Combining Cluster Sampling and Link-Tracing Sampling: Estimating the Size of a Hidden Population in the Presence of Heterogeneous Link-Probabilities Modeled by a Latent-Class Model

Martín H. Félix-Medina* Jesús A. Domínguez-Molina†

Abstract

In this work we proposed estimators of the size of a hidden population, such as sexual workers and drug users. Specifically, we derive unconditional and conditional maximum likelihood estimators to be used along with the variant of link-tracing sampling proposed by Félix-Medina and Thompson (Jour. Official Stat., 2004). In this variant, a sampling frame made up by sites where the members of the population can be found with high probabilities, such as bars and parks, is constructed. The population is not assumed to be completely covered by the frame. Then an initial simple random sample of sites is selected from the frame. The people in the sampled sites are identified and they are asked to name other members of the population. We say that there is a link between a site and a person if that person is named by at least one element in the site. Following an idea used by Pledger (Biometrics, 2000) in the context of capture-recapture, we derived maximum likelihood estimators under the assumption that the elements in the population can be grouped into a number of classes according to their susceptibility of being linked to a site in the initial sample. Elements in the same class have the same probability of being linked to a particular site, while elements in different classes have different link probabilities. This assumption allows us to model the heterogeneity of the link probabilities. The unconditional maximum likelihood estimator is obtained by using the ordinary maximum likelihood approach, whereas the conditional maximum likelihood estimator is obtained by using an approach proposed by Sanathanan (Annals of Math. Stat., 1972). The results of a simulation study indicate that the proposed estimators require relatively large sampling fractions to perform satisfactorily, otherwise they present problems of high variability and numerical instability.

Key Words: Capture-recapture, chain referral sampling, hard-to-detect population, latent class model, maximum likelihood estimator, snowball sampling

1. Introduction

Link-tracing sampling (LTS), also known as snowball sampling or chain referral sampling, has been proposed for sampling hidden or hard-to-detect populations, such as drug users, sex workers, HIV infected people and undocumented workers. In this method an initial sample of members of the target population is selected and the people in the initial sample are asked to name or to refer other members of the population to be included in the sample. The named people who are not in the initial sample might be asked to refer other persons, and the process might continue in this way until a specified stopping rule is satisfied.

Félix-Medina and Thompson (2004) proposed a variant of LTS in which the initial sample is a simple random sample without replacement (SRSWOR) of sites selected from a sampling frame made up by venues where the members of the population might be found with high probabilities, such as public parks, bars and blocks. The population is not assumed to be completely covered by the frame. The members of the population who belong

*Facultad de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México

†Facultad de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México

to a sampled site are identified and they are asked to name other members of the population. In order to obtain a maximum likelihood estimator (MLE) of the size of the population, those authors assumed that the probability that a person is named by any person in a particular sampled site, which we will call link probability, depends on the site, but not on the named person, that is, they assumed homogeneous link probabilities.

Later, Félix-Medina et al. (2009) extended the previous work to the case in which the link probabilities depend also on the named people, that is, heterogeneous link probabilities. They modeled the link probability between a site and a person by means of a mixed logistic normal model which is a function of two additive effects: a fixed effect associated with the site and a normally distributed random effect associated with the person. Those authors proposed a conditional MLE of of the population size. In a Monte Carlo study carried out by them, they found that their estimator performed reasonably well, but that it was not robust to some deviations from the assumptions under which it was derived. In particular, it was not robust to deviations from the normal distribution of the random effects.

In this work we use a latent class model, suggested by Pledger (2000) in the context of capture-recapture, to model the heterogeneity of the link probabilities. Our goal is to obtain a robust estimator of the population size. The structure of the paper is as follows. In Section 2 we describe the variant of LTS proposed by Félix-Medina and Thompson (2004). In Section 3 we present the models we propose to describe the sampling procedure; in Section 4 we derive unconditional and conditional MLEs of the population size; in Section 5 we present the results of a Monte Carlo study carried out to observe the performance of one of the proposed estimators, and finally in Section 6 we present conclusions and suggestions for future research.

2. Sampling Design

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

The sampling design is as follows. A SRSWOR S_A of n sites $A_1, \ldots, A_n$ is selected from the frame and the M_i members of the population who belong to the sampled site A_i are identified, $i = 1, \ldots, n$. Let S_0 be the set of people in the initial sample. Notice that the size of S_0 is $M = \sum_1^n M_i$. The people in each sampled site are asked to name other members of the population. We will say that a person and a site are linked if any of the people who belong to that site names him or her. For each named person we record the sites that are linked to him or her, and the portion of U: $U_1 - S_0$, a particular $A_i \in S_A$ or U_2, that contains him or her.

3. Probability Models

As in Félix-Medina and Thompson (2004), we will suppose that the numbers $M_1, \ldots, M_N$ of people who belong to the sites $A_1, \ldots, A_N$ are independent Poisson random variables

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

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

To model the heterogeneity of the link probabilities we will consider a latent class model proposed by Pledger (2000) in the context of capture-recapture. Thus, we will assume that each person in U_k belongs to only one of C classes according to his or her propensity to be linked to a site in S_A. The idea behind these classes is that people in the same class have the same probability of being linked to a site in S_A, but people in different classes have different link probabilities.

Let $p_c^{(k)}$ be the probability that a randomly selected person from U_k belongs to class c. We will suppose that $p_c^{(k)} > 0$, $c = 1, \ldots, C$, and that $\sum_1^C p_c^{(k)} = 1$. In addition, we will assume that the number C of classes is a fixed known number. It is worth noting that we are not assuming that we know the class to which a person belongs.

Let us define the link indicator variables $X_{ij}^{(k)}$s by $X_{ij}^{(k)} = 1$ if person j in $U_k - A_i$ is linked to site A_i, and $X_{ij}^{(k)} = 0$ if $j \in A_i$ or that person is not linked to A_i, $j = 1, \ldots, \tau_k$, $i = 1, \ldots, n$, and $k = 1, 2$. We will suppose that if person j in $U_k - A_i$ belongs to class c, then

$$\begin{aligned}\Pr\left(X_{ij}^{(k)} = 1 | j \in U_k - A_i \text{ belongs to class } c\right) &= \theta_{ic}^{(k)} \\ &= \frac{\exp[\mu^{(k)} + \alpha_i^{(k)} + \eta_c^{(k)} + (\alpha\eta)_{ic}^{(k)}]}{1 + \exp[\mu^{(k)} + \alpha_i^{(k)} + \eta_c^{(k)} + (\alpha\eta)_{ic}^{(k)}]}, \\ & \quad i = 1, \cdots, n;\ c = 1, \cdots, C.\end{aligned}$$

In this model, $\mu^{(k)}$ is a fixed general effect; $\alpha_i^{(k)}$ is a fixed effect associated with site A_i, which indicates the potential of A_i of forming links with people in $U_k - A_i$; $\eta_c^{(k)}$ is a random effect associated with class c, which indicates the propensity of the people in that class of being linked to the sites in S_A, and $(\alpha\eta)_{ic}^{(k)}$ is a random effect associated with site A_i and class c that indicates the interaction between them.

For person j in $U_k - S_0$ we will define the n-dimensional vector $\mathbf{X}_j^{(k)} = (X_{1j}^{(k)}, \ldots, X_{nj}^{(k)})$ of the link indicator variables associated with the j-th person. Notice that $\mathbf{X}_j^{(k)}$ indicates the sites $A_i \in S_A$ that are linked to that person. Let $\Omega = \{(x_1, \ldots, x_n) : x_i = 0, 1;\ i = 1, \ldots, n\}$, the set of all n-dimensional vectors such that each one of their elements is either 0 or 1. Then the probability that the vector of link indicator variables associated with a randomly selected person from $U_k - S_0$ equals $\mathbf{x} = (x_1, \ldots, x_n) \in \Omega$ is given by

$$\begin{aligned}\pi_{\mathbf{x}}^{(k)}(\mathbf{p}_k^*, \psi_k^*) &= \sum_{c=1}^C p_c^{(k)} \prod_{i=1}^n \left[\theta_{ic}^{(k)}\right]^{x_i} \left[1 - \theta_{ic}^{(k)}\right]^{1 - x_i} \\ &= \sum_{c=1}^C p_c^{(k)} \prod_{i=1}^n \frac{\exp\{x_i[\mu^{(k)} + \alpha_i^{(k)} + \eta_c^{(k)} + (\alpha\eta)_{ic}^{(k)}]\}}{1 + \exp[\mu^{(k)} + \alpha_i^{(k)} + \eta_c^{(k)} + (\alpha\eta)_{ic}^{(k)}]},\end{aligned}$$

where $\mathbf{p}_k^* = (p_1^{(k)}, \ldots, p_C^{(k)})$ and $\psi_k^* = (\mu^{(k)}, \{\alpha_i^{(k)}\}_1^n, \{\eta_c^{(k)}\}_1^C, \{(\alpha\eta)_{ic}^{(k)}\}_{1,1}^{n,C})$.

Similarly, for person j in $A_{i'} \in S_A$, we will define the $(n-1)$- dimensional vector $\mathbf{X}_j^{(A_{i'})} = (X_{1j}^{(A_{i'})}, \ldots, X_{i'-1j}^{(A_{i'})}, X_{i'+1j}^{(A_{i'})}, \ldots, X_{nj}^{(A_{i'})})$ of the link indicator variables associated with the j-th person. Let $\Omega_{i'} = \{(x_1, \ldots, x_{i'-1}, x_{i'+1}, \ldots, x_n) : x_i = 0, 1;\ i \neq i',\ i = 1, \ldots, n\}$. Then the probability that the vector of link indicator variables associated

with a randomly selected person from $A_{i'} \in S_A$ equals $\mathbf{x} = (x_1, \ldots, x_{i'-1}, x_{i'+1}, \ldots, x_n) \in \Omega_{i'}$ is given by

$$\begin{aligned}\pi_{\mathbf{x}}^{(A_{i'})}(\mathbf{p}_1^*, \psi_1^*) &= \sum_{c=1}^{C} p_c^{(1)} \prod_{i \neq i'}^{n} \left[\theta_{ic}^{(1)}\right]^{x_i} \left[1 - \theta_{ic}^{(1)}\right]^{1-x_i} \\ &= \sum_{c=1}^{C} p_c^{(1)} \prod_{i \neq i'}^{n} \frac{\exp\{x_i[\mu^{(1)} + \alpha_i^{(1)} + \eta_c^{(1)} + (\alpha\eta)_{ic}^{(1)}]\}}{1 + \exp[\mu^{(1)} + \alpha_i^{(1)} + \eta_c^{(1)} + (\alpha\eta)_{ic}^{(1)}]}.\end{aligned}$$

It is worth noting that the parameters $\mathbf{p}_k^*$ and ψ_k^* are not identifiable. To solve this problem we need to impose some constrains on them. For instance

$$\sum_{1}^{C} p_c^{(k)} = 1, \quad \alpha_n^{(k)} = 0, \quad \eta_C^{(k)} = 0 \quad \text{and} \quad (\alpha\eta)_{ic}^{(k)} = 0 \quad \text{if } i = n \text{ or } c = C, \quad k = 1, 2.$$

Another possibility is sum-to-zero constrains:

$$\sum_{1}^{C} p_c^{(k)} = 1, \quad \sum_{1}^{n} \alpha_i^{(k)} = 0, \quad \sum_{1}^{C} \eta_c^{(k)} = 0 \quad \text{and} \quad \sum_{1}^{n}\sum_{1}^{C} (\alpha\eta)_{ic}^{(k)} = 0, \quad k = 1, 2.$$

Thus, let $\mathbf{p}_k = (p_1^{(k)}, \ldots, p_{C-1}^{(k)})$ and ψ_k be the vector of the parameters $\mu^{(k)}$, $\alpha_i^{(k)}$, $\eta_c^{(k)}$ and $(\alpha\eta)_{ic}^{(k)}$ that are identifiable.

4. Likelihood Function

To construct the likelihood function we will factorize it into different components. One factor is associated with the probability of selecting the initial sample S_0, which is given by the multinomial distribution (1), that is,

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

Two other factors are associated with the probabilities of the configurations of links between the people in $U_k - S_0$, $k = 1, 2$, and the sites $A_i \in S_A$. To obtain these factors, for $\mathbf{x} = (x_1, \ldots, x_n) \in \Omega$, let $R_{\mathbf{x}}^{(k)}$ be the random variable that indicates the number of distinct people in $U_k - S_0$ whose vectors of link indicator variables are equal to $\mathbf{x}$. Finally, let R_k be the random variable that indicates the number of distinct people in $U_k - S_0$ that are linked to at least one site $A_i \in S_A$. Notice that $R_k = \sum_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} R_{\mathbf{x}}^{(k)}$, where $\mathbf{0}$ denotes the n-dimensional vector of zeros, and $R_{\mathbf{0}}^{(1)} = \tau_1 - M - R_1$ and $R_{\mathbf{0}}^{(2)} = \tau_2 - R_2$.

Because of the assumptions we made about the variables $X_{ij}^{(k)}$s, we have that given $M_1, \ldots, M_n$, the joint probability distribution of the variables $\{R_{\mathbf{x}}^{(1)}\}_{\mathbf{x} \in \Omega}$ is a multinomial distribution with parameter of size $\tau_1 - M$ and probabilities $\{\pi_{\mathbf{x}}^{(1)}(\mathbf{p}_1, \psi_1)\}_{\mathbf{x} \in \Omega}$, whereas that of the variables $\{R_{\mathbf{x}}^{(2)}\}_{\mathbf{x} \in \Omega}$ is a multinomial distribution with parameter of size τ_2 and probabilities $\{\pi_{\mathbf{x}}^{(2)}(\mathbf{p}_2, \psi_2)\}_{\mathbf{x} \in \Omega}$.

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

$$L_1(\tau_1, \mathbf{p}_1, \psi_1) \propto \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!} \prod_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} [\pi_{\mathbf{x}}^{(1)}(\mathbf{p}_1, \psi_1)]^{r_{\mathbf{x}}^{(1)}} [\pi_{\mathbf{0}}^{(1)}(\mathbf{p}_1, \psi_1)]^{\tau_1 - m - r_1}$$

and

$$L_2(\tau_2, \mathbf{p}_2, \psi_2) \propto \frac{\tau_2!}{(\tau_2 - r_2)!} \prod_{\mathbf{x} \in \Omega - \{\mathbf{0}\}} [\pi_{\mathbf{x}}^{(2)}(\mathbf{p}_2, \psi_2)]^{r_{\mathbf{x}}^{(2)}} [\pi_{\mathbf{0}}^{(2)}(\mathbf{p}_2, \psi_2)]^{\tau_2 - r_2}.$$

The last factor of the likelihood function is associated with the probability of the configuration of links between the people in S_0 and the sites $A_i \in S_A$. To obtain this factor, for $\mathbf{x} \in \Omega_{i'}$ let $R_{\mathbf{x}}^{(A_{i'})}$ be the random variable that indicates the number of distinct people in $A_{i'} \in S_A$ whose vectors of link indicator variables equal $\mathbf{x}$. Finally, let $R^{(A_{i'})}$ be the random variable that indicates the number of distinct people in $A_{i'} \in S_A$ that are linked to at least one site $A_i \in S_A$, $i \neq i'$. Notice that $R^{(A_{i'})} = \sum_{\mathbf{x} \in \Omega_{i'} - \{\mathbf{0}\}} R_{\mathbf{x}}^{(A_{i'})}$ and $R_0^{(A_{i'})} = M_{i'} - R^{(A_{i'})}$. Then, given $M_1, \ldots, M_n$, then the joint probability distribution of the variables $\{R_{\mathbf{x}}^{(A_{i'})}\}_{\mathbf{x} \in \Omega_{i'}}$ is a multinomial distribution with parameter of size $M_{i'}$ and probabilities $\{\pi_{\mathbf{x}}^{(A_{i'})}(\mathbf{p}_1, \psi_1)\}_{\mathbf{x} \in \Omega_{i'}}$.

Thus, the probability of the configuration of links between the people in S_0 and the sites in S_A is the product of the previous multinomial probabilities (one for each $A_{i'} \in S_A$), and consequently the factor of the likelihood function associated with that probability is

$$L_0(\mathbf{p}_1, \psi_1) \propto \prod_{i'=1}^{n} \prod_{\mathbf{x} \in \Omega_{i'} - \{\mathbf{0}\}} [\pi_{\mathbf{x}}^{(A_{i'})}(\mathbf{p}_1, \psi_1)]^{r_{\mathbf{x}}^{(A_{i'})}} [\pi_{\mathbf{0}}^{(A_{i'})}(\mathbf{p}_1, \psi_1)]^{m_{i'} - r^{(A_{i'})}}.$$

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

$$L(\tau_1, \tau_2, \mathbf{p}_1, \mathbf{p}_2, \psi_1, \psi_2) = L_{(1)}(\tau_1, \mathbf{p}_1, \psi_1) L_{(2)}(\tau_2, \mathbf{p}_2, \psi_2),$$

where

$$L_{(1)}(\tau_1, \mathbf{p}_1, \psi_1) = L_{MULT}(\tau_1) L_1(\tau_1, \mathbf{p}_1, \psi_1) L_0(\mathbf{p}_1, \psi_1) \quad \text{and} \tag{2}$$

$$L_{(2)}(\tau_2, \mathbf{p}_2, \psi_2) = L_2(\tau_2, \mathbf{p}_2, \psi_2). \tag{3}$$

5. Unconditional and Conditional Maximum Likelihood Estimators

5.1 Unconditional estimators

Numerical maximization of (2) and (3) yields the unconditional MLEs $\hat{\tau}_k^U$, $\hat{\mathbf{p}}_k^U$ and $\hat{\psi}_k^U$ of τ_k, $\mathbf{p}_k$ and ψ_k, $k = 1, 2$. Therefore, the unconditional MLE of $\tau = \tau_1 + \tau_2$ is $\hat{\tau}^U = \hat{\tau}_1^U + \hat{\tau}_2^U$.

Although we cannot obtain closed forms for $\hat{\tau}_k^U$, $\hat{\mathbf{p}}_k^U$ and $\hat{\psi}_k^U$, by computing the derivatives of the logarithms of (2) and (3) with respect to τ_1 and τ_2, respectively, equating those derivatives to zero, and solving the equations for τ_1 and τ_2, we get that the unconditional MLEs $\hat{\tau}_1^U$ and $\hat{\tau}_2^U$ can be expressed as

$$\hat{\tau}_1^U = \frac{M + R_1}{1 - (1 - n/N)\pi_{\mathbf{0}}^{(1)}(\hat{\mathbf{p}}_1^U, \hat{\psi}_1^U)} \quad \text{and} \quad \hat{\tau}_2^U = \frac{R_2}{1 - \pi_{\mathbf{0}}^{(2)}(\hat{\mathbf{p}}_2^U, \hat{\psi}_2^U)},$$

where $\hat{\mathbf{p}}_k^U$ and $\hat{\psi}_k^U$ are the unconditional MLEs of $\mathbf{p}_k$ and ψ_k, $k = 1, 2$.

5.2 Conditional estimators

Another approach to obtain estimators of τ_k, $\mathbf{p}_k$ and ψ_k is the conditional maximum likelihood estimation approach.This approach was proposed by Sanathanan (1972) in the context of capture-recapture estimation. Because the resulting conditional estimators are asymptotically equivalent to the unconditional MLEs, and they are easier to compute than the unconditional ones, several authors, such as Fienberg (1972) and Coull and Agresti (1999), have suggested this approach.

Thus, the idea is to factorize the probability mass function (pmf) of the multinomial distribution of the variables $\{R_{\mathbf{x}}^{(k)}\}_{\mathbf{x}\in\Omega}$ as follows

$$
\begin{aligned}
L_1(\tau_1,\mathbf{p}_1,\psi_1) &\propto f(\{r_{\mathbf{x}}^{(1)}\}_{\mathbf{x}\in\Omega}|m,\tau_1,\mathbf{p}_1,\psi_1)\\
&= f(\{r_{\mathbf{x}}^{(1)}\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}|r_1,m,\tau_1,\mathbf{p}_1,\psi_1)f(r_1|m,\tau_1,\mathbf{p}_1,\psi_1)\\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}\left[\frac{\pi_{\mathbf{x}}^{(1)}(\mathbf{p}_1,\psi_1)}{1-\pi_{\mathbf{0}}^{(1)}(\mathbf{p}_1,\psi_1)}\right]^{r_{\mathbf{x}}^{(1)}}\\
&\quad\times\frac{(\tau_1-m)!}{(\tau_1-m-r_1)!}[1-\pi_{\mathbf{0}}^{(1)}(\mathbf{p}_1,\psi_1)]^{r_1}[\pi_{\mathbf{0}}^{(1)}(\mathbf{p}_1,\psi_1)]^{\tau_1-m-r_1}\\
&= L_{11}(\mathbf{p}_1,\psi_1)L_{12}(\tau_1,\mathbf{p}_1,\psi_1) \quad \text{and}\\
L_2(\tau_2,\mathbf{p}_2,\psi_2) &\propto f(\{r_{\mathbf{x}}^{(2)}\}_{\mathbf{x}\in\Omega},\tau_2-r_2|\tau_2,\mathbf{p}_2,\psi_2)\\
&= f(\{r_{\mathbf{x}}^{(2)}\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}|r_2,\tau_2,\mathbf{p}_2,\psi_2)f(r_2|\tau_2,\mathbf{p}_2,\psi_2)\\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}\left[\frac{\pi_{\mathbf{x}}^{(2)}(\mathbf{p}_2,\psi_2)}{1-\pi_{\mathbf{0}}^{(2)}(\mathbf{p}_2,\psi_2)}\right]^{r_{\mathbf{x}}^{(2)}}\\
&\quad\frac{\tau_2!}{(\tau_2-r_2)!}[1-\pi_{\mathbf{0}}^{(2)}(\mathbf{p}_2,\psi_2)]^{r_2}[\pi_{\mathbf{0}}^{(2)}(\mathbf{p}_2,\psi_2)]^{\tau_2-r_2}\\
&= L_{21}(\mathbf{p}_2,\psi_2)L_{22}(\tau_2,\mathbf{p}_2,\psi_2).
\end{aligned}
$$

Notice that in each case the first factor $L_{k1}(\mathbf{p}_k,\psi_k)$ is proportional to the joint pmf of the variables $\{R_{\mathbf{x}}^{(k)}\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, which is the multinomial distribution with parameter of size R_k and probabilities $\{\pi_{\mathbf{x}}^{(k)}(\mathbf{p}_k,\psi_k)/[1-\pi_{\mathbf{0}}^{(k)}(\mathbf{p}_k,\psi_k)]\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, and that this distribution does not depend on τ_k. Notice also that the second factors $L_{12}(\tau_1,\mathbf{p}_1,\psi_1)$ and $L_{22}(\tau_2,\mathbf{p}_2,\psi_2)$ are proportional to the pmfs of the Bin($\tau_1-m, 1-\pi_{\mathbf{0}}^{(1)}(\mathbf{p}_1,\psi_1)$) and Bin($\tau_2, 1-\pi_{\mathbf{0}}^{(2)}(\mathbf{p}_2,\psi_2)$), respectively, where Bin(τ,θ) denotes the Binomial distribution with parameter of size τ and probability θ.

The conditional MLEs $\hat{\mathbf{p}}_k^C$ and $\hat{\psi}_k^C$ of $\mathbf{p}_k$ and ψ_k, $k=1,2$, are obtained by maximizing numerically

$$
L_{11}(\mathbf{p}_1,\psi_1)L_0(\mathbf{p}_1,\psi_1) \quad \text{and} \quad L_{21}(\mathbf{p}_2,\psi_2) \tag{4}
$$

with respect to $(\mathbf{p}_1,\psi_1)$ and $(\mathbf{p}_2,\psi_2)$, respectively. Notice that the factors in (4) do not depend on τ_k, $k=1,2$.

Finally, by plugging the estimates $\hat{\mathbf{p}}_k^C$ and $\hat{\psi}_k^C$ into the factors of the likelihood function that depend on τ_k, $k=1,2$, and maximizing these factors, that is, maximizing $L_{12}(\tau_1,\hat{\mathbf{p}}_1,\hat{\psi}_1)L_{MULT}(\tau_1)$ and $L_{22}(\tau_2,\hat{\mathbf{p}}_2,\hat{\psi}_2)$, with respect to τ_1 and τ_2, respectively, we get that the conditional MLEs $\hat{\tau}_1^C$ and $\hat{\tau}_2^C$ of τ_1 and τ_2 are given by

$$
\hat{\tau}_1^C = \frac{M+R_1}{1-(1-n/N)\pi_{\mathbf{0}}^{(1)}(\hat{\mathbf{p}}_1^C,\hat{\psi}_1^C)} \quad \text{and} \quad \hat{\tau}_2^C = \frac{R_2}{1-\pi_{\mathbf{0}}^{(2)}(\hat{\mathbf{p}}_2^C,\hat{\psi}_2^C)}.
$$

A conditional MLE of τ is $\hat{\tau}^C = \hat{\tau}_1^C + \hat{\tau}_2^C$.

Table 1: Characteristics of the four artificial populations used in the Monte Carlo study.

Population I	Population II
$N = 200$	$N = 200$
$M_i \sim$ Poisson	$M_i \sim$ Zero-truncated negative binomial
$E(M_i) = 8.0 \quad V(M_i) = 8.0$	$E(M_i) = 8.0 \quad V(M_i) = 24.0$
$\tau_1 = 1600 \quad \tau_2 = 800 \quad \tau = 2400$	$\tau_1 = 1701 \quad \tau_2 = 800 \quad \tau = 2501$
Link probabilities: Latent class model	Link probabilities: Latent class model
$C = 2$ classes	$C = 2$ classes
$p_1^{(k)} = 0.3 \quad p_2^{(k)} = 0.7,\ k = 1, 2$	$p_1^{(k)} = 0.3 \quad p_2^{(k)} = 0.7,\ k = 1, 2$
$\theta_{ic}^{(k)} = \frac{\exp[\mu^{(k)}+\alpha_i^{(k)}+\eta_c^{(k)}+(\alpha\eta)_{ic}^{(k)}]}{1+\exp[\mu^{(k)}+\alpha_i^{(k)}+\eta_c^{(k)}+(\alpha\eta)_{ic}^{(k)}]}$	$\theta_{ic}^{(k)} = \frac{\exp[\mu^{(k)}+\alpha_i^{(k)}+\eta_c^{(k)}+(\alpha\eta)_{ic}^{(k)}]}{1+\exp[\mu^{(k)}+\alpha_i^{(k)}+\eta_c^{(k)}+(\alpha\eta)_{ic}^{(k)}]}$
$\mu^{(k)} = -1.3 \quad \alpha_i^{(k)} = -12/(0.001 + \sqrt{m_i})$	$\mu^{(k)} = -1.3 \quad \alpha_i^{(k)} = -12/(0.001 + \sqrt{m_i})$
$\eta_1^{(k)} = 1.5 \quad \eta_2^{(k)} = 0.0$	$\eta_1^{(k)} = 1.5 \quad \eta_2^{(k)} = 0.0$

Population III	Population IV
$N = 200$	$N = 200$
$M_i \sim$ Poisson	$M_i \sim$ Poisson
$E(M_i) = 8.0 \quad V(M_i) = 8.0$	$E(M_i) = 8.0 \quad V(M_i) = 8.0$
$\tau_1 = 1600 \quad \tau_2 = 800 \quad \tau = 2400$	$\tau_1 = 1600 \quad \tau_2 = 800 \quad \tau = 2400$
Link probabilities: Latent class model	Link probabilities: Mixed logit model
$C = 3$ classes	with scaled Student's T random effects
$p_1^{(k)} = 0.5\, p_2^{(k)} = 0.3\, p_3^{(k)} = 0.2,\ k = 1, 2$	$\theta_{ij}^{(k)} = \frac{\exp[\alpha_i^{(k)}+\beta_j^{(k)}]}{1+\exp[\alpha_i^{(k)}+\beta_j^{(k)}]}$
$\theta_{ic}^{(k)} = \frac{\exp[\mu^{(k)}+\alpha_i^{(k)}+\eta_c^{(k)}+(\alpha\eta)_{ic}^{(k)}]}{1+\exp[\mu^{(k)}+\alpha_i^{(k)}+\eta_c^{(k)}+(\alpha\eta)_{ic}^{(k)}]}$	$\alpha_i^{(k)} = -6.3/(0.001 + m_i^{1/4}) \quad \beta_j^{(k)} \sim T_6/\sqrt{1.5}$
$\mu^{(k)} = -1.3 \quad \alpha_i^{(k)} = -12/(0.001 + \sqrt{m_i})$	
$\eta_1^{(k)} = 1.5 \quad \eta_2^{(k)} = -1.0 \quad \eta_3^{(k)} = 0.0$	

6. Monte Carlo Studies

To observe the performance of the proposed estimators and compare them with other estimators that have been proposed we carried out a simulation study. Thus, we constructed four artificial populations. A description of each one is presented in Table 1. Notice that in Population I, II and IV the $N = 200$ values of the M_i's were generated using a Poisson distribution, whereas in Population II they were generated using a zero-truncated negative binomial distribution. In Populations I, II and III the link probabilities were generated by using a latent class model with $C = 2$ classes in the case of the first two populations and $C = 3$ classes in the case of the last one. In Population IV the link probabilities were generated using a mixed logit model with fixed effects associated with the sampled sites and random effects associated with the people.

Since in each of the populations the proposed estimators were computing using $C = 2$ latent classes, we have that in Population I no misspecification problem was present; in Population II, the distribution of the M_is was misspecified; in Population III, the number of latent classes was misspecified, and in Population IV, the model of the link probabilities was misspecified.

The simulation study was carried out by repeatedly selecting $r = 5000$ samples from each of the populations by using the sampling design described in Section 2 with initial sample size $n = 20$. From each sample the following estimators of τ_1, τ_2 and τ were computed: the proposed conditional MLEs $\hat{\tau}_1^C$, $\hat{\tau}_2^C$ and $\hat{\tau}^C$ obtained using $C = 2$ classes;

Table 2: Relative biases and square roots of relative mean square errors of the estimators of the population sizes. Results based on 5000 samples.

		Population I		Population II		Population III		Population IV	
		f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2
		0.48	0.40	0.46	0.44	0.44	0.41	0.46	0.41
Estimator		r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$
Latent class	$\hat{\tau}_1^C$	$0.00^{(2)}$	$0.08^{(2)}$	$0.00^{(9)}$	$0.09^{(9)}$	$-0.02^{(1)}$	$0.10^{(1)}$	-0.23	0.23
($C = 2$)	$\hat{\tau}_2^C$	$0.11^{(26)}$	$0.54^{(26)}$	$0.07^{(32)}$	$0.66^{(32)}$	$-0.07^{(14)}$	$0.42^{(14)}$	$-0.31^{(0)}$	$0.32^{(0)}$
cond. MLEs	$\hat{\tau}^C$	$0.04^{(28)}$	$0.19^{(28)}$	$0.02^{(37)}$	$0.20^{(37)}$	$-0.04^{(15)}$	$0.16^{(15)}$	$-0.26^{(0)}$	$0.26^{(0)}$
Mixed logit	$\tilde{\tau}_1^C$	$-0.12^{(0)}$	$0.13^{(0)}$	$-0.15^{(1)}$	$0.17^{(1)}$	$-0.10^{(1)}$	$0.11^{(1)}$	$0.14^{(1)}$	$0.15^{(1)}$
normal	$\tilde{\tau}_2^C$	$-0.28^{(1)}$	$0.29^{(1)}$	$-0.32^{(0)}$	$0.33^{(0)}$	$-0.22^{(1)}$	$0.23^{(1)}$	$0.32^{(0)}$	$0.37^{(0)}$
cond. MLEs	$\tilde{\tau}^C$	$-0.17^{(1)}$	$0.18^{(1)}$	$-0.21^{(1)}$	$0.21^{(1)}$	$-0.14^{(1)}$	$0.15^{(1)}$	$0.20^{(1)}$	$0.21^{(1)}$
Homogeneous	$\check{\tau}_1$	-0.19	0.20	$-0.17_{(0)}$	$0.18_{(0)}$	-0.10	0.11	-0.32	0.32
link-prob.	$\check{\tau}_2$	-0.32	0.33	$-0.36_{(0)}$	$0.37_{(0)}$	-0.23	0.24	-0.45	0.45
MLEs	$\check{\tau}$	-0.23	0.24	$-0.23_{(0)}$	$0.24_{(0)}$	-0.14	0.15	-0.36	0.37

Notes: f_k, sampling fraction in U_k. Upper script in parentheses indicates the percentage of samples in which the estimator was not obtained because of numerical convergence problems or because its value was greater than 10,000.

the conditional MLEs $\tilde{\tau}_1^C$, $\tilde{\tau}_2^C$ and $\tilde{\tau}^C$ proposed by Félix-Medina et al. (2009) and derived from a mixed logit model with fixed effects for the sites and random normal effects for the people and no interaction effects, and the MLEs $\check{\tau}_1$, $\check{\tau}_2$ and $\check{\tau}$ proposed by Félix-Medina and Thompson (2004) and derived under the assumption of homogeneous link probabilities.

The performance of an estimator $\hat{\tau}$ of τ, say, was evaluated by means of its relative bias (r-bias) and the square root of its relative mean square error (r-mse) defined by r-bias $= \sum_1^r(\hat{\tau}_i - \tau)/(r\tau)$ and $\sqrt{\text{r-mse}} = \sqrt{\sum_1^r(\hat{\tau}_i - \tau)^2/(r\tau^2)}$, where $\hat{\tau}_i$ was the value of $\hat{\tau}$ obtained in the i-th sample.

It is worth noting that in many of the samples the proposed estimator $\hat{\tau}_2^C$, and consequently $\hat{\tau}^C$ were not computed because of problems of numerical convergence in the algorithm of maximization or overestimation problems (estimates greater than 10000). Therefore, the results of the simulation study for each estimator, which are shown in Table 2, were obtained using only the samples in which the estimator was computed. For each estimator the proportion of samples in which the estimator was not computed is shown in parentheses in Table 2.

From the results we can see that in Populations I, II and III, where the link probabilities were generated using a latent class model, the proposed estimators did not show serious problems of bias, although the estimator $\hat{\tau}_2^C$ presented problems of instability, which affected the stability of $\hat{\tau}^C$. The problems of instability, as well as those of convergence and overestimation were consequence of the not large enough sampling fractions used in U_2 (the sampling fraction f_2 was between 0.40 and 0.44). In these populations the other types of estimators showed biases of larger magnitudes than those of the proposed estimators; however, they were more stable than the proposed ones. In fact, in terms of the r-mse the best estimator of τ_1 was the proposed estimator $\hat{\tau}_1^C$, whereas the best estimators of τ_2 and τ were the ones obtained from the mixed logit normal model $\tilde{\tau}_2^C$ and $\tilde{\tau}^C$.

Finally, in the case of Population IV, every one of the estimators showed serious problems of biases that deteriorated its performance.

7. Conclusions and Suggestions for Further Research

Th results of the simulation study show that the proposed estimators of τ_2 and τ present serious problems of numerical stability and variability when the sampling fraction used in U_2 is not large enough. In addition, the proposed estimators seem to be not robust to deviations from the latent class model for the link probabilities. However, more comprehensive studies than that carried out here need to be implemented to analyze in detail the performance of the proposed estimators.

Acknowledgments

This research was supported by grant PROFAPI 2011/057 of the Universidad Autónoma de Sinaloa and grant OP/PIFI-2011-25MSU0013B-06-03 of SEP.

REFERENCES

Coull, B. A., and Agresti, A. (1999), "The use of mixed logit models to reflect heterogeneity in capture-recapture studies," *Biometrics*, 55, 294-3-01.

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

Félix-Medina, M. H., Monjardin, P. E., and Aceves-Castro, A.N. (2009), "Link-tracing sampling: estimating the size of a hidden population in presence of heterogeneous nomination probabilities," in *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association, pp. 4020–4033.

Fienberg, S. E. (1972), "The multiple recapture census for closed populations and incomplete 2^k contingency tables," *Biometrika*, 59, 591–603.

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

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