Journal of Official Statistics, Vol. 20, No. 1, 2004, pp. 19–38

Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations

Martín H. Félix-Medina[1] *and Steven K. Thompson*[2]

We present a variant of Link-Tracing Sampling which avoids the ordinary assumption of an initial Bernoulli sample of members of the target population. Instead of that, we assume that a portion of the target population is covered by a sampling frame of accessible sites, such as households, street blocks, or block venues, and that a simple random sample of sites is selected from the frame. As in ordinary Link-Tracing sampling, the people in the initial sample are asked to nominate other members of the population, but in this case we trace only the links between the sampled sites and the nominees. Maximum likelihood estimators of the population size are presented, and estimators of their variances that incorporate the initial sampling design are suggested. The results of a simulation study carried out in this research indicate that our proposed design is effective provided that the nomination probabilities are not too small.

Key words: Capture-recapture; design-based approach; finite population; hard-to-access population; maximum likelihood; model-based approach; sampling frame.

1. Introduction

Link-tracing sampling (LTS) has been proposed as an appropriate methodology for sampling hidden and hard-to-access human populations, such as drug users, homeless persons or undocumented worker populations. The basic idea behind this sampling methodology is to start with an initial sample of people from the population of interest, and then to increase the sample size by asking the people in the initial sample to nominate other members of the population. The nominated people might in turn be asked to nominate other members of the population, and so forth until a specified stopping rule is satisfied. (See Spreen (1992), and Thompson and Frank (2000) for descriptions and reviews of different variants of this sampling methodology.) For example, in a study of injecting drug users in relation to the risk of HIV infection, a drug user often can refer researchers to injecting and sexual partners and others in the at-risk population, so that starting from an initial sample the sample can be built up by following these social links.

[1] Escuela de Ciencias Físico Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México
[2] Department of Statistics, 326 Thomas Building, The Pennsylvania State University, University Park, PA 16802-2111, U.S.A.

Acknowledgments: The authors would like to thank Professor A. Valdez for having allowed them to use the information on which was based one of the simulation studies presented in this article, and the referees for their constructive comments which improved the presentation of the article. This research was supported by grant R01 DA09872 of the National Institutes of Health, National Institute of Drug Abuse, and grant DMS-9626102 of the National Science Foundation. The first author also acknowledges the support from the Consejo Nacional de Ciencia y Tecnologia, Mexico.

© Statistics Sweden

For such studies, link-tracing sampling tends to produce a larger number of individuals from a hidden population, in comparison with other sampling designs.

An attractive characteristic of LTS is that it allows the researcher to make valid model-based inferences about a number of population parameters. For instance, model-based estimation of the population size has been considered by Frank and Snijders (1994). These authors have derived a number of estimators of the size of a hidden population from the following two assumptions: (i) The initial sample is a Bernoulli sample; that is, persons are independently included in the initial sample and with equal inclusion probabilities. (ii) People are independently nominated by the persons in the initial sample and the nominations are made with equal probabilities. Other models and inferences about other parameters have been considered by other authors; for a review see Thompson and Frank (2000).

Although valid model-based inferences can be made using LTS, one problem is that model assumptions may not be realistic. For example, in real studies the assumption (i) of Frank and Snijders (1994) is frequently violated because researchers often carry out the initial recruitment by using health centers or police stations, so that members of the hidden population may not be encountered independently or with equal probabilities.

In this article, we develop a variant of LTS which avoids the assumption of an initial Bernoulli sample. We do that by supposing that a portion of the population of interest is covered by a sampling frame of accessible sites where members of the population can be found with high probability. An initial sample of sites (clusters) is selected by using an ordinary cluster sampling design and, as in an ordinary LTS, persons in the initial sample are asked to nominate other members of the population. However, because the sites are the sampling units, instead of tracing links between initial responders and their nominees, we follow the links between the clusters in the initial sample and the people nominated from these clusters. Here, a person will be meant to be nominated by a cluster if any person in the cluster nominates him or her.

The structure of the article is as follows. In Section 2, we describe the proposed sampling design and present some of the notation to be used throughout the article. Next, in Section 3, we describe a design-based estimator of the size of the population covered by the sampling frame and which does not use the nomination information. In Section 4, we present two models for the nomination probabilities, and under each model we derive maximum likelihood estimators (MLE's) of the population size, as well as model-based and design-based estimators of their variances. Then, in Section 5, we describe the results of two simulation studies carried out to explore the performance of the proposed sampling strategy. Finally, in Section 6, we present some final remarks and some possible extensions to our proposal.

2. Sampling Design and Notation

Let $U = \{u_1, \ldots, u_\tau\}$ be a finite hidden-human population of unknown size τ. We will assume that a portion of the population can be found in accessible sites, such as work places, parks, hospitals, city-blocks, or households, and that a list of N of those accessible sites can be constructed. We will also assume that we are able to define an operational rule which allows us to determine whether or not a person belongs to one of the sites on the list,

and in the affirmative case, to which site that person belongs (a person can belong to only one site). Let U_1 be the portion of U covered by the sampling frame (list), and let τ_1 be its size. Let A_i be the i-th cluster (site) on the list and let m_i be the number of members of the population who belong to A_i, $i = 1, \ldots, N$, so that $\tau_1 = \sum_1^N m_i$. Let $U_2 = U - U_1$ be the portion of U not covered by the sampling frame, and let $\tau_2 = \tau - \tau_1$ be its size (see Figure 1).

The sampling design is as follows. By using a simple random sampling without replacement (SRSWOR) design a sample $S_0 = \{A_1, \ldots, A_n\}$ of n clusters is selected from the sampling frame. (Although we are using as subscripts the integers $1, \ldots, n$, this does not mean that the first n clusters in the frame are the clusters in the sample.) We will assume that each of the m_i persons who belong to $A_i \in S_0$ is identified. Thus, the number of people in S_0 is $m = \sum_1^n m_i$. Next, the persons who belong to the cluster $A_i \in S_0$ are asked to nominate other members of the population outside of A_i; that is, in $U - A_i$. This nomination procedure is carried out in every cluster $A_i \in S_0$, and we will say that a person is nominated by a cluster if at least one of the members of the cluster nominates him or her. We will assume that the nominations from different clusters are carried out independently, but we will not assume that the same nomination strategy is used in every cluster. (For instance, in cluster A_i, the m_i members, as a group, might be asked to nominate other members; whereas, in cluster A_j, each of the m_j members might be separately asked to nominate other members.) For each nominated person, we will assume that the following information is obtained: the clusters that nominated him or her, and whether that person belongs to a cluster in S_0, or to a nonsampled cluster (a cluster in $U_1 - S_0$), or to the portion not covered by the sampling frame (U_2) (see Figure 1).

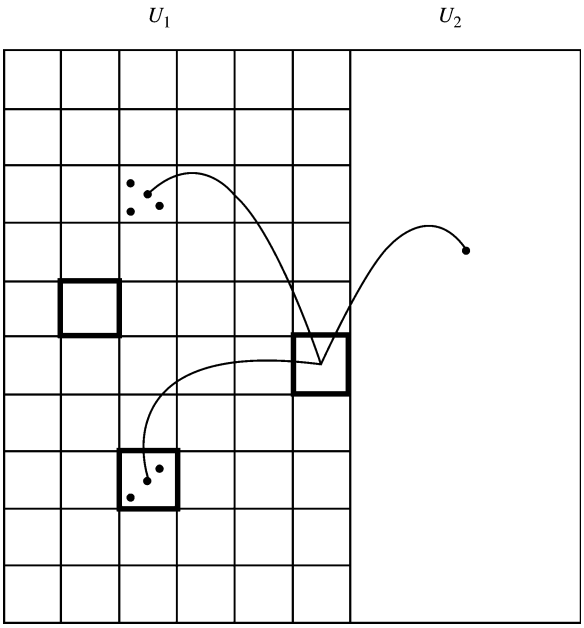

Fig. 1. Population U divided into U_1 and U_2. Bold squares represent sampled clusters. From a sampled cluster there might be three types of arcs: to a person in U_2, to a person in a sampled cluster ($A_i \in S_0$) *and to a person in $U_1 - S_0$*

It is worth noting that this sampling design resembles that of Multiple Capture-Recapture Sampling (MCRS). (See Otis et al. (1978) and the International Working Group for Disease Monitoring and Forecasting (1995a, b) for reviews of this methodology in the contexts of wildlife and human populations, respectively.) To see this, notice that in MCRS the population of interest is sampled on a specified number of occasions, and the elements captured (sampled) on any occasion are marked and then released to the population so that they can be captured on different occasions. Thus, a cluster in our sampling design corresponds to a sampling occasion in the context of MCRS. Similarly, the people nominated by a cluster correspond to the elements captured on a sampling occasion, and the probability that a person is nominated by a cluster corresponds to the probability that an element is captured on an occasion. Furthermore, models similar to those used in MCRS can be specified in our case, and consequently the estimators derived under those models will resemble those used in MCRS. However, in our design we have two additional complexities. The first one is that here the clusters are randomly selected, whereas in MCRS the sampling occasions are fixed. The second one is that here an initial sample of clusters is selected, and consequently a person can be included in the final sample if either he or she belongs to a sampled cluster or he or she is nominated from a sampled cluster, whereas in MCRS, an initial sample is not considered, and therefore an element is in the sample only if it is captured on a sampling occasion. Thus, these two factors introduce problems that are not found in MCRS.

We will end this section by introducing the matrix $\mathbf{x} = [x_{ij}]$ of indicator variables x_{ij}, where $x_{ij} = 1$ if person $u_j \in U$ is nominated by cluster A_i, and $x_{ij} = 0$ otherwise. Because we do not have a sampling frame of people, the labels of the individuals are not observable; consequently, the matrix $\mathbf{x}$ is known only up to permutations of its columns. For this reason, the x_{ij}'s will not be used for making inferences but only for defining models. Inferences will be based on the observable set of counts y_ω, $\omega \subseteq \Omega = \{1, \ldots, n\}$, of the people who are nominated by every sampled cluster A_i with i in the set $\omega \neq \emptyset$, but not otherwise. (For instance, if $\omega = \{1, 3, 9\}$, y_ω would be the number of people who are nominated by only A_1, A_3 and A_9.) The set of counts y_ω will be denoted by $\mathbf{y}$. Other variables will be used in this article, but they will be introduced as they are required.

3. A Design-based Estimator of τ_1

Because of the sampling design used to select the initial sample S_0, we have that $\check{\tau}_1 = Nm/n$ is a design-unbiased estimator of τ_1. The design-based variance of $\check{\tau}_1$ is

$$\mathbf{V}_p(\check{\tau}_1) = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N}\left(m_i - \frac{\tau_1}{N}\right)^2$$

and a design-unbiased estimator of $\mathbf{V}_p(\check{\tau}_1)$ is

$$\check{\mathbf{V}}_p(\check{\tau}_1) = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(m_i - \frac{\check{\tau}_1}{N}\right)^2$$

The estimators $\check{\tau}_1$ and $\check{\mathbf{V}}_p(\check{\tau}_1)$ have the attractive property of being free of model assumptions; that is, regardless of the stochastic process that generated the m_i's, $\check{\tau}_1$ and $\check{\mathbf{V}}_p(\check{\tau}_1)$ should be reasonable estimators of τ_1 and $\mathbf{V}_p(\check{\tau}_1)$, respectively. However, we do not

expect $\check{\tau}_1$ to be an efficient estimator (in terms of variance) of τ_1, because it does not incorporate the information about the nominations contained in the variables y_ω.

4. Maximum Likelihood Estimators of τ_1, τ_2, and τ

Our goal is to estimate τ_1, τ_2 and τ by using the information on both $\mathbf{m}_s = (m_1, \ldots, m_n)$ and $\mathbf{y}$. We will do that by assuming stochastic models for the distributions of these variables. Let us first consider the distribution of the cluster sizes. We will suppose that the number of persons m_i in A_i is a realization of a Poisson random variable M_i with mean λ, $i = 1, \ldots, N$, and that the M_i's are independently distributed. Although the assumed distribution for the M_i's might seem restrictive, we will later justify the robustness of the proposed likelihood estimators to deviations from the assumed model. To have τ_1 as a parameter, we will work with the conditional distribution of the M_i's given that $\sum_1^N M_i = \tau_1$. By a well-known property of the Poisson distribution, we have that the conditional joint distribution of $\mathbf{M}_s = (M_1, \ldots, M_n)$, given that $\sum_1^N M_i = \tau_1$, is a multinomial distribution with parameters τ_1 and $\{1/N\}_1^n$ [which will be denoted by $\text{Mult}(\tau_1; \{1/N\}_1^N)$]; that is,

$$f(m_1, \ldots, m_n|\tau_1) = \frac{\tau_1!}{(\tau_1 - m)!\Pi_1^n m_i!}\left(1 - \frac{n}{N}\right)^{\tau_1 - m}\left(\frac{1}{N}\right)^m \tag{1}$$

It is worth noting that under this model, the estimator $\check{\tau}_1$, which should now be written as $\check{\tau}_1 = NM/n$, where $M = \sum_1^n M_i$, is a maximum likelihood and an unbiased estimator of τ_1. The model-based variance of $\check{\tau}_1$ is

$$\mathbf{V}(\check{\tau}_1) = N^2\left(1 - \frac{n}{N}\right)\frac{\tau_1}{Nn} \tag{2}$$

and the MLE of this variance is $\check{\mathbf{V}}(\check{\tau}_1) = N^2(1 - n/N)\check{\tau}_1/Nn$.

Therefore, the assumed model gives rise to an MLE of τ_1 that is robust to the misspecification of the model. However, the model-based variance and variance estimator are not robust to the misspecification of the model. In fact, if the M_i's do not have the same mean λ, the value given by Expression (2) will be less than the actual variance.

We will now specify a model for the distribution of the indicator variables. First, we will assume that given $M_i = m_i$, x_{ij} is the realization of a Bernoulli random variable X_{ij} with mean p_{ij}. Furthermore, we will assume that given $\mathbf{M}_s = \mathbf{m}_s$, the X_{ij}'s are independently distributed. Second, we will reduce the dimensionality of the vector of probabilities p_{ij}'s by imposing an appropriate restriction on the p_{ij}'s. In this article we will consider the following two models for the p_{ij}'s:

Model I: $p_{ij} = p_i$ for every $u_j \in U - A_i$, and
Model II: $p_{ij} = p_i^{(1)}$ if $u_j \in U_1 - A_i$ and $p_{ij} = p_i^{(2)}$ if $u_j \in U_2$

Notice that in the first model the p_{ij}'s only depend on the clusters, whereas in the second one they depend on both the clusters and the regions in which the nominees are located.

Clearly, other models might be assumed. For instance, if we supposed that the persons in a cluster make independent nominations, each with probability p, then

$p_{ij} = 1 - (1 - p)^{m_i}$ would be a reasonable model. As another example, we might suppose that $p_{ij} = 1 - \exp(-\beta m_i)$, which is the ordinary model assumed in catch-effort studies (see Seber 1982, Ch. 7). However, because of the generality of Models I and II (they do not need the specification of functional forms for the p_{ij}'s), we will focus on these models.

4.1. Model I

The likelihood for τ_1, τ_2, and $\mathbf{p} = (p_1, \ldots, p_n)$ has two components: the conditional distribution of $\mathbf{M}_s$ given $\tau_1 = \sum_1^N m_i$, and the conditional distribution of $\mathbf{Y}$ given τ_1 and $\mathbf{M}_s = \mathbf{m}_s$. The first component is given by (1). The second component can be factorized into three factors which correspond to the three possible locations of the nominees: S_0, $U_1 - S_0$, and U_2. To obtain these factors, let $\mathbf{Y}_1, \mathbf{Y}_2$, and $\mathbf{Y}_{A_i}$, $A_i \in S_0$, be the sets of the variables Y_ω's that correspond to the counts of the people in $U_1 - S_0$, U_2, and $A_i \in S_0$, respectively. The sets of variables $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_{A_i}$, $A_i \in S_0$, are conditionally distributed, given $\mathbf{m}_s$, as $\text{Mult}(\tau_1 - m; \{P_\omega\}_{\omega \subseteq \Omega}, Q)$, $\text{Mult}(\tau_2; \{P_\omega\}_{\omega \subseteq \Omega}, Q)$, and $\text{Mult}(m_i; \{P_\omega\}_{\omega \subseteq \Omega - \{i\}}, Q/(1 - p_i))$, where $P_\omega = \Pi_{i \in \omega} p_i \Pi_{j \notin \omega}(1 - p_j)$, and $Q = \Pi_{i=1}^n (1 - p_i)$. Then, following Darroch's (1958) approach, we get that the factors of the second component of the likelihood that correspond to the locations S_0, $U_1 - S_0$, and U_2 are the following:

$$L_1(\tau_1, \mathbf{p} | \mathbf{y}_{A_1}, \ldots, \mathbf{y}_{A_n}, \mathbf{m}_s) \propto \prod_{i=1}^{n} p_i^{z_i^{(0)}} (1 - p_i)^{m - m_i - z_i^{(0)}} \tag{3}$$

$$L_2(\tau_1, \mathbf{p} | \mathbf{y}_1, \mathbf{m}_s) \propto \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!} \prod_{i=1}^{n} p_i^{z_i^{(1)}} (1 - p_i)^{\tau_1 - m - z_i^{(1)}} \quad \text{and} \tag{4}$$

$$L_3(\tau_2, \mathbf{p} | \mathbf{y}_2, \mathbf{m}_s) \propto \frac{\tau_2!}{(\tau_2 - r_2)!} \prod_{i=1}^{n} p_i^{z_i^{(2)}} (1 - p_i)^{r_2 - z_i^{(2)}} \tag{5}$$

where $z_i^{(0)}$, $z_i^{(1)}$ and $z_i^{(2)}$ are the observed values of the random variables, $Z_i^{(0)}$, $Z_i^{(1)}$, and $Z_i^{(2)}$, that count the number of nominees in $S_0 - A_i$, $U_1 - S_0$, and U_2, respectively, who are nominated by people in cluster $A_i \in S_0, i = 1, \ldots, n$; and r_1 and r_2 are the observed values of the random variables, R_1 and R_2, that count the total number of nominees in $U_1 - S_0$ and U_2, respectively.

Notice that the conditional distributions of $Z_i^{(0)}$, $Z_i^{(1)}$ and $Z_i^{(2)}$, given m_i, are $\text{bin}(m - m_i, p_i)$, $\text{bin}(\tau_1 - m, p_i)$ and $\text{bin}(\tau_2, p_i)$, respectively. Similarly, the conditional distributions of R_1 and R_2, given $\mathbf{m}_s$, are $\text{bin}(\tau_1 - m, 1 - Q)$ and $\text{bin}(\tau_2, 1 - Q)$, respectively.

From the previous results, and the independence of the nominations, we have that the likelihood function for τ_1, τ_2, and $\mathbf{p}$ is the product of (1), (3), (4), and (5).

To obtain the likelihood equations we will follow Darroch's (1958) approach; that is, the parameters τ_1, τ_2, and p_i, $i = 1, \ldots, n$, will be treated as continuous variables, and the partial derivatives of the log-likelihood with respect to these parameters will be computed (using the fact that for large x the derivative of $\ln x!$ is approximately $\ln x$) and will be set to zero. Doing this we obtain the following system of nonlinear

equations:

$$\hat{p}_i = \frac{Z_i}{\hat{\tau}_1 + \hat{\tau}_2 - M_i}, \quad i = 1, \ldots, n$$

$$\hat{\tau}_1 = \frac{M + R_1}{1 - (1 - n/N)\Pi_{i=1}^{n}(1 - \hat{p}_i)} \tag{6}$$

$$= \frac{M + R_1}{1 - (1 - n/N)\Pi_{i=1}^{n}[1 - Z_i/(\hat{\tau}_1 + \hat{\tau}_2 - M_i)]} \tag{7}$$

and

$$\hat{\tau}_2 = \frac{R_2}{1 - \Pi_{i=1}^{n}(1 - \hat{p}_i)}$$

$$= \frac{R_2}{1 - \Pi_{i=1}^{n}[1 - Z_i/(\hat{\tau}_1 + \hat{\tau}_2 - M_i)]} \tag{8}$$

where $Z_i = Z_i^{(0)} + Z_i^{(1)} + Z_i^{(2)}$ is the random variable that counts the number of nominees in $U - A_i$ that are nominated from A_i, $i = 1, \ldots, n$. (Notice that the conditional distribution of Z_i, given m_i, is bin$(\tau - m_i, p_i)$.)

The MLE's $\hat{\tau}_1$ and $\hat{\tau}_2$ of τ_1 and τ_2 are obtained by solving the previous system of equations, and the MLE of τ is $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

Notice that the likelihood equations are very natural. For instance, $\hat{p}_i$ is the ratio of the number of people in $U - A_i$ who are nominated by persons in A_i to the estimated number of people in $U - A_i$, and $\hat{\tau}_1$ is the ratio of the number of people in U_1 who are in the final sample to an estimate of the final sample inclusion-probability (see Equation (6)).

Now, if $\hat{\tau}$ is large enough so that $\hat{p}_i = Z_i/(\hat{\tau} - M_i) \approx Z_i/\hat{\tau}$, then from (7) we have that

$$\hat{\tau}_1 \approx \frac{(n/N)\check{\tau}_1 + R_1}{1 - (1 - n/N)\Pi_{i=1}^{n}(1 - Z_i/\hat{\tau})}$$

Therefore, $\hat{\tau}_1$ depends on the M_i's mainly through $\check{\tau}_1$, and since $\check{\tau}_1$ is robust to the misspecification of the distribution of the M_i's, we should expect $\hat{\tau}_1$ to have this property too. Similarly, since $\hat{\tau}_2$ and $\hat{\tau}$ depend on the M_i's through $\hat{\tau}_1$, we also expect these estimators to be robust to deviations from the assumed distribution of the M_i's.

Approximations to the model-based variances of $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}$ can be obtained by using the formula

$$\mathbf{V}(\hat{\tau}) = \mathbf{V}_\xi[\mathbf{E}_\xi(\hat{\tau}|\mathbf{m}_s)] + \mathbf{E}_\xi[\mathbf{V}_\xi(\hat{\tau}|\mathbf{m}_s)] \tag{9}$$

where $\mathbf{E}_\xi(\cdot\,|\mathbf{m}_s)$ and $\mathbf{V}_\xi(\cdot\,|\mathbf{m}_s)$ denote the conditional model-based expectation and variance operators, given $\mathbf{M}_s = \mathbf{m}_s$, and $\mathbf{E}_\xi(\cdot)$ and $\mathbf{V}_\xi(\cdot)$ denote the model-based expectation and variance operators computed with respect to the conditional distribution of the M_i's given that $\tau_1 = \sum_1^N m_i$.

From Equations (7) and (8), we have that $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}$ are functions of $w_s = (\mathbf{M}_s, \mathbf{Z}_s, R_1, R_2)$, where $\mathbf{Z}_s = (Z_1, \ldots, Z_n)$. Therefore, using the first-order Taylor approximations to these estimators about $\mathbf{E}_\xi(w_s)$ (the model-based conditional expectation of w_s given τ_1) and applying (9) to these approximations, we get

$$\mathbf{V}_\xi(\hat{\tau}_1) \approx E^{-1}(C - D), \quad \mathbf{V}_\xi(\hat{\tau}_2) \approx E^{-1}(B - D), \quad \text{and} \quad \mathbf{V}_\xi(\hat{\tau}) \approx E^{-1}(B + C)$$

where $E = B \times C - (B + C) \times D$,

$$B = \frac{1 - (1 - n/N)Q}{\tau_1(1 - n/N)Q}, \quad C = \frac{1 - Q}{\tau_2 Q}, \quad \text{and} \quad D = \frac{1}{\tau - \tau_1/N} \sum_1^n \frac{p_i}{1 - p_i}$$

Even though variance estimators can be obtained by replacing the unknown quantities in the expressions for the variances by their respective estimators, we will use the alternative estimators obtained by using the variant of the Delta method suggested by Binder (1996). In this variant, the derivatives that appear in the Taylor expansion are evaluated at the observed values of the variables instead of at their expected values; however, the derivatives are treated as constants (as in the ordinary Delta method). This approach yields variance estimators that are still less model dependent than those obtained by the ordinary Delta method.

Thus, using the Taylor approximations to $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}$, with the derivatives evaluated at w_s, replacing the unknown parameters by their estimators, and applying (9) to the approximations, we obtain that model-based estimators of the variances are

$$\hat{\mathbf{V}}_\xi(\hat{\tau}_1) = E_s^{-1}(C_s - D_s), \quad \hat{\mathbf{V}}_\xi(\hat{\tau}_2) = E_s^{-1}(B_s - D_s), \quad \text{and} \quad \hat{\mathbf{V}}_\xi(\hat{\tau}) = E_s^{-1}(B_s + C_s)$$

where $E_s = B_s \times C_s - (B_s + C_s) \times D_s$,

$$B_s = \frac{M + R_1}{\hat{\tau}_1(\hat{\tau}_1 - M - R_1)}, \quad C_s = \frac{R_2}{\hat{\tau}_2(\hat{\tau}_2 - R_2)}, \quad \text{and} \quad D_s = \sum_1^n \frac{\hat{p}_i}{1 - \hat{p}_i} \frac{1}{\hat{\tau} - M_i}$$

We expect $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}$ to be robust to the misspecification of the distribution of the M_i's; however, we do not expect their model-based variances and model-based variance estimators to be unaffected by deviations from this distribution. Therefore, our goal is to derive approximations to the variances of these estimators, as well as estimators of their variances, which are more robust to model misspecifications than the previous ones. Our strategy is to compute approximate variances and variance estimators by replacing, whenever possible, the assumption on the distribution of the M_i's by the design-based distribution used to select the initial sample S_0. This strategy is not new, and it has been used by Wolter (1986) in the context of estimating the census undercount.

The initial sampling design will be incorporated in the variances and variance estimators by replacing in (9) the model-based expectation and variance operators $\mathbf{E}_\xi(\cdot)$ and $\mathbf{V}_\xi(\cdot)$ by their corresponding design-based operators $\mathbf{E}_p(\cdot)$ and $\mathbf{V}_p(\cdot)$ which are computed with respect to the distribution used to select the initial sample.

Let us first derive an approximation to the variance of $\hat{\tau}_1$, as well as an estimator of its variance. Using the first-order Taylor approximation to $\hat{\tau}_1$ about $\mathbf{E}_\xi(w_s)$ we get that $\mathbf{E}_\xi(\hat{\tau}_1|\mathbf{m}_s) \approx a_1 m + c_1$, where c_1 does not depend on the m_i's, and $a_1 = [E^{-1}(C - D)]/[\tau_1(1 - n/N)]$. Then, treating a_1 and c_1 as constants with respect to the distribution used to select S_0, we have

$$\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_1|\mathbf{m}_s)] \approx n\left(1 - \frac{n}{N}\right) \frac{a_1^2}{N - 1} \sum_1^N \left(m_i - \frac{\tau_1}{N}\right)^2 \tag{10}$$

Using again the first-order Taylor approximation to $\hat{\tau}_1$, we get that

$$\mathbf{V}_{\xi}(\hat{\tau}_1|\mathbf{m}_s) \approx E^{-2}\left\{C^2\left[\frac{1}{(\tau-\tau_1/N)^2}\sum_1^n \frac{p_i}{1-p_i}(\tau-m_i) - 2D\frac{\tau_1-m}{\tau_1(1-n/N)}\right]\right.$$
$$\left. -C\times D^2\left[1-2\frac{\tau_1-m}{\tau_1(1-n/N)}\right] + (C-D)^2\times C\times\frac{\tau_2(\tau_1-m)}{[\tau_1(1-n/N)]^2}\right\} \qquad (11)$$

Therefore, by (9) an approximation to $\mathbf{V}(\hat{\tau}_1)$ is obtained by summing (10) and the design-based expectation of (11).

A design-based estimator of $\mathbf{V}(\hat{\tau}_1)$ is obtained by using Binder's (1996) approach. Following that strategy we obtain that an estimator of $\mathbf{E}_{\xi}(\hat{\tau}_1|\mathbf{m}_s) - c_1$ is $\hat{a}_1 m$, where $\hat{a}_1 = [E_s^{-1}(C_s - D_s)\hat{Q}]/(\hat{\tau}_1 - M - R_1)$, and $\hat{Q} = \Pi_1^n(1-\hat{p}_i)$. Consequently an estimator of $\mathbf{V}_p[\mathbf{E}_{\xi}(\hat{\tau}_1|\mathbf{m}_s)]$ is

$$\hat{\mathbf{V}}_{11} = n\left(1-\frac{n}{N}\right)\frac{\hat{a}_1^2}{n-1}\sum_1^n (m_i-\bar{m})^2 \qquad (12)$$

where $\bar{m} = m/n$, and an estimator of $\mathbf{E}_p[\mathbf{V}_{\xi}(\hat{\tau}_1|\mathbf{m}_s)]$ is

$$\hat{\mathbf{V}}_{12} = E_s^{-2}\left\{C_s^2 D_s\left(1-\frac{2(\hat{\tau}_1-m)\hat{Q}}{\hat{\tau}_1-m-R_1}\right) + D_s^2\frac{\hat{\tau}_2\hat{Q}(1-\hat{Q})}{(\hat{\tau}_2-R_2)^2}\right.$$
$$\left. +\ (C_s-D_s)^2\frac{(\hat{\tau}_1-m)\hat{Q}(1-\hat{Q})}{(\hat{\tau}_1-m-R_1)^2} + 2C_s D_s^2\hat{Q}\frac{\hat{\tau}_2 R_1 - (\hat{\tau}_1-m)R_2}{(\hat{\tau}_1-m-R_1)(\hat{\tau}_2-R_2)}\right\}$$

Therefore, $\hat{\mathbf{V}}(\hat{\tau}_1) = \hat{\mathbf{V}}_{11} + \hat{\mathbf{V}}_{12}$ is an estimator of $\mathbf{V}(\hat{\tau}_1)$.

An approximation to the variance of $\hat{\tau}_2$, as well as an estimator of its variance, can be derived using the same analysis as that used in the case of $\hat{\tau}_1$. Thus, an approximation to $\mathbf{V}_p[\mathbf{E}_{\xi}(\hat{\tau}_2|\mathbf{m}_s)]$ is given by (10) but replacing a_1 by $a_2 = (E^{-1}D)/[\tau_1(1-n/N)]$.

Similarly, an approximation to $\mathbf{V}_{\xi}(\hat{\tau}_2|\mathbf{m}_s)$ is

$$\mathbf{V}_{\xi}(\hat{\tau}_2|\mathbf{m}_s) \approx E^{-2}\left\{B^2\left[\frac{1}{(\tau-\tau_1/N)^2}\sum_1^n \frac{p_i}{1-p_i}(\tau-m_i) - 2D\right] + 2B\times D^2\right.$$
$$\left.\times\left[1-\frac{\tau_1-m}{\tau_1(1-n/N)}\right] + C\times D^2\times\frac{\tau_2(\tau_1-m)}{[\tau_1(1-n/N)]^2} + C(B-D)^2\right\} \qquad (13)$$

Therefore, by (9) an approximation to $\mathbf{V}(\hat{\tau}_2)$ is obtained by summing $\mathbf{V}_p[\mathbf{E}_{\xi}(\hat{\tau}_2|\mathbf{m}_s)]$ and the design-based expectation of (13).

An estimator $\hat{\mathbf{V}}_{21}$ of $\mathbf{V}_p[\mathbf{E}_{\xi}(\hat{\tau}_2|\mathbf{m}_s)]$ is given by (12) but replacing $\hat{a}_1$ by $\hat{a}_2 = (E_s^{-1}D_s\hat{Q})/(\hat{\tau}_1 - M - R_1)$.

Similarly, an estimator of $\mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_2|\mathbf{m}_s)]$ is

$$\hat{\mathbf{V}}_{22} = E_s^{-2}\left\{B_s^2 D_s\left(1 - \frac{2\hat{\tau}_2\hat{Q}}{\hat{\tau}_2 - R_2}\right) + D_s^2\frac{(\hat{\tau}_1 - m)\hat{Q}(1-\hat{Q})}{(\hat{\tau}_1 - m - R_1)^2}\right.$$
$$\left. +(B_s - D_s)^2\frac{\hat{\tau}_2\hat{Q}(1-\hat{Q})}{(\hat{\tau}_2 - R_2)^2} + 2B_s D_s^2\hat{Q}\frac{(\hat{\tau}_1 - m)R_2 - \hat{\tau}_2 R_1}{(\hat{\tau}_1 - m - R_1)(\hat{\tau}_2 - R_2)}\right\}$$

Thus $\hat{\mathbf{V}}(\hat{\tau}_2) = \hat{\mathbf{V}}_{21} + \hat{\mathbf{V}}_{22}$ is an estimator of $\mathbf{V}(\hat{\tau}_2)$.

Applying the previous approach to $\hat{\tau}$ we obtain that an approximation to $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}|\mathbf{m}_s)]$ is given by (10) but replacing a_1 by $a_1 + a_2$.

An approximation to $\mathbf{V}_\xi(\hat{\tau}|\mathbf{m}_s)$ is

$$\mathbf{V}_\xi(\hat{\tau}|\mathbf{m}_s) \approx \mathbf{V}_\xi(\hat{\tau}_1|\mathbf{m}_s) + \mathbf{V}_\xi(\hat{\tau}_2|\mathbf{m}_s) + 2E^{-2}\left\{\frac{B\times C}{(\tau - \tau_1/N)^2}\sum_1^n \frac{p_i}{1-p_i}(\tau - m_i)\right.$$
$$\left. -D(C-D)\frac{(\tau_1 - m)n/N}{[\tau_1(1-n/N)]^2} - D^2\left[B + C\times\frac{\tau_1 - m}{\tau_1(1-n/N)}\right]\right\} \tag{14}$$

Thus, an approximation to $\mathbf{V}(\hat{\tau})$ is obtained by summing $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}|\mathbf{m}_s)]$ and the design-based expectation of (14).

An estimator $\hat{\mathbf{V}}_1$ of $\mathbf{V}[\mathbf{E}(\hat{\tau}|\mathbf{m}_s)]$ is given by (12) but replacing $\hat{a}_1$ by $\hat{a}_1 + \hat{a}_2$. Similarly, an estimator of $\mathbf{V}_\xi(\hat{\tau}|\mathbf{m}_s)$ is

$$\hat{\mathbf{V}}_2 = \hat{\mathbf{V}}_{12} + \hat{\mathbf{V}}_{22} + 2E_s^{-2}\left\{B_s C_s D_s - D_s(C_s - D_s)\frac{(\hat{\tau}_1 - m)\hat{Q}}{\hat{\tau}_1 - m - R_1}\left(B_s - \frac{1-\hat{Q}}{\hat{\tau}_1 - m - R_1}\right)\right.$$
$$\left. -D_s(B_s - D_s)\frac{\hat{\tau}_2\hat{Q}}{\hat{\tau}_2 - R_2}\left(C_s - \frac{1-\hat{Q}}{\hat{\tau}_2 - R_2}\right) - D_s^2\hat{Q}\left(\frac{B_s\hat{\tau}_2}{\hat{\tau}_2 - R_2} + \frac{C_s(\hat{\tau}_2 - m)}{\hat{\tau}_1 - m - R_1}\right)\right\}$$

Thus $\hat{\mathbf{V}}(\hat{\tau}) = \hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2$ is an estimator of $\mathbf{V}(\hat{\tau})$.

4.2. *Model II*

In this case the parameters of interest are τ_1, τ_2, $\mathbf{p}^{(1)} = (p_1^{(1)}, \ldots, p_n^{(n)})$ and $\mathbf{p}^{(2)} = (p_1^{(2)}, \ldots, p_n^{(2)})$. The likelihood function for these parameters can be constructed using the same approach as that used in the case of Model I. Using that approach, we obtain that the three factors of the second component of the likelihood function are still given by (3), (4), and (5) but replacing p_i by $p_i^{(1)}$ in (3) and (4), and p_i by $p_i^{(2)}$ in (5). Notice that now the conditional distributions of $Z_i^{(0)}$, $Z_i^{(1)}$ and $Z_i^{(2)}$ are bin$(m - m_i, p_i^{(1)})$, bin$(\tau_1 - m, p_i^{(1)})$ and bin$(\tau_2, p_i^{(2)})$, respectively. Similarly, the conditional distributions of R_1 and R_2, given $\mathbf{m}_s$, are bin$(\tau_1 - m,\ 1 - Q_1)$ and bin$(\tau_2,\ 1 - Q_2)$, respectively, where $Q_1 = \Pi_{i=1}^n\left(1 - p_i^{(1)}\right)$ and $Q_2 = \Pi_{i=1}^n\left(1 - p_i^{(2)}\right)$.

As in the case of Model I, the likelihood function for τ_1, τ_2, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ is given by the product of (1), (3), (4), and (5). Likewise, using Darroch's approach we obtain the

following likelihood equations:

$$\tilde{p}_i^{(1)} = \frac{Z_i^{(01)}}{\tilde{\tau}_1 - M_i}, \quad \tilde{p}_i^{(2)} = \frac{Z_i^{(2)}}{\tilde{\tau}_2}, \quad i = 1, \ldots, n$$

$$\tilde{\tau}_1 = \frac{M + R_1}{1 - (1 - n/N)\Pi_{i=1}^{n}\left[1 - Z_i^{(01)}/(\tilde{\tau}_1 - M_i)\right]} \tag{15}$$

and

$$\tilde{\tau}_2 = \frac{R_2}{1 - \Pi_{i=1}^{n}\left(1 - Z_i^{(2)}/\tilde{\tau}_2\right)} \tag{16}$$

where $Z_i^{(01)} = Z_i^{(0)} + Z_i^{(1)}$.

The MLE's $\tilde{\tau}_1$ and $\tilde{\tau}_2$ of τ_1 and τ_2 are obtained by solving Equations (15) and (16), respectively, and the MLE of τ is $\tilde{\tau} = \tilde{\tau}_1 + \tilde{\tau}_2$.

Similarly to the case of Model I, the likelihood equations are very natural. Also, from (15) we can see that if $\tilde{\tau}_1$ is large enough so that $\tilde{p}_i^{(1)} = Z_i^{(01)}/(\tilde{\tau}_1 - M_i) \approx Z_i^{(01)}/\tilde{\tau}_1$, then we should expect $\tilde{\tau}_1$, and consequently $\tilde{\tau}$, to be robust to deviations from the assumed distribution of the M_i's. Furthermore, since $\tilde{\tau}_2$ is not a function of the M_i's, it is also robust to deviations from the hypothesized joint distribution of the M_i's.

Approximations to the model-based variances of $\tilde{\tau}_1$ and $\tilde{\tau}_2$ can be obtained using the same approach as that used in the case of Model I. From (15) and (16), we have that $\tilde{\tau}_1$ and $\tilde{\tau}_2$ are functions of $w_s^{(1)} = (\mathbf{M}_s, \mathbf{Z}_s^{(01)}, R_1)$ and $w_s^{(2)} = (\mathbf{Z}_s^{(2)}, R_2)$, respectively, where $\mathbf{Z}_s^{(01)} = (Z_1^{(01)}, \ldots, Z_n^{(01)})$ and $\mathbf{Z}_s^{(2)} = (Z_1^{(2)}, \ldots, Z_{\mathrm{n}}^{(2)})$. Therefore, using the first-order Taylor approximations to $\tilde{\tau}_1$ and $\tilde{\tau}_2$ about $\mathbf{E}_\xi\left(w_s^{(1)}\right)$ and $\mathbf{E}_\xi\left(w_s^{(2)}\right)$, respectively, and applying (9) to these approximations, we get that

$$\mathbf{V}_\xi(\tilde{\tau}_1) \approx K_1^{-1}\tau_1 \quad \text{and} \quad \mathbf{V}_\xi(\tilde{\tau}_2) \approx K_2^{-1}\tau_2$$

where $K_1 = F_1 - G_1$, $K_2 = F_2 - G_2$, $F_1 = [1 - (1 - n/N)Q_1]/[(1 - n/N)Q_1]$, $F_2 = (1 - Q_2)/Q_2$,

$$G_1 = \frac{N}{N-1}\sum_1^n \frac{p_i^{(1)}}{1 - p_i^{(1)}} \quad \text{and} \quad G_2 = \sum_1^n \frac{p_i^{(2)}}{1 - p_i^{(2)}}$$

Since we have that $\mathbf{E}_\xi(\tilde{\tau}_2|\mathbf{m}_s) \approx \tau_2$, then $Cov_\xi(\tilde{\tau}_1, \tilde{\tau}_2) \approx 0$, and consequently $\mathbf{V}_\xi(\tilde{\tau}) \approx \mathbf{V}_\xi(\tilde{\tau}_1) + \mathbf{V}_\xi(\tilde{\tau}_2)$.

Estimators of these variances obtained using Binder's (1996) approach are

$$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1) = K_{1s}^{-1}, \quad \tilde{\mathbf{V}}_\xi(\tilde{\tau}_2) = K_{2s}^{-1}\tilde{\tau}_2, \quad \text{and} \quad \tilde{\mathbf{V}}_\xi(\tilde{\tau}) = \tilde{\mathbf{V}}_\xi(\tilde{\tau}_1) + \tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$$

where $K_{1s} = F_{1s} - G_{1s}$, $K_{2s} = F_{2s} - G_{2s}$, $F_{1s} = (M + R_1)/[\tilde{\tau}_1(\tilde{\tau}_1 - M - R_1)]$, $F_2 = R_2/(\tilde{\tau}_2 - R_2)$,

$$G_{1s} = \sum_1^n \frac{\tilde{p}_i^{(1)}}{1 - \tilde{p}_i^{(1)}} \frac{1}{\tilde{\tau}_1 - M_i} \quad \text{and} \quad G_{2s} = \sum_1^n \frac{\tilde{p}_i^{(2)}}{1 - \tilde{p}_i^{(2)}}$$

We can obtain approximations to the variances of $\tilde{\tau}_1$, $\tilde{\tau}_2$, and $\tilde{\tau}$, as well as variance estimators, which are more robust to the misspecification of the distribution of the M_i's

than the previous ones by using the same approach as that used in Model I. Thus, using the first order Taylor approximation to $\tilde{\tau}_1$ about $\mathbf{E}_\xi(w_s^{(1)})$, we obtain that an approximation to $\mathbf{V}_p[\mathbf{E}_\xi(\tilde{\tau}_1|\mathbf{m}_s)]$ is given by (10) but replacing a_1 by $b_1 = N/[K_1(N-n)]$, and that an approximation to $\mathbf{V}_\xi(\tilde{\tau}_1|\mathbf{m}_s)$ is given by

$$\mathbf{V}_\xi(\tilde{\tau}_1|\mathbf{m}_s) \approx K_1^{-2}\left\{\frac{(\tau_1 - m)(1-Q_1)}{(1-n/N)^2 Q_1} - \left[\frac{2(\tau_1 - m)}{1-n/N} - \tau_1\right]G_1\right\} \tag{17}$$

Therefore, an approximation to $\mathbf{V}(\tilde{\tau}_1)$ is obtained by summing $\mathbf{V}_p[\mathbf{E}_\xi(\tilde{\tau}_1|\mathbf{m}_s)]$ and the design-based expectation of (17).

An estimator $\tilde{\mathbf{V}}_{11}$ of $\mathbf{V}_p[\mathbf{E}_\xi(\tilde{\tau}_1|\mathbf{m}_s)]$ is given by (12) but replacing $\hat{a}_1$ by $\tilde{b}_1 = \tilde{Q}_1/[K_{1s} \times (\tilde{\tau}_1 - M - R_1)]$, where $\tilde{Q}_1 = \Pi_{i=1}^n[1 - \tilde{p}_i^{(1)}]$.

Similarly, we have that an estimator of $\mathbf{E}_p[\mathbf{V}_\xi(\tilde{\tau}_1|\mathbf{m}_s)]$ is

$$\tilde{\mathbf{V}}_{12} = K_{1s}^{-2}\left\{\frac{(\tilde{\tau}_1 - m)\tilde{Q}_1(1-\tilde{Q}_1)}{(\tilde{\tau}_1 - m - R_1)^2} - \left[\frac{2(\tilde{\tau}_1 - m)\tilde{Q}_1}{\tilde{\tau}_1 - m - R_1} - 1\right]G_{1s}\right\}$$

Thus, $\tilde{\mathbf{V}}(\tilde{\tau}_1) = \tilde{\mathbf{V}}_{11} + \tilde{\mathbf{V}}_{12}$ is an estimator of $\mathbf{V}(\tilde{\tau}_1)$.

In the case of $\tilde{\tau}_2$, we have that $\mathbf{V}_p[\mathbf{E}(\tilde{\tau}_2|\mathbf{m}_s)] \approx 0$, and $\mathbf{V}_\xi(\tilde{\tau}_2|\mathbf{m}_s) = K_2^{-1}\tau_2$, which is exactly the same as the model-based variance $\mathbf{V}_\xi(\tilde{\tau}_2)$. Therefore, $\mathbf{V}(\tilde{\tau}_2)$ is the design-based expectation of $\mathbf{V}_\xi(\tilde{\tau}_2|\mathbf{m}_s)$, and a variance estimator of $\mathbf{V}(\tilde{\tau}_2)$ is $\tilde{\mathbf{V}}(\tilde{\tau}_2) = K_{2s}^{-1}\tilde{\tau}_2$, which is the same as the model-based estimator $\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$.

An approximation to the variance of $\tilde{\tau}$ is $\mathbf{V}(\tilde{\tau}) \approx \mathbf{V}(\tilde{\tau}_1) + \mathbf{V}(\tilde{\tau}_2)$, and a design-based estimator of $\mathbf{V}(\tilde{\tau})$ is $\tilde{\mathbf{V}}(\tilde{\tau}) = \tilde{\mathbf{V}}(\tilde{\tau}_1) + \tilde{\mathbf{V}}(\tilde{\tau}_2)$.

5. Monte Carlo Studies

In order to observe the performances of the estimators derived in the previous section, two simulation studies were carried out. The first study was based on data obtained from a real study. The second one, which was more extensive than the first one, was based on data obtained from simulated populations.

5.1. Monte Carlo study based on the Nuevo Laredo sex worker population

In the Nuevo Laredo study on high-risk behavior in relation to HIV/AIDS transmission (Valdez 2000), a sampling frame of $N = 107$ venues (bars, clubs, and other establishments) where sex workers can be found with high probability was constructed. The sampling frame was divided into eleven strata, which were formed taking into account the characteristics and locations of the venues. A stratified sample of $n = 27$ venues was selected, and an average of about two sex workers were interviewed in each sampled venue. The median of the numbers of people in the target population nominated by the interviewed sex workers was 20. It is worth noting that the sampling design used in this study was not the same as that considered in this article. In particular, in the study the sample of sites was stratified and the responders only indicated the number of sex workers known by them, but the nominees were not identified. However, the information contained in the study allowed us to set realistic values to the population parameters used in this numerical study.

From the results of the study we set the following values to the parameters used in this simulation study: $N = 107$, $\{m_i\}_1^N = \{5, \ldots, 5$ (18 times), $6, \ldots, 6$ (16 times), $8, \ldots, 8$ (33 times), $9, 9, 9, 10, \ldots, 10$ (5 times), $11, 11, 11, 15, 15, 22, \ldots, 22$ (6 times), $23, \ldots, 23$ (5 times), $25, 25, 25, 26, 26, 26, 27, \ldots, 27$ (5 times), $36, 36, 36, 37, 37\}$, $\tau_1 = 1307$, $\tau_2 = 1193$, $\tau = 2500$, $n = 27$, and $p_i^{(1)} = p_i^{(2)} = .016$, $i = 1, \ldots, n$.

The simulation study was executed as follows. From the finite population of $N = 107$ values of the m_i's, $r = 10{,}000$ samples of $n = 27$ values were selected using an SRSWOR design. For cluster A_i, in the sample, the values of the indicator variables X_{ij}'s, were generated using τ_2 independent identically distributed Bernoulli random variables with mean $p_i^{(2)}$, and $\tau_1 - m_i$ independent identically distributed Bernoulli random variables with mean $p_i^{(1)}$.

The six estimators $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}$, $\tilde{\tau}_1$, $\tilde{\tau}_2$, and $\tilde{\tau}$, their respective model-based and design-based variance estimators, and their corresponding 95% normal-based confidence interval estimators were considered. The performances of an estimator $\hat{\tau}$ and a variance estimator $\hat{\mathbf{V}}(\hat{\tau})$ were evaluated by their simulation relative-biases (r-bias) and the square root of their simulation relative mean squared error (r-mse), defined as $\text{r}-\text{bias} = \sum_1^r(\hat{\theta}_i - \theta)/(r\theta)$ and $\sqrt{\text{r}-\text{mse}} = \sqrt{\sum_1^r(\hat{\theta}_i - \theta)^2/(r\theta^2)}$, where $\hat{\theta}_i$, is the value of $\hat{\tau}$ or $\hat{\mathbf{V}}(\hat{\tau})$ obtained in the i-th replication, and θ is the value of τ or that of the simulation variance of $\hat{\mathbf{V}}(\hat{\tau})$. Finally, the performance of a confidence interval estimator $\hat{\tau} \pm 1.96\sqrt{\hat{\mathbf{V}}(\hat{\tau})}$ was evaluated by its simulation relative frequency of coverage, and by the simulation mean of its semi-length.

The results of the numerical study (Tables 1 and 2) indicate that every one of the estimators of the population size performed very well. They all are practically unbiased and the squared roots of their mean squared errors are less than 0.1. Notice that even though the simulation study was carried out using the assumption that $p_i^{(1)} = p_i^{(2)}$, the performances of the estimators derived under the assumption that the probabilities are not necessarily equal were almost as good as those derived under the assumption of equal probabilities.

With respect to the performances of the variance estimators, we have that the model-based variance estimators behaved very badly, whereas the design-based estimators performed very well. The model-based estimators greatly underestimated the actual variances (except $\tilde{\mathbf{V}}_{\xi}(\tilde{\tau}_2)$, which is also a design-based estimator), and their biases affected the performances of the confidence intervals. The poor behaviors of these estimators were consequences of the fact that the M_i's were not distributed as Poisson random variables. On the other hand, the design-based estimators were practically unbiased, and the relative frequencies of coverage of the confidence intervals were close to 95. Thus, according to

Table 1. Simulation results for the population of sex workers in Nuevo Laredo: Estimators of population size. First entry in each cell is r-bias, second entry is $\sqrt{\text{r-mse}}$

Cluster-element link probability	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$
$\mathbf{E}[p_i^{(1)}] = .016$	.002	.002	.002	.002	.008	.005
$\mathbf{E}[p_i^{(2)}] = .016$	.065	.067	.060	.071	.092	.057

Table 2. Simulation results for the population of sex workers in Nuevo Laredo: Variance estimators

	Model based	r-bias	$\sqrt{\text{r} - \text{mse}}$	Coverage	Semi-length	Design based	r-bias	$\sqrt{\text{r} - \text{mse}}$	Coverage	Semi-length
	$\hat{\mathbf{V}}_{\xi}(\hat{\tau}_1)$	−.735	.736	.686	85.2	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.007	.307	.930	163.2
	$\hat{\mathbf{V}}_{\xi}(\hat{\tau}_2)$	−.319	.338	.889	128.9	$\hat{\mathbf{V}}(\hat{\tau}_2)$	.001	.236	.943	155.7
$\mathbf{E}[p_i^{(1)}] = .016$	$\hat{\mathbf{V}}_{\xi}(\hat{\tau})$	−.616	.619	.776	181.5	$\hat{\mathbf{V}}(\hat{\tau})$	−.002	.290	.931	290.3
$\mathbf{E}[p_i^{(2)}] = .016$	$\tilde{\mathbf{V}}_{\xi}(\tilde{\tau}_1)$	−.755	.755	.668	89.5	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.008	.320	.925	177.8
	$\tilde{\mathbf{V}}_{\xi}(\tilde{\tau}_2)$	−.004	.322	.950	211.9					
	$\tilde{\mathbf{V}}_{\xi}(\tilde{\tau})$	−.314	.367	.895	230.5	$\tilde{\mathbf{V}}(\tilde{\tau})$	−.004	.230	.946	278.4

the results of this study, we have that the model-based variance estimators are very sensitive to deviations from the Poisson distribution.

5.2. Monte Carlo study based on simulated populations

This study was more extensive than the previous one, but because of the limited number of situations considered, the character of the study was still exploratory. Four finite populations of $N = 250$ values of M_i's were generated. A description of each of those finite populations is presented in Table 3.

The nomination probabilities $p_i^{(j)}$, $j = 1, 2$, were obtained by means of the model $p_i^{(j)} = 1 - \exp(-\beta_j m_i)$, where the values of β_j were set so that the specified values of $\mathbf{E}(p_i^{(j)})$ were obtained. This Monte Carlo study was carried out similarly to the previous one. Although several situations were considered in this study, because of limitations of space only some selected results are shown in Tables 4 to 6.

A summary of the results of the estimators of the population sizes (Table 4) follows:

- The use of the Negative binomial distribution as the distribution of the cluster sizes did not have a serious effect on the performances of the estimators.
- The violation of the assumption $p_i^{(1)} = p_i^{(2)}$ affected the unbiasedness of the estimators derived under this assumption. The biases of these estimators were large enough to affect the coverage properties of their corresponding confidence intervals.
- The average value of the nomination probabilities had a great effect on the performance of the estimator $\tilde{\tau}_2$, and a moderate effect on the performances of the other estimators. When both the probabilities and the initial sample size were small ($p_i^{(2)} \approx .01$ and $n = 20$), $\tilde{\tau}_2$ was highly variable and excessively overestimated τ_2. Its performance improved considerably when the sample size was increased ($n = 50$), but it was not good enough to yield good estimates of τ_2. Finally, when the probabilities were large (about .05), its performance was good regardless of the sample size. The behavior of $\tilde{\tau}$ was affected by the behavior of $\tilde{\tau}_2$. The performances of the other estimators were not greatly affected by the size of the probabilities.
- The fraction of coverage of the sampling frame did not have a great effect on the performances of the estimators $\hat{\tau}$ and $\tilde{\tau}$. However, these estimators performed slightly better in the case in which the fraction of coverage of the sampling frame was large, $\tau_1/\tau \approx 0.8$, (and omitting the case in which $\tilde{\tau}_2$ performed very badly), than in the case in which the fraction of coverage was small, $\tau_1/\tau \approx 0.4$.

Table 3. Simulated finite populations

Population I	Population II	Population III	Population IV
M_i Poisson	M_i Neg. binomial	M_i Poisson	M_i Neg. binomial
$\mathbf{E}(M_i) = 8$	$\mathbf{E}(M_i) = 8$	$\mathbf{E}(M_i) = 4$	$\mathbf{E}(M_i) = 4$
$\mathbf{V}(M_i) = 8$	$\mathbf{V}(M_i) = 29.33$	$\mathbf{V}(M_i) = 4$	$\mathbf{V}(M_i) = 12$
$\tau_1 = 2{,}002$	$\tau_1 = 2{,}023$	$\tau_1 = 980$	$\tau_1 = 1{,}011$
$\tau_2 = 500$	$\tau_2 = 500$	$\tau_2 = 1{,}500$	$\tau_2 = 1{,}500$
$\tau = 2{,}502$	$\tau = 2{,}523$	$\tau = 2{,}480$	$\tau = 2{,}511$
$\tau_1/\tau = .8$	$\tau_1/\tau = .8$	$\tau_1/\tau = .4$	$\tau_1/\tau = .4$

Table 4. Simulation results for the estimators of the population sizes. First entry in each cell is r-bias, second entry is $\sqrt{\text{r-mse}}$

		Population I						Population II					
	n	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$
$\mathbf{E}[p_i^{(1)}] = .01$	20	−.002	.002	−.001	−.002	.407	.080	−.007	−.006	−.007	−.007	1.32	.256
		.063	.118	.065	.064	14.08	2.81	.104	.138	.104	.107	31.3	6.19
$\mathbf{E}[p_i^{(2)}] = .01$	50	−.001	−.000	−.001	−.001	.013	.002	−.003	−.002	−.002	−.003	.013	.000
		.030	.063	.030	.030	.122	.034	.043	.068	.043	.045	.123	.043
$\mathbf{E}[p_i^{(1)}] = .05$	20	−.001	−.000	−.001	−.001	.002	−.000	−.004	−.003	−.004	−.004	.001	−.003
		.022	.038	.021	.023	.054	.021	.025	.039	.024	.027	.055	.024
$\mathbf{E}[p_i^{(2)}] = .05$	50	−.000	.000	−.000	−.000	.000	−.000	−.000	−.000	−.000	−.000	.000	−.000
		.006	.013	.006	.006	.015	.006	.006	.013	.006	.007	.014	.006
$\mathbf{E}[p_i^{(1)}] = .02$	20	.020	−.436	−.071	−.002	.530	.105	.013	−.440	−.077	−.008	1.46	.283
		.053	.441	.085	.048	21.2	4.24	.073	.446	.102	.071	35.7	7.07
$\mathbf{E}[p_i^{(2)}] = .01$	50	.020	−.361	−.056	−.001	.009	.001	.019	−.360	−.056	−.002	.012	.001
		.028	.363	.059	.019	.122	.029	.030	.362	.061	.023	.124	.031
$\mathbf{E}[p_i^{(1)}] = .07$	20	.019	−.144	−.013	−.000	.003	.000	.017	−.144	−.015	−.002	.002	−.001
		.025	.148	.020	.016	.054	.017	.024	.149	.023	.017	.055	.018
$\mathbf{E}[p_i^{(2)}] = .05$	50	.005	−.046	−.005	−.000	.000	−.000	.004	−.046	−.005	−.000	.000	−.000
		.006	.048	.007	.003	.015	.004	.006	.048	.007	.004	.015	.004

		Population III						Population IV					
	n	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$
$\mathbf{E}[p_i^{(1)}] = .01$	20	−.005	−.000	−.002	−.004	.038	.021	−.011	−.008	−.009	−.009	.040	.021
		.086	.109	.092	.095	.216	.134	.118	.125	.115	.134	.228	.142
$\mathbf{E}[p_i^{(2)}] = .01$	50	−.001	−.000	−.001	−.001	.004	.002	−.003	−.002	−.002	−.004	.005	.002
		.040	.050	.040	.044	.071	.046	.049	.052	.045	.057	.068	.046
$\mathbf{E}[p_i^{(1)}] = .05$	20	−.002	−.001	−.001	−.002	.001	−.001	−.004	−.002	−.003	−.006	.000	−.002
		.028	.027	.023	.034	.032	.023	.031	.028	.024	.038	.032	.024
$\mathbf{E}[p_i^{(2)}] = .05$	50	−.000	−.000	−.000	−.000	−.000	−.000	−.001	−.000	−.000	−.001	−.000	−.000
		.009	.008	.006	.009	.009	.006	.009	.008	.006	.009	.009	.006
$\mathbf{E}[p_i^{(1)}] = .02$	20	.098	−.381	−.192	−.005	.035	.020	.089	.385	−.194	−.013	.040	.019
		.123	.386	.201	.072	.221	.135	.132	.391	.208	.092	.230	.139
$\mathbf{E}[p_i^{(2)}] = .01$	50	.095	−.295	−.141	−.001	.005	.003	.092	−.293	−.138	−.003	.004	.001
		.100	.296	.143	.028	.072	.045	.097	.295	.141	.031	.069	.043
$\mathbf{E}[p_i^{(1)}] = .07$	20	.075	−.094	−.028	−.001	−.000	−.001	.070	−.094	−.028	−.004	.001	−.001
		.078	.097	.034	.023	.032	.021	.074	.098	.035	.025	.032	.022
$\mathbf{E}[p_i^{(2)}] = .05$	50	.020	−.031	−.011	−.000	−.000	−.000	.018	−.029	−.010	−.000	.000	−.000
		.021	.032	.012	.005	.009	.006	.019	.031	.012	.005	.009	.006

Table 5. Simulation results for Population I: Variance estimators

	Model based	r-bias	$\sqrt{\text{r} - \text{mse}}$	Coverage	Semi-length	Design based	r-bias	$\sqrt{\text{r} - \text{mse}}$	Coverage	Semi-length
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.040	.080	.945	241.4	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.014	.216	.942	243.4
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	.011	.186	.948	115.9	$\hat{\mathbf{V}}(\hat{\tau}_2)$	.016	.192	.948	116.3
$\mathbf{E}[p_i^{(1)}] = .01$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.033	.095	.946	312.6	$\hat{\mathbf{V}}(\hat{\tau})$	−.007	.204	.944	315.4
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.043	.082	.943	245.8	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.018	.223	.942	247.7
$\mathbf{E}[p_i^{(2)}] = .01$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	85.5	5,891	.932	2,626					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	85.6	5,897	.959	2,711	$\tilde{\mathbf{V}}(\tilde{\tau})$	85.6	5,897	.959	2,713
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.020	.160	.948	84.8	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.030	.157	.947	84.4
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.030	.154	.944	36.7	$\hat{\mathbf{V}}(\hat{\tau}_2)$	−.031	.153	.944	36.7
$\mathbf{E}[p_i^{(1)}] = .05$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.021	.169	.948	102.3	$\hat{\mathbf{V}}(\hat{\tau})$	−.027	.164	.948	102.0
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.025	.163	.947	88.9	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.030	.158	.946	88.7
$\mathbf{E}[p_i^{(2)}] = .05$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	−.003	.304	.949	51.8					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	−.014	.177	.951	103.0	$\tilde{\mathbf{V}}(\tilde{\tau})$	−.017	.172	.950	102.9
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.039	.093	.930	188.2	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.044	.143	.929	187.5
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.211	.249	.000	54.3	$\hat{\mathbf{V}}(\hat{\tau}_2)$	−.212	.249	.000	54.3
$\mathbf{E}[p_i^{(1)}] = .02$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.085	.122	.619	215.1	$\hat{\mathbf{V}}(\hat{\tau})$	−.087	.156	.615	214.5
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.033	.094	.944	186.9	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.027	.144	.944	187.1
$\mathbf{E}[p_i^{(2)}] = .01$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	97.4	5,804	.932	4,217					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	97.4	5,804	.958	4,269	$\tilde{\mathbf{V}}(\tilde{\tau})$	97.4	5,804	.957	4,269
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	.061	.203	.779	62.3	$\hat{\mathbf{V}}(\hat{\tau}_1)$	.049	.198	.776	61.9
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.358	.372	.002	25.2	$\hat{\mathbf{V}}(\hat{\tau}_2)$	−.359	.372	.002	25.2
$\mathbf{E}[p_i^{(1)}] = .07$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.080	.190	.834	72.3	$\hat{\mathbf{V}}(\hat{\tau})$	−.088	.191	.833	72.0
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.016	.189	.950	60.4	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.017	.187	.949	60.4
$\mathbf{E}[p_i^{(2)}] = .05$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	.001	.304	.951	52.0					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	−.026	.210	.949	79.8	$\tilde{\mathbf{V}}(\tilde{\tau})$	−.026	.209	.949	79.8

Table 6. Simulation results for Population II: Variance estimators

	Model based	r-bias	$\sqrt{\text{r} - \text{mse}}$	Coverage	Semi-length	Design based	r-bias	$\sqrt{\text{r} - \text{mse}}$	Coverage	Semi-length
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.655	.656	.750	241.4	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.020	.310	.922	402.2
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.278	.309	.894	114.7	$\hat{\mathbf{V}}(\hat{\tau}_2)$	.008	.245	.935	135.2
$\mathbf{E}[p_i^{(1)}] = .01$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.627	.629	.772	311.7	$\hat{\mathbf{V}}(\hat{\tau})$	−.014	.303	.926	501.7
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.663	.663	.747	245.7	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.019	.317	.922	413.9
$\mathbf{E}[p_i^{(2)}] = .01$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	107.9	3,695	.934	11,766					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	108.1	3,700	.884	11,851	$\tilde{\mathbf{V}}(\tilde{\tau})$	108.1	3,700	.951	11,968
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.244	.320	.916	85.3	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.091	.273	.938	93.5
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.038	.250	.944	36.7	$\hat{\mathbf{V}}(\hat{\tau}_2)$	−.019	.254	.946	37.0
$\mathbf{E}[p_i^{(1)}] = .05$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.227	.318	.920	102.9	$\hat{\mathbf{V}}(\hat{\tau})$	−.087	.283	.940	111.8
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.263	.331	.915	89.4	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.093	.272	.939	99.1
$\mathbf{E}[p_i^{(2)}] = .05$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	.001	.475	.948	52.5					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	−.202	.317	.925	103.8	$\tilde{\mathbf{V}}(\tilde{\tau})$	−.067	.291	.943	112.3
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.566	.568	.795	187.7	$\hat{\mathbf{V}}(\hat{\tau}_1)$	−.117	.244	.923	266.2
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.350	.367	.000	53.7	$\hat{\mathbf{V}}(\hat{\tau}_2)$	−.250	.287	.000	57.7
$\mathbf{E}[p_i^{(1)}] = .02$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.575	.576	.569	214.3	$\hat{\mathbf{V}}(\hat{\tau})$	−.157	.256	.736	299.9
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.557	.559	.809	186.3	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.081	.241	.921	266.7
$\mathbf{E}[p_i^{(2)}] = .01$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	117.0	3,837	.932	13,849					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	117.1	3,841	.922	13,901	$\tilde{\mathbf{V}}(\tilde{\tau})$	117.1	3,841	.953	13,948
	$\hat{\mathbf{V}}_\xi(\hat{\tau}_1)$	−.047	.327	.773	62.7	$\hat{\mathbf{V}}(\hat{\tau}_1)$	.046	.364	.798	65.7
$n = 20$	$\hat{\mathbf{V}}_\xi(\hat{\tau}_2)$	−.476	.498	.002	25.1	$\hat{\mathbf{V}}(\hat{\tau}_2)$	−.473	.496	.002	25.2
$\mathbf{E}[p_i^{(1)}] = .02$	$\hat{\mathbf{V}}_\xi(\hat{\tau})$	−.237	.356	.819	72.7	$\hat{\mathbf{V}}(\hat{\tau})$	−.170	.338	.836	75.8
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_1)$	−.163	.337	.934	60.9	$\tilde{\mathbf{V}}(\tilde{\tau}_1)$	−.073	.339	.945	64.0
$\mathbf{E}[p_i^{(2)}] = .01$	$\tilde{\mathbf{V}}_\xi(\tilde{\tau}_2)$	.103	.492	.950	52.5					
	$\tilde{\mathbf{V}}_\xi(\tilde{\tau})$	−.105	.364	.937	80.5	$\tilde{\mathbf{V}}(\tilde{\tau})$	−.050	.371	.945	82.9

A summary of the results of the variance estimators and confidence interval estimators (Tables 5 and 6) follows:

- When the cluster sizes were distributed as Poisson random variables, both model-based and design-based variance estimators and the confidence intervals associated with them behaved reasonably well, provided that their corresponding estimators of the population size behaved well.
- When the cluster sizes were distributed as Negative binomial random variables, the model-based variance estimators underestimated the actual variances, and the biases affected the coverage properties of the confidence intervals. The distortion of the relative frequencies was more serious when the probabilities were small than when they were large. The design-based variance estimators and their corresponding confidence intervals behaved reasonably well.

6. Conclusions and Directions for Future Research

In this article we have developed two sets of estimators of population sizes: one based on the assumption $p_i^{(1)} = p_i^{(2)}$, and another on the assumption $p_i^{(1)} \neq p_i^{(2)}$. For each estimator, model-based and design-based estimators of its variance have been developed. From two simulation studies carried out in this research we obtained the following results. Firstly, the performance of the estimator $\tilde{\tau}_2$ which does not use the information about the cluster sizes strongly depends on the average size of the nomination probabilities: with small probabilities the estimator is not reliable, whereas with large probabilities it behaves reasonably well under its assumed model. (A similar result have been reported by Otis et al. (1978) for the well-known Schnabel estimator used in MCRS). The performances of the other estimators are not greatly affected by the size of the probabilities. Secondly the estimators derived under the assumption $p_i^{(1)} = p_i^{(2)}$ are not robust to deviations from that assumption. Thirdly and finally, the estimators are robust to deviations from the assumed Poisson distribution of the cluster sizes. This property is shared by the design-based variance estimators, but not by the model-based variance estimators.

In future research, the study of other sampling strategies obtained by using other initial sampling designs should be a topic of interest. In addition, the development of estimators that perform reasonably well with small nomination probabilities should be considered. An alternative might be the use of estimators obtained by means of the Bayesian approach. (See Fienberg, Johnson, and Junker 1999, for a review of the Bayesian approach in the context of MCRS.) Also, the development of design-based variance estimators that are not completely based on asymptotic expansions, like those presented here, should be a topic of study. For instance, the use of bootstrap variance estimators might be considered. (See Buckland 1984, for the use of bootstrap in MCRS.) Finally, the development of estimators that take into account the effect of heterogeneous nomination probabilities should also be considered. (See Chao et al. 1992, and Fienberg, Johnson, and Junker 1999, for descriptions of this type of estimator in the context of MCRS.) This is important because studies carried out by Otis et al. (1978) in the context of capture-recapture sampling indicate that the Schnabel estimator is not robust to deviations from the assumption of homogeneous probabilities.

7. References

Binder, D.A. (1996). Linearization Methods for Single-Phase and Two-Phase Samples: A Cookbook Approach. Survey Methodology, 22, 17–22.

Buckland, S.T. (1984). Monte Carlo Confidence Intervals. Biometrics, 40, 811–817.

Chao, A., Lee, S.-M., and Jeng, S.-L. (1992). Estimating Population Size for Capture-recapture Data When Capture Probabilities Vary by Time and Individual Animal. Biometrics, 48, 201–216.

Darroch, J.N. (1958). The Multiple-recapture Census I: Estimation of a Closed Population. Biometrika, 45, 343–359.

Fienberg, S.E., Johnson, M.S., and Junker, B.W. (1999). Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists. Journal of the Royal Statistical Society, Series A, 162, 383–405.

Frank, O. and Snijders, T.A.B. (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. Journal of Official Statistics, 10, 53–67.

International Working Group for Disease Monitoring and Forecasting (1995a). Mark-recapture and Multiple Record Systems: I, History and Theoretical Development. American Journal of Epidemiology, 142, 1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Mark-recapture and Multiple Record Systems: II, Applications in Human Diseases. American Journal of Epidemiology, 142, 1059–1068.

Otis, D.L., Burnham, K.P., White, G.C., and Anderson, D.R. (1978). Statistical Inference from Capture Data on Closed Animal Populations. Wildlife Monographs, 62, 1–135.

Seber, G.A.F. (1982). The Estimation of Animal Abundance. 2nd edition. London: Griffin.

Spreen, M. (1992). Rare Populations, Hidden Populations, and Link-tracing Designs: What and Why? Bulletin de Methodologie Sociologique, 36, 34–58.

Thompson, S. and Frank, O. (2000). Model-based Estimation with Link-tracing Sampling Designs. Survey Methodology, 26, 87–98.

Valdez, A. (2000). Personal Communication.

Wolter, K. (1986). Some Coverage Errors Models for Census Data. Journal of the American Statistical Association, 81, 338–346.

Received February 2001
Revised December 2003