Graubard, B.I. and Korn, E.L. (1996). Survey Inference for Subpopulations. American Journal of Epidemiology, 144, 102–106.
Hansen, M.H., Hurwitz, W.H., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol. II. New York: Wiley.
Joanes, D.N. and Gill, C.A. (1998). Comparing Measures of Sample Skewness and Kurtosis. The Statistician, 47, 183–189.
Johnson, E. and Rust, K. (1992). Population Inferences and Variance Estimation for NAEP Data. Journal of Educational Statistics, 17, 175–190.
Johnson, E.G., Rust, K.F., and Hansen, M.H. (1988). Weighting Procedures and Estimation of Sampling Variance. Chapter 8 in E.G. Johnson and R. Zwick, Focusing the New Design: The NAEP 1988 Technical Report. Washington DC: U.S. Department of Education.
Korn, E.L. and Graubard, B.I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni *t* Statistics. The American Statistician, 44, 270–276.
Korn, E.L. and Graubard, B.I. (1998). Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. Survey Methodology, 24, 193–201.
Korn, E.L. and Graubard, B.I. (1999). Analysis of Health Surveys. New York: John Wiley.
Kott, P.S. and Liu, Y. (2009). One-Sided Coverage Intervals for a Proportion Estimated from a Stratified Simple Random Sample. International Statistical Review, 77, 251–265.
Research Triangle Institute (2004). SUDAAN User's Manual, Release 9.0. Research Triangle Park, NC: Research Triangle Institute.
Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys, unpublished Ph.D. dissertation, Ann Arbor MI: University of Michigan.
Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. Journal of Official Statistics, 1, 381–397.
Rust, K. (1986). Efficient Replicated Variance Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 81–87.
Rust, K. and Kalton, G. (1987). Strategies for Collapsing Strata for Variance Estimation. Journal of Official Statistics, 3, 69–81.
Rust, K.F. and Rao, J.N.K. (1996). Variance Estimation for Complex Estimators in Sample Surveys. Statistical Methods in Medical Research, 5, 381–397.
Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
Satterthwaite, F.W. (1946). An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin, 2, 110–114.
Stata Corporation (2005). Survey Data Reference Manual, Release 9. College Station: Stata Press.
Westat (2000). WesVar 4.0 User's Guide, available at www.westat.com/wesvar. Rockville MD: Westat.
Wolter, K.M. (2007). Introduction to Variance Estimation, (Second edition). New York: Springer-Verlag.

Received October 2008
Revised January 2010

Journal of Official Statistics, Vol. 26, No. 4, 2010, pp. 603–631

Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations

Martín H. Félix-Medina[1] *and Pedro E. Monjardin*[1]

Félix-Medina and Thompson (2004) proposed a variant of link-tracing sampling in which it is assumed that only a portion of a hidden population, such as drug users or sex workers, is covered by a frame of sites where the members of the population can be found with high probability. A sample of sites is selected and the people on those sites are asked to nominate other members of the population to be included in the sample. We consider this sampling design, and propose several types of Horvitz-Thompson-like estimators of the total and the mean of a response variable, such as monthly drug expenses or number of sexual partners. We also propose Horvitz-Thompson-like estimators of the variances of the estimators of the total and the mean, as well as Wald confidence intervals for these parameters. The results of several simulation studies with real and artificial data indicate that point and interval estimators of the total and mean perform well as long as all the assumptions about the stated models are satisfied and the number of nominees in the portion of the population not covered by the frame is not small, but that their performance deteriorates as the number of nominees decreases. The results also indicate that the proposed estimators are robust to deviations from the model that describes the numbers of people found on the sites, but not to deviations from the assumption that every member of the population has the same probability of being nominated by a particular site. However, in this case, the proposed estimators still yield estimates of the parameters of the correct order of magnitude.

Key words: Capture-recapture; design-based approach; finite population; hard-to-access population; Horvitz-Thompson estimator; model-based approach; snowball sampling.

1. Introduction

Sampling hidden or hard-to-access human populations, such as drug-users, sex workers, homeless people and illegal workers, is a challenging problem because of the lack of appropriate sampling frames. Although several sampling methods have been proposed (see Magnani et al. 2005 and Kalton 2009 for recent reviews and references), according to Heckathorn (2002) two types of methods are the most commonly used in practical situations. These are location sampling, which is also known as time-and-space sampling, aggregation point sampling or intercept point sampling, and snowball sampling, which is also known as link-tracing sampling (LTS) or chain-referral sampling.

[1] Both at the Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán, Sinaloa 80010, México. Email: mhfelix@uas.uasnet.mx; pemo@uas.uasnet.mx
Acknowledgments: This research was supported by PROMEP/SEP Grant UASIN-EXB-01-01 and PIFI/SEP Grant 2004-25-07-1-15. We thank John Potterat and Steve Muth for allowing us to use the data from the Colorado Springs study and Steve Muth for his advice on the use of those data. We also thank the referees for their helpful suggestions and comments which improved this work.

In location sampling, a frame of primary units is constructed, where these units are formed as combinations of places and time segments in which the elements of the population tend to gather. A probability sample of primary units is selected and from each of them a sort of systematic sample of people is selected. For descriptions of this method, see MacKellar et al. (1996), Kalton (2001), Munhib et al. (2001), and McKenzie and Mistiaen (2009). Although design unbiased estimators of different characteristics of the population can be constructed, the main drawback of this method is that inferences are valid only for the part of the population covered by the frame. Clearly, if the sampled portion has very particular characteristics, the results will not be applicable to the whole population.

Snowball sampling consists in selecting an initial sample of members of the target population and asking them to nominate their friends who belong to the population. The nominated people who are not in the initial sample are added to the sample and they might be asked to nominate their friends who belong to the population. The sampling process continues in this way until a specified stopping rule is satisfied. For a review of different variants of LTS see Spreen (1992), Thompson and Frank (2000), and Heckathorn (2002).

Respondent-Driven Sampling (RDS) is a variant of snowball sampling that has recently been used in many studies of hidden populations completed in different countries. It was proposed by Heckathorn (1997), and improved by Heckathorn (2002), Salganik and Heckathorn (2004) and Volz and Heckathorn (2008). The particular characteristic of this method is that after purposively selecting some members of the population (initial seeds), the other participating members are recruited by previously recruited participants and not by the researchers. By modeling the recruitment process as an irreducible Markov chain, those authors use the stationary distribution of the chain to construct asymptotically unbiased estimators of means and proportions. In recent works, Gile and Handcock (2009) and Lu et al. (2010) have indicated that some of the assumptions might not be easy to satisfy in practical situations and that deviations from those assumptions might bias the estimators. Additionally, RDS is not appropriate for estimating the size of the population nor the total of a variable of interest unless the size of the population is known.

One variant of snowball sampling that allows the sampler to estimate the size of the population is the one proposed by Frank and Snijders (1994). In this one-wave snowball sampling variant the initial sample is assumed to be a Bernoulli sample, that is, the inclusions of people in the initial sample are supposed to be independent and equally probable. Furthermore, the probability that a particular person in the population is nominated by a specific person in the initial sample, which is called nomination probability, is assumed constant, that is, it does not depend on the nominator nor on the nominee. This premise is called homogeneity assumption. Clearly, both suppositions are difficult to satisfy in practical situations. Frank and Snijders (1994) reported that this method yielded a reasonable estimate of the number of heroin users in Groningen, but Dávid and Snijders (2002) reported an underestimate of the number of homeless in Budapest. The latter authors indicate that the underestimate might be a consequence of deviations from the assumption of a Bernoulli initial sample.

The problem of selecting, in practical situations, an initial sample that approximately satisfies the assumptions of a Bernoulli sample motivated Félix-Medina and Thompson (2004) to develop a variant of LTS in which the initial sample is selected from a sampling frame. Thus, those authors assume that the sampler can construct a frame of sites, such as

parks, bars and hospitals, where the members of the population can be found with high probability. They do not suppose that the frame covers the whole population, but only a portion. Then, a simple random sample without replacement of sites is selected from the frame and the people who belong to each sampled site are identified. Finally, as in ordinary LTS, the persons in the initial sample are asked to nominate their friends who belong to the population.

Those authors proposed maximum likelihood estimators (MLEs) of the population size derived from a probability model that describes the number of elements found in each site, and a model that considers homogeneous nomination probabilities. They found that the MLEs perform well provided that the nomination probabilities are not small. However, when these probabilities are small their proposed estimators have serious problems of bias.

Later, Félix-Medina and Monjardin (2006) proposed estimators of the population size derived under the Bayesian approach. They found that their estimators perform similarly to the MLEs when the nomination probabilities are not small and that there are no serious problems of bias when the nomination probabilities are relatively small. It is worth noting that those authors used the Bayesian approach to assist themselves in the construction of their estimators, but they used a frequentist design-based approach to make inferences that is, inferences were based on the probability distribution used to select the sample, as in finite population sampling, and not on the final distribution, as in the Bayesian approach.

This variant of LTS has not been applied in any practical situation, nor has the performance of the proposed estimators been analyzed under deviations from the homogeneity assumption. However, because the proposed estimators resemble those used in capture-recapture methodology, we should expect underestimation of the population size when this assumption is not satisfied.

In this article we consider the problem of estimating the total and the mean of a variable of interest from a sample selected by the variant of LTS proposed by Félix-Medina and Thompson (2004). Examples of population characteristics that we are interested in estimating are the total and the mean of monthly drug expenses, the number of people who consume more than one type of drug, and the average number of weekly clients in a sex worker population. To estimate these parameters we propose two classes of Horvitz-Thompson-like estimators. One class is based on the MLEs of the population size proposed by Félix-Medina and Thompson (2004), and the other is based on the Bayesian estimators of the population size proposed by Félix-Medina and Monjardin (2006). The proposed estimators are not real Horvitz-Thompson estimators because they use estimates of the model-based conditional inclusion probabilities as these probabilities are unknown. For each of the proposed estimators of the total and mean we derive an expression for a Horvitz-Thompson-like estimator of its variance. In addition, for interval estimation we propose Wald confidence intervals. Finally, we explore the performance of the proposed estimators and their robustness to deviations from the assumed models by means of three Monte Carlo studies.

2. Sampling Design and Notation

We will use the sampling design proposed by Félix-Medina and Thompson (2004). Thus, we will assume a finite hidden human population U of an unknown number τ of persons.

We will suppose that a portion $U_1 \subset U$ of the population can be covered by a sampling frame $A_1, \ldots, A_N$ of N sites where the members of the population can be found with high probability. We will assume that the researcher has a criterion that allows him or her to decide whether or not a person in U belongs to a site in the frame and in the affirmative case to assign that person to only one site. This does not mean that a person cannot be found on several sites, but that, as in ordinary cluster sampling, he or she is assigned to only one of those sites. We will denote by M_i the number of people who belong to the site A_i, $i = 1, \ldots, N$. Thus, the unknown numbers of people in U_1 and in $U_2 = U - U_1$ are $\tau_1 = \sum_1^N M_i$ and $\tau_2 = \tau - \tau_1$. We will suppose that associated with person j in U_k is the value $y_j^{(k)}$ of a nonrandom variable of interest y. The totals and means of the y-values in U_k and in U are $Y_k = \sum_{j \in U_k} y_j^{(k)}$ and $\bar{Y}_k = Y_k/\tau_k$, $k = 1, 2$, and $Y = Y_1 + Y_2$ and $\bar{Y} = Y/\tau$.

The sampling design is as follows. A simple random sample without replacement (SRSWOR) S_A of n sites $A_1, \ldots, A_n$ is selected. The members who belong to each sampled site are identified and their associated y-values are recorded. Let $M = \sum_1^n M_i$ be the number of people in the initial sample, that is, in $S_0 = \{\text{people in } A_i : A_i \in S_A\}$. Then, as in ordinary LTS, the people on each sampled site are asked to nominate other members of the population and the y-value associated with each nominated person is recorded. Let $X_{ij}^{(k)} = 1$ if person $j \in U_k - A_i$ is nominated by site $A_i \in S_A$ and $X_{ij}^{(k)} = 0$ if $j \in A_i$ or j is not nominated by A_i, $k = 1, 2$, where we say that a person is nominated by a site if at least one member of that site nominates him or her. We will suppose that the variables $X_{ij}^{(k)}$ are jointly independent, that is, that the nominations are made independently. Let $p_i^{(k)} = \Pr(X_{ij}^{(k)} = 1)$ be the probability that person j in $U_k - A_i$ is nominated by the site A_i, $i = 1, \ldots, n$, $k = 1, 2$. These probabilities are called nomination probabilities. Let $Z_i^{(k)} = \sum_{j \in U_k - A_i} X_{ij}^{(k)}$ be the number of people in $U_k - A_i$ nominated by the site A_i, and let R_1 and R_2 be the numbers of distinct people in $U_1 - S_0$ and U_2, respectively, that are nominated in the study. Notice that the nomination probabilities are assumed to be homogeneous, that is, that they do not depend on the people, but only on the sites. Furthermore, the model for $p_i^{(k)}$ implies that every person in $U_k - A_i$ can be nominated by site A_i. Clearly, these assumptions are difficult to satisfy in practical situations, but we expect the estimators we propose in this article to yield estimates of the parameters of the correct orders of magnitude.

3. Estimators of the Population Sizes

In addition to proposing the previous sampling variant, Félix-Medina and Thompson (2004) propose MLEs of the population sizes τ_1, τ_2 and τ. To obtain those estimators they suppose that the variables M_i, $i = 1, \ldots, N$, are independent identically distributed Poisson random variables with mean λ_1. This implies that given that $\tau_1 = \sum_1^N M_i$, the joint conditional distribution of the vector of variables $M_S = (M_1, \ldots, M_n, \tau_1 - M)$, where $M = \sum_{i=1}^n M_i$, is multinomial with parameter τ_1 and vector of probabilities $(1/N, \ldots, 1/N, 1 - n/N)$. Additionally, they assume that the conditional distribution of $X_{ij}^{(k)}$ given M_i is binomial with parameters 1 and $p_i^{(k)}$, $i = 1, \ldots, n$, $k = 1, 2$, which will be denoted by $X_{ij}^{(k)} | M_i \sim \text{bin}(1, p_i^{(k)})$. Notice that these assumptions imply that $Z_i^{(1)} | M_i \sim \text{bin}(\tau_1 - M_i, p_i^{(1)})$, $Z_i^{(2)} | M_i \sim \text{bin}(\tau_2, p_i^{(2)})$, $R_1 | M_s \sim \text{bin}(\tau_1 - M, 1 - Q_1)$ and $R_2 | M_s \sim \text{bin}(\tau_2, 1 - Q_2)$, where $Q_k = \prod_1^n (1 - p_i^{(k)})$, $k = 1, 2$. Using these assumptions

the authors show that the MLEs $\tilde{\tau}_1, \tilde{\tau}_2, \tilde{p}_i^{(1)}$ and $\tilde{p}_i^{(2)}$ of $\tau_1, \tau_2, p_i^{(1)}$ and $p_i^{(2)}$, $i = 1, \ldots, n$, are obtained as the solutions to the following equations:

$$\tilde{\tau}_1 = \frac{M + R_1}{1 - (1 - n/N)\prod_{i=1}^n \left(1 - \tilde{p}_i^{(1)}\right)}, \quad \tilde{p}_i^{(1)} = Z_i^{(1)}/(\tilde{\tau}_1 - M_i), \quad i = 1, \ldots, n$$

$$\tilde{\tau}_2 = \frac{R_2}{1 - \prod_{i=1}^n \left(1 - \tilde{p}_i^{(2)}\right)} \quad \text{and} \quad \tilde{p}_i^{(2)} = Z_i^{(2)}/\tilde{\tau}_2, \quad i = 1, \ldots, n \tag{1}$$

Thus, the MLE of τ is $\tilde{\tau} = \tilde{\tau}_1 + \tilde{\tau}_2$.

Later, Félix-Medina and Monjardin (2006) propose estimators of the population sizes derived under the Bayesian approach. They do that by assuming the previous models, and defining prior distributions for τ_k and $\alpha_i^{(k)} = \ln[p_i^{(k)}/(1 - p_i^{(k)})]$, $i = 1, \ldots, n$, $k = 1, 2$. Those authors consider the following three types of prior distributions for the τ_k's: (i) uniform distributions: $\pi(\tau_k) \propto 1$, $k = 1, 2$; (ii) Jeffreys' distributions: $\pi(\tau_k) \propto 1/\tau_k$, $k = 1, 2$; and (iii) Poisson-Gamma distributions: $\pi(\tau_1|\lambda_1) \propto (N\lambda_1)^{\tau_1}/\tau_1!$ and $\pi(\lambda_1) \propto \lambda_1^{a_1 - 1} e^{-b_1\lambda_1}$, and $\pi(\tau_2|\lambda_2) \propto \lambda_2^{\tau_2}/\tau_2!$ and $\pi(\lambda_2) \propto \lambda_2^{a_2 - 1} e^{-b_2\lambda_2}$, where a_1, b_1, a_2 and b_2 are known constants. They indicate that the first two distributions can be obtained as limit cases of the last distribution by setting $a_k = 1$, $b_k = 0$, $k = 1, 2$, for the Uniform distribution and $a_k = 0$, $b_k = 0$, $k = 1, 2$, for the Jeffreys' distribution. In the case of the $\alpha_i^{(k)}$'s the authors consider the following two-stage normal prior distribution: $\alpha_i^{(k)} | \theta_k \sim N(\theta_k, \sigma_k^2)$, $i = 1, \ldots, n$, and $\theta_k \sim N(\mu_k, \gamma_k^2)$, where μ_k, σ_k^2 and γ_k^2, $k = 1, 2$, are assumed known, and $N(\phi, \psi^2)$ denotes the normal distribution with mean ϕ and variance ψ^2. They also suppose that all the random vectors (τ_k, λ_k) and (α_k, θ_k), where $\alpha_k = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, $k = 1, 2$, are mutually independent.

The authors propose that τ_k and $\alpha_i^{(k)}$ be estimated by the mode of their joint posterior distribution. Thus, they find that the estimators $\hat{\tau}_k$ and $\hat{p}_i^{(k)} = \exp(\hat{\alpha}_i^{(k)})/[1 + \exp(\hat{\alpha}_i^{(k)})]$ of τ_k and $p_i^{(k)}$, $i = 1, \ldots, n$; $k = 1, 2$, are given as the solutions to the following equations:

$$\hat{\tau}_1 = \frac{M + R_1 + (1 - n/N)[N(a_1 - 1)/(N + b_1)]\prod_{i=1}^n \left(1 - \hat{p}_i^{(1)}\right)}{1 - (1 - n/N)[N/(N + b_1)]\prod_{i=1}^n \left(1 - \hat{p}_i^{(1)}\right)};$$

$$\hat{p}_i^{(1)} = \frac{\exp\left\{\hat{\alpha}_i^{(1)}\right\}}{1 + \exp\left\{\hat{\alpha}_i^{(1)}\right\}} = \frac{Z_i^{(1)}}{\hat{\tau}_1 - M_i} - \frac{\hat{\alpha}_i^{(1)} - \hat{\bar{\alpha}}^{(1)}}{(\hat{\tau}_1 - M_i)\sigma_1^2} - \frac{\hat{\bar{\alpha}}^{(1)} - \mu_1}{n(\hat{\tau}_1 - M_i)\nu_1}; i = 1, \ldots, n;$$

$$\hat{\tau}_2 = \frac{R_2 + [(a_2 - 1)/(1 + b_2)]\prod_{i=1}^n \left(1 - \hat{p}_i^{(2)}\right)}{1 - [1/(1 + b_2)]\prod_{i=1}^n \left(1 - \hat{p}_i^{(2)}\right)} \quad \text{and} \tag{2}$$

$$\hat{p}_i^{(2)} = \frac{\exp\left\{\hat{\alpha}_i^{(2)}\right\}}{1 + \exp\left\{\hat{\alpha}_i^{(2)}\right\}} = \frac{Z_i^{(2)}}{\hat{\tau}_2} - \frac{\hat{\alpha}_i^{(2)} - \hat{\bar{\alpha}}^{(2)}}{\hat{\tau}_2\sigma_2^2} - \frac{\hat{\bar{\alpha}}^{(2)} - \mu_2}{n\hat{\tau}_2\nu_2}; \quad i = 1, \ldots, n;$$

where $\nu_k = \gamma_k^2 + \sigma_k^2$ and $\hat{\bar{\alpha}}^{(k)} = \sum_1^n \hat{\alpha}_i^{(k)}/n$, $k = 1, 2$. They propose that τ be estimated by $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

4. Horvitz-Thompson-like Estimators of Totals and Means

To construct Horvitz-Thompson estimators (HTEs) of the population totals Y_1, Y_2 and $Y = Y_1 + Y_2$, we need to compute the inclusion probability of each element.

Unfortunately, we cannot compute (nor estimate) the inclusion probabilities because we do not have information about the nomination probabilities $p_i^{(k)}$'s associated with the sites A_i's that are not in the initial sample S_A. However, from a model-based approach we can compute the conditional inclusion probability of an element given that a particular set of sites $A_1, \ldots, A_n$ are in S_A. To obtain that probability for a person j in U_1, we will suppose that the people in U_1 are uniformly distributed over the N sites $A_1, \ldots, A_N$ in the frame. Notice that this assumption is in agreement with the assumed multinomial distribution of M_s. Therefore, given that $A_1, \ldots, A_n$ are in S_A, the conditional inclusion probability of a person $j \in U_1$ is

$$\begin{aligned}\pi_j^{(1)} = \pi^{(1)} &= 1 - \Pr(j \text{ is in none of the } A_i\text{'s} \in S_A) \\ &\quad \times \Pr(j \text{ is not nominated by any of the } A_i\text{'s} \in S_A \mid j \text{ is in none of the } A_i\text{'s} \in S_A) \\ &= 1 - (1 - n/N)Q_1\end{aligned}$$

Similarly, the conditional inclusion probability of a person $j \in U_2$ is $\pi_j^{(2)} = \pi^{(2)} = 1 - Q_2$.

Since Q_1 and Q_2 are unknown, $\pi_j^{(1)}$ and $\pi_j^{(2)}$ are unknown, but we can estimate them by

$$\check{\pi}^{(1)} = 1 - (1 - n/N)\check{Q}_1 \quad \text{and} \quad \check{\pi}^{(2)} = 1 - \check{Q}_2 \tag{3}$$

where $\check{Q}_k = \prod_{i=1}^{n}\left(1 - \check{p}_i^{(k)}\right)$, $k = 1, 2$, and $\check{p}_i^{(k)}$ denotes either the MLE or a Bayesian estimator of $p_i^{(k)}$. Therefore, Horvitz-Thompson-like estimators of Y_1, Y_2 and Y are

$$\check{Y}_1 = \frac{1}{\check{\pi}^{(1)}}\sum_{j \in S_1} y_j^{(1)}, \quad \check{Y}_2 = \frac{1}{\check{\pi}^{(2)}}\sum_{j \in S_2} y_j^{(2)} \quad \text{and} \quad \check{Y} = \check{Y}_1 + \check{Y}_2,$$

where S_k is the set of distinct elements of U_k, $k = 1, 2$, that are in the sample.

Notice that in $\check{Y}_k$ we are using an estimate of the conditional inclusion probability $\pi^{(k)}$ given that the sites $A_1, \ldots, A_n$ are in S_A. The idea is that if $\pi^{(k)}$ were known, $\sum_{j \in S_k} y_j^{(k)}/\pi^{(k)}$ would be a model-based conditional unbiased estimator of Y_k given that the sites $A_1, \ldots, A_n$ are in S_A. Consequently, it would also be unconditionally unbiased, that is, it would be unbiased with respect to the joint distribution formed by the model-based conditional distribution given the sites in S_A, that models both the numbers of people that are in the sites and the numbers of people nominated by the sites, and the design-based distribution that models the selection of sites in S_A.

The proposed estimators are not "real" HTEs because we are not using the actual conditional inclusion probabilities but estimates of them. Thus, we will call these estimators "Horvitz-Thompson-like estimators" (HTLEs). This type of HTLE of a total has been considered by Pollock, Turner, and Brown (1994), Haines and Pollock (1998) and Haines, Pollock, and Pantula (2000) in the context of estimation from incomplete list frames.

If the interest is to estimate the means $\bar{Y}_1$, $\bar{Y}_2$ and $\bar{Y}$, then the estimators will be

$$\check{\bar{Y}}_1 = \frac{\check{Y}_1}{\check{\tau}_1}, \quad \check{\bar{Y}}_2 = \frac{\check{Y}_2}{\check{\tau}_2} \quad \text{and} \quad \check{\bar{Y}} = \frac{\check{Y}}{\check{\tau}}$$

where $\check{\tau}_1$, $\check{\tau}_1$ and $\check{\tau}$ denote either the MLEs or any of the Bayesian estimators of τ_1, τ_2 and τ.

As was indicated earlier, we can estimate $\pi_k^{(k)}$ either using the MLEs or any of the Bayesian estimators of the $p_i^{(k)}$'s. Thus, we have two classes of estimators of totals and means: the estimators $\tilde{Y}_1$, $\tilde{Y}_2$, $\tilde{Y}$, $\tilde{\bar{Y}}_1$, $\tilde{\bar{Y}}_2$ and $\tilde{\bar{Y}}$ obtained by using the MLEs $\tilde{\tau}_k$ and $\tilde{p}_i^{(k)}$, $i = 1, \ldots, n$, $k = 1, 2$, and the estimators $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}$, $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}$ obtained by using the Bayesian estimators $\hat{\tau}_k$ and $\hat{p}_i^{(k)}$, $i = 1, \ldots, n$, $k = 1, 2$. Note that within this last class of estimators we have three sets of estimators: $\left\{\hat{Y}_1^{(a)}, \hat{Y}_2^{(a)}, \hat{Y}^{(a)}, \hat{\bar{Y}}_1^{(a)}, \hat{\bar{Y}}_2^{(a)}, \hat{\bar{Y}}^{(a)}\right\}$, where $a = U, J$ or P indicate that the corresponding set is obtained by using as prior distribution for the τ_k's the Uniform (U), Jeffreys' (J) or Poisson-Gamma (P) distribution, respectively.

5. Horvitz-Thompson-like Estimators of the Variances of the Estimators of Totals

5.1. General Form of the Variance Estimators

The results that are presented in this subsection are valid for the estimators $\tilde{Y}_k$ and $\tilde{Y}$, as well as for $\hat{Y}_k$ and $\hat{Y}$. The results that are presented in the other subsections depend on the type of estimator, and consequently each type of estimator will be considered separately.

As was done previously, let $\check{Y}$ denote either $\tilde{Y}$ or $\hat{Y}$, and $\check{p}^{(k)}$ denote either $\tilde{p}^{(k)}$ or $\hat{p}^{(k)}$. An expression for an HTLE of the variance of $\check{Y}$ can be obtained by the Delta method. To do that, notice that $\check{Y}$ can be expressed as

$$\check{Y} = \frac{\pi^{(1)}}{\pi^{(1)}(\check{\mathbf{p}}^{(1)})}\check{Y}_1^* + \frac{\pi^{(2)}}{\pi^{(2)}(\check{\mathbf{p}}^{(2)})}\check{Y}_2^* = f(\check{Y}_1^*, \check{Y}_2^*, \check{\mathbf{p}}^{(1)}, \check{\mathbf{p}}^{(2)}), \quad \text{say}$$

where $\pi^{(k)}(\check{\mathbf{p}}^{(k)}) = \check{\pi}^{(k)}$, [this notation emphasizes that $\check{\pi}^{(k)}$ is a function of $\check{\mathbf{p}}^{(k)} = (\check{p}_1^{(k)}, \ldots, \check{p}_n^{(k)})'$], and $\check{Y}_k^* = \sum_{j \in S_k} y_j^{(k)}/\pi^{(k)}$ is a random variable whose form is that of an HTE of Y_k. Since for samples of large sizes we would expect that $\check{Y}_k^* \approx Y_k$ and $\check{\mathbf{p}}^{(k)} \approx \mathbf{p}^{(k)}$, and consequently that $\pi^{(k)}(\check{\mathbf{p}}^{(k)}) \approx \pi^{(k)}$, we have that the Taylor linear approximation to $\check{Y}$ about $\theta = (Y_1, Y_2, \mathbf{p}^{(1)}, \mathbf{p}^{(2)})'$ is

$$\check{Y} \approx Y + \sum_{k=1}^{2}\left\{\left(\check{Y}_k^* - Y_k\right) - \frac{Y_k}{\pi^{(k)}}\left[\frac{\partial\pi^{(k)}(\check{\mathbf{p}}^{(k)})}{\partial\check{\mathbf{p}}^{(k)}}\right]_{\mathbf{p}^{(k)}}'(\check{\mathbf{p}}^{(k)} - \mathbf{p}^{(k)})\right\}$$

where $\left[\partial\pi^{(k)}(\check{\mathbf{p}}^{(k)})/\partial\check{\mathbf{p}}^{(k)}\right]_{\mathbf{p}^{(k)}}$ is the vector of derivatives of $\pi^{(k)}(\check{\mathbf{p}}^{k})$ evaluated at $\mathbf{p}^{(k)}$. Consequently, an estimator of the variance of $\check{Y}$ is

$$\begin{aligned}\check{V}(\check{Y}) &= \sum_{k=1}^{2}\left\{\check{V}\left(\check{Y}_k^*\right) + \left[\check{Y}_k/\check{\pi}^{(k)}\right]^2\left[\frac{\partial\pi^{(k)}(\check{\mathbf{p}}^{(k)})}{\partial\check{\mathbf{p}}^{(k)}}\right]'\check{V}(\check{\mathbf{p}}^{(k)})\left[\frac{\partial\pi^{(k)}(\check{\mathbf{p}}^{(k)})}{\partial\check{\mathbf{p}}^{(k)}}\right]\right\} \\ &= \check{V}(\check{Y}_1) + \check{V}(\check{Y}_2), \quad \text{say}\end{aligned} \tag{4}$$

where $\check{V}(\check{Y}_k^*)$ is an HTE of the variance of $\check{Y}_k^*$, and $\check{V}(\check{\mathbf{p}}^{(k)})$ is an estimator of the covariance matrix of $\check{\mathbf{p}}^{(k)}$.

To obtain an expression for $\check{V}(\check{Y}_k^*)$ we first need to get the second-order conditional inclusion probabilities $\pi_{jj'}$. Although we can obtain these probabilities from the assumption used to obtain the first-order conditional inclusion probabilities $\pi^{(k)}$, this supposition implies that inclusions of people in the initial sample are independent even if they belong to the same site, and this contradicts the fact that two persons on the same site are included in the sample if that site is selected. Therefore, for the people in U_1, we will suppose that the N groups of people of sizes $m_1, \ldots, m_N$ are independently and uniformly distributed over the N sites $A_1, \ldots, A_N$. Notice that although this assumption is just partly in agreement with the assumed multinomial distribution of M_s, the first-order conditional inclusion probabilities obtained from the previous assumption can also be obtained from this one. Therefore, given that $A_1, \ldots, A_N$ are in S_A, the second-order conditional inclusion probability of persons j and j' is

$$\begin{aligned}\pi_{jj'} &= \Pr(j \text{ and } j' \text{ are in } S)\\ &= 1 - \Pr(j \text{ is not in } S) - \Pr(j' \text{ is not in } S) + \Pr(j \text{ and } j' \text{ are not in } S)\\ &= \pi_j + \pi_{j'} - 1 + \Pr(j \text{ and } j' \text{ are not in } S)\end{aligned} \tag{5}$$

where S is the final sample and π_j and $\pi_{j'}$ are the first-order conditional inclusion probabilities of j and j', respectively.

Because of the assumption of independent nominations, the last term of (5) is equal to $(1-\pi_j)(1-\pi_{j'})$ if both j and j' are in U_2 or if j is in U_1 and j' is in U_2 or conversely, whereas if both j and j' are in U_1 then

$$\begin{aligned}&\Pr(j \text{ and } j \text{ are not in } S)\\ &\quad= \Pr(j \text{ and } j' \text{ are in none of the } A_i\text{'s in } S_A) \times \Pr(j \text{ and } j' \text{ are not nominated by any}\\ &\qquad \text{of the } A_i\text{'s} \in S_A \mid j \text{ and } j' \text{ are in none of the } A_i\text{'s} \in S_A)\\ &\quad= \begin{cases} [(1-n/N)Q_1]^2 & \text{if } j \text{ and } j' \text{ are on different sites}\\ (1-n/N)Q_1^2 & \text{if } j \text{ and } j' \text{ are on the same site}\end{cases}\end{aligned}$$

Consequently, $\pi_{jj'} - \pi_j\pi_{j'} = 0$ except when both j and j' are in U_1 and on the same site. Thus, we have the following estimators of the variances of $\check{Y}_1^*$ and $\check{Y}_2^*$:

$$\check{V}(\check{Y}_1^*) = \frac{1-\check{\pi}^{(1)}}{(\check{\pi}^{(1)})^2}\sum_{j\in S_1}(y_j^{(1)})^2 + \frac{\check{\pi}^{(1,1)} - (\check{\pi}^{(1)})^2}{\check{\pi}^{(1,1)}(\check{\pi}^{(1)})^2}\left[\sum_{i=1}^{n}(Y_{i\bullet}^{(1)})^2 - \sum_{i=1}^{n}\sum_{j\in A_i}\left(y_j^{(1)}\right)^2\right] \text{ and}$$

$$\check{V}(\check{Y}_2^*) = \frac{1-\check{\pi}^{(2)}}{(\check{\pi}^{(2)})^2}\sum_{j\in S_2}\left(y_j^{(2)}\right)^2$$

where $Y_{i\bullet}^{(1)} = \sum_{j\in A_i} y_j^{(1)}$ and $\check{\pi}^{(1,1)} = 1 - (1-n/N)\check{Q}_1(2-\check{Q}_1)$ is an estimator of $\pi_{jj'}$ when both j and j' are on the same site.

To obtain the second component of $\check{V}(\check{Y}_k)$, that is, the quadratic form, we need to obtain the vector of partial derivatives of $\pi^{(k)}(\check{\mathbf{p}}^{(k)})$ and an estimator of the covariance matrix of $\check{\mathbf{p}}^{(k)}$, $k = 1, 2$. The elements of the vector of derivatives are

$$\frac{\partial \pi^{(1)}(\check{\mathbf{p}}^{(1)})}{\partial \check{p}_j^{(1)}} = \frac{(1-n/N)\check{Q}_1}{1-\check{p}_j^{(1)}}, \quad j = 1, \ldots, n; \quad \text{and}$$

$$\frac{\partial \pi^{(2)}(\check{\mathbf{p}}^{(2)})}{\partial \check{p}_j^{(2)}} = \frac{\check{Q}_2}{1-\check{p}_j^{(2)}}, \quad j = 1, \ldots, n.$$

To get a partly design-based estimator of the covariance matrix $V(\check{\mathbf{p}}^{(k)})$, we will use the same strategy as that used by Félix-Medina and Thompson (2004) and Félix-Medina and Monjardin (2006). They compute the variances by replacing the multinomial distribution of the M_s by the distribution of the sampling design used to select the initial sample S_A. We will carry this out by computing the entries of the estimated covariance matrix $\check{V}(\check{\mathrm{p}}^{(k)})$ by means of the formulas:

$$\begin{aligned}V(\check{p}_i^{(k)}) &= V_p\left[E_\xi(\check{p}_i^{(k)}|m_s)\right] + E_p\left[V_\xi(\check{p}_i^{(k)}|m_s)\right] \quad \text{and}\\ Cov\left(\check{p}_i^{(k)}, \check{p}_j^{(k)}\right) &= Cov_p\left[E_\xi(\check{p}_i^{(k)}|m_s), E_\xi(\check{p}_j^{(k)}|m_s)\right] + E_p\left[Cov_\xi(\check{p}_i^{(k)}, \check{p}_j^{(k)}|m_s)\right]\end{aligned} \tag{6}$$

where $E_\xi\left(\check{p}_i^{(k)}|m_s\right)$, $V_\xi\left(\check{p}_i^{(k)}|m_s\right)$ and $Cov_\xi\left(\check{p}_i^{(k)}, \check{p}_j^{(k)}|m_s\right)$ denote the model-based conditional expectation, variance and covariance operators, given that $M_S = m_s$; and $E_p(\cdot)$, $V_p(\cdot)$ and $\mathrm{Cov}_p(\cdot,\cdot)$ denote the design-based expectation, variance and covariance operators.

Since the results for the MLEs $\tilde{p}_i^{(k)}$'s are different from those for the Bayesian estimators $\hat{p}_i^{(k)}$'s, we will consider each case separately. Notice that once we have calculated the estimator $\tilde{V}(\tilde{\mathbf{p}}^{(k)})$ of the covariance matrix of $\tilde{\mathbf{p}}^{(k)}$ or the estimator $\hat{V}(\hat{\mathbf{p}}^{(k)})$ of the covariance matrix of $\hat{\mathbf{p}}^{(k)}$, we can compute the quadratic forms that appear in (4) and consequently the estimators $\tilde{V}(\tilde{Y}_1)$, $\tilde{V}(\tilde{Y}_2)$ and $\tilde{V}(\tilde{Y})$ or the estimators $\hat{V}(\hat{Y}_1)$, $\hat{V}(\hat{Y}_2)$ and $\hat{V}(\hat{Y})$.

5.2. *Estimator of the Covariance Matrix of the Maximum Likelihood Estimator* $\tilde{\mathbf{p}}^{(k)}$

In the case of the MLEs, from (1) we have that $\tilde{p}_i^{(1)} = f_i^{(1)}(c_s^{(1)})$ and $\tilde{p}_i^{(2)} = f_i^{(2)}(c_s^{(2)})$, where $c_S^{(1)} = \left(M_S, Z_S^{(1)}, R_1\right)$, $c_S^{(2)} = \left(Z_S^{(2)}, R_2\right)$, $Z_S^{(k)} = \left(Z_1^{(k)}, \ldots, Z_n^{(k)}\right)$, and $f_i^{(k)}(\cdot)$ denotes the functional relationship between $c_s^{(k)}$ and $\tilde{p}_i^{(k)}$, $k = 1, 2$. Applying (6) to the first-order Taylor approximations to $\tilde{p}_i^{(1)}$ and $\tilde{p}_i^{(2)}$ about $E_\xi(c_s^{(1)})$ and $E_\xi\left(c_s^{(2)}\right)$, respectively, we obtain that

$$\begin{aligned}\tilde{V}\left(\tilde{p}_i^{(1)}\right) &= n(1-n/N)\left(\tilde{F}_i^{(1)}\right)^2\frac{1}{n-1}\sum_{j=1}^{n}(M_j - \bar{M})^2\\ &\quad + \frac{\tilde{p}_i^{(1)}(1-\tilde{p}_i^{(1)})}{\tilde{\tau}_1 - M_i} + \frac{1}{\tilde{A}_1}\left(\frac{\tilde{p}_i^{(1)}}{\tilde{\tau}_1 - M_i}\right)^2\tilde{H}_1\\ \widetilde{Cov}\left(\tilde{p}_i^{(1)}, \tilde{p}_j^{(1)}\right) &= n(1-n/N)\tilde{F}_i^{(1)}\tilde{F}_j^{(1)}\frac{1}{n-1}\sum_{j=1}^{n}(M_j - \bar{M})^2\\ &\quad + \frac{1}{\tilde{A}_1}\frac{\tilde{p}_i^{(1)}\tilde{p}_j^{(1)}}{(\tilde{\tau}_1 - M_i)(\tilde{\tau}_1 - M_j)}\tilde{H}_1;\\ \tilde{V}(\tilde{p}_i^{(2)}) &= \frac{\tilde{p}_i^{(2)}(1-\tilde{p}_i^{(2)})}{\tilde{\tau}_2} + \frac{1}{\tilde{A}_2}\left(\frac{\tilde{p}_i^{(2)}}{\tilde{\tau}_2}\right)^2\tilde{H}_2 \quad \text{and}\\ \widetilde{Cov}\left(\tilde{p}_i^{(2)}, \tilde{p}_j^{(2)}\right) &= \frac{\tilde{p}_i^{(2)}\tilde{p}_j^{(2)}}{\tilde{A}_2\tilde{\tau}_2^2}\tilde{H}_2\end{aligned}$$

where

$$\bar{M} = \sum_{k=1}^{n} M_k/n, \qquad \tilde{F}_i^{(1)} = \frac{\tilde{Q}_1}{\tilde{A}_1(\tilde{\tau}_1 - M - R_1)} \frac{\tilde{p}_i^{(1)}}{\tilde{\tau}_1 - M_i}$$

$$\tilde{A}_1 = \sum_{k=1}^{n} \frac{\tilde{p}_k^{(1)}}{1 - \tilde{p}_k^{(1)}} \frac{1}{\tilde{\tau}_1 - M_k} - \frac{M + R_1}{\tilde{\tau}_1(\tilde{\tau}_1 - M - R_1)}$$

$$\tilde{H}_1 = \frac{1}{\tilde{A}_1} \sum_{k=1}^{n} \frac{\tilde{p}_k^{(1)}}{1 - \tilde{p}_k^{(1)}} \frac{1}{\tilde{\tau}_1 - M_k} \left[1 - 2\frac{(\tilde{\tau}_1 - M)\tilde{Q}_1}{\tilde{\tau}_1 - M - R_1}\right] + \frac{1}{\tilde{A}_1} \frac{(\tilde{\tau}_1 - M)\tilde{Q}_1(1 - \tilde{Q}_1)}{(\tilde{\tau}_1 - M - R_1)^2}$$
$$+ \frac{2(\tilde{\tau}_1 - M)\tilde{Q}_1}{\tilde{\tau}_1 - M - R_1} - 2;$$

$$\tilde{A}_2 = \frac{1}{\tilde{\tau}_2}\left[\sum_{k=1}^{n} \frac{\tilde{p}_k^{(2)}}{1 - \tilde{p}_k^{(2)}} - \frac{R_2}{\tilde{\tau}_2 - R_2}\right] \quad \text{and}$$

$$\tilde{H}_2 = \frac{1}{\tilde{A}_2\tilde{\tau}_2} \sum_{k=1}^{n} \frac{\tilde{p}_k^{(2)}}{1 - \tilde{p}_k^{(2)}} \left[1 - 2\frac{\tilde{\tau}_2\tilde{Q}_2}{\tilde{\tau}_2 - R_2}\right] + \frac{1}{\tilde{A}_2} \frac{\tilde{\tau}_2\tilde{Q}_2(1 - \tilde{Q}_2)}{(\tilde{\tau}_2 - R_2)^2} + \frac{2\tilde{\tau}_2\tilde{Q}_2}{\tilde{\tau}_2 - R_2} - 2$$

5.3. *Estimator of the Covariance Matrix of the Bayesian Estimator* $\hat{\mathbf{p}}^{(k)}$

In the case of the Bayesian estimators $\hat{p}_i^{(k)}$, using the previous strategy we get that

$$\hat{V}(\hat{p}_i^{(1)}) = n(1 - n/N)(\hat{F}_i^{(1)})^2 \frac{1}{n - 1} \sum_{j=1}^{n} (M_j - \bar{M})^2$$
$$+ \left[\frac{\hat{E}_i^{(1)}}{\hat{B}_i^{(1)}}\right]^2 \left[(\hat{p}_i^{(1)} - \hat{D}_1)^2 \hat{J}_1 + \hat{K}_1 + (\hat{\tau}_1 - M_i)\hat{E}_i^{(1)} + 2(\hat{p}_i^{(1)} - \hat{D}_1)\hat{L}_1\right.$$
$$-2\frac{\hat{G}_1}{n} \frac{(\hat{\tau}_1 - M_i)\hat{E}_i^{(1)}}{\hat{B}_i^{(1)}} + 2\frac{(\hat{\tau}_1 - M)\hat{Q}_1}{\hat{A}_1(\hat{\tau}_1 - M - R_1)} \hat{p}_i^{(1)}(\hat{p}_i^{(1)} - \hat{D}_1)$$
$$\left. -2\frac{1}{\hat{A}_1} \frac{(\hat{\tau}_1 - M_i)\hat{E}_i^{(1)}(\hat{p}_i^{(1)} - \hat{D}_1)^2}{\hat{B}_i^{(1)}}\right]$$

$$\widehat{Cov}(\hat{p}_i^{(1)}, \hat{p}_j^{(1)}) = n(1 - n/N)\hat{F}_i^{(1)}\hat{F}_j^{(1)} \frac{1}{n-1} \sum_{j=1}^{n} (M_j - \bar{M})^2$$
$$+ \left[\frac{\hat{E}_i^{(1)}}{\hat{B}_i^{(1)}}\right]\left[\frac{\hat{E}_j^{(1)}}{\hat{B}_j^{(1)}}\right]\left\{(\hat{p}_i^{(1)} - \hat{D}_1)(\hat{p}_j^{(1)} - \hat{D}_1)\hat{J}_1 + \hat{K}_1 + \left[(\hat{p}_i^{(1)} - \hat{D}_1) + (\hat{p}_j^{(1)} - \hat{D}_1)\right]\hat{L}_1\right.$$
$$-\frac{\hat{G}_1}{n}\left[\frac{(\hat{\tau}_1 - M_i)\hat{E}_i^{(1)}}{\hat{B}_i^{(1)}} + \frac{(\hat{\tau}_1 - M_j)\hat{E}_j^{(1)}}{\hat{B}_j^{(1)}}\right]$$
$$+ \frac{(\hat{\tau}_1 - M)\hat{Q}_1}{\hat{A}_1(\hat{\tau}_1 - M - R_1)}\left[\hat{p}_i^{(1)}(\hat{p}_j^{(1)} - \hat{D}_1) + \hat{p}_j^{(1)}(\hat{p}_i^{(1)} - \hat{D}_1)\right]$$
$$\left. -\frac{1}{\hat{A}_1}\left[\frac{(\hat{\tau}_1 - M_i)\hat{E}_i^{(1)}(\hat{p}_i^{(1)} - \hat{D}_1)}{\hat{B}_i^{(1)}}(\hat{p}_j^{(1)} - \hat{D}_1) + \frac{(\hat{\tau}_1 - M_j)\hat{E}_j^{(1)}(\hat{p}_j^{(1)} - \hat{D}_1)}{\hat{B}_j^{(1)}}(\hat{p}_i^{(1)} - \hat{D}_1)\right]\right\};$$

$$\hat{V}(\hat{p}_i^{(2)}) = \left[\frac{\hat{E}_i^{(2)}}{\hat{B}_i^{(2)}}\right]^2 \left[(\hat{p}_i^{(2)} - \hat{D}_2)^2\hat{J}_2 + \hat{K}_2 + \hat{\tau}_2\hat{E}_i^{(2)} + 2(\hat{p}_i^{(2)} - \hat{D}_2)\hat{L}_2\right.$$
$$\left. -2\frac{\hat{G}_2}{n}\frac{\hat{\tau}_2\hat{E}_i^{(2)}}{\hat{B}_i^{(2)}} + 2\frac{\hat{\tau}_2\hat{Q}_2}{\hat{A}_2(\hat{\tau}_2 - R_2)}\hat{p}_i^{(2)}(\hat{p}_i^{(2)} - \hat{D}_2) - 2\frac{1}{\hat{A}_2}\frac{\hat{\tau}_2\hat{E}_i^{(2)}(\hat{p}_i^{(2)} - \hat{D}_2)^2}{\hat{B}_i^{(2)}}\right] \text{and}$$

$$\widehat{Cov}(\hat{p}_i^{(2)}, \hat{p}_j^{(2)}) = \left[\frac{\hat{E}_i^{(2)}}{\hat{B}_i^{(2)}}\right]\left[\frac{\hat{E}_j^{(2)}}{\hat{B}_j^{(2)}}\right]\left\{(\hat{p}_i^{(2)} - \hat{D}_2)(\hat{p}_j^{(2)} - \hat{D}_2)\hat{J}_2 + \hat{K}_2\right.$$
$$+ \left[(\hat{p}_i^{(2)} - \hat{D}_2) + (\hat{p}_j^{(2)} - \hat{D}_2)\right]\hat{L}_2 - \frac{\hat{G}_2}{n}\left[\frac{\hat{\tau}_2\hat{E}_i^{(2)}}{\hat{B}_i^{(2)}} + \frac{\hat{\tau}_2\hat{E}_j^{(2)}}{\hat{B}_j^{(2)}}\right]$$
$$+ \frac{\hat{\tau}_2\hat{Q}_2}{\hat{A}_2(\hat{\tau}_2 - R_2)}\left[\hat{p}_i^{(2)}(\hat{p}_j^{(2)} - \hat{D}_2) + \hat{p}_j^{(2)}(\hat{p}_i^{(2)} - \hat{D}_2)\right]$$
$$\left. -\frac{1}{\hat{A}_2}\left[\frac{\hat{\tau}_2\hat{E}_i^{(2)}(\hat{p}_i^{(2)} - \hat{D}_2)}{\hat{B}_i^{(2)}}(\hat{p}_j^{(2)} - \hat{D}_2) + \frac{\hat{\tau}_2\hat{E}_j^{(2)}(\hat{p}_j^{(2)} - \hat{D}_2)}{\hat{B}_j^{(2)}}(\hat{p}_i^{(2)} - \hat{D}_2)\right]\right\}$$

where

$$\hat{F}_i^{(1)} = \frac{\hat{Q}_1}{\hat{A}_1(\hat{\tau}_1 - M - R_1)} \frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}} \hat{E}_i^{(1)}, \quad \hat{E}_i^{(1)} = \hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)})$$

$$\hat{B}_i^{(1)} = (\hat{\tau}_1 - M_i)\hat{E}_i^{(1)} + \sigma_1^{-2}$$

$$\hat{A}_1 = \sum_{k=1}^{n} (\hat{p}_k^{(1)})^2/\hat{B}_k^{(1)} - \hat{C}_1 + \frac{1}{\hat{\tau}_1 + a_1 - 1} - \frac{1}{\hat{\tau}_1 - M - R_1}$$

$$\hat{C}_1 = \frac{\left(\nu_1^{-1} - n\sigma_1^{-2}\right)\left[n^{-1}\sum_{k=1}^{n}\hat{p}_k^{(1)}/\hat{B}_k^{(1)}\right]^2}{1 + n^{-1}\left(\nu_1^{-1} - n\sigma_1^{-2}\right)n^{-1}\sum_{k=1}^{n}1/\hat{B}_k^{(1)}}$$

$$\hat{D}_1 = \frac{n^{-1}(\nu_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{k=1}^{n}\hat{p}_k^{(1)}/\hat{B}_k^{(1)}}{1 + n^{-1}(\nu_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{k=1}^{n}1/\hat{B}_k^{(1)}}$$

$$\hat{J}_1 = \frac{1}{\hat{A}_1^2}\left\{\sum_{k=1}^{n}\left[\frac{\hat{p}_k^{(1)} - \hat{D}_1}{\hat{B}_k^{(1)}}\right]^2(\hat{\tau}_1 - M_k)\hat{E}_k^{(1)} + \frac{(\hat{\tau}_1 - M)\hat{Q}_1(1 - \hat{Q}_1)}{(\hat{\tau}_1 - M - R_1)^2}\right.$$
$$\left. -2\frac{(\hat{\tau}_1 - M)\hat{Q}_1}{\hat{\tau}_1 - M - R_1}\sum_{k=1}^{n}\left[\frac{\hat{p}_k^{(1)} - \hat{D}_1}{\hat{B}_k^{(1)}}\right]\hat{p}_k^{(1)}\right\}$$

$$\hat{G}_1 = \frac{n^{-1}\left(\nu_1^{-1} - n\sigma_1^{-2}\right)}{1 + n^{-1}\left(\nu_1^{-1} - n\sigma_1^{-2}\right)n^{-1}\sum_{k=1}^{n}1/\hat{B}_k^{(1)}}, \quad \hat{K}_1 = \frac{\hat{G}_1^2}{n^2}\sum_{k=1}^{n}\frac{(\hat{\tau}_1 - M_k)\hat{E}_k^{(1)}}{(\hat{B}_k^{(1)})^2}$$

$$\hat{L}_1 = \frac{1}{\hat{A}_1}\left\{\frac{\hat{G}_1}{n}\sum_{k=1}^{n}\left[\frac{\hat{p}_k^{(1)} - \hat{D}_1}{(\hat{B}_k^{(1)})^2}\right](\hat{\tau}_1 - M_k)\hat{E}_k^{(1)} - \frac{(\hat{\tau}_1 - M)\hat{Q}_1}{\hat{\tau}_1 - M - R_1}\frac{\hat{G}_1}{n}\sum_{k=1}^{n}\frac{\hat{p}_k^{(1)}}{\hat{B}_k^{(1)}}\right\};$$

$$\hat{E}_i^{(2)} = \hat{p}_i^{(2)}(1 - \hat{p}_i^{(2)}), \quad \hat{B}_i^{(2)} = \hat{\tau}_2\hat{E}_i^{(2)} + \sigma_2^{-2}$$

$$\hat{A}_2 = \sum_{k=1}^{n}(\hat{p}_k^{(2)})^2/\hat{B}_k^{(2)} - \hat{C}_2 + \frac{1}{\hat{\tau}_2 + a_2 - 1} - \frac{1}{\hat{\tau}_2 - R_2}$$

$$\hat{C}_2 = \frac{\left(\nu_2^{-1} - n\sigma_2^{-2}\right)\left[n^{-1}\sum_{k=1}^{n}\hat{p}_k^{(2)}/\hat{B}_k^{(2)}\right]^2}{1 + n^{-1}\left(\nu_2^{-1} - n\sigma_2^{-2}\right)n^{-1}\sum_{k=1}^{n}1/\hat{B}_k^{(2)}}$$

$$\hat{D}_2 = \frac{n^{-1}\left(\nu_2^{-1} - n\sigma_2^{-2}\right)n^{-1}\sum_{k=1}^{n}\hat{p}_k^{(2)}/\hat{B}_k^{(2)}}{1 + n^{-1}\left(\nu_2^{-1} - n\sigma_2^{-2}\right)n^{-1}\sum_{k=1}^{n}1/\hat{B}_k^{(2)}}$$

$$\hat{J}_2 = \frac{1}{\hat{A}_2^2}\left\{\sum_{k=1}^{n}\left[\frac{\hat{p}_k^{(2)} - \hat{D}_2}{\hat{B}_k^{(2)}}\right]^2\hat{\tau}_2\hat{E}_k^{(2)} + \frac{\hat{\tau}_2\hat{Q}_2(1 - \hat{Q}_2)}{(\hat{\tau}_2 - R_2)^2} - 2\frac{\hat{\tau}_2\hat{Q}_2}{\hat{\tau}_2 - R_2}\sum_{k=1}^{n}\left[\frac{\hat{p}_k^{(2)} - \hat{D}_2}{\hat{B}_k^{(2)}}\right]\hat{p}_k^{(2)}\right\}$$

$$\hat{G}_2 = \frac{n^{-1}\left(\nu_2^{-1} - n\sigma_2^{-2}\right)}{1 + n^{-1}\left(\nu_2^{-1} - n\sigma_2^{-2}\right)n^{-1}\sum_{k=1}^{n}1/\hat{B}_k^{(2)}}, \quad \hat{K}_2 = \frac{\hat{G}_2^2}{n^2}\sum_{k=1}^{n}\frac{\hat{\tau}_2\hat{E}_k^{(2)}}{\left(\hat{B}_k^{(2)}\right)^2} \quad \text{and}$$

$$\hat{L}_2 = \frac{1}{\hat{A}_2}\left\{\frac{\hat{G}_2}{n}\sum_{k=1}^{n}\left[\frac{\hat{p}_k^{(2)} - \hat{D}_2}{\left(\hat{B}_k^{(2)}\right)^2}\right]\hat{\tau}_2\hat{E}_k^{(2)} - \frac{\hat{\tau}_2\hat{Q}_2}{\hat{\tau}_2 - R_2}\frac{\hat{G}_2}{n}\sum_{k=1}^{n}\frac{\hat{p}_k^{(2)}}{\hat{B}_k^{(2)}}\right\}.$$

6. Horvitz-Thompson-like Estimators of the Variances of the Estimators of Means

Since estimators of means are ratios of two estimators, for instance $\hat{\bar{Y}} = \hat{Y}/\hat{\tau}$, we estimate their variances by estimating the variances of Taylor linear approximations to the corresponding ratios about the parameters estimated by the numerator and denominator, say Y and τ. This strategy yields that

$$\hat{V}(\hat{\bar{Y}}) = \hat{V}(\hat{Y} - \bar{Y}\hat{\tau})/\hat{\tau}^2$$

In the case of the estimators of the means based on the MLEs, from (1) we have that

$$\tilde{\tau} = \sum_{j\in S_1}1/\tilde{\pi}_1 + \sum_{j\in S_2}1/\tilde{\pi}_2$$

Thus

$$\tilde{Y} - \bar{Y}\tilde{\tau} = \frac{1}{\tilde{\pi}_1}\sum_{j\in S_1}\left(y_j^{(1)} - \bar{Y}\right) + \frac{1}{\tilde{\pi}_2}\sum_{j\in S_2}\left(y_j^{(2)} - \bar{Y}\right)$$

that is, $\tilde{Y} - \bar{Y}\tilde{\tau}$ has the same form as that of $\tilde{Y}$, but the y-values $y_j^{(k)}$ in $\tilde{Y}$ are replaced by the values $y_j^{(k)} - \bar{Y}$. Therefore, $\tilde{V}(\tilde{\bar{Y}})$ is obtained by dividing (4) by $\tilde{\tau}^2$ and replacing $y_j^{(k)}$ by $y_j^{(k)} - \tilde{\bar{Y}}$. Similarly, $\tilde{V}(\tilde{\bar{Y}}_k)$ is obtained by dividing $\tilde{V}(\tilde{Y}_k)$ by $\tilde{\tau}_k^2$ and replacing $y_j^{(k)}$ by $y_j^{(k)} - \tilde{\bar{Y}}_k$.

In the case of the estimators based on the Bayesian estimators, from (2) we have that

$$\hat{\tau} = \hat{W}_{11}\hat{T}_1 + \hat{W}_{12} + \hat{W}_{21}\hat{T}_2 + \hat{W}_{22}$$

where $\hat{T}_k = \sum_{j\in S_k}1/\hat{\pi}^{(k)}$ is an HTLE of τ_k,

$$\hat{W}_{11} = \frac{1 - (1 - n/N)\hat{Q}_1}{1 - (1 - n/N)[N/(N + b_1)]\hat{Q}_1}, \quad \hat{W}_{12} = \frac{(1 - n/N)[N(a_1 - 1)/(N + b_1)]\hat{Q}_1}{1 - (1 - n/N)[N/(N + b_1)]\hat{Q}_1}$$

$$\hat{W}_{21} = \frac{1 - \hat{Q}_2}{1 - [1/(1 + b_2)]\hat{Q}_2} \quad \text{and} \quad \hat{W}_{22} = \frac{[(a_2 - 1)/(1 + b_2)]\hat{Q}_2}{1 - [1/(1 + b_2)]\hat{Q}_2}$$

Thus

$$\begin{aligned}\hat{Y} - \bar{Y}\hat{\tau} &= [\hat{Y}_1 - \hat{W}_{11}\bar{Y}\hat{T}_1 - \hat{W}_{12}\bar{Y}] + [\hat{Y}_2 - \hat{W}_{21}\bar{Y}\hat{T}_2 - \hat{W}_{22}\bar{Y}] \\ &= f_1(\hat{Y}_1, \hat{T}_1, \hat{Q}_1) + f_2(\hat{Y}_2, \hat{T}_2, \hat{Q}_2)\end{aligned}$$

Obtaining the first-order Taylor approximation to $f_k(\hat{Y}_k, \hat{T}_k, \hat{Q}_k)$, $k = 1, 2$, and estimating the variances of the first-order approximations, we get that

$$\hat{V}(\hat{Y} - \bar{Y}\hat{\tau}) = \sum_{k=1}^{2}\left\{\hat{V}\left(\hat{Y}_k - \hat{W}_{k1}\bar{Y}\hat{T}_k\right) + \hat{W}_{k3}^2\hat{V}(\hat{Q}_k)\right\} \tag{7}$$

where $\hat{V}\left(\hat{Y}_k - \hat{W}_{k1}\bar{Y}\hat{T}_k\right)$ has the same form as that of $\hat{V}(\hat{Y}_k)$, except that $y_j^{(k)}$ is replaced by $y_j^{(k)} - \hat{W}_{k1}\hat{\bar{Y}}$,

$$\hat{W}_{13} = \frac{[\hat{\tau}_1 b_1 - N(a_1 - 1)]\hat{\hat{Y}}/(N + b_1)}{\left\{1 - (1 - n/N)[N/(N + b_1)]\hat{Q}_1\right\}^2} \quad \text{and}$$

$$\hat{W}_{23} = \frac{[\hat{\tau}_2 b_2 - (a_2 - 1)]\hat{\hat{Y}}/(1 + b_2)}{\left\{1 - [1/(1 + b_2)]\hat{Q}_2\right\}^2}$$

Because of Equations (3), we have that

$$\hat{V}(\hat{Q}_1) = \frac{1}{(1 - n/N)^2}\left[\frac{\partial \pi^{(1)}(\hat{\mathbf{p}}^{(1)})}{\partial \hat{\mathbf{p}}^{(1)}}\right]' \hat{V}(\hat{\mathbf{p}}^{(1)})\left[\frac{\partial \pi^{(1)}(\hat{\mathbf{p}}^{(1)})}{\partial \hat{\mathbf{p}}^{(1)}}\right] \quad \text{and}$$

$$\hat{V}(\hat{Q}_2) = \left[\frac{\partial \pi^{(2)}(\hat{\mathbf{p}}^{(2)})}{\partial \hat{\mathbf{p}}^{(2)}}\right]' \hat{V}(\hat{\mathbf{p}}^{(2)})\left[\frac{\partial \pi^{(2)}(\hat{\mathbf{p}}^{(2)})}{\partial \hat{\mathbf{p}}^{(2)}}\right]$$

Thus, $\hat{V}(\hat{\hat{Y}})$ is obtained by dividing (7) by $\hat{\tau}^2$. Similarly, $\hat{V}(\hat{\hat{Y}}_k)$ is obtained by dividing the kth term of (7) by $\hat{\tau}_k^2$ and replacing $\hat{\hat{Y}}$ by $\hat{\hat{Y}}_k$ in any place where $\hat{\hat{Y}}$ appears.

7. Wald Confidence Intervals

Though in this work we will not justify theoretically that the proposed estimators of the population totals and means are asymptotically normally distributed, we will suppose that the normal distribution is a reasonable approximation to the distributions of the estimators. Thus, we suggest that $100(1 - \alpha)\%$ design-based Wald confidence intervals for the population totals and means are used. These intervals have the form $\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta})}$, where $z_{1-\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution, $\hat{V}(\hat{\theta})$ is a partly design-based estimator of the variance of $\hat{\theta}$, and $\hat{\theta}$ denotes an estimator either of a total or of a mean.

8. Monte Carlo Studies

In order to observe the performance of the proposed estimators, three simulation studies were carried out. 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 used data from the Colorado Springs study on transmission of HIV/AIDS to construct two populations. In the second study we constructed two artificial populations in which all of the model assumptions set in Section 3 were satisfied, and two populations in which only the assumption of the Poisson distribution of the M_i's was not satisfied. Finally, in the third study we constructed an artificial population in which only the assumption of homogeneous nomination probabilities was not satisfied.

8.1. *Study Based on the Colorado Springs Study on HIV/AIDS Transmission*

In the Colorado Springs study on heterosexual transmission of HIV/AIDS, described in Potterat et al. (1993), Rothenberg et al. (1995) and Potterat et al. (2004), among other papers, a set of 595 persons presumably at high risk of acquiring and transmitting HIV were enrolled through a sexually transmitted disease clinic, a drug clinic, self-referral and outreach. Those people were interviewed about their demographic characteristics and their knowledge and practices with regard to HIV/AIDS. They were also asked for a complete enumeration of their personal contacts, defined as those persons with whom they had social (sharing meals or lodging), sexual, or drug-associated relations. The interviewees named 7,379 contacts who were not in the set of the 595 interviewees and 367 contacts in that set. The 7,379 contacts were also interviewed and asked to nominate their contacts, but in our study we omitted the information about their contacts.

We defined U_1 as the set of the 595 original interviewees and U_2 as the set of the 7,379 contacts who were not original interviewees. We defined as the response variable a binary variable which took on the value 1 if the person was a sex worker and on the value 0 in other case. Thus, $\tau_1 = 595$, $\tau_2 = 7{,}379$, $\tau = 7{,}974$, $Y_1 = 135$, $Y_2 = 417$, and $Y = 552$. Since no sampling frame of sites was defined in the Colorado Springs study, we constructed one by forming $N = 105$ clusters (groups) with the 595 interviewees. The sizes m_i's of the clusters were generated by sampling from a zero-truncated negative binomial distribution with parameter of size 2.5 and probability 2/3. The sample mean and variance of the 105 m_i's were 5.67 and 15.03, respectively. The clusters were formed by putting people located in the same or similar places in the same cluster. For instance, the people located at the drug clinic were assigned to several groups which were different from the groups to which the people located at the sexually transmitted disease clinic were assigned. We assumed that a person was nominated by a group if that person was not in the group and was nominated by at least one of the members of the group. The average values of the nomination probabilities were $\bar{p}^{(1)} = 0.02$ and $\bar{p}^{(2)} = 0.01$ for people in U_1 and U_2, respectively.

It is worth noting that 6,924 persons in U_2 (94%) were named by only one group, 273 by only two groups, 81 by three, 26 by four, 14 by five, 1 by six, 5 by seven, 2 by eight, 1 by twelve and 1 by thirteen. Since this high percentage of the people in U_2 who were nominated by only one group was expected to cause serious overestimation of τ_2 (estimators of the population size of the type used in capture-recapture have serious problems of overestimation when most of the sampled elements are captured only one time), and consequently to affect the performance of the proposed estimators, we defined a reduced population in which U_1 was defined as in the previous case (complete population) and U_2 as the set formed by all the nominees that were named by at least two groups (415 people) plus the 379 sex workers who were named by only one group. Thus, in this reduced population, $\tau_1 = 595$, $\tau_2 = 794$, $\tau = 1{,}389$, $Y_1 = 135$, $Y_2 = 417$ and $Y = 552$. The average value of the nomination probabilities was $\bar{p}^{(1)} = \bar{p}^{(2)} = 0.02$.

The simulation experiment was carried out by replicating $r = 10{,}000$ times the following procedure. For each population of $N = 105$ values of m_i's, a SRSWOR of size n was selected, where $n = 40$ in the case of the complete population and $n = 30$ in the case of the reduced population. From cluster A_i in the sample, the people in $U_k - A_i$, $k = 1, 2$, named by that cluster were included in the sample. The values of the variables M, $Z_i^{(k)}$ and R_k, $k = 1, 2$, were calculated and those data were used to compute the following estimators of the totals and proportions of sex workers: the sets of estimators $\{\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}\}$ and $\{\tilde{\tilde{Y}}_1, \tilde{\tilde{Y}}_2, \tilde{\tilde{Y}}\}$ obtained from the set of MLE's $\{\tilde{\tau}_2, \tilde{\tau}_2, \tilde{\tau}\}$; and the three pairs of sets of estimators $\left\{\hat{Y}_1^{(a)}, \hat{Y}_2^{(a)}, \hat{Y}^{(a)}\right\}$ and $\left\{\hat{\hat{Y}}_1^{(a)}, \hat{\hat{Y}}_2^{(a)}, \hat{\hat{Y}}^{(a)}\right\}$, $a = U, J, P$, obtained from the corresponding sets of Bayesian estimators $\left\{\hat{\tau}_1^{(a)}, \hat{\tau}_2^{(a)}, \hat{\tau}^{(a)}\right\}$, $a = U, J, P$, which use as prior distributions for the τ_k's the Uniform (U), Jeffreys' (J) and Poisson-Gamma (P) distributions, respectively. In addition, estimators of the variances of the estimators of the population totals and proportions, and 95% confidence intervals for these parameters, were also computed.

To obtain each of the sets of estimators $\{\hat{Y}_1^{(a)}, \hat{Y}_2^{(a)}, \hat{Y}^{(a)}\}$ and $\{\hat{\bar{Y}}_1^{(a)}, \hat{\bar{Y}}_2^{(a)}, \hat{\bar{Y}}^{(a)}\}$, $a = U, J, P$, the parameters of the initial distributions for the logits $\alpha_i^{(k)} = \ln[p_i^{(k)}/(1 - p_i^{(k)})]$ were set to the following values: $\mu_k = -3.5$, $\sigma_k^2 = \gamma_k^2 = 9$, $k = 1, 2$. The parameters of the Poisson-Gamma distributions of τ_1 were $a_1 = 1$ and $b_1 = 0.1$ ($E(\lambda_1) = 10$ and $V(\lambda_1) = 100$). The parameters of the distribution of τ_2 were $a_2 = 42.25$ and $b_2 = 0.0065$ ($E(\lambda_2) = 6,500$ and $V(\lambda_2) = 10^6$) in the case of the complete population, whereas $a_2 = 8$ and $b_2 = 0.01$ ($E(\lambda_2) = 800$ and $V(\lambda_2) = 80,000$) in the case of the reduced population. The values set for the parameters of the prior distributions implied that these distributions were well dispersed over relatively long intervals that contained the parameters of interest.

The performance of an estimator $\hat{Y}^{(a)}$ of Y, say, was evaluated by its relative bias and the square root of its relative mean squared error, defined as r-bias $= \sum_1^r(\hat{Y}_i^{(a)} - Y)/(rY)$ and $\sqrt{\text{r-mse}} = \sqrt{\sum_1^r(\hat{Y}_i^{(a)} - Y)^2/(rY^2)}$, where $\hat{Y}_i^{(a)}$ was the value of $\hat{Y}^{(a)}$ obtained at the ith replication. The performance of a variance estimator was also evaluated by those parameters, which were similarly defined as those of the estimator $\hat{Y}^{(a)}$, but using the empirically determined variance instead of the real variance. Finally, the performance of a 95% confidence interval for Y, say, was evaluated by its coverage probability and its relative length defined as the proportion of replications in which the replicated intervals contained Y and the average length of the replicated intervals divided by Y, respectively.

The results of the simulation study are shown in Table 1. We can see that in the case of the complete population, everyone of the estimators of Y_1 performed well in terms of bias and mean squared error. However, as was expected, the estimators of Y_2 had serious problems of overestimation. The bad performance of the estimators of Y_2 deteriorated the performance of the estimators of Y. In the case of the reduced population, every one of the estimators of Y_1 and Y performed acceptably in terms of bias, whereas each of the estimators of Y_2 showed slight problems of bias. In terms of mean squared error, the estimators of Y_1 performed well, whereas the estimators of Y_2 and Y showed some problems of instability ($\sqrt{\text{r-mse}} > 0.2$).

The estimators of the variances of the estimators of Y_1 performed well in terms of bias in both populations, although they showed some problems of instability. (Results for variance estimators are shown in parentheses in Table 1.) However, the estimators of the variances of the estimators of Y_2 and Y had serious problems of subestimation.

The 95% confidence intervals for Y_1 performed moderately well in the complete population (coverage probabilities and relative lengths were about 0.91 and 0.23, respectively), and acceptably well in the reduced population (coverage probabilities and relative lengths were about 0.93 and 0.31, respectively). (Results for confidence intervals are not shown.) However, in both populations the confidence intervals for Y_2 and Y had coverage probabilities close to zero. Their bad performance was a consequence of the biases of the estimators of Y_2 and Y and the great subestimation problems of the estimators of their variances.

With respect to the estimators of proportions, the estimators of $\bar{Y}_1$ showed moderate positive biases in both populations. The estimators of $\bar{Y}_2$ performed acceptably in the

Table 1. Colorado Springs study: Relative biases and squared roots of relative mean square errors of the estimators of the numbers and proportions of sex workers and of the estimators of their variances

Estimators of the numbers of sex workers					Estimators of the proportions of sex workers				
	Complete population		Reduced population			Complete population		Reduced population	
	$n = 40, \bar{M} = 226.7$ $\bar{R}_1 = 142.3$ $\bar{R}_2 = 2944.5$		$n = 30, \bar{M} = 169.9$ $\bar{R}_1 = 138.9$ $\bar{R}_2 = 344.3$			$n = 40, \bar{M} = 226.7$ $\bar{R}_1 = 142.3$ $\bar{R}_2 = 2944.5$		$n = 30, \bar{M} = 169.9$ $\bar{R}_1 = 138.9$ $\bar{R}_2 = 344.3$	
	r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$		r-bias	$\sqrt{\text{r-mse}}$	r-bias	$\sqrt{\text{r-mse}}$
$\tilde{Y}_1$	−.02 (−.03)	.06 (.28)	.01 (−.03)	.08 (.33)	$\tilde{\bar{Y}}_1$	.21 (−.25)	.22 (.29)	.27 (−.24)	.29 (.30)
$\tilde{Y}_2$	3.7 (−.77)	3.9 (.79)	−.10 (−.80)	.27 (.83)	$\tilde{\bar{Y}}_2$	.04 (−.69)	.14 (.69)	−.27 (−.81)	.28 (.81)
$\tilde{Y}$	2.8 (−.77)	3.0 (.79)	−.08 (−.80)	.21 (.83)	$\tilde{\bar{Y}}$	−.10 (−.68)	.15 (.68)	−.11 (−.80)	.14 (.80)
$\hat{Y}_1^{(U)}$	−.02 (−.03)	.06 (.28)	.01 (−.03)	.08 (.33)	$\hat{\bar{Y}}_1^{(U)}$	.21 (−.25)	.22 (.29)	.27 (−.24)	.29 (.30)
$\hat{Y}_2^{(U)}$	3.7 (−.77)	3.9 (.79)	−.11 (−.80)	.27 (.83)	$\hat{\bar{Y}}_2^{(U)}$	.04 (−.69)	.14 (.69)	−.27 (−.81)	.28 (.81)
$\hat{Y}^{(U)}$	2.8 (−.77)	2.9 (.79)	−.08 (−.80)	.21 (.83)	$\hat{\bar{Y}}^{(U)}$	−.10 (−.68)	.15 (.68)	−.11 (−.80)	.14 (.80)
$\hat{Y}_1^{(J)}$	−.02 (−.03)	.06 (.28)	.00 (−.03)	.08 (.33)	$\hat{\bar{Y}}_1^{(J)}$	.21 (−.25)	.22 (.29)	.27 (−.24)	.29 (.30)
$\hat{Y}_2^{(J)}$	3.7 (−.77)	3.9 (.79)	−.11 (−.80)	.27 (.83)	$\hat{\bar{Y}}_2^{(J)}$	.04 (−.69)	.14 (.69)	−.27 (−.81)	.28 (.81)
$\hat{Y}^{(J)}$	2.8 (−.77)	2.9 (.79)	−.08 (−.81)	.21 (.83)	$\hat{\bar{Y}}^{(J)}$	−.10 (−.68)	.15 (.68)	−.11 (−.80)	.14 (.81)
$\hat{Y}_1^{(P)}$	−.02 (−.03)	.06 (.27)	.01 (−.03)	.08 (.33)	$\hat{\bar{Y}}_1^{(P)}$	.21 (−.25)	.22 (.29)	.27 (−.24)	.29 (.30)
$\hat{Y}_2^{(P)}$	1.9 (−.73)	2.0 (.74)	−.13 (−.81)	.26 (.82)	$\hat{\bar{Y}}_2^{(P)}$	.07 (−.71)	.15 (.71)	−.27 (−.82)	.28 (.82)
$\hat{Y}^{(P)}$	1.5 (−.74)	1.5 (.74)	−.10 (−.81)	.20 (.82)	$\hat{\bar{Y}}^{(P)}$	−.06 (−.71)	.12 (.71)	−.11 (−.81)	.14 (.81)

Notes: Results for variance estimators in parentheses; $\tilde{Y}_k$ and $\tilde{\bar{Y}}_k$ estimators of the number and proportion of sex workers based on the MLEs of the population sizes; $\hat{Y}_k^{(U)}$ and $\hat{\bar{Y}}_k^{(U)}$, $\hat{Y}_K^{(J)}$ and $\hat{\bar{Y}}_K^{(J)}$, and $\hat{Y}_K^{(P)}$ and $\hat{\bar{Y}}_K^{(P)}$ estimators of the number and proportion of sex workers obtained from the Bayesian estimators of the population sizes based on the prior Uniform, Jeffreys and two-stage Poisson-Gamma distributions, respectively. Results based on 10^4 trials.

complete population, but showed moderate negative biases in the reduced population. Finally, the estimators of $\bar{Y}$, the main proportion of interest, performed moderately well in both populations (r-bias≈ -0.1 and $\sqrt{\text{r-mse}} \approx 0.15$).

The estimators of the variances showed problems of subestimation and instability. As a result of the biases of the estimators of the proportions and the estimators of their variances, the 95% confidence intervals had very small coverage probabilities.

It is worth noting that because of the relatively large sample sizes used in this study, no practical differences regarding the performance of the different estimators were observed.

8.2. *Study Based on Artificial Populations with Homogeneous Nomination Probabilities*

We constructed four artificial populations of y-values. The populations were generated by considering two probability distributions for the M_i's: Poisson and negative binomial, and two distributions for the y_j's: chi-square (χ^2) and Bernoulli. The characteristics of the populations are described in Table 2. The nomination probabilities $p_i^{(k)}$, $i = 1, \ldots, N$, $k = 1, 2$, were generated using the model $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, where the values of β_k were set so that the following values of $\bar{p}^{(k)} = \sum_1^N p_i^{(k)}/N$ were obtained. Case A: $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.03, 0.02)$ and Case B: $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.01, 0.006)$.

The simulation experiment was carried out as in the previous study, except that when the cluster A_i was included in the initial sample, the values $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ were generated by drawing samples of sizes $\tau_1 - m_i$ and τ_2 from Bernoulli distributions with means $p_i^{(1)}$ and $p_i^{(2)}$, respectively. The size of the initial sample was $n = 25$. The values of the parameters of the initial distributions used to construct the Bayesian estimators were the same as those set in the reduced population defined in the previous study.

Because of restrictions of space we present a selection of the results in Tables 3 to 6, but the following comments refer to the complete set of results. We can see that when the $p_i^{(k)}$'s were not so small (Case A), the estimators of the totals and means did not present problems of bias regardless of the distribution of the M_i's, although the estimators of the total Y_2 showed slight problems of instability, particularly when the $y_j^{(k)}$'s were Bernoulli distributed. The 95% confidence intervals for the totals and means had good coverage probabilities, with the exception of the interval for Y_1 which showed slightly low

Table 2. Parameters of simulated populations

$N = 250$		$N = 250$	
$M_i \sim$Poisson $E(M_i) = 7.2$ $V(M_i) = 7.2$ $\tau_1 = 1{,}800$ $\tau_2 = 700$ $\tau = 2{,}500$		$M_i \sim$ neg. bin. $E(M_i) = 7.2$ $V(M_i) = 24.4$ $\tau_1 = 1{,}758$ $\tau_2 = 700$ $\tau = 2{,}458$	
Population I $y_i \sim \chi^2(10)$	Population II $y_i \sim$Bernoulli(0.2)	Population III $y_i \sim \chi^2(10)$	Population IV $y_i \sim$Bernoulli(0.2)
$Y_1 = 18{,}234.4$, $\bar{Y}_1 = 10.13$ $Y_2 = 6{,}855.6$, $\bar{Y}_2 = 9.79$ $Y = 25{,}090.0$, $\bar{Y} = 10.04$	$Y_1 = 366$, $\bar{Y}_1 = 0.20$ $Y_2 = 142$, $\bar{Y}_2 = 0.20$ $Y = 508$, $\bar{Y}_1 = 0.20$	$Y_1 = 17{,}798.8$, $\bar{Y}_1 = 10.12$ $Y_2 = 6{,}862.6$, $\bar{Y}_2 = 9.80$ $Y = 24{,}661.5$, $\bar{Y}_1 = 10.03$	$Y_1 = 361$, $\bar{Y}_1 = 0.21$ $Y_2 = 143$, $\bar{Y}_2 = 0.20$ $Y_1 = 504$, $\bar{Y}_1 = 0.21$

Table 3. Relative biases and square roots of relative mean squared errors of the total and mean estimators: $Y_i^{(k)} \sim \chi^2(10)$

Estimators of the population totals

	Population I				Population III			
$\tilde{M}$	180.1		180.1		175.7		175.7	
$\tilde{R}_1$	862.4		359.7		825.2		342.9	
$\tilde{R}_2$	277.4		97.8		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006	
	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$
$\tilde{Y}_1$	−.00	.03	−.00	.06	−.01	.04	−.01	.09
$\tilde{Y}_2$	.01	.11	L_1	L_1	.01	.11	L_1	L_1
$\tilde{Y}$	.00	.04	L_2	L_1	−.00	.04	L_1	L_1
$\hat{Y}_1^{(U)}$	−.00	.03	−.00	.06	−.01	.04	−.01	.09
$\hat{Y}_2^{(U)}$	.01	.11	.14	.72	.01	.11	.13	.65
$\hat{Y}^{(U)}$	.00	.04	.04	.20	−.00	.04	.03	.19
$\hat{Y}_1^{(J)}$	−.00	.03	−.00	.06	−.01	.04	−.01	.09
$\hat{Y}_2^{(J)}$	.00	.10	−.02	.38	.00	.11	−.04	.38
$\hat{Y}^{(J)}$	−.00	.04	−.01	.11	−.00	.04	−.02	.13
$\hat{Y}_1^{(P)}$	−.00	.03	−.00	.06	−.01	.04	−.01	.09
$\hat{Y}_2^{(P)}$	.01	.10	.00	.20	.01	.10	.00	.20
$\hat{Y}^{(P)}$	.00	.04	−.00	.07	−.00	.04	−.00	.09

Estimators of the population means

	Population I				Population III			
	180.1		180.1		175.7		175.7	
	862.4		359.7		825.2		342.9	
	277.4		97.8		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006	
	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$
$\tilde{\bar{Y}}_1$	−.00	.01	.00	.02	−.00	.01	−.00	.02
$\tilde{\bar{Y}}_2$	.00	.02	−.00	.04	−.00	.02	.00	.04
$\tilde{\bar{Y}}$	−.00	.01	−.00	.02	−.00	.01	−.00	.02
$\hat{\bar{Y}}_1^{(U)}$	−.00	.01	−.00	.02	−.00	.01	−.00	.02
$\hat{\bar{Y}}_2^{(U)}$	.00	.02	.00	.04	−.00	.02	−.00	.04
$\hat{\bar{Y}}^{(U)}$	−.00	.01	−.00	.02	−.00	.01	−.00	.02
$\hat{\bar{Y}}_1^{(J)}$	.00	.01	.00	.02	.00	.01	.00	.02
$\hat{\bar{Y}}_2^{(J)}$	.00	.02	.01	.04	.00	.02	.01	.04
$\hat{\bar{Y}}^{(J)}$	.00	.01	.00	.02	.00	.01	.00	.02
$\hat{\bar{Y}}_1^{(P)}$	.00	.01	.00	.02	.00	.01	.00	.02
$\hat{\bar{Y}}_2^{(P)}$	.00	.02	−.00	.04	.00	.02	.00	.05
$\hat{\bar{Y}}^{(P)}$	.00	.01	.00	.02	.00	.01	.00	.02

Notes: rβ, relative bias; rε^2, relative mean squared error; $\tilde{Y}_k$ and $\tilde{\bar{Y}}_k$ estimators of the population total and mean based on the MLEs of the population sizes; $\hat{Y}_k^{(U)}$ and $\hat{\bar{Y}}_k^{(U)}$, $\hat{Y}_k^{(J)}$ and $\hat{\bar{Y}}_k^{(J)}$, and $\hat{Y}_k^{(P)}$ and $\hat{\bar{Y}}_k^{(P)}$ estimators of the population total and mean obtained from the Bayesian estimators of the population sizes based on the prior Uniform, Jeffreys and two-stage Poisson-Gamma distributions, respectively. L_1 and L_2 indicate values greater than 10^6 and 10^4, respectively. Results based on 10^4 trials.

coverage probabilities in Population III. The lengths of the intervals were reasonable. The estimators of the variances of the estimators of the totals and means did not show problems of bias, except those of the estimators of Y_1, which presented problems of subestimation when the M_i's were negative binomial distributed. Each of the estimators of the variances of the estimators of the totals had problems of instability. The most serious problems were observed for those of the estimators of Y_2 ($0.34 < \sqrt{\text{r-mse}} < 0.57$), followed by those of the estimators of Y ($\sqrt{\text{r-mse}} < 0.35$) and then those of the estimators of Y_1 ($\sqrt{\text{r-mse}} < 0.2$). The estimators of the variances of the estimators of the means had modest problems of instability, ($\sqrt{\text{r-mse}} < 0.21$).

On the other hand, when the $p_i^{(k)}$'s were small (Case B), every one of the estimators of Y_1 performed well. With respect to the estimators of Y_2, we have that $\tilde{Y}_2$ performed very badly, $\hat{Y}_2^{(U)}$ badly, $\hat{Y}_2^{(J)}$ poorly, and $\hat{Y}_2^{(P)}$ modestly. In the case of the estimators of the main parameter, the total Y, $\tilde{Y}$ performed very badly, $\hat{Y}^{(U)}$ modestly, and $\hat{Y}^{(J)}$ and $\hat{Y}^{(P)}$ well. Every one of the confidence intervals for Y_1 performed well. Although the intervals for Y_2 had good coverage probabilities (with the exception of the interval based on $\hat{Y}_2^{(J)}$), their lengths were long. The interval for Y based on $\tilde{Y}$ performed very badly (it had huge length), those based on $\hat{Y}^{(U)}$ and $\hat{Y}^{(J)}$ performed modestly, and that based on $\hat{Y}^{(P)}$ performed well. Every one of the estimators of the variances of the estimators of the totals had problems of bias and instability. The best performance was achieved by the estimators of the variances of $\hat{Y}_1^{(P)}$, $\hat{Y}_2^{(P)}$, and $\hat{Y}^{(P)}$, which performed modestly. Finally, every one of the estimators of the means performed well. The estimators of their variances and confidence intervals also performed well, except the estimators of the variances of $\tilde{\tilde{Y}}_2$ and $\tilde{\tilde{Y}}$, which had huge biases and variances, and the intervals based on them, which had huge lengths.

It is worth noting that, in general, the estimators of the totals and means, the estimators of their variances and the confidence intervals performed better when the M_i's were Poisson distributed than when they were negative binomial distributed, as well as when the $y_j^{(k)}$'s were χ^2-distributed than when they were Bernoulli distributed. In addition, when the $p_i^{(k)}$'s were not so small, no difference among the performance of the different types of estimators of the totals and means was observed, but when the $p_i^{(k)}$'s were small, the sets of estimators $\left\{\hat{Y}_1^{(P)}, \hat{Y}_2^{(P)}, \hat{Y}^{(P)}\right\}$ and $\left\{\hat{\hat{Y}}_1^{(P)}, \hat{\hat{Y}}_2^{(P)}, \hat{\hat{Y}}^{(P)}\right\}$, as well as the estimators of their variances and confidence intervals based on them, had the best performance.

8.3. *Study Based on Artificial Populations with Heterogeneous Nomination Probabilities*

To analyze the sensitivity of the proposed estimators to deviations from the homogeneity assumption, we carried out a simulation study in which we used the populations I and III defined in the previous study, but generating the $p_{ij}^{(k)}$'s by means of the following Rasch model: $p_{ij}^{(k)} = \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right) / \left[1 + \exp(\alpha_i^{(k)} + \beta_j^{(k)})\right]$, where $p_{ij}^{(k)}$ is the probability that person j in U_k is nominated by site A_i; $\alpha_i^{(k)}$ is a fixed (nonrandom) effect that represents the potential that the site A_i has of nominating a person in $U_k - A_i$, and $\beta_j^{(k)}$ is a random effect, with distribution $N(0, \sigma_k^2)$, that represents the propensity that person $j \in U_k$ has of being nominated. It is worth noting that this model was used by Coull and Agresti (1999) in the context of capture-recapture sampling. In this model, 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.

Table 4. Relative biases and square roots of relative mean squared errors of the total and mean estimators: $Y_i^{(k)} \sim$ Bernoulli(0.20)

Estimators of the population totals									Estimators of the population means								
	Population II				Population IV					Population II				Population IV			
$\bar{M}$	180.1		180.1		175.7		175.7			180.1		180.1		175.7		175.7	
$\bar{R}_1$	862.3		359.8		825.2		343.0			862.3		359.8		825.2		343.0	
$\bar{R}_2$	277.4		97.7		271.7		95.5			277.4		97.7		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01		$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006		$\bar{p}_2$	.02		.006		.02		.006	
	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$		rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$	rβ	$\sqrt{\text{r}\varepsilon^2}$
$\tilde{Y}_1$	−.00	.05	−.00	.09	−.01	.06	−.01	.12	$\tilde{\tilde{Y}}_1$	−.00	.04	−.00	.07	−.00	.04	−.00	.07
$\tilde{Y}_2$	.01	.14	L_2	L_1	.01	.15	L_1	L_1	$\tilde{\tilde{Y}}_2$	−.00	.09	.00	.19	.00	.09	.00	.19
$\tilde{Y}$	.00	.05	L_2	L_1	−.00	.06	L_1	L_1	$\tilde{\tilde{Y}}$	−.00	.04	.00	.08	.00	.04	−.00	.08
$\hat{Y}_1^{(U)}$	−.00	.05	−.00	.09	−.01	.06	−.01	.12	$\hat{\hat{Y}}_1^{(U)}$	.00	.04	.00	.07	−.00	.04	−.00	.07
$\hat{Y}_2^{(U)}$	.01	.14	.14	.69	.01	.15	.13	.73	$\hat{\hat{Y}}_2^{(U)}$	−.00	.09	.00	.19	.00	.09	.00	.19
$\hat{Y}^{(U)}$	.00	.05	.04	.20	−.00	.06	.03	.22	$\hat{\hat{Y}}^{(U)}$	−.00	.04	.00	.08	.00	.04	−.00	.08
$\hat{Y}_1^{(J)}$	−.00	.05	−.00	.09	−.01	.06	−.01	.12	$\hat{\hat{Y}}_1^{(J)}$	.00	.04	.00	.07	.00	.04	.00	.07
$\hat{Y}_2^{(J)}$	−.00	.14	−.03	.41	.00	.14	−.04	.43	$\hat{\hat{Y}}_2^{(J)}$	.00	.09	.01	.19	.00	.09	.01	.19
$\hat{Y}^{(J)}$	−.00	.05	−.01	.13	−.00	.06	−.02	.15	$\hat{\hat{Y}}^{(J)}$	.00	.04	.00	.07	.00	.04	.01	.08
$\hat{Y}_1^{(P)}$	−.00	.05	−.00	.09	−.01	.06	−.01	.12	$\hat{\hat{Y}}_1^{(P)}$	.00	.04	.00	.07	.00	.04	−.00	.07
$\hat{Y}_2^{(P)}$	.01	.13	.00	.27	.01	.14	.00	.27	$\hat{\hat{Y}}_2^{(P)}$	−.00	.09	−.00	.19	.00	.09	.00	.19
$\hat{Y}^{(P)}$	−.00	.05	−.00	.10	−.00	.06	−.01	.12	$\hat{\hat{Y}}^{(P)}$	.00	.04	.00	.07	.00	.04	.00	.08

Notes: rβ, relative bias; $r\varepsilon^2$, relative mean squared error; $\tilde{Y}_k$ and $\tilde{\tilde{Y}}_k$ estimators of the population total and mean based on the MLEs of the population sizes; $\hat{Y}_k^{(U)}$ and $\hat{\hat{Y}}_k^{(U)}$, $\hat{Y}_K^{(J)}$ and $\hat{\hat{Y}}_K^{(J)}$, and $\hat{Y}_K^{(P)}$ and $\hat{\hat{Y}}_K^{(P)}$ estimators of the population total and mean obtained from the Bayesian estimators of the population sizes based on the prior Uniform, Jeffreys and two-stage Poisson-Gamma distributions, respectively. L_1 and L_2 indicate values greater than 10^6 and 10^4, respectively. Results based on 10^4 trials.

Table 5. Coverage probabilities and relative lengths of 95% confidence intervals: $Y_i^{(k)} \sim \chi^2(10)$

Intervals for the population totals

	Population I				Population III			
$\tilde{M}$	180.1		180.1		175.7		175.7	
$\tilde{R}_1$	862.4		359.7		825.2		342.9	
$\tilde{R}_2$	277.4		97.8		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006	
	cp	rl	cp	rl	cp	rl	cp	rl
$\hat{Y}_1$	.95	.13	.98	.28	.92	.14	.93	.34
$\hat{Y}_2$	.95	.41	.93	L_1	.95	.43	.92	L_1
$\hat{Y}$	.96	.15	.96	L_1	.94	.16	.95	L_1
$\hat{Y}_1^{(U)}$	.95	.13	.98	.28	.92	.14	.93	.34
$\hat{Y}_2^{(U)}$	.95	.41	.92	1.9	.95	.43	.91	1.9
$\hat{Y}^{(U)}$	.96	.15	.96	.57	.94	.16	.95	.61
$\hat{Y}_1^{(J)}$	.95	.13	.98	.28	.92	.14	.93	.34
$\hat{Y}_2^{(J)}$	.94	.40	.87	1.4	.95	.42	.86	1.4
$\hat{Y}^{(J)}$	.95	.14	.92	.44	.94	.16	.90	.48
$\hat{Y}_1^{(P)}$	.95	.13	.98	.28	.92	.14	.93	.34
$\hat{Y}_2^{(P)}$	.95	.39	.95	.86	.95	.40	.96	.87
$\hat{Y}^{(P)}$	.96	.14	.97	.31	.94	.15	.95	.34

Intervals for the population means

	Population I				Population III			
	180.1		180.1		175.7		175.7	
	862.4		359.7		825.2		342.9	
	277.4		97.8		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006	
	cp	rl	cp	rl	cp	rl	cp	rl
$\hat{\bar{Y}}_1$	.94	.03	.95	.06	.95	.04	.95	.06
$\hat{\bar{Y}}_2$	.95	.08	.94	.16	.95	.08	.94	79
$\hat{\bar{Y}}$	.95	.03	.96	.07	.95	.04	.96	77
$\hat{\bar{Y}}_1^{(U)}$	.94	.03	.95	.06	.95	.04	.95	.06
$\hat{\bar{Y}}_2^{(U)}$	.95	.08	.94	.16	.95	.08	.94	.17
$\hat{\bar{Y}}^{(U)}$	.95	.03	.96	.07	.95	.04	.96	.07
$\hat{\bar{Y}}_1^{(J)}$	.95	.03	.95	.06	.95	.04	.95	.06
$\hat{\bar{Y}}_2^{(J)}$	.95	.08	.94	.16	.95	.08	.94	.17
$\hat{\bar{Y}}^{(J)}$	.95	.03	.95	.07	.95	.04	.95	.07
$\hat{\bar{Y}}_1^{(P)}$	.94	.03	.95	.06	.95	.04	.95	.06
$\hat{\bar{Y}}_2^{(P)}$	.95	.08	.95	.17	.95	.08	.95	.17
$\hat{\bar{Y}}^{(P)}$	.95	.03	.95	.06	.95	.04	.95	.07

Notes: cp, coverage probability; rl, relative length; $\hat{Y}_k$ and $\hat{\bar{Y}}_k$ estimators of the population total and mean based on the MLEs of the population sizes; $\hat{Y}_k^{(U)}$ and $\hat{\bar{Y}}_k^{(U)}$, $\hat{Y}_k^{(J)}$ and $\hat{\bar{Y}}_k^{(J)}$, and $\hat{Y}_k^{(P)}$ and $\hat{\bar{Y}}_k^{(P)}$ estimators of the population total and mean obtained from the Bayesian estimators of the population sizes based on the prior Uniform, Jeffreys and two-stage Poisson-Gamma distributions, respectively. L_1 indicates values greater than 10^6. Results based on 10^4 trials.

Table 6. Relative biases and square roots of relative mean squared errors of the variance estimators: $Y_i^{(k)} \sim \chi^2(10)$

Variance estimators: est. of totals

	Population I				Population III			
$\tilde{M}$	180.1		180.1		175.7		175.7	
$\tilde{R}_1$	862.4		359.7		825.2		342.9	
$\tilde{R}_2$	277.4		97.8		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\hat{V}(\hat{Y}_1)$	.02	.12	.47	.49	−.18	.23	−.12	.22
$\hat{V}(\hat{Y}_2)$	.02	.43	L_1	L_1	−.00	.57	L_1	L_1
$\hat{V}(\hat{Y})$	.03	.27	L_1	L_1	−.07	.35	L_1	L_1
$\hat{V}\left(\hat{Y}_1^{(U)}\right)$	.02	.12	.46	.50	−.18	.23	−.11	.22
$\hat{V}\left(\hat{Y}_2^{(U)}\right)$	.02	.42	.46	19	−.00	.56	.55	12
$\hat{V}(\hat{Y}^{(U)})$	.03	.27	.47	18	−.07	.34	.52	11
$\hat{V}\left(\hat{Y}_1^{(J)}\right)$	.05	.13	.48	.51	−.19	.23	−.11	.22
$\hat{V}\left(\hat{Y}_2^{(J)}\right)$	.03	.42	.26	3.7	.03	.54	.34	3.5
$\hat{V}(\hat{Y}^{(J)})$	.03	.26	.27	3.1	−.09	.32	.15	2.4
$\hat{V}\left(\hat{Y}_1^{(P)}\right)$	.02	.12	.46	.49	−.19	.23	−.12	.22
$\hat{V}\left(\hat{Y}_2^{(P)}\right)$	.04	.36	.26	.47	.03	.44	.31	.50
$\hat{V}(\hat{Y}^{(P)})$	.04	.23	.34	.42	−.07	.27	.04	.18

Variance estimators: est. of means

	Population I				Population III			
	180.1		180.1		175.7		175.7	
	862.4		359.7		825.2		342.9	
	277.4		97.8		271.7		95.5	
$\bar{p}_1$	.03		.01		.03		.01	
$\bar{p}_2$	.02		.006		.02		.006	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\hat{V}(\hat{\bar{Y}}_1)$	−.03	.10	.00	.11	−.02	.17	−.00	.16
$\hat{V}(\hat{\bar{Y}}_2)$	−.01	.14	.42	.34	−.00	.20	L_1	L_1
$\hat{V}(\hat{\bar{Y}})$	−.02	.12	2.4	185	.00	.20	L_1	L_1
$\hat{V}\left(\hat{\bar{Y}}_1^{(U)}\right)$	−.03	.10	−.02	.11	−.02	.17	−.01	.15
$\hat{V}\left(\hat{\bar{Y}}_2^{(U)}\right)$	−.01	.14	−.03	.21	−.01	.20	−.01	.25
$\hat{V}(\hat{\bar{Y}}^{(U)})$	−.02	.12	.05	.41	.00	.20	.05	.50
$\hat{V}\left(\hat{\bar{Y}}_1^{(J)}\right)$	−.04	.11	.02	.11	−.02	.17	−.02	.15
$\hat{V}\left(\hat{\bar{Y}}_2^{(J)}\right)$	−.00	.14	−.03	.21	.01	.20	−.02	.26
$\hat{V}(\hat{\bar{Y}}^{(J)})$	−.00	.12	.06	.31	.01	.20	.07	.39
$\hat{V}\left(\hat{\bar{Y}}_1^{(P)}\right)$	−.03	.10	−.01	.10	−.02	.17	−.02	.15
$\hat{V}\left(\hat{\bar{Y}}_2^{(P)}\right)$	−.01	.14	.02	.22	−.01	.20	.02	.27
$\hat{V}(\hat{\bar{Y}}^{(P)})$	−.02	.12	.01	.19	−.00	[illegible]	.01	.27

Notes: $r\beta$, relative bias; $r\varepsilon^2$, relative mean squared error; $\hat{Y}_k$ and $\hat{\bar{Y}}_k$ estimators of the population total and mean based on the MLEs of the population sizes; [illegible] and $\hat{\bar{Y}}_k^{(U)}$, $\hat{Y}_K^{(J)}$ and $\hat{\bar{Y}}_K^{(J)}$, and $\hat{Y}_K^{(P)}$ and $\hat{\bar{Y}}_K^{(P)}$ estimators of the population total and mean obtained from the Bayesian estimators of the population sizes based on the prior Uniform, [illegible] and two-stage Poisson-Gamma distributions, respectively. L_1 indicates values greater than 10^6. Results based on 10^4 trials.

Table 7. Relative biases and square roots of relative mean squared errors of the total and mean estimators: $Y_i^{(k)} \sim \chi^2(10)$

	Population I				Population III			
$\bar{M}$	180.1		180.1		175.7		175.7	
$\bar{R}_1$	856.4		791.3		801.1		743.4	
$\bar{R}_2$	278.0		258.1		264.5		245.9	
σ	.4		.64		.4		.64	
$\bar{p}_1$	.03		.03		.03		.03	
$\bar{p}_2$	.02		.02		.02		.02	
Estimators of the population totals								
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\hat{Y}_1^{(P)}$	−.07	.07	−.15	.15	−.07	.08	−.15	.16
$\hat{Y}_2^{(P)}$	−.10	.13	−.23	.24	−.01	.14	−.24	.25
$\hat{Y}^{(P)}$	−.07	.08	−.17	.17	−.08	.09	−.18	.18
Variance estimators: est. of totals								
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\hat{V}\left(\hat{Y}_1^{(P)}\right)$	−.01	.12	−.11	.16	−.21	.25	−.31	.33
$\hat{V}\left(\hat{Y}_2^{(P)}\right)$	−.03	.37	−.10	.33	−.03	.42	−.10	.41
$\hat{V}(\hat{Y}^{(P)})$	−.02	.23	−.12	.21	−.13	.27	−.25	.32
95% confidence intervals for the totals								
	cp	rl	cp	rl	cp	rl	cp	rl
$\hat{Y}_1^{(P)}$	.39	.11	.00	.10	.46	.13	.00	.12
$\hat{Y}_2^{(P)}$	.71	.32	.14	.26	.71	.34	.16	.28
$\hat{Y}^{(P)}$	.34	.12	.00	.10	.39	.13	.00	.11

	Population I				Population III			
$\bar{M}$	180.1		180.1		175.7		175.7	
$\bar{R}_1$	856.4		791.3		801.1		743.4	
$\bar{R}_2$	278.0		258.1		264.5		245.9	
σ	.4		.64		.4		.64	
$\bar{p}_1$	.03		.03		.03		.03	
$\bar{p}_2$	.02		.02		.02		.02	
Estimators of the population means								
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\hat{\bar{Y}}_1^{(P)}$	.00	.01	.01	.01	.00	.01	.00	.01
$\hat{\bar{Y}}_2^{(P)}$	.00	.02	.00	.02	−.00	.02	−.01	.02
$\hat{\bar{Y}}^{(P)}$	.00	.01	.01	.01	.00	.01	.00	.01
Variance estimators: est. of means								
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\hat{V}\left(\hat{\bar{Y}}_1^{(P)}\right)$	−.03	.10	−.09	.13	−.05	.15	−.09	.17
$\hat{V}\left(\hat{\bar{Y}}_2^{(P)}\right)$	−.04	.14	−.11	.17	−.02	.20	−.06	.21
$\hat{V}(\hat{\bar{Y}}^{(P)})$	−.04	.12	−.10	.14	−.02	.18	−.07	.18
95% confidence intervals for the means								
	cp	rl	cp	rl	cp	rl	cp	rl
$\hat{\bar{Y}}_1^{(P)}$	.93	.03	.89	.03	.94	.04	.93	.04
$\hat{\bar{Y}}_2^{(P)}$	.94	.08	.93	.08	.94	.08	.92	.08
$\hat{\bar{Y}}^{(P)}$	.92	.03	.88	.03	.95	.03	.94	.03

Notes: $r\beta$, relative bias; $r\varepsilon^2$, relative mean squared error; $\hat{Y}_k^{(P)}$ and $\hat{\bar{Y}}_k^{(P)}$ estimators of the population total and mean obtained from the Bayesian estimators of the population sizes based on the prior two-stage Poisson-Gamma distribution. Results based on 10^4 trials.

For each of the two populations we considered two levels of heterogeneity of the $p_{ij}^{(k)}$'s. Case A: Small degree of heterogeneity, which was obtained using the following values of the parameters: $\alpha_i^{(k)} = c_k/(m_i^{1/4} + d_k)$, $c_1 = -5.7$, $c_2 = -6.4$, $d_1 = d_2 = 0.0001$ and $\sigma_1 = \sigma_2 = 0.4$. Case B: Great degree of heterogeneity, which was obtained using the following values of the parameters: $\alpha_i^{(k)}$ defined as in the previous case with $c_1 = -6.0$, $c_2 = -6.7$, $d_1 = d_2 = 0.0001$ and $\sigma_1 = \sigma_2 = 0.64$. These values of the parameters implied that in both Cases A and B the average values of the $p_{ij}^{(1)}$'s and $p_{ij}^{(2)}$'s were $\bar{p}_1 = 0.03$ and $\bar{p}_2 = 0.02$, respectively. In addition, in Case A the average values of the ratios $\max_j p_{ij}^{(1)}/\min_j p_{ij}^{(1)}$ and $\max_j p_{ij}^{(2)}/\min_j p_{ij}^{(2)}$ were 17.3 and 12.7, respectively, whereas in Case B those ratios were 67.6 and 51.9 (for the m_i's greater than 0).

The simulation study was carried out in the same way as the previous one, but we only considered the estimators $\hat{Y}_1^{(P)}$, $\hat{Y}_2^{(P)}$ and $\hat{Y}^{(P)}$. The results of the study are shown in Table 7. We can see that in Case A the estimators of the totals performed acceptably, although they showed a tendency to subestimate those parameters. On the other hand, in Case B, the estimators presented problems of subestimation which increased their mean square errors, although the magnitudes of the r-bias and $\sqrt{\text{r-mse}}$ of the estimator of Y, the main parameter, were not very large (both were less than 0.2). In general, the estimators of the variances presented problems of instability. In Case A and M_i's with Poisson distribution, the estimators did not show problems of bias, although slight tendencies to subestimate the variances were observed; however, when the M_i's were negative binomial distributed the tendencies to subestimate the variances of $\hat{Y}_1^{(P)}$ and $\hat{Y}^{(P)}$ were greater than in the previous case. In Case B, the biases of the estimations of the variances were greater than in Case A. In particular, when the M_i's were negative binomial distributed, the magnitudes of the r-biases of $\hat{V}(\hat{Y}_1^{(P)})$ and $\hat{Y}^{(P)}$ were 0.31 and 0.25, respectively. Finally, the 95% confidence intervals for the totals showed very low coverage probabilities and very short lengths. These results were consequences of the subestimation problems of the estimators of the totals and the estimators of their variances.

With respect to the estimators of the means, every one of them performed very well. The estimators of their variances did not show serious problems of bias, but they presented some problems of instability (particularly when the M_i's were negative binomial distributed). The confidence intervals for the means also worked well, except in Case B and M_i's with Poisson distribution, where the intervals for $\bar{Y}_1$ and $\bar{Y}$ showed coverage probabilities slightly below 0.9.

9. Conclusions

From the results of our simulation studies, we can conclude that the main factors that determine the performance of the proposed estimators are the initial sample size n, the average size of the $p_i^{(k)}$'s, which along with n determine the numbers of nominees r_k, and the degree of heterogeneity of the $p_i^{(k)}$'s. In the context of capture-recapture studies, Xi, Watson, and Yip (2008) have found that the minimum value of the capture proportion (MCP) that yields reliable estimates of the population size depends mainly on the size of the population and the degree of heterogeneity. They encountered that the MCP decreases as the population size increases and it increases as the degree of heterogeneity increases. For populations of sizes about 1,000, they found that the MCP is between 0.3 and 0.5, and we think that similar values are required for the proportion of nominees in U_2 to obtain reliable estimates of Y_2 (the proportion of nominees in U_1 does not need to be so great because the estimators of τ_1 and Y_1 use also the information of the people in S_0). Thus, when the assumption of homogeneous nomination probabilities is satisfied and the combination of the value of n and that of the average size of the $p_i^{(k)}$'s is such that the number of nominees r_2 in U_2 is not small, say between 30% and 50% of the size of U_2, the estimators and confidence intervals for the totals work well regardless of the distributions of the M_i's and $y_j^{(k)}$'s, and no differences among the performance of the distinct types of estimators and intervals are observed. However, as the number of nominees decreases (say below 30% of τ_2), the performance of the estimators and intervals for the totals deteriorates, although the estimator $\hat{Y}^{(P)}$ and the interval obtained from it could still work well when r_2 is relatively small. With respect to the estimators and confidence intervals for the means, they all work well when the homogeneity assumption is satisfied, regardless of the value of r_2, except the confidence intervals based on the estimators obtained from the MLE's of the τ's, which could work very badly when r_2 is small and the M_i's are not Poisson distributed.

On the other hand, when the homogeneity assumption is not satisfied, the estimators of the totals and means have small to moderate problems of subestimation and instability. The estimators of their variances also have problems of subestimation and instability which range from small to relatively large. As a consequence of these problems of subestimation the confidence intervals present low coverage probabilities, as well as smaller lengths than what one would expect.

The very good performance of every one of the estimators of the means in the case of homogeneous nomination probabilities and regardless of the size of the $p_i^{(k)}$'s deserves an explanation. From Equations (1) we have that $\tilde{\tau}_1 = (M + R_1)/\tilde{\pi}^{(1)}$ and $\tilde{\tau}_2 = R_2/\tilde{\pi}^{(2)}$, and consequently

$$\tilde{\bar{Y}}_1 = \frac{1}{M + R_1}\sum_{j \in S_1} y_j^{(1)}, \quad \tilde{\bar{Y}}_2 = \frac{1}{R_2}\sum_{j \in S_2} y_j^{(2)}, \quad \text{and}$$

$$\tilde{\bar{Y}} = \frac{\tilde{\pi}^{(2)}(M + R_1)}{\tilde{\pi}^{(2)}(M + R_1) + \tilde{\pi}^{(1)}R_2}\tilde{\bar{Y}}_1 + \frac{\tilde{\pi}^{(1)}R_2}{\tilde{\pi}^{(2)}(M + R_1) + \tilde{\pi}^{(1)}R_2}\tilde{\bar{Y}}_2$$

Therefore, $\tilde{\bar{Y}}_1$ and $\tilde{\bar{Y}}_2$ are the sample means of the y-values associated with the elements in S_1 and S_2, respectively. Notice that given these samples, $\tilde{\bar{Y}}_1$ and $\tilde{\bar{Y}}_2$ do not depend on the $p_i^{(k)}$'s. In addition, since every element in U_k has the same probability of being included in S_k, it follows that $\tilde{\bar{Y}}_k$ is a good estimator of $\bar{Y}_k$, regardless of the size of the $p_i^{(k)}$'s, $k = 1, 2$. Furthermore, since in our simulation study $\bar{Y}_1 \approx \bar{Y}_2$, it follows that $\tilde{\bar{Y}} \approx \tilde{\bar{Y}}_1 \approx \tilde{\bar{Y}}_2$, and consequently $\tilde{\bar{Y}}$ is also a good estimator regardless of the size of the $p_i^{(k)}$'s. Notice also from the expression for $\tilde{\bar{Y}}$ that even if the values of $\tilde{\bar{Y}}_1$ and $\tilde{\bar{Y}}_2$ were very different from each other, and the $p_i^{(2)}$'s were small (which would imply that $\tilde{\pi}^{(2)}$ would also be small), $\tilde{\bar{Y}}$ would not have as serious problems of overestimation as $\bar{Y}$ would have. With respect to the estimators $\hat{\bar{Y}}_k^{(a)}$ and $\hat{\bar{Y}}^{(a)}$, $k = 1, 2$; $a = U, J, P$, by considering the values of the parameters a_k and b_k, $k = 1, 2$, used in the simulation study, and carrying out a similar analysis to the previous one, we can show that those estimators have similar performance to that of $\tilde{\bar{Y}}_k$ and $\tilde{\bar{Y}}$.

In the case of the artificial populations with heterogeneous probabilities, the estimators of the means also performed very well because the $y_j^{(k)}$'s were not associated with the $p_i^{(k)}$'s, and consequently the sample mean of the $y_j^{(k)}$'s was a good estimator of $\bar{Y}_k$. However, in the case of the reduced population obtained from the Colorado Springs study data, the performance of the estimators of the means was not very good because the $y_j^{(k)}$'s were associated with the $p_i^{(k)}$'s. For instance, the elements in U_2 with $y_j^{(2)} = 0$ were linked, on average, to 2.6 sites, whereas those with $y_j^{(2)} = 1$ to 1.2 sites; therefore, the sample mean of the $y_j^{(2)}$ associated with the elements in S_2 tended to subestimate $\bar{Y}_2$. It is worth noting that we carried out a small simulation study with the reduced population, but replacing the original y-values with values obtained by sampling from a chi-square distribution with one degree of freedom ($\chi^2(1)$) and also from a Bernoulli distribution with mean 0.1, so that the $y_j^{(k)}$'s were not associated with the $p_i^{(k)}$'s. The results, which are not shown, indicated a very good performance of the estimators of the means.

We want to end this section with the following remarks. (1) In the variant of LTS sampling proposed by Félix-Medina and Thompson (2004) it is assumed that each person in U_1 is assigned to only one site in the frame. Although this assumption reduces the efficiency of the sampling design, its relaxation would make the derivation of estimators more difficult, and we consider that this is a topic for future research. (2) The results of the simulation studies indicate that when the number of nominees is not so small, the proposed estimators of the totals and means are robust to deviations from the assumed Poisson distribution for the M_i's. (3) The simulation results also indicate that when the homogeneity assumption is not satisfied, the estimators yield estimates of the parameters of the correct order of magnitude. (4) The previous analysis shows that even in presence of heterogeneity, the estimators of the means perform well if the y-values are not associated with the nomination probabilities. (5) To reduce the effect of the heterogeneity, one could divide the population into subpopulations defined according to the values of an appropriate categorical variable, such as race, socio-economic status or gender. Then one could estimate the total of the variable of interest for each subpopulation and the sum of those estimates would be an estimate of the population total. An estimate of the variance of this estimator could be obtained by summing the estimates of the variances of the estimators of the subpopulation totals. (6) The previous remarks imply that our proposed sampling strategy is a reasonable alternative when the size of the population is unknown and the researcher is interested in estimating that parameter and additionally in estimating means and totals of some response variables (although inferences based on confidence intervals might not be reliable). (7) If the researcher's interest is only in estimating means and proportions, he or she has other alternatives, such as RDS. RDS has the advantage of being more economical and easier to perform than the LTS variant employed by Félix-Medina and Thompson, but the construction of the frame of sites gives the latter variant the advantage of producing good estimates of the means of any characteristics of the elements in U_1. Thus, if U_1 is a great portion of U, those estimates could be used as estimates of the means of corresponding characteristics of the elements in U. Regardless of this fact, it is not clear which alternative is the best from a statistical point of view, and consequently, further research needs to be carried out to answer this question. (8) Other topics that require to be researched are the development of procedures for testing the presence of heterogeneity and the development of estimators that take into account this characteristic.

10. References

Coull, B.A. and Agresti, A. (1999). The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-recapture Studies. Biometrics, 55, 294–301.

Dávid, B. and Snijders, T.A.B. (2002). Estimating the Size of the Homeless Population in Budapest, Hungary. Quality & Quantity, 36, 291–303.

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.

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

Gile, K.J. and Handcock, M.S. (2009). Respondent-driven Sampling: An Assessment of Current Methodology. arXiv:0904.1855v1 [stat.AP]

Haines, D.E. and Pollock, K.H. (1998). Combining Multiple Frames to Estimate Population Size and Totals. Survey Methodology, 24, 79–88.

Haines, D.E., Pollock, K.H., and Pantula, S.G. (2000). Population Size and Total Estimation when Sampling from Incomplete List Frames with Heterogeneous Inclusion Probabilities. Survey Methodology, 26, 121–129.

Heckathorn, D.D. (1997). Respondent-driven sampling: A New Approach to the Study of Hidden Populations. Social Problems, 44, 174–199.

Heckathorn, D.D. (2002). Respondent-driven Sampling II: Deriving Valid Population Estimates from Chain-referral Samples of Hidden Populations. Social Problems, 49, 11–34.

Kalton, G. (2001). Practical Methods for Sampling Rare and Mobile Populations. Proceedings of the American Statistical Association, Section on Survey Research Methods, Available from http://www.amstat.org/sections/srms/Proceedings/y2001/Proceed/00454.pdf.

Kalton, G. (2009). Methods for Oversampling Rare Populations in Social Surveys. Survey Methodology, 35, 125–141.

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Jun-Kim, B., Thorson, A., and Liljeros, F. (2010). The Sensitivity of Respondent-driven Sampling Method. arXiv:1002.2426v3 [stat.AP]

MacKellar, D., Valleroy, L., Karon, J., Lemp, G., and Janssen, R. (1996). The Young Men's Survey: Methods for Estimating HIV Sero-prevalence and Risk Factors Among Young Men Who Have Sex With Men. Public Health Reports, 111 (Supplement 1), 138–144.

McKenzie, D.J. and Mistiaen, J. (2009). Surveying Migrant Households: A Comparison of Census-based, Snowball and Intercept Point Surveys. Journal of the Royal Statistical Society Series *A*, 172, 339–360.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. (2005). Review of Sampling Hard-to-reach and Hidden Populations for HIV Surveillance. AIDS, 19, S67–S72.

Munhib, F.B., Lin, L.S., Stueve, A., Miller, R.L., Ford, W.L., Johnson, W.D., and Smith, P. (2001). A Venue-based Method for Sampling Hard-to-reach Populations. Public Health Reports, 116 (Supplement 1), 216–222.

Pollock, K.H., Turner, S.C., and Brown, C.A. (1994). Use of Capture-recapture Techniques to Estimate Population Size and Population Totals when a Complete Frame is Unavailable. Survey Methodology, 20, 117–124.

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, M. Morris (ed.). New York: Oxford University Press, 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, R.H. Needle, S.G. Genser, and R.T. II Trotter (eds). NIDA Research Monograph 151, Rockville, MD: National Institute of Drug Abuse, 3–19.

Salganik, M. and Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations using Respondent-driven Sampling. Sociological Methodology, 34, 193–239.

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

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

Volz, E. and Heckathorn, D.D. (2008). Probability Based Estimation Theory for Respondent-driven Sampling. Journal of Official Statistics, 24, 79–97.

Xi, L., Watson, R., and Yip, P.S.F. (2008). The Minimum Capture Proportion for Reliable Estimation in Capture-recapture Models. Biometrics, 64, 242–249.

Received September 2008
Revised June 2010