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

Martín H. Félix-Medina*

Abstract

We propose Horvitz-Thompson type of estimators of the total and mean of the values of a variable of interest associated with the elements of a hard-to-reach population sampled by the variant of link-tracing sampling proposed by Félix-Medina and Thompson (2004). As examples of this type of population are drug users, homeless people and sex workers. In this sampling variant a frame of sites where the members of the population tend to gather, such as parks and bars, is constructed. The frame is not assumed to cover the whole population. A cluster sample of elements is selected from the frame, where the clusters are the sites, and the sampled elements are asked to named other members of the population. The proposed estimators do not use design-based inclusion probabilities, but model-based inclusion probabilities which are heterogeneous, that is, they depend on the sampled people. These probabilities are derived from a model proposed by Félix-Medina et al. (2009) and are estimated by maximum likelihood estimators. The performance of the proposed estimators is evaluated by simulation studies and the results show that their performance is at least acceptable.

Key Words: chain-referral sampling, capture-recapture, Horvitz-Thompson estimator, inclusion probabilities, 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 that 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 that is not assumed to cover the whole population. The sites are venues where the members of the population might be found with high probabilities, such as public parks, bars and blocks. The members of the population who belong 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) derived MLEs of the population size under the assumption that the link probabilities depend on the named persons, that is,

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

that they are heterogeneous. In this work we use the model proposed by these authors and consider the problem of estimating the total and the mean of a variable of interest, such as monthly drug expenses, age of first drug use, number of drug user partners and presence of HIV. It is worth noting that Félix-Medina and Monjardin (2010) also considered the problem addressed in this work, but they proposed estimators derived under the assumption of homogeneous link probabilities.

The structure of this paper is as follows. In Section 2 we introduce the LTS variant proposed by Félix-Medina and Thompson (2004). In Section 3 we present the models and MLEs of the population sizes proposed by Félix-Medina et al. (2009). In Section 4 we present the proposed Horvitz-Thompson-like estimators of the total and the mean. In Section 5 we present the results of two simulation studies, and finally, in Section 6 we present some conclusions and suggestions for future research.

2. Sampling design and notation

Félix-Medina and Thompson (2004) proposed the following variant of LTS. Let U be a finite population of an unknown number τ of people. We assume that a portion U_1 of U is covered by a sampling frame of N sites $A_1, \ldots, A_N$, where the members of the population can be found with high probability. We suppose that we have a criterion that allows us to assign a person in U_1 to only one site in the frame. Notice that we are not assuming that a person could not be found in different sites, but that, as in ordinary cluster sampling, we are able to assign him or her to only one site, for instance, the site where he or she spends most of his or her time. 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. The M_i members of the population who belong to the sampled site A_i are identified and their associated y-values of the variable of interest y are recorded, $i = 1, \ldots, n$. 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. Let $X_{ij}^{(k)} = 1$ if person $j \in U_k - A_i$ is linked to site $A_i \in S_A$ and $X_{ij}^{(k)} = 0$ if $j \in A_i$ or j is not linked to A_i, $i = 1, \ldots, n$, $k = 1, 2$. For each named person we record the value of the variable of interest y associated with him or her, the sampled sites that are linked to him or her, and the subset of U: $U_1 - S_0$, a specific $A_i \in S_A$ or U_2, that contains him or her.

3. MLEs of the population sizes

Félix-Medina et al. (2009) proposed MLEs of the population sizes τ_1, τ_2 and τ, which derived from the following assumptions. The variables M_i, $i = 1, \ldots, N$, are supposed to be independent identically distributed Poisson random variables with mean λ_1. Notice that this implies that the joint conditional distribution of the vector of variables $\mathbf{M}_s = (M_1, \ldots, M_n, \tau_1 - M)$, where $M = \sum_1^n M_i$, given that $\tau_1 = \sum_1^N M_i$, is multinomial with parameter of size τ_1 and vector of probabilities $(1/N, \ldots, 1/N, 1 - n/N)$. The link indicator variables $X_{ij}^{(k)}$s are supposed to be

independent Bernoulli random variables with means $p_{ij}^{(k)}$s, where the means or link probabilities $p_{ij}^{(k)}$s are given by the following Rasch model:

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

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

Let $\mathbf{X}_j^{(k)} = (X_{1j}^{(k)}, \ldots, X_{nj}^{(k)})$ be the n-dimensional vector of link indicator variables $X_{ij}^{(k)}$ associated with the j-th person in $U_k - S_0$. The conditional probability that $\mathbf{X}_j^{(k)}$ equals $\mathbf{x} = (x_1, \ldots, x_n)$ given $\beta_j^{(k)}$, that is, the probability that the j-th person in $U_k - S_0$ is linked to only the sites $A_i \in S_A$ such that the i-th element x_i of $\mathbf{x}$ equals 1, is

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

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

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

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

Félix-Medina et al. (2009), following Coull and Agresti (1999), approximated $\pi_{\mathbf{x}}^{(k)}(\alpha^{(k)}, \sigma_k)$ by using the Gaussian quadrature method, that is by

$$\tilde{\pi}_{\mathbf{x}}^{(k)}(\alpha^{(k)}, \sigma_k) = \sum_{t=1}^{q} \prod_{i=1}^{n} \frac{\exp[x_i(\alpha_i^{(k)} + \sigma_k z_t)]}{1 + \exp(\alpha_i^{(k)} + \sigma_k z_t)} \nu_t, \tag{2}$$

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

Similarly, the Gaussian quadrature approximation to the probability $\pi_{\mathbf{x}}^{(A_i)}(\alpha_{-i}^{(1)}, \sigma_1)$ that the vector of link indicator variables associated with a randomly person selected from the sampled site A_i equals an $(n-1)$-dimensional vector $\mathbf{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ whose elements are zeros and ones is

$$\tilde{\pi}_{\mathbf{x}}^{(A_i)}(\alpha_{-i}^{(1)}, \sigma_1) = \sum_{t=1}^{q} \prod_{j \neq i}^{n} \frac{\exp[x_j(\alpha_j^{(1)} + \sigma_1 z_t)]}{1 + \exp(\alpha_j^{(1)} + \sigma_1 z_t)} \nu_t, \tag{3}$$

where $\alpha_{-i}^{(1)} = (\alpha_1^{(1)}, \ldots, \alpha_{i-1}^{(1)}, \alpha_{i+1}^{(1)}, \ldots, \alpha_n^{(1)})$, $i = 1, \ldots, n$.

Also, Félix-Medina et al. (2009), following Coull and Agresti (1999), used Sanathanan's (1972) approach to derive conditional MLEs $\hat{\alpha}^{(k)}$ and $\hat{\sigma}_k$ of $\alpha^{(k)}$ and σ_k, $k = 1, 2$, given the number of distinct people in $U_k - S_0$ that are linked to at least one site in S_A. The values of these estimators are obtained by maximizing numerically the corresponding conditional likelihood functions. Then, they obtained the conditional MLEs $\hat{\tau}_1$ and $\hat{\tau}_2$ of τ_1 and τ_2 by replacing $\alpha^{(k)}$ and σ_k by $\hat{\alpha}^{(k)}$ and $\hat{\sigma}_k$ in the remaining parts of the likelihood function and maximizing these parts with respect to τ_1 and τ_2. Thus, the estimators $\hat{\tau}_1$ and $\hat{\tau}_2$ proposed by Félix-Medina et al. (2009) are

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

where $\hat{\hat{\pi}}_{\mathbf{0}}^{(k)}(\hat{\alpha}^{(k)}, \hat{\sigma}_k)$ is an estimator of the probability $\pi_{\mathbf{0}}^{(k)}(\alpha^{(k)}, \sigma_k)$ that a randomly selected person from $U_k - S_0$ is not linked to any site in S_A. The conditional MLE of τ is $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

4. Estimators of the total and mean

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

Since we cannot compute the design-based inclusion probabilities of the sampled elements because we do not know the sites in the frame that are linked to each sampled person, we compute conditional model-based inclusion probabilities given the sites $A_i \in S_A$. These probabilities are given by

$$\pi_j^{(1)}(\alpha^{(1)}, \sigma_1, \beta_j^{(1)}) = 1 - (1 - n/N)\prod_{i=1}^{n}(1 - p_{ij}^{(1)}) \quad \text{if } j \in U_1 \tag{4}$$

$$\pi_j^{(2)}(\alpha^{(2)}, \sigma_2, \beta_j^{(2)}) = 1 - \prod_{i=1}^{n}(1 - p_{ij}^{(2)}) \quad \text{if } j \in U_2. \tag{5}$$

The probabilities $\pi_j^{(k)}(\alpha^{(k)}, \sigma_k, \beta_j^{(k)})$s are not known because depend on unknown parameters. However, we could estimate them by estimating those parameters and replacing in (4) and (5) the parameters by their estimates. Estimators of $\alpha^{(k)}$ and σ_k have already been derived by Félix-Medina et al. (2009). We will next derive a predictor of the random effect $\beta_j^{(k)}$.

Thus, given the subset U_2, $U_1 - S_0$ or $A_{i'} \in S_A$ that contains the element j, we have that the conditional joint probability density function of the vector $\mathbf{X}_j^{(k)}$ of link indicator variables associated with that element and the random effect $\beta_j^{(k)}$ is

$$f(\mathbf{x}_j^{(k)}, \beta_j^{(k)} | j \in U_k - S_0) = \Pr(\mathbf{X}_j^{(k)} = \mathbf{x}_j^{(k)} | \beta_j^{(k)}, j \in U_k - S_0) f(\beta_j^{(k)})$$

$$\propto \prod_{i=1}^{n} [p_{ij}^{(k)}]^{x_{ij}^{(k)}} [1-p_{ij}^{(k)}]^{1-x_{ij}^{(k)}} \exp[-(\beta_j^{(k)})^2/2\sigma_k^2]$$

$$\text{if } j \in U_k - S_0,\ k=1,2, \text{ and}$$

$$f(\mathbf{x}_j^{(1)}, \beta_j^{(1)} | j \in A_{i'} \in S_A) \propto \prod_{i \neq i'}^{n} [p_{ij}^{(1)}]^{x_{ij}^{(1)}} [1-p_{ij}^{(1)}]^{1-x_{ij}^{(1)}} \exp[-(\beta_j^{(1)})^2/2\sigma_1^2]$$

$$\text{if } j \in A_{i'} \in S_A,\ i'=1,\ldots,n.$$

We will propose as a prediction of $\beta_j^{(k)}$ the value $\hat{\beta}_j^{(k)}$ that maximizes the conditional joint probability density function of $\mathbf{X}_j^{(k)}$ and $\beta_j^{(k)}$. This procedure yields that $\hat{\beta}_j^{(k)}$ is given as the solution to the following equation:

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

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

Notice that this equation implies that the predictor $\hat{\beta}_j^{(k)}$ of $\beta_j^{(k)}$ depends on the number of sites that are linked to the element j, but not on the particular sites to which that element is linked. Thus, if two persons j and j' in $U_k - S_0$ are linked to the same number of sites in S_A, the predictors $\hat{\beta}_j^{(k)}$ and $\hat{\beta}_{j'}^{(k)}$ are equal one another. The same happens for two persons in $A_i \in S_A$.

Thus, model-based Horvitz-Thompson-like estimators (HTLEs) of the totals Y_k, $k=1,2$, and Y are

$$\hat{Y}_k = \sum_{j \in S_k^*} y_{kj}/\hat{\pi}_j^{(k)}(\hat{\alpha}^{(k)}, \hat{\sigma}_k, \hat{\beta}_j^{(k)}), \quad k=1,2, \text{ and}$$
$$\hat{Y} = \hat{Y}_1 + \hat{Y}_2.$$

Similarly, model-based HTLEs of the means $\bar{Y}_k$ and $\bar{Y}$ are

$$\hat{\bar{Y}}_k = \hat{Y}_k/\hat{\tau}_k,\ k=1,2, \quad \text{and} \quad \hat{\bar{Y}} = \hat{Y}/\hat{\tau}.$$

5. Monte Carlo studies

We carried out two simulation studies to explore the performance of the proposed estimators of the population totals and means. In each of the studies we constructed populations from which samples were repeatedly selected using the sampling design described in Section 2. In the first study we constructed two artificial populations, whereas in the second study we used data from the Colorado Springs study on transmission of HIV/AIDS to construct a population.

5.1 First simulation study

We considered two finite populations of $N=150$ values M_i. In Population I the values were generated from a Poisson distribution with mean 8.0, whereas in Population II from a zero truncated negative binomial distribution with mean 8.0 and variance 24.0. In Table 1 are displayed the characteristics of each population. Notice that in

Table 1: Characteristics of the artificial populations generated for the first study.

Population I	Population II
$N = 150$	$N = 150$
$M_i \sim$ Poisson	$M_i \sim$ Zero-truncated neg. bin.
$E(M_i) = 8.0 \quad V(M_i) = 8.0$	$E(M_i) = 8.0 \quad V(M_i) = 24.0$
$\tau_1 = 1209 \quad \tau_2 = 400 \quad \tau = 1609$	$\tau_1 = 1208 \quad \tau_2 = 400 \quad \tau = 1608$
$Y_1 = 50135.6 \quad Y_2 = 16793.7$	$Y_1 = 50244.8 \quad Y_2 = 16876.9$
$Y = 66929.4$	$Y = 67121.7$
$\bar{Y}_1 = 41.5 \quad \bar{Y}_2 = 42.0 \quad \bar{Y} = 41.6$	$\bar{Y}_1 = 41.6 \quad \bar{Y}_2 = 42.2 \quad \bar{Y} = 41.7$

both populations the value of τ_2 was set to 400. The values of the link probabilities $p_{ij}^{(k)}$'s were obtained by means of Rasch model (1), where $\alpha_i^{(k)} = c_k/(M_i^{1/4} + 0.001)$, $c_1 = -5.7$ and $c_2 = -6.5$. The $\beta_j^{(k)}$'s were obtained by sampling from the $N(0, 1)$ distribution. The values of these parameters were such that the average values of $p_{ij}^{(1)}$ and $p_{ij}^{(2)}$ were about 0.04 and 0.03. In addition, since we used an initial sample of size $n = 15$, the sampling fractions in U_1 and U_2 were about 0.4 and 0.3. The value $y_j^{(k)}$ of the variable of interest Y was obtained by sampling from the noncentral chi-square distribution with two degrees of freedom and noncentrality parameter $\theta_j^{(k)} = 0.4 \exp(\beta_j^{(k)})/[1 + \exp(\beta_j^{(k)})]$. The use of these distributions yielded that the y-values and the inclusion probabilities were positively correlated with correlation coefficients $\rho_1 = 0.73$ and $\rho_2 = 0.71$ for the elements in U_1 and U_2, respectively.

The simulation study was carried out by replicating $r = 10000$ times the following procedure. From each population of $N = 150$ values of M_is a SRSWOR of $n = 15$ values was selected. From the i-th selected value, $i = 1, \ldots, n$, the values of $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ were obtained from Bernoulli distributions with means $p_{ij}^{(1)}$, $j = 1, \ldots, \tau_1 - M_i$, and $p_{ij}^{(2)}$, $j = 1, \ldots, \tau_2$, respectively. These data on the M_is and the $X_{ij}^{(k)}$s were used to compute the estimates of the population totals and means.

In this study we considered the estimators $\{\hat{Y}_1, \hat{Y}_2, \hat{Y}\}$ and $\{\hat{\hat{Y}}_1, \hat{\hat{Y}}_2, \hat{\hat{Y}}\}$ proposed in this paper, and the following two types of estimators proposed by Félix-Medina and Monjardin (2010). The estimators $\{\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}\}$ and $\{\tilde{\tilde{Y}}_1, \tilde{\tilde{Y}}_2, \tilde{\tilde{Y}}\}$ based on the MLEs of the population sizes proposed by Félix-Medina and Thompson (2004) and derived under the assumption of homogeneous $p_{ij}^{(k)}$s, and the estimators $\{\check{Y}_1, \check{Y}_2, \check{Y}\}$ and $\{\check{\check{Y}}_1, \check{\check{Y}}_2, \check{\check{Y}}\}$ based on the Bayesian-assisted estimators of the population sizes proposed by Félix-Medina and Monjardin (2006), also derived under the assumption of homogeneous $p_{ij}^{(k)}$s, and using the following initial distributions for τ_1, τ_2 and $\alpha_i^{(k)} = \ln[p_i^{(k)}/(1 - p_i^{(k)})]$, where $p_i^{(k)}$ is given by (1), but setting $\beta_j^{(k)} = 0$; $\xi(\tau_1|\lambda_1) \propto (N\lambda_1)^{\tau_1}/\tau_1!$ and $\xi(\lambda_1) \propto \lambda_1^{a_1-1} \exp(-b_1\lambda_1)$; $\xi(\tau_2|\lambda_2) \propto \lambda_2^{\tau_2}/\tau_2!$ and $\xi(\lambda_2) \propto \lambda_2^{a_2-1} \exp(-b_2\lambda_2)$, and $\xi(\alpha_i^{(k)}|\theta_k) \propto \exp[-(\alpha_i^{(k)} - \theta_k)^2/2\sigma_k^2]$ and $\xi(\theta_k) \propto \exp[-(\theta_k - \mu_k)^2/2\gamma_k^2]$, where $a_1 = 1.0$, $b_1 = 0.1$, $a_2 = 6.0$, $b_2 = 0.01$, $\mu_k = -3.5$ and $\sigma_k^2 = \gamma_k^2 = 9.0$. These values assigned to the parameters of the initial distributions made them practically non-informative. The Gaussian quadrature approximations (2) and (3) to the probabilities $\pi_{\mathbf{x}}^{(k)}(\alpha^{(k)}, \sigma_k)$ and $\pi_{\mathbf{x}}^{(A_i)}(\alpha_{-i}^{(1)}, \sigma_1)$ were computed using $q = 20$ terms. The performance of an estimator $\hat{Y}$ say, of Y was evaluated by

Table 2: Populations constructed using artificial data: relative biases and square roots of relative mean square errors of estimators of the population totals and means. Results based on 10000 samples.

	Population I		Population II			Population I		Population II	
Esti-					Esti-				
mator	r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$	mator	r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$
$\hat{Y}_1$	-0.03	0.08	-0.04	0.08	$\hat{\tilde{Y}}_1$	-0.03	0.04	-0.04	0.04
$\hat{Y}_2$	-0.16(35)	0.26(35)	-0.18(22)	0.25(22)	$\hat{\tilde{Y}}_2$	-0.06(35)	0.17(35)	-0.09(22)	0.15(22)
$\hat{Y}$	-0.06(35)	0.10(35)	-0.08(22)	0.10(22)	$\hat{\tilde{Y}}$	-0.05(35)	0.06(35)	-0.05(22)	0.06(22)
$\tilde{Y}_1$	-0.15	0.15	-0.18	0.19	$\tilde{\tilde{Y}}_1$	0.17	0.17	0.18	0.18
$\tilde{Y}_2$	-0.31	0.32	-0.33	0.34	$\tilde{\tilde{Y}}_2$	0.23	0.24	0.22	0.22
$\tilde{Y}$	-0.19	0.19	-0.22	0.22	$\tilde{\tilde{Y}}$	0.18	0.18	0.19	0.19
$\check{Y}_1$	-0.15	0.15	-0.18	0.19	$\check{\tilde{Y}}_1$	0.17	0.17	0.18	0.18
$\check{Y}_2$	-0.28	0.29	-0.30	0.31	$\check{\tilde{Y}}_2$	0.21	0.21	0.19	0.20
$\check{Y}$	-0.18	0.18	-0.21	0.21	$\check{\tilde{Y}}$	0.18	0.18	0.18	0.18

Notes: Number in parentheses indicates the number of samples in which the estimator was not obtained. $\hat{Y}_k$ and $\hat{\tilde{Y}}_k$, proposed estimators. $\tilde{Y}_k$ and $\tilde{\tilde{Y}}_k$, as well as $\check{Y}_k$ and $\check{\tilde{Y}}_k$, estimators proposed by Félix-Medina and Monjardin (2010).

means of its relative bias (r-bias) and the square root of its relative mean square error (r-mse) defined by r-bias $= \sum_1^r (\hat{Y}_i - Y)/(rY)$ and $\sqrt{\text{r-mse}} = \sqrt{\sum_1^r (\hat{Y}_i - Y)^2/(rY^2)}$, where $\hat{Y}_i$ was the value of $\hat{Y}$ obtained in the i-th trial.

The results are shown in Table 2. The following are the main aspects of the results. First, the proposed estimators of the totals and means showed smaller values of r-bias and $\sqrt{\text{r-mse}}$ than the values showed by the estimators derived under the assumption of homogeneous link probabilities. Second, the estimators of the total Y_2, including the proposed estimator, showed problems of bias that increased the values of the square roots of their r-mse. Third, the estimators of the means performed better than the corresponding estimators of the totals. Finally, the performance of every one of the estimators was practically not affected by the type of population. So the proposed estimators are robust to deviations from the assumption of the Poisson distribution of the M_is. It is worth noting that in some of the samples the proposed estimators were not computed because convergence problems precluded the calculation of the corresponding estimators of the population sizes and inclusion probabilities.

5.2 Second simulation study

In this simulation study we constructed a population using data from the Colorado Springs study on heterosexual transmission of HIV/AIDS, described by Potterat et al. (1993), Rothenberg et al. (1995) and Potterat et al. (2004), among others. That epidemiological research was focused on a population of people who lived in the Colorado Springs metropolitan area from 1982-1992 and who were at high risk of acquiring and transmitting HIV. That population included drug users, sex workers and their personal contacts, defined as those persons with whom they had close

social, sexual or drug-associated relations. In that study, 595 initial responders were selected in a non-random fashion and they were asked for a complete enumeration of their personal contacts. A total of 7379 contacts who were not in the set of the initial responders were named and included in the study. In our simulation study the set U_1 was defined as the set of the 595 initial responders and, as in Félix-Medina and Monjardin (2010), they were grouped into $N = 105$ clusters of sizes M_is generated by sampling from a zero-truncated negative binomial distribution with parameter of size 2.5 and probability 2/3. The sample mean and variance of the M_is were 5.67 and 15.03, respectively. A person was defined to be linked to a cluster if he or she was a personal contact of at least one element in that cluster. Since, approximately 95% of the 7379 contacts of the initial responders were linked to only one cluster, and this could affect the performance of the proposed estimators, in our study we considered the subset of the 7379 contacts formed by the 415 persons who were linked to at least two clusters plus the 379 sex workers who were linked to only one cluster. In our simulation study, the set U_2 was defined as that subset of 794 contacts. Thus, $\tau_1 = 595$, $\tau_2 = 794$ and $\tau = 1389$. The variable of interest was a binary variable which equaled 1 if a person was a sex worker and equaled 0 otherwise. It is worth noting that this population is the same as the one called "reduced population" by Félix-Medina and Monjardin (2010). We used an initial sample of size $n = 20$, which yielded that the sampling fractions for U_1 and U_2 were about 0.4 and 0.3.

The simulation experiment was carried out as the previous one, except that any time that the i-th cluster was included in an initial sample, every one of the people linked to that cluster was included in the sample. The results of the study are shown in Table 3. The following are the main aspects of the results. First, every one of the estimators of Y_1 performed acceptably well and similarly. Second, the proposed estimator $\hat{\bar{Y}}_1$ of $\bar{Y}_1$ was the one of the best performance, although its performance was only moderate. Third, in estimating the other parameters, the estimators derived under the assumption of homogeneous link probabilities were the ones of the best performance, although their performance was hardly acceptable. Fourth, in some of the samples the proposed estimators were not computed because of convergence problems.

6. Conclusions and suggestions for future research

Based on the results of the Monte Carlo studies we have the following conclusions. The proposed estimators seem to work acceptably well when every of the assumptions is satisfied or when only the assumption of the Poisson distribution of the M_is is not satisfied. However, they do not seem to be robust to deviations from the other assumptions, as we can see from the results of the second simulation study. Unfortunately, at this moment we do not know which assumptions need to be satisfied in order that the estimators perform acceptably well. Thus, additional Monte Carlo studies need to be carried out to obtain more information about this problem of lack of robustness. The performance of the estimators derived under the assumption of homogeneous link probabilities is not good when this assumption is not satisfied, as we can see from the results of the first simulation study. Therefore, it is surprising that in the second simulation study those estimators performed better than the proposed estimators. A possible explanation for this result is that the small degree of heterogeneity of the $p_{ij}^{(2)}$s ($\hat{\sigma}_2 = 0.12$) caused that the estimators $\tilde{Y}_2$ and $\check{Y}_2$ had negative r-biases of moderate magnitudes which were canceled out by

Table 3: Population constructed using data from the Colorado Springs study: Relative biases and square roots of relative mean square errors of estimators of the population totals and means. Results based on 10000 samples.

Estimator	r-bias	$\sqrt{\text{r-mse}}$	Estimator	r-bias	$\sqrt{\text{r-mse}}$
$\hat{Y}_1$	0.04(29)	0.14(29)	$\hat{\bar{Y}}_1$	0.16(29)	0.20(29)
$\hat{Y}_2$	-0.23(29)	0.35(29)	$\hat{\bar{Y}}_2$	-0.33(29)	0.36(29)
$\hat{Y}$	-0.17(29)	0.27(29)	$\hat{\bar{Y}}$	-0.20(29)	0.23(29)
$\tilde{Y}_1$	0.06	0.14	$\tilde{\bar{Y}}_1$	0.37	0.39
$\tilde{Y}_2$	-0.12	0.41	$\tilde{\bar{Y}}_2$	-0.30	0.32
$\tilde{Y}$	-0.08	0.31	$\tilde{\bar{Y}}$	-0.12	0.17
$\check{Y}_1$	0.06	0.14	$\check{\bar{Y}}_1$	0.37	0.39
$\check{Y}_2$	-0.19	0.33	$\check{\bar{Y}}_2$	-0.29	0.32
$\check{Y}$	-0.13	0.25	$\check{\bar{Y}}$	-0.12	0.16

Notes: Number in parentheses indicates the number of samples in which the estimator was not obtained. $\hat{Y}_k$ and $\hat{\bar{Y}}_k$, proposed estimators. $\tilde{Y}_k$ and $\tilde{\bar{Y}}_k$, as well as $\check{Y}_k$ and $\check{\bar{Y}}_k$, estimators proposed by Félix-Medina and Monjardin (2010). $\hat{\sigma}_1 = 1.12$, $\hat{\sigma}_2 = 0.12$.

the positive biases of $\tilde{Y}_1$ and $\check{Y}_1$. This did not happen with the proposed estimators because the magnitude of the r-bias of $\hat{Y}_2$ was not small enough to be canceled out by the positive r-bias of $\hat{Y}_1$.

Regardless of the only fair performance of the proposed estimators, we consider that they are a better alternative than that based on the estimators derived under the homogeneity assumption. Obviously, our proposal still need to be improved. As we indicated earlier, additional simulation studies need to be carried out to obtain information about the lack of robustness of the proposed estimators. That information could be used to modify the estimators so that they have better robustness properties than the ones proposed in this work. Other problem that we have not considered is interval estimation. A possible solution to this problem is to use bootstrap to construct confidence intervals or, if the previous solution is computationally very expensive, to use bootstrap to construct variance estimators and then to obtain Wald confidence intervals.

Acknowledgments

This research was supported by grant PROFAPI 2011/057 of the Universidad Autónoma de Sinaloa.

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., and Monjardin, P. E. (2006), "Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: a Bayesian assisted approach," *Survey Methodology*, 32, 187–195.

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.

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

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

Potterat, J. J., Woodhouse, D. E., Muth, S. Q., Rothenberg, R. B., Darrow, W. W., Klovdahl, A. S., and Muth, J. B. (2004), "Network dynamism: History and lessons of the Colorado Springs study," in *Network epidemiology: a handbook for survey design and data collection*, ed. M. Morris, New York: Oxford University Press, pp. 87–114.

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

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