1$, $\nu_2$ and $\nu$ that are used in the CIs for $\tau_1$, $\tau_2$ and $\tau$ are $m + r_1, r_2$ and $m + r_1 + r_2$, respectively. This type of CI was considered by Chao (1987), in the context of capture-recapture studies, as an alternative to the ordinary Wald CI to take into account the usually right skewed distribution of the estimator of the size and avoid the problem of obtaining CIs with lower bounds smaller than the number of sampled elements. She found that this CI generally performs better than the Wald CI. (See Williams et al., 2002, subsec. 14.2). (2) If the parameter is $\hat{Y}_k$, $k = 1, 2$, or $\hat{Y}$, and it is a proportion, that is, the $y$-value associated with an element is equal to one if the element has a characteristic of interest and is equal to zero otherwise, then we will assume that the number of sampled elements with the characteristic of interest has a binomial distribution and a CI for $\hat{Y}$ will be constructed using the proposal of Korn and Graubard (1998), which is an adaptation of the Clopper-Pearson CI for a proportion in the case of complex samples. Thus, a CI for $\hat{Y}_k$ is $\left( \frac{\hat{v}(a)}{v(k)}F_{\hat{d}(a),v(k),v(a)/2} + \frac{1}{v(k)}(1 - \frac{1}{v(k)}) \right)$, where $\hat{d}(a) = 2\hat{d}(a)$ and $\hat{d}(a) = 2(n_k(a) - y_k(a) + 1)$, $v(k) = 2(n_k(a) - y_k(a))$, $y_k(a) = n_k(a)\hat{Y}_k$, $n_k(a) = \hat{Y}_k(1 - \hat{Y}_k)/\hat{V}(\hat{Y}_k)$, $\hat{Y}_k$ is an estimator of $\hat{Y}_k$, $\hat{V}(\hat{Y}_k)$ is an estimate of the variance of $\hat{Y}_k$, and $F_{d_1,d_2}(\beta)$ is the $\beta$ quantile of the $F$ distribution with $d_1$ and $d_2$ degrees of freedom. A CI for $\hat{Y}$ is built analogously. (3) If the parameter is $\hat{Y}_k$, $k = 1, 2$, or $\hat{Y}$, and it is the mean of the $y$-values a continuous variable of interest or if it is $Y_k$, $k = 1, 2$, or $Y$, that is, it is the total of the
y-values of a continuous or a binary variable of interest, we will assume that
the estimator \( \hat{Y}_k^{(a)} \) (or \( \tilde{Y}_k^{(a)} \)) is normally distributed. Thus, a CI for \( \hat{Y}_k \) is
\[
\left( \hat{Y}_k - z_{\alpha/2} \sqrt{V(\hat{Y}_k)}, \hat{Y}_k + z_{\alpha/2} \sqrt{V(\hat{Y}_k)} \right),
\]
where \( z_{\alpha/2} \) is the upper \( \alpha/2 \) point of the standard normal distribution and \( V(\hat{Y}_k) \) is an estimate of the variance of \( \hat{Y}_k^{(a)} \). CIs for \( Y_k, \bar{Y} \)
and \( Y \) are built analogously.

6. Monte Carlo Studies

In order to observe the performance of the proposed estimators and CIs and to compare
their performance with the ones proposed by Félix-Medina and Monjardin (2010), which
were derived under the assumption of homogeneity of the link probabilities, we carried out
two numerical studies. In the first study, we used artificial data to construct two
populations with specific characteristics, whereas in the second one we used data from the
National Longitudinal Study of Adolescent Health collected during the 1994–1995 school
year to construct a population. Both studies were carried out using the R software
environment for statistical computing (R Core Team, 2018).

6.1. Populations Constructed Using Artificial Data

We constructed two populations whose characteristics are described in Table 1.
The difference between the two populations is that in Population I the link probabilities
were generated by using Expression (1), that is, under the assumed model, whereas in
Population II they were generated by the following latent-class model used by Pledger
(2000) in the context of capture-recapture studies:
\[
p_{ij}^{(k)} = \exp \left[ \mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)} \right]/\left\{1 + \exp \left[ \mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)} \right]\right\}, i = 1, \ldots, n; j = 1, 2, \text{ and } k = 1, 2.
\]
In this model, the people in \( U_k \) are divided into two classes according to their propensities to be
linked to the sample clusters. The probability that a person in \( U_k \) is in class \( j \) is \( p_{ij}^{(k)} \) and it is the
same for each person in \( U_k \). The values of the parameters that appear in each of the two
expressions of the link probabilities were set so that when the size of the initial sample of
clusters is \( n = 15 \), in both populations the sampling fractions were \( f_1 = 0.5 \) in \( U_1 \) and \( f_2 = 0.4 \)
in \( U_2 \). Note from Table 1 that associated with each element of each population, there are two
values of two response variables. One variable is a continuous variable whose value associated
with the \( j \)-th element of \( U_k \) was obtained by sampling from a non-central chi-square
distribution with two degrees of freedom and non-centrality parameter \( \psi_{ij}^{(k)} \). The other variable
is a binary variable whose value associated with that element was obtained from a Bernoulli
distribution with mean \( \varphi_{ij}^{(k)} \). The values of the parameters that appear in the expressions of \( \psi_{ij}^{(k)} \)
and \( \varphi_{ij}^{(k)} \) were set so that the values of the population means of the continuous variable in both
populations were \( \bar{Y}_1 = 50 \) and \( \bar{Y}_2 = 40 \), whereas the corresponding values of the binary
variable were \( \bar{Y}_1 = 0.3 \) and \( \bar{Y}_2 = 0.2 \). Furthermore, they were set so that for \( n = 15 \), in
Population I the values of the Pearson correlation coefficients between the values of the
continuous response variable and those of the inclusion probabilities were \( \rho(y^{(1)}, \pi_{(1)}) \approx 0.8 \)
and \( \rho(y^{(2)}, \pi_{(2)}) \approx 0.7 \), whereas the corresponding values for the binary response variable
were \( \rho(y^{(1)}, \pi_{(1)}) \approx 0.3 \) and \( \rho(y^{(2)}, \pi_{(2)}) \approx 0.27 \). In the case of Population II and continuous
response variable those values were \( \rho(y^{(1)}, \pi_{(1)}) \approx 0.15 \) and \( \rho(y^{(2)}, \pi_{(2)}) \approx 0.1 \), whereas the
corresponding values for the binary variable were \( \rho(y^{(1)}, \pi_{(1)}) \approx 0 \) and \( \rho(y^{(2)}, \pi_{(2)}) \approx 0.1 \).
The Monte Carlo study was carried out by repeatedly selecting \( r \) samples from the population \( U \) using the sampling design described in Section 2. Thus, a SRSWOR of \( n = 15 \) values \( m_i \) was selected from the population of \( N = 150 \) values. For each selected value \( m_i \), the value \( x_{ij}^{(k)} \) of the link indicator variable \( X_{ij}^{(k)} \) was generated from the Bernoulli distribution with mean \( p^{(k)}_{ij} \) (see its expression in Table 1). From each selected sample, the estimators indicated in Table 2 were computed. The performance of an estimator \( \hat{u} \) of a parameter \( u \) was evaluated by its relative bias (r-bias), the square root of its relative mean square error (rMSE), and the confidence interval (CI) coverage.

**Table 1. Parameters of simulated populations**

| Population I | Population II |
|--------------|--------------|
| \( N = 150 \) | \( N = 150 \) |
| \( M_i \sim \text{zero truncated neg. binom. distribution} \) | \( M_i \sim \text{zero truncated neg. binom. distribution} \) |
| \( E(M_i) = 8 \), \( V(M_i) = 24 \) | \( E(M_i) = 8 \), \( V(M_i) = 24 \) |
| \( \tau_1 = 1208 \), \( \tau_2 = 400 \), \( \tau = 1608 \) | \( \tau_1 = 1208 \), \( \tau_2 = 400 \), \( \tau = 1608 \) |
| Rasch model for \( p_{ij}^{(k)} \): | Rasch model for \( p_{ij}^{(k)} \): |
| \( p_{ij}^{(k)} = \frac{\exp(a_i^{(k)} + b_j^{(k)})}{1 + \exp(a_i^{(k)} + b_j^{(k)})} \) | \( p_{ij}^{(k)} = \frac{\exp(\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha \beta)^{ij})}{1 + \exp(\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha \beta)^{ij})} \) |
| \( i = 1, \ldots, N, \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) | \( i = 1, \ldots, N, \ j = 1, 2, \ k = 1, 2 \) |
| \( a_i^{(k)} = \frac{c_k}{0.001 + M_i^{(k)}}, \ i = 1, \ldots, N, \ k = 1, 2 \) | \( a_i^{(k)} = \frac{c_k}{0.001 + M_i^{(k)}}, \ i = 1, \ldots, N, \ k = 1, 2 \) |
| \( c_1 = -5.45 \), \( c_2 = -5.85 \) | \( c_1 = -12 \), \( c_2 = -12 \) |
| \( \beta_j^{(k)} \sim N(0, 1), \ j = 1, \ldots, \tau_k \) | \( \beta_j^{(k)} = 1.5, \ \beta_j^{(k)} = 0, \ k = 1, 2 \) |
| Continuous response variable: | Continuous response variable: |
| \( Y_j^{(k)} \sim \chi^2_j(Y_j^{(k)}), \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) | \( Y_j^{(k)} \sim \chi^2_j(Y_j^{(k)}), \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) |
| \( \psi_j^{(k)} = 5 + \frac{d_k \exp(\beta_j^{(k)})}{1 + \exp(\beta_j^{(k)})}, \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) | \( \psi_j^{(k)} = 5 + \frac{d_k \exp(\beta_j^{(k)})}{1 + \exp(\beta_j^{(k)})}, \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) |
| \( d_1 = 87 \), \( d_2 = 65 \) | \( d_1 = 65.05 \), \( d_2 = 50.05 \) |
| \( Y_1 = 60390.34, \ Y_2 = 15945.89, \ Y = 76336.22 \) | \( Y_1 = 60289.03, \ Y_2 = 16112.78, \ Y = 76401.81 \) |
| \( \tilde{Y}_1 = 49.99, \tilde{Y}_2 = 39.87, \tilde{Y} = 47.47 \) | \( \tilde{Y}_1 = 49.91, \tilde{Y}_2 = 40.28, \tilde{Y} = 47.51 \) |
| \( \rho(y^{(1)}, \pi_{(1)}) = 0.79 \), \( \rho(y^{(2)}, \pi_{(2)}) = 0.72 \) | \( \rho(y^{(1)}, \pi_{(1)}) = 0.16 \), \( \rho(y^{(2)}, \pi_{(2)}) = 0.11 \) |
| Binary response variable: | Binary response variable: |
| \( Y_j^{(k)} \sim \text{Bernoulli}(\phi_j^{(k)}), \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) | \( Y_j^{(k)} \sim \text{Bernoulli}(\phi_j^{(k)}), \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) |
| \( \phi_j^{(k)} = \frac{g_k \exp(\beta_j^{(k)})}{1 + \exp(\beta_j^{(k)})}, \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) | \( \phi_j^{(k)} = \frac{g_k \exp(\beta_j^{(k)})}{1 + \exp(\beta_j^{(k)})}, \ j = 1, \ldots, \tau_k, \ k = 1, 2 \) |
| \( g_1 = 0.6 \), \( g_2 = 0.39 \) | \( g_1 = 0.46 \), \( g_2 = 0.33 \) |
| \( Y_1 = 366, \ Y_2 = 82, \ Y = 448 \) | \( Y_1 = 365, \ Y_2 = 79, \ Y = 444 \) |
| \( \tilde{Y}_1 = 0.303, \tilde{Y}_2 = 0.205, \tilde{Y} = 0.279 \) | \( \tilde{Y}_1 = 0.302, \tilde{Y}_2 = 0.198, \tilde{Y} = 0.276 \) |
| \( \rho(y^{(1)}, \pi_{(1)}) = 0.29 \), \( \rho(y^{(2)}, \pi_{(2)}) = 0.27 \) | \( \rho(y^{(1)}, \pi_{(1)}) = -0.01 \), \( \rho(y^{(2)}, \pi_{(2)}) = 0.09 \) |

Note: Correlations coefficients \( \rho(y^{(k)}, \pi_{(k)}) \) were computed assuming a sample with initial size \( n = 15 \).

The Monte Carlo study was carried out by repeatedly selecting \( r \) samples from the population \( U \) using the sampling design described in Section 2. Thus, a SRSWOR of \( n = 15 \) values \( m_i \) was selected from the population of \( N = 150 \) values. For each selected value \( m_i \), the value \( x_{ij}^{(k)} \) of the link indicator variable \( X_{ij}^{(k)} \) was generated from the Bernoulli distribution with mean \( p_{ij}^{(k)} \) (see its expression in Table 1). From each selected sample, the estimators indicated in Table 2 were computed. The performance of an estimator \( \hat{u} \) of a parameter \( u \) was evaluated by its relative bias (r-bias), the square root of its relative mean square error (rMSE), and the confidence interval (CI) coverage.
Table 2. Point estimators and standard deviation estimators included in the Monte Carlo studies.

| Estimators of population sizes | Type of estimator | Notation | Type of standard deviation estimator | Proposed by |
|-------------------------------|-------------------|----------|-------------------------------------|-------------|
| UMLEs                         | \( \hat{t}_1^{(U)}, \hat{t}_2^{(U)} \) and \( \hat{t}^{(U)} \) | Bootstrap | Félix-Medina et al. (2015)         |
| CMLEs                         | \( \hat{t}_1^{(C)}, \hat{t}_2^{(C)} \) and \( \hat{t}^{(C)} \) | Bootstrap | This work                           |
| HTLEs based on UMLEs of inclusion probabilities | \( \hat{t}_{HT,1}, \hat{t}_{HT,2}^{(U)} \) and \( \hat{t}_{HT}^{(U)} \) | Bootstrap | This work                           |
| HTLEs based on CMLEs of inclusion probabilities | | | |
| MLEs derived under the homogeneity assumption | \( \hat{t}_{ML,1}^{(H)}, \hat{t}_{ML,2}^{(H)} \) and \( \hat{t}_{ML}^{(H)} \) | Linearization | Félix-Medina and Monjardin (2006) |
| Bayesian-assisted estimators derived under the homogeneity assumption | \( \hat{t}_{BA,1}^{(H)}, \hat{t}_{BA,2}^{(H)} \) and \( \hat{t}_{BA}^{(H)} \) | Linearization | Félix-Medina and Monjardin (2006) |

| Estimators of population totals and means | Type of estimator | Notation | Type of standard deviation estimator | Proposed by |
|-----------------------------------------|-------------------|----------|-------------------------------------|-------------|
| HTLEs based on UMLEs of inclusion probabilities | \( \hat{Y}_{HT,1}^{(U)}, \hat{Y}_{HT,2}^{(U)} \) and \( \hat{Y}_{HT}^{(U)} \) | Bootstrap | This work                           |
| HTLEs based on CMLEs of inclusion probabilities | | | |
| HKLEs based on UMLEs of inclusion probabilities | \( \hat{Y}_{HK,1}^{(U)}, \hat{Y}_{HK,2}^{(U)} \) and \( \hat{Y}_{HK}^{(U)} \) | Bootstrap | This work                           |
| HKLEs based on CMLEs of inclusion probabilities | | | |
| HTLEs based on MLEs and derived under the homogeneity assumption | \( \hat{Y}_{ML,1}^{(H)}, \hat{Y}_{ML,2}^{(H)} \) and \( \hat{Y}_{ML}^{(H)} \) | Bootstrap | Félix-Medina and Monjardin (2010) |
| HTLEs based on Bayesian assisted estimators derived under the homogeneity assumption | | | |
square error (\(\sqrt{r\text{-mse}}\)), the median of its relative estimation error (mdre), and the median of its absolute relative estimation error (mdare) defined as 
\[
r\text{-bias} = \sum (\hat{\theta}_i - \theta)/r\theta, \\
\sqrt{r\text{-mse}} = \sqrt{\sum (\hat{\theta}_i - \theta)^2/(r\theta^2)}, \\
\text{mdre} = \text{median} \{(\hat{\theta}_i - \theta)/\theta\} \quad \text{and} \quad \text{mdare} = \text{median} \{|(\hat{\theta}_i - \theta)/\theta|\},
\]
respectively, where \(\hat{\theta}_i\) is the value of \(\hat{\theta}\) obtained in the \(i\)-th sample, \(i = 1, \ldots, r\). In the case of the point estimators of the population sizes, totals and means their performance was evaluated using \(r = 5,000\) samples.

We also computed estimators of the standard deviation of the point estimators, which are indicated in Table 2. The performance of a standard deviation estimator \(\hat{sd}(\hat{\theta})\) of the standard deviation \(sd(\hat{\theta})\) of \(\hat{\theta}\) was also evaluated by its \(r\)-bias, \(\sqrt{r\text{-mse}}\), mdre and mdare, where \(sd(\hat{\theta})\) was computed by the sample standard deviation of the \(\hat{\theta}_i\), \(i = 1, \ldots, r\). Because of the time required to compute the bootstrap standard deviation estimators, we used bootstrap samples of size \(B = 50\) and their performance was evaluated using \(r = 500\) samples, whereas in the case of the linearization standard deviation estimators using \(r = 5,000\) samples.

From each point estimator and its associated standard deviation estimator a 95% CI was compute for the corresponding parameter. In the case of the estimators based on the assumption of heterogeneous link probabilities the CIs were computed as was described in Section 5, whereas in the case of those based on the homogeneity assumption the CIs were Wald type CIs. The performance of a CI was evaluated by its coverage probability (cp) defined as the proportion of samples in which the parameter is inside the interval, and by both its mean relative length (mrl) and median relative length (mdrl) defined as the sample mean and median of the lengths of the \(r\) intervals divided by the value of the parameter, respectively. In the case of the CIs based on point estimators derived under the assumption of heterogeneous link probabilities, their performance was evaluated using \(r = 500\) samples, whereas in the case of the estimators derived under the assumption of homogeneous link probabilities using \(r = 5,000\) samples.

In this and in the following study, and in the case of the estimators that were derived under the assumption of heterogeneous link probabilities, we only present the outcomes corresponding to the estimators based on the UMLEs of these probabilities because their performance was very similar to that of the estimators based on the CMLEs of the link probabilities. In the case of the estimators derived under the assumption of homogeneous link probabilities, we only present the outcomes corresponding to the estimators based on the MLEs of these probabilities because their performance was very similar to that of the estimators based on the Bayesian assisted estimators of the link probabilities. In addition, in the descriptions of the results of the numerical studies we will use the convention that the performance of a point estimator will be considered as acceptable if both its \(r\)-bias (or mdre) and its \(\sqrt{r\text{-mse}}\) (or mdare) are around or are lesser than 0.1. Similarly, we will use the convention that the performance of a 95% CI is acceptable if its cp is around or greater than 0.9 and its mrl (or mdrl) is around or is lesser than 0.4 (\(= 4 \times 0.1\)).

The results of the study about the inferences on the population sizes are shown in Figure 1 and in Table 3. We see that in both populations, the distributions of the UMLE \(\hat{\tau}^{(U)}_1\) were symmetrical, and those of \(\hat{\tau}^{(U)}_2\) and \(\hat{\tau}^{(U)}\) were skewed to the right with long tails. In Population I, the UMLEs did not show bias problems, but in Population II, they presented slight negative biases, except for the estimator \(\hat{\tau}^{(U)}_2\) which presented a
moderate negative bias. In both populations, the bootstrap estimators of the standard deviations of the UMLEs showed problems of bias. The confidence intervals based on these estimators had good coverage probabilities in Population I, although the interval for $\tau_2$ was very long, whereas in Population II their coverage probabilities were small. The HTLEs derived under the heterogeneity assumption exhibited approximately symmetric distributions, except for the estimators $\hat{\tau}^{HT.1}$ which exhibited right skewed distributions in Population II. In both populations these estimators underestimated, with moderate biases, their corresponding parameters. The bootstrap estimators of their standard deviations exhibited bias problems, and the confidence intervals based on these estimators presented low coverage probabilities, especially in the case of Population II, where they were very small. Finally, the MLEs derived under the homogeneity assumption seriously underestimated their corresponding sizes. Their large biases affected the performance of the confidence intervals based on these estimators in such a way that their coverage probabilities were practically null. Thus, in summary, the UMLEs had the best performance, which was good in Population I, and regular in Population II. The MLEs derived under the homogeneity assumption had the worst performance, which was poor.

The results of the study on the point estimators of totals and means are shown in Figure 2 and in Table 4. We see that in Population I, and regardless of the type of variable, continuous or binary, the distributions of the HTLEs of the totals were relatively symmetrical, whereas in Population II their distributions were skewed to the right, except for that of $\hat{\tau}^{HT.1}$ which was approximately symmetric. These estimators did not have problems of bias in Population I, and presented moderate negative biases in Population II. The distributions of the HKLEs of the totals were right skewed, except for

Fig. 1. Boxplots of standardized values of the estimators of sizes in Populations I and II.
Table 3. Simulation results on point estimators, standard deviation estimators and 95% confidence intervals for population sizes.

| Results on point estimators | UMLEs | HTLEs | Homogeneous MLEs |
|-----------------------------|-------|-------|------------------|
|                             | $\hat{\tau}_1$ | $\hat{\tau}_2$ | $\hat{\tau}$ | $\hat{\tau}_{HT,1}$ | $\hat{\tau}_{HT,2}$ | $\hat{\tau}_{HT}$ | $\hat{\tau}_{ML,1}$ | $\hat{\tau}_{ML,2}$ | $\hat{\tau}_{ML}$ |
| **Population rates**        |       |       |                  |
|                             |       |       |                  |
| **I**                       |       |       |                  |
| $f_1 = 0.5$                 |       |       |                  |
| $f_2 = 0.4$                 |       |       |                  |
|                             |       |       |                  |
| **Population rates**        |       |       |                  |
|                             |       |       |                  |
| **II**                      |       |       |                  |
| $f_1 = 0.5$                 |       |       |                  |
| $f_2 = 0.4$                 |       |       |                  |

Sampling rates | rbias | $\sqrt{\text{rmse}}$ | $\sqrt{\text{mse}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ |
|----------------|-------|-----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
|                |       |                       |                       |                       |                       |                       |                       |                       |                       |

Sampling rates | $\sqrt{\text{mse}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ | $\sqrt{\text{me}}$ |
|----------------|-----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
|                |                       |                       |                       |                       |                       |                       |                       |                       |                       |
Table 3. Continued.

| Results on 95% confidence intervals | $\tilde{\tau}_1^{(U)}$ | $\tilde{\tau}_2^{(U)}$ | $\tilde{\tau}_1^{(H)}$ | $\tilde{\tau}_2^{(H)}$ | $\tilde{\tau}^{(H)}_1$ | $\tilde{\tau}^{(H)}_2$ | $\tilde{\tau}^{(H)}$ |
|-------------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| Population                          | cp                   | mrl                  | mdrl                  | cp                   | mrl                  | mdrl                  | cp                   |
| Sampling rates: cp                   | .95                  | .97                  | .98                  | .78                  | .85                  | .78                  | .00                  |
| $f_1 = 0.5$                         | mrl                  | .37                  | 5.6                  | 1.4                  | .31                  | 1.2                  | .39                  |
| $f_2 = 0.4$                         | mdrl                 | .36                  | 1.8                  | .52                  | .30                  | .85                  | .33                  |
| Sampling rates: cp                   | .76                  | .51                  | .63                  | .44                  | .36                  | .28                  | .05                  |
| $f_1 = 0.5$                         | mrl                  | .29                  | 1.7                  | .54                  | .24                  | .66                  | .27                  |
| $f_2 = 0.4$                         | mdrl                 | .29                  | .44                  | .26                  | .24                  | .34                  | .21                  |
| Notes: Results on point estimators are based on 5,000 samples. In Population I the percentages of samples in which the estimators of $\tau_1$, $\tau_2$ and $\tau$ derived under the heterogeneity assumption that were not computed because of numerical convergence problems were 0%, 0.02%, and 0.02%, respectively, whereas in Population II the corresponding percentages were 0.36%, 8.7%, and 9.0%. Bootstrap standard deviation estimators corresponding to estimators of sizes derived under the heterogeneity assumption were computed using 50 bootstrap samples, and their results, as well as those of the 95% confidence intervals are based on 500 replicated samples. In Population II the percentages of samples in which the bootstrap standard deviation estimators and confidence intervals were not computed because of convergence problems were 0.8%, 8.2% and 8.8% in $U_1$, $U_2$ and $U$, respectively, whereas in Population I the respective percentages were all 0%. Results on point estimators, standard deviation estimators and confidence intervals derived under the homogeneity assumption are based on 5,000 samples and no convergence problems occurred.
those of \( \hat{Y}_{HK,1} \) which were more or less symmetrical. The distributions presented this shape regardless of the population and type of variable. These estimators showed moderate positive biases in Population I, and moderate negative biases in Population II. Note that in Population II the magnitudes of the biases of these estimators were smaller than those of the HTLEs. The HTLEs of the totals derived under the homogeneity assumption exhibited serious underestimation problems in both populations and with both types of response variables.

In the case of the estimators of the population means (see Figure 3 and Table 4), we have that the HTLEs exhibited more or less symmetric distributions, except for \( \hat{Y}_{HT} \) whose
Table 4. Relative biases, square roots of relative mean square errors, medians of relative errors and medians of absolute relative errors of the point estimators of the totals and means.

| Population  | I            | II           | I            | II           |
|-------------|--------------|--------------|--------------|--------------|
|             | f₁ = 0.5     | f₂ = 0.4     | f₁ = 0.5     | f₂ = 0.4     |
| Sampling rates |             |              |              |              |
| f₁           | 0.5          | 0.4          | 0.5          | 0.4          |
| f₂           |              |              |              |              |
| Response variable | Continuous |              | Binary       |              |
| Estimator    | rbias        | √rmse        | mdre         | mdare        | rbias        | √rmse        | mdre         | mdare        | rbias        | √rmse        | mdre         | mdare        |
| HTLEs of totals |            |              |              |              |              |              |              |              |              |              |              |              |
| ^Y^(U)jt₁   | -0.0         | 0.06         | -0.0         | 0.04         | 0.01         | 0.07         | 0.01         | 0.05         | -0.14        | 0.17         | -0.14        | 0.14         | -0.16        | 0.19         | -0.16        | 0.16         |
| ^Y^(U)jt₂   | -0.06        | 0.17         | -0.08        | 0.12         | 0.03         | 0.19         | 0.01         | 0.11         | -0.27        | 0.34         | -0.30        | 0.31         | -0.21        | 0.32         | -0.25        | 0.26         |
| ^Y^(U)jt₃   | -0.01        | 0.06         | -0.01        | 0.04         | 0.02         | 0.07         | 0.02         | 0.05         | -0.17        | 0.18         | -0.17        | 0.17         | -0.16        | 0.19         | -0.17        | 0.17         |
| HKLEs of totals |            |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| ^Y^(U)jk₁   | 0.11         | 0.14         | 0.11         | 0.13         | 0.16         | 0.13         | 0.13         | 0.12         | 0.07         | 0.07         | 0.08         | 0.09         | 0.14         | 0.09         | 0.10         |              |
| ^Y^(U)jk₂   | 0.24         | 0.47         | 0.15         | 0.18         | 0.35         | 0.25         | 0.25         | 0.25         | -0.15        | 1.9          | -0.26        | 0.27         | -0.06        | 2.7          | -0.20        | 0.22         |
| ^Y^(U)jk₃   | 0.15         | 0.18         | 0.14         | 0.18         | 0.21         | 0.17         | 0.17         | 0.08         | 0.45         | -0.09        | 0.10         | -0.08        | 0.49         | -0.09        | 0.10         |              |
| Homo. HTLEs totals |            |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| ^Y^(H)j₁    | -0.19        | -0.19        | -0.19        | -0.17        | -0.17        | -0.17        | -0.17        | -0.24        | -0.25        | -0.24        | -0.24        | -0.26        | -0.28        | -0.26        | 0.26         |              |
| ^Y^(H)j₂    | -0.29        | -0.29        | -0.29        | -0.21        | -0.23        | -0.21        | -0.21        | -0.33        | 0.35         | -0.35        | 0.35         | -0.27        | 0.31         | -0.29        | 0.30         |              |
| ^Y^(H)j₃    | -0.21        | -0.21        | -0.21        | -0.18        | -0.18        | -0.18        | -0.18        | -0.26        | 0.27         | -0.26        | -0.26        | -0.26        | 0.27         | -0.26        | 0.26         |              |
| HTLEs of means |            |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| ^Y^(U)j₁    | 0.00         | 0.03         | 0.00         | 0.02         | 0.02         | 0.05         | 0.02         | 0.03         | -0.06        | 0.07         | -0.07        | 0.07         | -0.08        | 0.09         | -0.08        | 0.08         |
| ^Y^(U)j₂    | -0.08        | 0.14         | -0.07        | 0.08         | 0.01         | 0.16         | 0.01         | 0.10         | -0.06        | 0.10         | -0.05        | 0.05         | 0.02         | 0.14         | 0.03         | 0.09         |
| ^Y^(U)j₃    | -0.02        | 0.08         | -0.01        | 0.03         | 0.01         | 0.09         | 0.01         | 0.04         | -0.06        | 0.11         | -0.06        | 0.06         | -0.06        | 0.15         | -0.06        | 0.06         |
Table 4. Continued.

| Sampling rates | 1 | 1 | II | II |
|----------------|---|---|----|----|
|                | f₁ = 0.5 | f₁ = 0.5 | f₂ = 0.4 | f₂ = 0.4 |
| Response variable | Continuous | Continuous | Binary | Binary |
| Estimator | rbias | \( \text{rmse} \) | mdare | rbias | \( \text{rmse} \) | mdare | rbias | \( \text{rmse} \) | mdare |
| HKLEs \( \hat{Y}_{HKL} \) \( (H) \) | .12 | .17 | .13 | .12 | .17 | .13 | .13 | .13 | .13 | .13 |
| Y \( \hat{Y}_{HKL} \) \( (U) \) | .12 | .17 | .13 | .12 | .17 | .13 | .13 | .13 | .13 | .13 |
| ML | .12 | .17 | .13 | .12 | .17 | .13 | .13 | .13 | .13 | .13 |
| Homo. \( \hat{Y}_{H} \) \( (H) \) | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| Y \( \hat{Y}_{H} \) \( (U) \) | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| HTLEs \( \hat{Y}_{HTL} \) \( (H) \) | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| Y \( \hat{Y}_{HTL} \) \( (U) \) | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| ML | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| Notes: Results are based on 5,000 samples. In Population I the percentages of samples in which the estimators of the totals and means of \( U_1, U_2 \) and \( U \) derived under the heterogeneity assumption that were not computed because of numerical convergence problems were 0.0%, 0.3%, and 0.1%, respectively, whereas in Population II the corresponding percentages were 0.5%, 8.7%, and 9.0%. In the computation of the estimators derived under the homogeneity assumption no convergence problems occurred.
distributions were skewed to the left. These estimators did not present problems of bias, regardless of the type of response variable and population. The HKLEs showed symmetrical distributions. In the case of Population I, they exhibited moderate positive biases, whereas in Population II, they were practically unbiased. Finally, the HTLEs derived under the homogeneity assumption performed very similarly to the HKLEs, but the magnitudes of their biases were slightly larger than those of the HKLEs.

Thus, in summary, the HTLEs of totals and means derived under the heterogeneity assumption had the best performance in Population I, and their performance was good, whereas in Population II, the HKLEs had the best performance and their performance was acceptable. These results are the consequence of the high correlations between the response variables and the inclusion probabilities in Population I, which favors the HTLEs, and the small correlations in Population II, which favors the HKLEs. (See Thompson 2012, subsec. 6.2; Särndal et al. 1992, subsec. 5.7). The HTLEs derived under
the homogeneity assumption performed very badly, except in the case of estimating the means of the Population II, where their performance was fairly good. The small biases of these estimators could be explained by the fact that the biases of this type of estimator when totals were estimated were practically the same as the biases of this type of estimator when sizes were estimated, and consequently, their biases were canceled out when the quotients were computed to form the estimators of the means.

The results of the estimators of the standard deviations of the estimators of the totals and means are shown in Table 5. In general, each one of the estimators exhibited problems of bias. The estimators of the standard deviations of the HKLEs had the best performance. Estimators of the standard deviations of the estimators of means performed better than those of the estimators of totals. In addition, their performance was better in the case of the binary response variable than in the case of the continuous response variable. Furthermore, their performance was better in Population I than in Population II.

The results on the confidence intervals for the totals and means are shown in Table 6. In Population I, the confidence intervals for the totals based on the HTLEs had good values of the coverage probabilities and relative lengths regardless of the type of response variable, except for the interval for \(Y_2\), which had large relative lengths. In Population II, the coverage probabilities of these intervals were low. This result is a consequence of the moderate biases exhibited by the HTLEs of the totals. The intervals for the totals based on the HKLEs had, in general, relatively low coverage probabilities regardless of the type of response variable and population, except for the interval for \(Y_2\), which showed large coverage probabilities, as well as large lengths, in Population I. The intervals for the totals based on the HTLES derived under the homogeneity assumption showed very low coverage probabilities, which was a consequence of the large biases exhibited by these estimators. In the case of the confidence intervals for the means, we have that the intervals based on the HTLEs derived under the heterogeneity assumption had good performance in Population I. In Population II, the coverage probabilities of the intervals for \(\hat{Y}_1\) and \(\hat{Y}\) exhibited low coverage probabilities. The intervals based on the HKLEs had very low coverage probabilities in Population I. In Population II, and particularly in the case of the binary response variable, these intervals had good performance. Finally, the intervals based on the HTLEs derived under the homogeneity assumption showed very low coverage probabilities in each of the situations that were considered in this study, except in Population II and binary response variable, where the intervals performed acceptably.

Thus, in summary, the intervals for the totals and means based on the HTLEs derived under the heterogeneity assumption had the best performance, but their performance was good only in the case of Population I (although the intervals for \(Y_2\) were very long). The intervals based on the HKLEs and on the HTLEs derived under the homogeneity assumption did not perform well, except in the case of Population II and binary response variable, where their performance was acceptable.

6.2. Populations Constructed Using Data from the National Longitudinal Study of Adolescent Health

In this Monte Carlo study, we used data from the National Longitudinal Study of Adolescent Health (Add Health) to construct a population. The Add Health is a longitudinal study of a
Table 5. Relative biases, square roots of relative mean square errors, medians of relative errors and medians of absolute relative errors of the standard deviation estimators of the estimators of the totals and means.

| Response variable | Continuous | Binary | Continuous | Binary |
|-------------------|------------|--------|------------|--------|
| Sampling rates    | I          | II     | I          | II     |
| \( f_1 = 0.5 \)   | \( f_2 = 0.4 \) | \( f_1 = 0.5 \) | \( f_2 = 0.4 \) |
| \( \sqrt{\text{bias}} \) | \( \sqrt{\text{rmse}} \) | \( \text{mdre} \) | \( \text{mdare} \) | \( \sqrt{\text{bias}} \) | \( \sqrt{\text{rmse}} \) | \( \text{mdre} \) | \( \text{mdare} \) | \( \sqrt{\text{bias}} \) | \( \sqrt{\text{rmse}} \) | \( \text{mdre} \) | \( \text{mdare} \) |
| \( \tilde{sd_B}(\hat{Y}_{HT}) \) | .20 | .32 | .16 | .19 | .12 | .26 | .10 | .15 | -.29 | .35 | -.30 | .30 | -.23 | .30 | -.24 | .25 |
| \( \tilde{sd_B}(\hat{Y}_{HT,2}) \) | .78 | 1.5 | .33 | .45 | .63 | 1.3 | .26 | .41 | -.09 | 1.9 | -.47 | .54 | .02 | 1.6 | -.35 | .44 |
| \( \tilde{sd_B}(\hat{Y}_{HT}) \) | .57 | .89 | .40 | .40 | .46 | .72 | .34 | .34 | -.18 | .82 | -.31 | .34 | -.13 | .59 | -.22 | .26 |
| \( \tilde{sd_B}(\hat{Y}_{HK}) \) | .20 | .32 | .16 | .19 | .12 | .26 | .09 | .15 | -.18 | .27 | -.19 | .21 | -.16 | .26 | -.16 | .19 |
| \( \tilde{sd_B}(\hat{Y}_{HK,2}) \) | 1.9 | 6.3 | .14 | .56 | 1.6 | 5.9 | .09 | .52 | .26 | 10.3 | -.59 | .63 | .32 | 9.6 | -.50 | .57 |
| \( \tilde{sd_B}(\hat{Y}_{HK}) \) | 1.9 | 5.6 | .34 | .37 | 1.7 | 5.2 | .33 | .37 | .22 | 6.7 | -.27 | .30 | .18 | 5.2 | -.21 | .26 |
| \( \tilde{sd_B}(\hat{Y}_{ML}) \) | -.24 | .25 | -.24 | .24 | -.18 | .19 | -.18 | .18 | -.69 | .71 | -.72 | .72 | -.62 | .64 | -.65 | .65 |
| \( \tilde{sd_B}(\hat{Y}_{ML,2}) \) | -.11 | .21 | -.14 | .17 | -.08 | .19 | -.10 | .14 | -.47 | .65 | -.58 | .59 | -.31 | .54 | -.42 | .45 |
| \( \tilde{sd_B}(\hat{Y}_{ML}) \) | -.25 | .27 | -.26 | .26 | -.17 | .19 | -.18 | .18 | -.66 | .68 | -.69 | .69 | -.58 | .60 | -.61 | .61 |
| \( \tilde{sd_B}(\hat{Y}_{HT}) \) | .01 | .18 | -.01 | .11 | .16 | .23 | .15 | .16 | -.50 | .54 | -.51 | .51 | -.11 | .27 | -.13 | .17 |
| \( \tilde{sd_B}(\hat{Y}_{HT,2}) \) | .35 | .65 | .22 | .35 | .24 | .44 | .17 | .21 | -.20 | .67 | -.38 | .44 | .06 | .50 | -.06 | .24 |
| \( \tilde{sd_B}(\hat{Y}_{HT}) \) | .84 | 1.8 | .20 | .43 | .65 | 1.3 | .19 | .28 | -.35 | .73 | -.50 | .52 | -.02 | .60 | -.16 | .22 |
| Table 5. Continued |
|---------------------|
| **Population**      |
|                     |
| **I**               |
| **II**              |
| **Sampling rates**  |
| $f_1 = 0.5$         |
| $f_2 = 0.4$         |
| $f_1 = 0.5$         |
| $f_2 = 0.4$         |
| **Response variable** | **Continuous** | **Binary** | **Continuous** | **Binary** |
| **Standard deviation estimator** | rbias | $\sqrt{\text{rmse}}$ | mdre | mdare | rbias | $\sqrt{\text{rmse}}$ | mdre | mdare | rbias | $\sqrt{\text{rmse}}$ | mdre | mdare |
| HTLEs of means      |
| $\hat{s}_d(U_{1})$  | -.12  | .18  | -1.13 | .14  | .11  | .18  | .10  | .12  | -.10 | .20  | -.12 | .15  | .05  | .23  | .04  | .15  |
| $\hat{s}_d(U_{2})$  | .09   | .23  | .06  | .15  | .11  | .25  | .09  | .15  | .07  | .35  | .00  | .20  | .03  | .33  | -.03 | .20  |
| $\hat{s}_d(U_{3})$  | .06   | .30  | -.14 | .14  | .16  | .29  | .13  | .14  | -.13 | .47  | -.24 | .29  | .08  | .32  | .06  | .16  |
| Homo. HTLEs of means |
| $\hat{s}_d(U_{1})$  | -.38  | .39  | -.39 | .39  | -.16 | .18  | -.16 | .16  | -.21 | .29  | -.23 | .24  | -.09 | .25  | -.11 | .18  |
| $\hat{s}_d(U_{2})$  | -.19  | .22  | -.20 | .20  | -.11 | .15  | -.12 | .13  | -.12 | .29  | -.16 | .22  | -.12 | .29  | -.15 | .22  |
| $\hat{s}_d(U_{3})$  | -.37  | .38  | -.38 | .38  | -.16 | .18  | -.17 | .17  | -.26 | .36  | -.30 | .31  | -.13 | .25  | -.15 | .19  |

Notes: Bootstrap standard deviation estimators $\hat{s}_d$ were computed using 50 bootstrap samples and their results are based on 500 replicated samples. In Population II the percentages of replicated samples in which the estimators $\hat{s}_d$ were not computed because of convergence problems were 0.8%, 8.2% and 8.8% in $U_1$, $U_2$ and $U_3$, respectively, whereas in Population I the respective percentages were all 0%. Results on standard deviation estimators derived under the homogeneity assumption are based on 5000 replicated samples and no convergence problems occurred.
Table 6. Coverage probabilities and means and medians of the relative lengths of the 95% confidence intervals for the totals and means.

| Population | I | II |
|------------|---|----|
| Response variable | Continuous | Binary | Continuous | Binary |
| 95% CI | cp | mrl | mdrl | cp | mrl | mdrl | cp | mrl | mdrl |
| HTLEs CI | CI(\(\hat{Y}_{HT,I}^{(U)}\)) | .96 | .29 | .28 | .96 | .32 | .32 | .42 | .24 | .23 | .53 | .29 | .28 |
| CI(\(\hat{Y}_{HT,I}^{(U)}\)) | .90 | 1.1 | .81 | .98 | 1.4 | 1.0 | .28 | .56 | .33 | .63 | .82 | .49 |
| CI(\(\hat{Y}_{HT,I}^{(U)}\)) | .96 | .35 | .31 | .98 | .38 | .35 | .27 | .25 | .21 | .47 | .30 | .26 |
| HKLEs CI | CI(\(\hat{Y}_{HK,I}^{(U)}\)) | .80 | .34 | .33 | .66 | .38 | .37 | .74 | .28 | .28 | .78 | .33 | .33 |
| CI(\(\hat{Y}_{HK,I}^{(U)}\)) | .98 | 4.1 | 1.6 | .99 | 6.4 | 2.0 | .45 | 1.3 | .43 | .75 | 1.9 | .60 |
| CI(\(\hat{Y}_{HK,I}^{(U)}\)) | .94 | 1.1 | .50 | .79 | 1.3 | .54 | .64 | .44 | .26 | .75 | .53 | .31 |
| Homo. CI | CI(\(\hat{Y}_{ML,I}^{(H)}\)) | .00 | .10 | .10 | .01 | .14 | .14 | .09 | .12 | .11 | .10 | .15 | .14 |
| CI(\(\hat{Y}_{ML,I}^{(H)}\)) | .00 | .21 | .20 | .32 | .32 | .32 | .10 | .25 | .20 | .26 | .38 | .32 |
| CI(\(\hat{Y}_{ML,I}^{(H)}\)) | .00 | .09 | .09 | .00 | .13 | .13 | .03 | .11 | .10 | .06 | .15 | .14 |
| HTLEs CI | CI(\(\hat{Y}_{HT,I}^{(U)}\)) | .95 | .10 | .09 | .96 | .19 | .19 | .19 | .08 | .07 | .55 | .18 | .18 |
| CI(\(\hat{Y}_{HT,I}^{(U)}\)) | .98 | .57 | .52 | .96 | .79 | .75 | .91 | .23 | .18 | .95 | .58 | .51 |
| CI(\(\hat{Y}_{HT,I}^{(U)}\)) | .98 | .33 | .21 | .96 | .41 | .29 | .38 | .11 | .08 | .84 | .22 | .19 |
Table 6. Continued.

| Population | Sampling rates | Response variable | Continuous | Binary | Continuous | Binary |
|------------|----------------|------------------|------------|--------|------------|--------|
|            | $f_1 = 0.5$    | $f_2 = 0.4$      |            |        |            |        |
| I          |                |                  |            |        |            |        |
| II         |                |                  |            |        |            |        |
| HKLEs CI   |                |                  | cp         | mrl    | mdrl       | cp     | mrl    | mdrl   |
|            | CI($\hat{F}_{HK,1}$) | .00   | .07   | .07   | .17   | .19   | .19   | .48   | .03   | .03   | .96   | .17   | .17   |
|            | CI($\hat{F}_{HK,2}$) | .00   | .13   | .12   | .48   | .55   | .54   | .77   | .07   | .07   | .84   | .47   | .44   |
|            | CI($\hat{H}_{HK}$) | .00   | .08   | .08   | .08   | .21   | .20   | .30   | .04   | .04   | .92   | .18   | .17   |
| Homo. CI   |                |                  | cp         | mrl    | mdrl       | cp     | mrl    | mdrl   |
|            | CI($\hat{F}_{ML,1}$) | .00   | .00   | .04   | .00   | .13   | .13   | .17   | .03   | .03   | .91   | .14   | .13   |
|            | CI($\hat{F}_{ML,2}$) | .00   | .08   | .08   | .11   | .39   | .39   | .66   | .06   | .06   | .75   | .40   | .38   |
|            | CI($\hat{H}_{ML}$) | .00   | .04   | .04   | .00   | .13   | .13   | .13   | .03   | .03   | .86   | .14   | .13   |

Notes: Results on confidence intervals derived under the assumption of heterogeneous link probabilities are based on 500 samples. In Population II the percentages of replicated samples in which the CIs were not computed because of convergence problems were 0.8%, 8.2% and 8.8% in $U_1$, $U_2$ and $U$, respectively, whereas in Population I the respective percentages were all 0%. Results on CIs derived under the homogeneity assumption are based on 5,000 replicated samples and no convergence problems occurred.
representative sample of more than 90,000 adolescents who in the years 1994—1995 were in grades 7–12 in the United States. The participants were followed through adolescence and the transition to adulthood with the goal of helping to explain the causes of adolescent health and health behavior. The sample of students was selected by a stratified probability proportional to size cluster sampling design, where the clusters were the high schools and the strata were defined in terms of region, urbanicity, school type and so on. For each of the 84 selected high schools, one of its feeder middle schools was selected with probability proportional to the number of contributed students to the high school.

Each student in the representative sample was asked to named up to five male and five female friends within his or her high school or in the feeder school, and in addition, to complete an in-school questionnaire. Thus, the collected information can be modeled as a directed network, where the nodes are the sampled students and their referred friends, and a directed arc from node i to node j is considered to exist if student i names student j as a friend. See Harris (2013) for a description of this study.

A subset of the data obtained in the Add Health study is contained in Linton Freeman’s web page: See Freeman. In our numerical study we used data from this subset corresponding to the high school and its feeder in Community 50 to construct a population $U$ of $\tau = 2,497$ elements divided into subpopulations $U_1$ and $U_2$ of sizes $\tau_1 = 1,800$ and $\tau_2 = 697$, respectively. The elements assigned to $U_1$ were those at positions labeled with odd numbers in the data file and that named at least one friend plus a simple random sample of elements at positions labeled with even numbers and that named at least one friend. These elements were grouped into $N = 150$ clusters of sizes $m_i$, $i = 1, \ldots, N$, obtained by sampling from a negative binomial distribution with mean and variance equal to 12 and 24, respectively. The elements assigned to $U_2$ were the remaining elements in the data file that were named as a friend by at least one element in $U_1$. Once the subpopulation $U_k$ was constructed, the $N \times \tau_k$ matrix $X_k$ of values $x_{ij}^{(k)}$s of the link indicator variables $X_{ij}^{(k)}$s was constructed, $k = 1, 2$. We considered as response variables the following: “Number of friends” (named by each element) and “Sex” (1 = male, 0 = female). The totals and means of the variable “Number of friends” were $(Y_1, Y_2, Y) = (10,101, 2,729, 12,830)$ and $(\bar{Y}_1, \bar{Y}_2, \bar{Y}) = (5.612, 3.915, 5.138)$, and those of the variable “Sex” were $(Y_1, Y_2, Y) = (838, 361, 1,200)$ and $(\bar{Y}_1, \bar{Y}_2, \bar{Y}) = (0.466, 0.518, 0.481)$. For an initial sample size $n = 20$, which was the size used in this study, the values of the Pearson correlation coefficients between the values of the variable “Number of friends” and those of the inclusion probabilities associated with subpopulations $U_1$ and $U_2$ were $\rho(y^{(1)}, \pi^{(1)}) \approx 0.36$ and $\rho(y^{(2)}, \pi^{(2)}) \approx 0.29$, respectively, whereas the corresponding values for the variable “Sex” were $\rho(y^{(1)}, \pi^{(1)}) \approx -0.04$ and $\rho(y^{(2)}, \pi^{(2)}) \approx -0.07$.

The Monte Carlo study was carried out as in the previous study, except that the size of the initial sample of clusters was $n = 20$ and that for each selected value $m_i$, the values $x_{ij}^{(k)}$s of the link indicator variables $X_{ij}^{(k)}$, $j = 1, \ldots, \tau_k$ were obtained from the matrix $X_k$, $k = 1, 2$. Furthermore, for each element $j \in U_k$ that was sampled, the values $y_j^{(k)}$s of both response variables were recorded.

The results about the inferences on the population sizes are shown in Figure 4 and in Table 7. We can see that the UMLES performed well: they did not have problems of bias or instability. The estimators of their standard deviations presented problems of bias, but the confidence intervals based on these estimators performed acceptably. The HTLEs also
had acceptable performance, although the magnitudes of their biases were slightly greater than those of the UMLEs, and these biases affected the coverage probabilities of the intervals based on these estimators. Finally, the MLEs derived under the homogeneity assumption had large biases that affected their performance and that of the confidence intervals based on these estimators. Thus, in summary, the UMLEs had the best performance, which was good; the HTLEs performed acceptably, whereas the MLEs derived under the homogeneity assumption performed poorly.

The results of the point estimators of the totals and means and of the estimators of their standard deviations are shown in Table 8 and Figure 5. We can see that the HTLEs performed well regardless of the response variable: they did not present problems of bias or instability. In the case of the variable “Sex” the magnitudes of their biases were slightly greater than in the case of the variable “Number of friends”. The HKLEs also performed well, although slightly less well than the HTLEs, except in the estimation of the mean of the variable “Sex”, where the HKLEs performed better than the HTLEs. The HTLEs of the totals derived under the homogeneity assumption presented problems of underestimation, although in the case of the variable “Number of friends” the magnitudes of their biases were not very great. These estimators performed well in the estimation of the mean, especially of the variable “Sex”, where these estimators, along with the HKLEs, had the best performance. The estimators of the standard deviations of the point estimators considered in this study presented problems of biases, with some exceptions. However, the magnitudes of their biases were not so great.

The results of the confidence intervals for the totals and means are shown in Table 9. We can see that the confidence intervals for the totals of the variable “Number of friends” based on the HTLEs performed well. In the case of the variable “Sex” the coverage probabilities of these intervals were somewhat low. The intervals for the totals based on the HKLEs had
Table 7. Simulation results obtained for the point estimators, standard deviation estimators and 95% confidence intervals for the sizes in a population constructed using data from the national longitudinal study of adolescent health.

|                  | Point estimators | Standard deviation estimators | Confidence intervals |
|------------------|------------------|-------------------------------|----------------------|
|                  | rbias            | √rmse                        | mdre                 | mdare |
| UMLEs of sizes   |                  |                               |                      |       |
| \( \hat{\tau}_1^{(U)} \) | -.01             | .06                           | -.01                 | .04   |
| \( \hat{\tau}_2^{(U)} \) | .06              | .25                           | .00                  | .13   |
| \( \hat{\tau}_2^{(U)} \) | .01              | .08                           | .00                  | .05   |
| HTLEs of sizes   |                  |                               |                      |       |
| \( \hat{\tau}_{HT,1}^{(U)} \) | -.04             | .07                           | -.04                 | .05   |
| \( \hat{\tau}_{HT,2}^{(U)} \) | -.05             | .15                           | -.07                 | .10   |
| \( \hat{\tau}_{HT,2}^{(U)} \) | -.05             | .08                           | -.05                 | .05   |
| Homo. MLEs of sizes |     |                               |                      |       |
| \( \hat{\tau}_{ML,1}^{(H)} \) | -.14             | .15                           | -.14                 | .14   |
| \( \hat{\tau}_{ML,2}^{(H)} \) | -.24             | .25                           | -.24                 | .24   |
| \( \hat{\tau}_{ML,2}^{(H)} \) | -.17             | .18                           | -.17                 | .17   |

Notes: Results for point estimators are based on 5,000 samples; those for standard deviation estimators and confidence intervals derived under the heterogeneity assumption are based on 500 samples and those under the homogeneity assumption are based on 5,000 samples. Average sampling fractions are \( f_1 = 0.46 \) and \( f_2 = 0.40 \). No convergence problems were observed, except in one sample in which the estimators \( \hat{\tau}_2^{(U)} \) and \( \hat{\tau}_{HT,2}^{(U)} \) could not be computed.
Table 8. Simulation results obtained for the point and standard deviation estimators of totals and means in a population constructed using data from the National Longitudinal Study of Adolescent Health.

| Response variable | Point estimators |                           |                           |                           | Standard deviation estimators |                           |                           |
|------------------|------------------|---------------------------|---------------------------|---------------------------|-----------------------------|---------------------------|---------------------------|
|                  |                  | Number of friends         | Sex                       |                           | Number of friends            | Sex                       |                           |
|                  | Estimator        | rbias | √rmse | mdre | mdare | rbias | √rmse | mdre | mdare | rbias | √rmse | mdre | mdare | rbias | √rmse | mdre | mdare |
|                  |                  |       |       |      |       |       |       |      |       |       |       |      |       |       |       |      |       |
| HTLEs of totals  | \( \hat{Y}_{HT,1}^{(U)} \) | .00  | .06   | .01  | .04   | -.07  | .09   | -.07 | .07   | -.14  | .22   | -.16  | .17   | -.15  | .22   | -.16  | .17   |
|                  | \( \hat{Y}_{HT,2}^{(U)} \) | .07  | .17   | .05  | .09   | -.06  | .16   | -.08 | .11   | .12   | .75   | -.10  | .31   | .19   | .81   | -.04  | .29   |
|                  | \( \hat{Y}_{HT}^{(U)} \) | .02  | .06   | .02  | .04   | -.07  | .09   | -.07 | .07   | -.07  | .36   | -.16  | .21   | -.02  | .45   | -.13  | .23   |
| HKLEs of totals  | \( \hat{Y}_{HK,1}^{(U)} \) | .04  | .08   | .04  | .05   | -.03  | .08   | -.03 | .05   | -.15  | .23   | -.17  | .18   | -.16  | .23   | -.18  | .19   |
|                  | \( \hat{Y}_{HK,2}^{(U)} \) | .19  | .33   | .13  | .15   | .04   | .24   | -.01 | .13   | .05   | 1.2   | .30   | .43   | .12   | 1.3   | .26   | .40   |
|                  | \( \hat{Y}_{HK}^{(U)} \) | .08  | .11   | .07  | .07   | -.01  | .09   | -.02 | .06   | .03   | .89   | -.22  | .30   | .06   | .96   | -.21  | .31   |
| Homo. HTLEs      | \( \hat{Y}_{ML,1}^{(H)} \) | -.09 | .11   | -.09 | .09   | -.16  | .17   | -.16 | .16   | -.20  | .21   | -.19  | .19   | -.22  | .23   | -.22  | .22   |
| totals           | \( \hat{Y}_{ML,2}^{(H)} \) | -.13 | .16   | -.14 | .14   | -.25  | .26   | -.25 | .25   | -.17  | .23   | -.19  | .20   | -.19  | .25   | -.21  | .22   |
|                  | \( \hat{Y}_{ML}^{(H)} \) | -.10 | .11   | -.10 | .10   | -.19  | .20   | -.19 | .19   | -.28  | .28   | -.27  | .27   | -.31  | .31   | -.31  | .31   |
| HTLEs of means   | \( \hat{Y}_{HT,1}^{(U)} \) | .01  | .02   | .01  | .02   | -.06  | .07   | -.06 | .06   | -.15  | .19   | -.15  | .16   | -.03  | .13   | -.03  | .09   |
|                  | \( \hat{Y}_{HT,2}^{(U)} \) | .03  | .10   | .04  | .08   | -.10  | .13   | -.09 | .09   | -.14  | .42   | -.27  | .32   | .03   | .42   | -.10  | .24   |
|                  | \( \hat{Y}_{HT}^{(U)} \) | .01  | .08   | .02  | .03   | -.07  | .06   | -.07 | .07   | -.09  | .70   | -.33  | .40   | .10   | .67   | -.09  | .20   |
Table 8. Continued.

| Response variable | Point estimators |             |             |             | Standard deviation estimators |             |             |             |
|-------------------|------------------|-------------|-------------|-------------|-------------------------------|-------------|-------------|-------------|
|                   | Number of friends| Sex         |             |             | Number of friends             | Sex         |             |             |
|                   | rbias            | $\sqrt{\text{rmse}}$ | mdre | mdare | rbias            | $\sqrt{\text{rmse}}$ | mdre | mdare | rbias            | $\sqrt{\text{rmse}}$ | mdre | mdare |
| HKLEs             |                  |             |             |             |                 |             |             |             |
| of means          | $\hat{Y}_{HK,1}$ | .05         | .05         | .05         | -.02            | .04         | -.02 | .03 | -.12            | .16         | -.13 | .13 |
|                   | $\hat{Y}_{HK,2}$ | .13         | .13         | .13         | -.02            | .04         | -.02 | .03 | .01            | .16         | .00  | .10 |
|                   | $\hat{Y}_{HK}$   | .07         | .07         | .07         | -.02            | .03         | -.02 | .02 | -.09            | .29         | -.15 | .20 |
| Homo.             |                  |             |             |             |                 |             |             |             |
|                | $\hat{Y}_{ML,1}$ | .06         | .06         | .06         | -.02            | .04         | -.02 | .03 | -.18            | .19         | -.18 | .18 |
|                | $\hat{Y}_{ML,2}$ | .14         | .14         | .14         | -.02            | .04         | -.02 | .03 | -.10            | .12         | -.11 | .11 |
|                | $\hat{Y}_{ML}$   | .08         | .02         | .08         | -.02            | .03         | -.02 | .03 | -.25            | .26         | -.26 | .26 |
| HTLEs             |                  |             |             |             |                 |             |             |             |
| means             | $\hat{Y}_{ML,1}$ | .06         | .06         | .06         | -.02            | .04         | -.02 | .03 | -.18            | .19         | -.18 | .18 |
|                   | $\hat{Y}_{ML,2}$ | .14         | .14         | .14         | -.02            | .04         | -.02 | .03 | -.10            | .12         | -.11 | .11 |
|                   | $\hat{Y}_{ML}$   | .08         | .02         | .08         | -.02            | .03         | -.02 | .03 | -.25            | .26         | -.26 | .26 |

Notes: Results for point estimators are based on 5,000 samples; those for standard deviations estimators derived under the heterogeneity assumption are based on 500 samples and those the homogeneity assumption are based on 5,000 samples. Average sampling fractions are $f_1 = 0.46$ and $f_2 = 0.40$. No convergence problems were observed, except in one sample in which the estimators of $Y_2$ and $\bar{Y}$ derived under the heterogeneity assumption could not be computed.
slightly low coverage probabilities, regardless of the response variable. In addition, the lengths of the intervals for $Y_2$ were very large. The intervals for the totals based on the HTLEs derived under the homogeneity assumption showed very low coverage probabilities, especially in the case of the variable “Sex”, where the values of the probabilities were about 0.1. In the case of the intervals for the means, we have that the intervals based on the HTLEs derived under the heterogeneity assumption presented coverage probabilities that were somewhat low in the case of the variable “Number of friends”, and clearly low in the case of the variable “Sex”, except for the interval for $\hat{Y}_2$. These low probabilities were a consequence of the moderate biases exhibited by these estimators. The confidence intervals based on the HKLEs showed very low coverage probabilities in the case of the variable “Number of friends”, which was also a consequence.

Fig. 5. Boxplots of standardized values of the estimators of totals and means in the population constructed using data from the National Longitudinal Study of Adolescent Health.
| Response variable | CIs based on HTLEs of totals | CIs based on HKLEs of totals | CIs based on homogeneous HTLEs of totals |
|-------------------|--------------------------------|--------------------------------|------------------------------------------|
|                   | \( \hat{Y}_{HT,1} \)       | \( \hat{Y}_{HT,2} \)       | \( \hat{Y}_{HT} \)                       |
| Number             | cp                             | .90                           | .97                                      | .94                                      | .84                                      | .98                                      | .88                                      | .42                                      | .49                                      | .28                                      |
|                   | mrl                            | .20                           | .68                                      | .23                                      | .22                                      | 1.1                                      | .33                                      | .16                                      | .28                                      | .14                                      |
|                   | mdrf                           | .20                           | .55                                      | .21                                      | .22                                      | .71                                      | .25                                      | .16                                      | .27                                      | .14                                      |
| Sex               | cp                             | .73                           | .85                                      | .74                                      | .84                                      | .92                                      | .88                                      | .17                                      | .25                                      | .02                                      |
|                   | mrl                            | .22                           | .66                                      | .26                                      | .24                                      | 1.0                                      | .37                                      | .17                                      | .25                                      | .14                                      |
|                   | mdrf                           | .22                           | .54                                      | .23                                      | .23                                      | .68                                      | .28                                      | .17                                      | .25                                      | .14                                      |
|                   | Number of friends              | \( \hat{Y}_{HT,1} \)       | \( \hat{Y}_{HT,2} \)       | \( \hat{Y}_{HT} \)                       |
|                   | \( Y_{HT,1} \)                 | \( Y_{HT,2} \)               | \( Y_{HT} \)                         |
|                   | \( Y_{HK,1} \)                 | \( Y_{HK,2} \)               | \( Y_{HK} \)                         |
|                   | \( Y_{ML,1} \)                 | \( Y_{ML,2} \)               | \( Y_{ML} \)                         |
|                   | \( Y_{HT,1} \)                 | \( Y_{HT,2} \)               | \( Y_{HT} \)                         |
|                   | \( Y_{HK,1} \)                 | \( Y_{HK,2} \)               | \( Y_{HK} \)                         |
|                   | \( Y_{ML,1} \)                 | \( Y_{ML,2} \)               | \( Y_{ML} \)                         |

Notes: Results for confidence intervals derived under the heterogeneity assumption are based on 500 samples and those derived under the homogeneity assumption are based on 5,000 samples. Sampling fractions are \( f_1 = 0.46 \) and \( f_2 = 0.40 \). No convergence problems were observed.
of the biases exhibited by these estimators (see Figure 5). However, in the case of the variable “Sex” the coverage probabilities of these intervals were just somewhat low. The intervals based on the HTLEs derived under the homogeneity assumption performed very similarly to those based on the HKLEs, that is, they showed very low coverage probabilities in the case of the variable “Number of friends” and somewhat low coverage probabilities in the case of the variable “Sex”, although in both cases the coverage probabilities were slightly lower than those of the intervals based on the HKLEs. Thus, in summary, the best intervals for the total and mean of the variable “Number of friends” were the ones based on the HTLEs. Their performance was good in the case of the total and acceptable in the case of the mean. On the other hand, the best intervals for the total and mean of the variable “Sex” were the ones based on the HKLEs, and their performance was acceptable.

7. Conclusions and Suggestions for Future Research

In this article, we have considered the link-tracing sampling variant proposed by Félix-Medina and Thompson (2004) and have proposed Horvitz-Thompson-like and Hájek-like estimators of population totals and means. This work extends that of Félix-Medina and Monjadin (2010) by assuming heterogeneous, rather than homogeneous, link probabilities which are modeled by a Rasch model used by Félix-Medina et al. (2015). The variances of the proposed estimators are estimated by a variant of bootstrap which extends the variant used by the previously cited authors by estimating the variances of estimators of totals and means, in addition to the variances of estimators of population sizes. This variant of bootstrap allows the estimation of variances when the response variable is either continuous, discrete or binary. Large sample confidence intervals for the population parameters are constructed by assuming that the estimators of the parameters have normal distributions, except for the estimators of the population sizes, which are assumed to have log-normal distributions. In addition, confidence intervals for proportions are constructed using Korn and Graubard’s (1998) proposal.

We evaluated the performance of the proposed estimators by means of two Monte Carlo studies. In the first study, two populations were constructed using simulated data, and the results show that erroneous inferences might be obtained if some model assumptions were not satisfied. In particular, we found that if any of the assumptions are satisfied, then reliable inferences about population sizes, totals and means are obtained. Furthermore, we found that the assumption of the Poisson distribution of the sizes $M_i$ of the venues $A_i$ does not need to be satisfied to obtain reliable inferences. Nevertheless, we also found out that severe deviations from the Rasch model of the link probabilities lead to erroneous inferences, and that inferences about population sizes and totals are affected in greater extent than inferences about population means. In fact, inferences about the population means seem to be robust to deviations from the assumed models. In addition, we came upon that in any situation, the performance of the proposed bootstrap variance estimators is at most just good. However, in the second study, in which a finite population was constructed using data from the Add Health study, the results are promising. Thus, if this population were more or less representative of the populations that could be found in applications of this methodology, then reliable inferences would be expected to be obtained.
In the light of these results, we consider that the following issues are worthy of future research: (1) To develop a model for the link probabilities \( p_{ij}^{(k)} \)’s that is robust to deviations from the assumed model. For instance, to model \( p_{ij}^{(k)} \) as a quadratic function of \( \alpha_{ij}^{(k)} \) and \( \beta_{ij}^{(k)} \), or to assume that the \( \beta_{ij}^{(k)} \)’s are T-Student distributed instead of normally distributed, or to change the Rasch model by the latent classes model proposed by Pledger (2000) and which was used in the first Monte Carlo study to generate the values \( x_{ij}^{(k)} \)’s in Population II. (2) To improve the proposed bootstrap variance estimators. For example, to predict the values of the response variable associated with the non-sampled elements by using a quadratic or a nonparametric regression model instead of a simple linear regression model. (3) To enhance the proposed CIs for the population sizes, totals and means. For instance, using the bootstrap percentile method which does not require assuming a probability distribution for the estimator of the parameter of interest.

8. Appendix: Bootstrap Procedure

In this section, we will describe the bootstrap variant that is proposed to construct a sample of values of an estimator \( \hat{\theta} \) of a parameter \( \theta \) (population size, total or mean), from which an estimate of the variance and/or the standard deviation of \( \hat{\theta} \), as well as a confidence interval for \( \theta \) can be computed. Thus, hereinafter, we will denote by \( [x] \), the greatest integer less than or equal to \( x \in \mathbb{R} \). The steps of the proposed bootstrap procedure are as follows. (1) Construct a population vector \( m_{\text{Boot}} \) of \( N \) values of \( m_i \)s by means of the following procedure. If \( N/n \) is an integer, repeat \( N/n \) times the observed sample of \( n \) cluster sizes \( m_k = \{m_1, \ldots, m_n\} \). 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 \( m_k \) and add to this set a SRSWOR of \( b \) values of \( m_i \)s selected from \( m_k \). If the sum of the elements of the vector \( m_{\text{Boot}} \) is greater than the value \( t_{1}^*(a) \), \( a = U, C \), (depending of the type of estimator that is being considered), delete one element at a time from \( m_{\text{Boot}} \) starting from the \( N \)-th element until the sum is less than or equal to \( t_{1}^*(a) \). Let \( N_{\text{Boot}} \) be the final number of elements in \( m_{\text{Boot}} \). Note that this procedure for constructing the vector \( m_{\text{Boot}} \) avoids the assumption that the \( M_i \)s be independent and identically distributed Poisson random variables. Therefore, the resulting bootstrap variance estimates and confidence intervals are robust to deviations from this assumption. In addition, the condition that the sum of the elements in \( m_{\text{Boot}} \) be less than or equal to \( t_{1}^{(a)} \) guarantees that every bootstrap initial sample \( m_1', \ldots, m_n' \), satisfies \( \sum m_i' < t_{1}^{(a)} \), and consequently, that no initial bootstrap sample contains the whole bootstrap population \( U_{\text{Boot}} \) of size \( t_{1}^{(a)} \). See step (6) of the procedure. (2) For each \( k = 1, 2 \), construct a population vector \( \alpha_{(k)\text{Boot}}^{(a)} \) of length \( N_{\text{Boot}} \) whose elements are the estimates \( \hat{\alpha}_{ij}^{(a)} \)’s of the \( \alpha_{ij}^{(k)} \)’s associated with the clusters whose sizes \( m_i \)s are in \( m_{\text{Boot}} \). (3) For each \( k = 1, 2 \), construct a population vector \( \beta_{(k)\text{Boot}}^{(a)} \) of length \( \hat{\tau}_k^{(a)} \) whose first \( m + r_1 \) elements in the case of \( k = 1 \), or whose first \( r_2 \) elements in the case of \( k = 2 \), are the estimates \( \hat{\beta}_{(k)\text{Boot}}^{(a)} \)’s of the \( \beta_{ij}^{(k)} \)’s associated with the people in \( S_k^* \), and each one of the remaining elements is the estimate \( \hat{\beta}_{i(k)j}^{(a)} \) of \( \beta_{ij}^{(k)} \) obtained using Equation (7) with \( x_i = 0 \), \( i = 1, \ldots, n \). (4) For each \( k = 1, 2 \), construct a population vector \( \gamma_{(k)\text{Boot}}^{(a)} \) of length \( \hat{\tau}_k^{(a)} \) whose first \( m + r_1 \) elements in the case of \( k = 1 \), or whose first \( r_2 \) elements in the case of \( k = 2 \), are the \( y \)-values associated with the elements in \( S_k^* \), and the remaining elements are estimates of the \( y \)-values associated with the people in \( U_k - S_k^* \) and obtained using the following procedure. If the variable of interest \( y \) is continuous, then fit a simple linear
regression model to the data \( (\pi^{(a)}_{(k)}(\hat{\alpha}^{(a)}_k, \hat{\sigma}^{(a)}_k, \hat{\beta}^{(a)}_{(k)j}), y^{(k)}_j) j \in S^*_k \). Next, predict the y-value associated with the j-th element in \( U_k - S^*_k \) by using a value sampled from the normal distribution with mean equals to the quantity obtained by evaluating the fitted model at the estimate \( \pi^{(a)}_{(k)}(\hat{\alpha}^{(a)}_k, \hat{\sigma}^{(a)}_k, \hat{\beta}^{(a)}_{(k)j}) \) of the inclusion probability of an element in \( U_k - S^*_k \) and variance equals to the estimate of the variance of the error terms of the regression model. If the design matrix is numerically singular, then predict the y-value associated with \( j \in U_k - S^*_k \) by a value sampled from the normal distribution with mean and variance given by the sample mean and sample variance, respectively, of the y-values associated with the elements in \( S^*_k \). On the other hand, if the variable of interest y is binary, then fit a simple logistic regression model to the data \( (\pi^{(a)}_{(k)}(\hat{\alpha}^{(a)}_k, \hat{\sigma}^{(a)}_k, \hat{\beta}^{(a)}_{(k)j}), y^{(k)}_j) j \in S^*_k \). Next, predict the y-value associated with the j-th element in \( U_k - S^*_k \) by using a value sampled from the Bernoulli distribution with success probability equals to the quantity obtained by evaluating \( \pi^{(a)}_{(k)}(\hat{\alpha}^{(a)}_k, \hat{\sigma}^{(a)}_k, \hat{\beta}^{(a)}_{(k)j}) \). If the design matrix is numerically singular, then predict the y-value associated with \( j \in U_k - S^*_k \) by a value sampled from the Bernoulli distribution with success probability equals to the sample mean of the y-values of the elements in \( S^*_k \). (5) Select a SRSWOR of \( n \) values \( m_i \) from \( m_{Boot} \). Let \( S^{Boot}_A = \{i_1, \ldots, i_n\} \) be the set of indices of the \( m_i \) in the sample. In addition, let \( A^{Boot}_i = (\sum_{t=1}^{i-1}m_t, \sum_{t=1}^{i}m_t) \cap \mathbb{Z} \) be the set of indices \( j \) associated with the elements in the cluster whose index is \( i \in S^{Boot}_A \), where \( m_i \) is the t-th element of \( m_{Boot} \) and \( \mathbb{Z} \) is the set of the integer numbers. Finally, let \( S^{Boot}_0 = \bigcup_{i \in S^{Boot}_A} A^{Boot}_i \) be the set of indices \( j \) associated with the elements in the clusters whose indices are in \( S^{Boot}_A \). (6) For each \( k = 1, 2, i \in S^{Boot}_A \) and \( j \in \{1, \ldots, [\pi^{(a)}_{(1)}]\} - A^{Boot}_i \) in the case of \( k = 1 \), or \( j \in \{1, \ldots, [\pi^{(a)}_{(2)}]\} \) in the case of \( k = 2 \), generate a value \( y^{(k)}_{ij} \) by sampling from the Bernoulli distribution with success probability equals to the value obtained by evaluating (1) at the \( i \)-th element of the vector \( \hat{\alpha}^{(a)}_{(k)Boot} \) and the \( j \)-th element of the vector \( \hat{\beta}^{(a)}_{(k)Boot} \). (7) Compute the estimates of the sizes \( \tau_1, \tau_2 \) and \( \tau \), those of the totals \( Y_1, Y_2 \) and \( Y \), and those of the means \( \bar{Y}_1, \bar{Y}_2 \) and \( \bar{Y} \) using the same procedure as that used to compute the original estimates. (8) Repeat the steps (5)-(7) a large enough number \( B \) of times.

It is worth noting that in the proposed variant of bootstrap, we are using a simple linear regression model or a simple logistic regression model to roughly approximate a potential relation between the y-values and the estimated inclusion probabilities. This does not mean that there exists such a relation. However, if it existed, we would expect that the employed simple regression models yield predicted y-values that allow us to compute acceptable variance estimators and confidence intervals. Note that if there were not a relation, the predicted y-values would be basically obtained, in the case of a quantitative response variable, by sampling from a normal distribution with mean and variance equal to the sample mean and variance of the observed y-values, and in the case of a binary response variable, by sampling from a Bernoulli distribution with mean equals to the sample mean of the y-values.

9. References

Abramowitz, M., and I.A. Stegun. 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Ninth Dover printing, tenth GPO printing. New York: Dover.
Bernard, H.R., T. Hallett, A. Iovita, E.C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M.J. Salganik, T. Saliu, O. Scutelniciu, G.A. Shelley, P. Sirinirund, S. Weir, and D.F. Stroup. 2010. “Counting hard-to-count populations: The network scale-up method for public health.” *Sexually Transmitted Infections* 86 (Suppl. 2): 11–15. DOI: http://dx.doi.org/10.1136/sti.2010.044446.

Booth, J.G., R.W. Butler, and P. Hall. 1994. “Bootstrap methods for finite populations.” *Journal of the American Statistical Association* 89: 1282–1289. DOI: https://doi.org/10.1080/01621459.1994.10476868.

Burnham, K.P., and W.S. Overton. 1978. “Estimation of the size of a closed population when capture probabilities vary among animals.” *Biometrika* 65: 625–633. DOI: https://doi.org/10.1093/biomet/65.3.625.

Chao, A. 1987. “Estimating the population size for capture-recapture data with unequal catchability.” *Biometrics* 43: 783–791. DOI: https://doi.org/10.2307/2531532.

Cheng, S., D.J. Eck, and F.W. Crawford. 2020. “Estimating the size of a hidden finite set: Large-sample behavior of estimators.” *Statistics Surveys* 14: 1–31. DOI: https://doi.org/10.1214/19-SS127.

Chow, M., and S.K. Thompson. 2003. “Estimation with link-tracing sampling designs – A Bayesian approach.” *Survey Methodology* 29: 197–205. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9553-eng.pdf?st=6cXXjDD2 (accessed April 2020).

Coull, B.A., and A. Agresti. 1999. “The use of mixed logit models to reflect heterogeneity in capture-recapture studies.” *Biometrics* 55: 294–301. DOI: https://doi.org/10.1111/j.0006-341X.1999.00294.x.

Crawford, F.W., J. Wu, and R. Heimer. 2018. “Hidden population size estimation from respondent-driven sampling: A network approach.” *Journal of the American Statistical Association* 113: 755–766. DOI: https://doi.org/10.1080/01621459.2017.1285775.

Dávid, B., and T.A.B. Snijders. 2002. “Estimating the Size of the Homeless Population in Budapest, Hungary.” *Quality & Quantity* 36: 291–303. DOI: https://doi.org/10.1023/A:1016080606287.

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

Dombrowski, K., B. Khan, T. Wendel, K. McLean, E. Misshula, and R. Curtis. 2012. “Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques.” *Advances in Applied Sociology* 2: 245–252. DOI: https://doi.org/10.4236/aasoci.2012.24032.

Félix-Medina, M.H., and P.E. Monjardin. 2006. “Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A Bayesian assisted approach.” *Survey Methodology* 32: 187–195. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9553-eng.pdf?st=6cXXjDD2 (accessed April 2020).

Félix-Medina, M.H., and P.E. Monjardin. 2010. “Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations.” *Journal of Official Statistics* 26(4): 603–631. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bb455be7f3/20200206/felix-medina.pdf (accessed April 2020).
Félix-Medina, M.H., P.E. Monjardin, and A.N. Aceves-Castro. 2015. “Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities.” *Survey Methodology*: 349–376. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14238-eng.pdf?st=ki7rx2GZ (accessed April 2020).

Félix-Medina, M.H., and S.K. Thompson. 2004. “Combining Cluster Sampling and Link-Tracing Sampling to Estimate the Size of Hidden Populations.” *Journal of Official Statistics* 20(1): 19–38. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/combining-link-tracing-sampling-and-cluster-sampling-to-estimate-the-size-of-hidden-populations.pdf (accessed April 2020).

Fienberg, S.E., M.S. Johnson, and B.W. Junker. 1999. “Classical multilevel and Bayesian approaches to population size estimation using multiple lists.” *Journal of the Royal Statistical Society. Series A* 162: 383–405. DOI: https://doi.org/10.1111/1467-985X.00143.

Frank, O., and T. Snijders. 1994. “Estimating the Size of Hidden Populations Using Snowball Sampling.” *Journal of Official Statistics* 10 (1): 53–67. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/estimating-the-size-of-hidden-populations-using-snowball-sampling.pdf (accessed April 2020).

Freeman L.C. (n.d.) Network data sets repository. Available at: http://moreno.ss.uci.edu/data (accessed June 2018).

Giner, G., and G.K. Smyth. 2016. “statmod: Probability calculations for the inverse Gaussian distribution.” *The R Journal* 8: 339–351. DOI: https://doi.org/10.32614/RJ-2016-024.

Handcock, M.S., K.J. Gile, and C.M. Mar. 2014. “Estimating hidden population size using respondent-driven sampling data.” *Electronic Journal of Statistics* 8: 1491–1521. DOI: https://doi.org/10.1214/14-EJS923.

Harris, K.M. 2013. “The add health study: Design and accomplishments.” Available at: www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf (accessed September 2017).

Heckathorn, D.D. 1997. “Respondent driven sampling: a new approach to the study of hidden samples.” *Social Problems* 44: 174–199. DOI: https://doi.org/10.2307/3096941.

Heckathorn, D.D. 2002. “Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations.” *Social Problems* 49: 11–34. DOI: https://doi.org/10.1525/sp.2002.49.1.11.

Heckathorn, D.D., and C.J. Cameron. 2017. “Network sampling: From snowball and multiplicity to respondent-driven sampling.” *Annual Review of Sociology* 43: 101–119. DOI: https://doi.org/10.1146/annurev-soc-060116-053556.

Hwang, W.-H., and R. Huggins. 2005. “An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data.” *Biometrika* 92: 229–233. DOI: https://doi.org/10.1093/biomet/92.1.229.

Johnston, L.G., D. Prybylski, H.F. Raymond, A. Mirzazadeh, C. Manopaiboon, and W. McFarland. 2013. “Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: Case studies from around the world.” *Sexually Transmitted Diseases* 40: 304–310. DOI: https://doi.org/10.1097/OLQ.0b013e31827fd650.
Kalton, G. 2009. “Methods for oversampling rare populations in social surveys.” *Survey Methodology* 35: 125–141. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11036-eng.pdf?st=PyvQkTH4 (accessed April 2020).

Khan, B., H.-W. Lee, I. Fellows, and K. Dombrowski. 2018. “One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity.” *PLoS ONE* 13 (4): 1–39. DOI: https://doi.org/10.1371/journal.pone.0195959.

Killworth, P., E. Johnsen, C. McCarty, G. Shelley, and H. Bernard. 1998a. “A social network approach to estimating seroprevalence in the United States.” *Social Networks* 20: 23–50. DOI: https://doi.org/10.1016/S0378-8733(96)00305-X.

Killworth, P., C. McCarty, H. Bernard, G. Shelley, and E. Johnsen. 1998b. “Estimation of seroprevalence, rape and homelessness in the United States using a social network approach.” *Evaluation Review* 22: 289–308. DOI: https://doi.org/10.1177/0193841X9802200205.

Klodsvahl, A.S. 1989. “Urban Social Networks: Some Methodological Problems and Possibilities.” In *The Small World*, edited by M. Kochen, 176–210. Norwood, NJ: Ablex.

Korn, E.L., and B.I. Graubard. 1998. “Confidence intervals for proportions with small expected number of positive counts estimated from survey data.” *Survey Methodology* 24: 193–201.

Lee, S., J. Wagner, R. Valliant, and S. Heeringa. 2014. “Recent developments of sampling hard-to-survey populations: An assessment.” In *Hard-to-Survey Populations*, edited by R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, and N. Bates (Eds.), 424–444. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139381635.025.

MacKellar, D., L. Valleroy, J. Karon, G. Lemp, and R. Janssen. 1996. “The Young Men’s Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men.” *Public Health Reports* 111(Suppl. 1): 138–44. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1382056/pdf/pubhealthrep00044-0140.pdf (accessed April 2020).

Magnani, R., K. Sabin, T. Saidel, and D. Heckathorn. 2005. “Review of sampling hard-to-reach populations for HIV surveillance.” *AIDS* 19: S67–S72. DOI: https://doi.org/10.1097/01.aids.0000172879.20628.e1.

Maltiel, R., A.E. Raftery, T.H. McCormick, and A.J. Baraff. 2015. “Estimating population size using the network scale up method.” *The Annals of Applied Statistics* 9: 1247–1277. DOI: https://doi.org/10.1214/15-AOAS827.

Marpsat, M., and N. Razafindratsima. 2010. “Survey methods for hard-to-reach populations: Introduction to the special issue.” *Methodological Innovations Online* 5: 3–16. DOI: https://doi.org/10.4256/mio.2010.0014.

McCormick, T.H., M.J. Salganik, and T. Zheng. 2010. “How many people do you know? Efficiently estimating personal network size.” *Journal of the American Statistical Association* 105: 59–70. DOI: https://doi.org/10.1198/jasa.2009.ap08518.

Meng, V.Y., and P. Gustafsson. 2017. “Inferring population size: extending the multiplier method to incorporate multiple traits with a likelihood-based approach.” *Stat* 6: 4–13. DOI: https://doi.org/10.1002/sta4.131.
Pledger, S. 2000. “Unified maximum likelihood estimates for closed capture-recapture models using mixtures.” *Biometrics* 56: 434–442. DOI: https://doi.org/10.1111/j.0006-341X.2000.00434.x.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (accessed April 2020).

Reiser, M. 2019. “Goodness of fit testing in sparse contingency tables when the number of variables is large.” *Wiley Interdisciplinary Reviews: Computational Statistics* 11(6): e1470. DOI: https://doi.org/10.1002/wics.1470.

Sanathanan, L. 1972. “Estimating the size of a multinomial population.” *Annals of Mathematical Statistics* 43: 142–152. DOI: https://doi.org/10.1214/aoms/1177692709.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Spreen, M. 1992. “Rare populations, hidden populations and link-tracing designs: What and why?” *Bulletin de Méthodologie Sociologique* 36: 34–58. DOI: https://doi.org/10.1177/075910639203600103.

Spreen, M., and S. Bogaerts. 2015. “B-Graph Sampling to Estimate the Size of a Hidden Population.” *Journal of Official Statistics* 31: 723–736. DOI: https://doi.org/10.1515/jos-2015-0042.

Staudte, R.G., and S.J. Sheather. 1990. *Robust Estimation and Testing*. New York: Wiley.

St. Clair, K., and D. O’Connell. 2012. “A Bayesian model for estimating population means using a link-tracing sampling design.” *Biometrics* 68: 165–173. DOI: https://doi.org/10.1111/j.1541-0420.2011.01631.x.

Thompson, S.K. 2012. *Sampling, Third edition*. New Jersey: Wiley.

Thompson, S.K., and O. Frank. 2000. “Model-based estimation with link-tracing sampling designs.” *Survey Methodology* 26: 87–98. Available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000001/article/5181-eng.pdf?st=dkg9U6pj (accessed April 2020).

Tourangeau, R. 2014. “Defining hard-to-survey populations.” In *Hard-to-Survey Populations*, edited by R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, and N. Bates (Eds.), 3–20. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139381635.003.

UNAIDS/WHO (World Health Organization Working Group on Global HIV/AIDS, and STI Surveillance). 2010. *Guidelines on estimating the size of populations most at risk to HIV*. UNAIDS– Joint United Nations Programme on HIV/AIDS. Available at: https://data.unaids.org/pub/manual/2010/guidelines_popnestimationsize_en.pdf. (accessed June 2015).

Volz, E., and D. Heckathorn. 2008. “Probability based estimation theory for respondent driven sampling.” *Journal of Official Statistics* 24: 79–97. Available at: http://www.sverigesiffror.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/probability-based-estimation-theory-for-respondent-driven-sampling.pdf. (accessed April 2020).

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

Received June 2020
Revised April 2021
Accepted June 2021