Catalogue no. 11-522-X
ISSN 1709-8211

Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society

Bayesian model assisted design-based estimators of the size, total and mean of a hard-to-reach population from a link-tracing sample with initial cluster sample

by Martín H. Félix-Medina

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

Bayesian Model Assisted Design-Based Estimators of the Size, Total and Mean of a Hard-to-Reach Population from a Link-Tracing Sample with Initial Cluster Sample

Martín H. Félix-Medina[1]

Abstract

We present design-based Horvitz-Thompson and multiplicity estimators of the population size, as well as of the total and mean of a response variable associated with the elements of a hidden population to be used with the link-tracing sampling variant proposed by Félix-Medina and Thompson (2004). Since the computation of the estimators requires to know the inclusion probabilities of the sampled people, but they are unknown, we propose a Bayesian model which allows us to estimate them, and consequently to compute the estimators of the population parameters. The results of a small numeric study indicate that the performance of the proposed estimators is acceptable.

Key Words: Gibbs sampling; Hidden population; Horvitz-Thompson estimators; Snowball sampling.

1. Introduction

Conventional sampling methods are not appropriate for sampling hidden or hard-to-detect populations, such as drug users, sex workers and homeless people, because of factors such as lack of appropriate sampling frames, rareness of those population and elusiveness of their members to be sampled. See Tourangeau (2014) for a discussion about these and other factors. For this reason, several methods have been proposed for sampling this type of population. One of these is link-tracing sampling (LTS). The idea behind this method is to select an initial sample from the hidden population, and then ask the people in the initial sample to name their contacts who are also members of the population. The named people who are not in the initial sample are included in the sample and might be asked to name their contacts who also belong to the population. This process might continue in this way until specified stopping rule is satisfied.

Félix-Medina and Thompson (2004) proposed a variant link-tracing sampling (LTS) to estimate the size of a hard-to-reach population. In their proposed sampling design, a portion of the population is covered by a sampling frame of venues, such as bars, parks, and block streets, where the members of the population tend to gather. A simple random sample without replacement of venues is selected from the frame and the members of the population who belong to any of the sampled venues are included in the sample. Next, from each sampled venue its members are asked to name their contacts who also belong to the population.

Estimators of the population size, as well as estimators of the population total and mean of a variable of interest, such as weekly drug spending of a drug user and weekly number of clients of a sex worker, which have been proposed for use with this sampling design and derived under different models appear in Félix-Medina and Thompson (2004), Félix-Medina and Monjardin (2006), Félix-Medina and Monjardin (2010), Félix-Medina et al. (2015) and Félix-Medina (2021). In this work, we present estimators of the population size, total and mean which are derived under a design-based approach, but assisted in a Bayesian model. Thus, we expect that the proposed estimators perform acceptably under a wide range of conditions.

2. Sampling design and notation

[1]Martín H. Félix-Medina, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacan Sinaloa, Mexico, 80013 (mhfelix@uas.edu.mx)

The variant of LTS proposed by Félix-Medina and Thompson (2004) is as follows. Let U be a finite population of τ elements. A portion U_1 of U is assumed that is covered by a sampling frame of N venues $A_1, \ldots, A_N$, where members of the population tend to gather. Let m_i be the number of elements of U_1 that belong to A_i, $i = 1, \ldots, N$. As in conventional cluster sampling, a person in U_1 is assumed to belong to only one venue; thus, the number of people in U_1 is $\tau_1 = \sum_1^N m_i$. Let $U_2 = U - U_1$ be the portion of U that is not covered by the frame. Notice that the size of U_2 is $\tau_2 = \tau - \tau_1$. A simple random sample without replacement $S_A = \{A_1, \ldots, A_n\}$ of n venues is selected from the frame. The m_i elements of U_1 that belong to each $A_i \in S_A$ are included in the sample. Let S_0 be the set of members of U_1 that belong to the sampled venues and let $m = \sum_1^n m_i$ be the size of S_0. In each sampled venue its members are asked to name their contacts who also belong to the population. A named person is said to be linked to a venue if any of the members of that venue name him or her. Let $x_{ij}^{(k)} = 1$ if person $j \in U_k - A_i$ is linked to venue $A_i \in S_A$, and $x_{ij}^{(k)} = 0$ otherwise, $k = 1, 2$. For sampled person $j \in U_k$, the following information associated with him or her is recorded: the value $y_j^{(k)}$ of the variable of interest y; the values $x_{ij}^{(k)}$, $i = 1, \ldots, n$, of the link-indicator variables, and to which of the following subsets of U the person belongs: $U_1 - S_0$, a specific $A_i \in S_A$ or U_2. We will denote by S_1 and S_2 the sets of r_1 and r_2 people in $U_1 - S_0$ and U_2 that are linked to at least one venue $A_i \in S_A$. Finally, let $S_1^* = S_0 \cup S_1$ and $S_2^* = S_2$ be the sets of $m + r_1$ and r_2 sampled people from U_1 and U_2, respectively.

3. Estimators of the size, total and mean

Let $Y_k = \sum_1^{\tau_k} y_j^{(k)}$ and $\bar{Y}_k = Y_k/\tau_k$ be the population total and mean of the y-values associated with the elements of U_k, $k = 1, 2$, and let $Y = Y_1 + Y_2$ and $\bar{Y} = Y/\tau$ be the corresponding total and mean of the elements of U. Notice that if $y_j^{(k)} = 1$ for each $j \in U_k$, then $Y_k = \tau_k$, $k = 1, 2$, and $Y = \tau$. Horvitz-Thompson estimators of τ_k, Y_k, τ and Y are $\hat{\tau}_k = \sum_{j \in S_k^*} 1/\pi_j^{(k)}$, $\hat{Y}_k = \sum_{j \in S_k^*} y_j^{(k)}/\pi_j^{(k)}$, $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$ and $\hat{Y} = \hat{Y}_1 + \hat{Y}_2$, where $\pi_j^{(k)}$ is the inclusion probability of the element $j \in U_k$, and which will be derived later. Estimators of the means $\bar{Y}_k$ and $\bar{Y}$ based on Horvitz-Thompson estimators are $\hat{\bar{Y}}_k = \hat{Y}_k/\hat{\tau}_k$, $k = 1, 2$, and $\hat{\bar{Y}} = \hat{Y}/\hat{\tau}$.

To compute the inclusion probability $\pi_j^{(k)}$ we will define the variables $N_j^{(k)} = \sum_{i=1}^N x_{ij}^{(k)}$ and $X_j^{(k)} = \sum_{i=1}^n x_{ij}^{(k)}$ which count the numbers of sites in the frame and in the sample S_A, respectively, that are linked to person $j \in U_k$, $k = 1, 2$. Notice that $N_j^{(k)}$ is unknown, whereas $X_j^{(k)}$ is known. Since the sample S_A is selected by a SRSWOR design, it follows that $X_j^{(2)}$ has a Hypergeometric distribution $\left(N, N_j^{(2)}, n\right)$. Therefore, $\pi_j^{(2)} = 1 - \Pr\left(X_j^{(2)} = 0 \middle| N_j^{(2)}\right) = 1 - \binom{N - N_j^{(2)}}{n} \Big/ \binom{N}{n}$, $j \in U_2$. In the case of person $j \in U_1$, the conditional distribution of $X_j^{(1)}$ given that $j \notin S_0$ is Hypergeometric distribution $\left(N - 1, N_j^{(1)}, n\right)$, because a person in U_1 cannot be linked to the venue to which he or she belongs. Thus, the inclusion probability $\pi_j^{(1)}$ of person $j \in U_1$, is given by

$$\pi_j^{(1)} = 1 - \Pr(j \notin S_0)\Pr(j \notin S_1 | j \notin S_0) = 1 - \left(1 - \frac{n}{N}\right)\Pr\left(X_j^{(1)} = 0 \middle| N_j^{(1)}, j \notin S_0\right) = 1 - \binom{N - N_j^{(1)}}{n} \Big/ \binom{N}{n}.$$

Notice that the inclusion probability $\pi_j^{(k)}$ depends on $N_j^{(k)}$ and consequently is unknown. Therefore, we will estimate it by estimating $N_j^{(k)}$ using a Bayesian approach.

We will firstly define a two-stage initial distribution of $N_j^{(k)}$. In the case of $j \in U_2$, we will suppose that $Pr\left(N_j^{(2)} \geq 1\right) = 1$, because otherwise $j \in U_2$ cannot be sampled. Thus, we will define as the initial distribution of $N_j^{(2)} | \alpha_2, \beta_2$ the truncated at zero Beta-Binomial (N, α_2, β_2), that is

$$g_{N_j^{(2)}|N,\alpha_2,\beta_2}\left(n_j^{(2)} | N, \alpha_2, \beta_2\right) = \binom{N}{n_j^{(2)}} \frac{B(\alpha_2 + n_j^{(2)}, N + \beta_2 - n_j^{(2)})}{B(\alpha_2, \beta_2)} \frac{1}{1 - B(\alpha_2, N + \beta_2)/B(\alpha_2, \beta_2)}, \quad (1)$$

$n_j^{(2)} = 1, \ldots, N$. In the case of $j \in U_1$, we will define as the initial distribution of $N_j^{(1)}|\alpha_1, \beta_1$ the truncated at zero Beta-Binomial $(N-1, \alpha_1, \beta_1)$, that is

$$g_{N_j^{(1)}|N,\alpha_1,\beta_1}\left(n_j^{(1)}|N,\alpha_1,\beta_1\right) = \binom{N-1}{n_j^{(1)}} \frac{B(\alpha_1 + n_j^{(1)}, N-1+\beta_1-n_j^{(1)})}{B(\alpha_1,\beta_1)} \frac{1}{1-B(\alpha_1, N-1+\beta_1)/B(\alpha_1,\beta_1)}, \quad (2)$$

$n_j^{(1)} = 1, \ldots, N-1$. Notice that we used $N-1$ instead of N. The reason for this is because a person in U_1 cannot be linked to the venue to which he or she belongs; therefore, an upper bound of his or her number of links is $N-1$. To end the specification of the initial distribution we need to define the joint distribution of (α_k, β_k). However, instead of specifying a distribution for these parameters, we will use a reparameterization suggested by Lee and Sabavala (1987), who define the parameters $\mu_k = \alpha_k/(\alpha_k+\beta_k)$ and $\rho_k = 1/(\alpha_k+\beta_k)$. Notice that the inverse transformation is $\alpha_k = \mu_k(1-\rho_k)/\rho_k$ and $\rho_k = (1-\mu_k)(1-\rho_k)/\rho_k$. Since the parameters μ_k and ρ_k are between zero and one, they define as the initial distributions of μ_k and ρ_k the beta distributions with parameters $\left(a^{(\mu_k)}, b^{(\mu_k)}\right)$ and $\left(a^{(\mu_k)}, b^{(\mu_k)}\right)$, respectively. The joint probability density function $g(\mu_k, \rho_k)$ of (μ_k, ρ_k) is defined as the product of the density functions $g_{\mu_k}(\mu_k)$ and $g_{\rho_k}(\rho_k)$ of μ_k and ρ_k, respectively, $k = 1, 2$.

The likelihood function is based on the variables $X_j^{(k)}$ associated with people in the sample. We have that if $j \in S_2$, then $\Pr\left(X_j^{(2)} \geq 1 \middle| j \in S_2\right) = 1$, and consequently, the conditional distribution of $X_j^{(2)}$, given that $j \in S_2$ and $N_j^{(2)}$, is the truncated at zero Hypergeometric distribution $\left(N, N_j^{(2)}, n\right)$, that is

$$f_{X_j^{(2)}|j\in S_2,N_j^{(2)}}\left(x_j^{(2)}|j \in S_2, n_j^{(2)}\right) = \left[\binom{n_j^{(2)}}{x_j^{(2)}}\binom{N-n_j^{(2)}}{n-x_j^{(2)}} \Big/ \binom{N}{n}\right]\left[1-\binom{N-n_j^{(2)}}{n} \Big/ \binom{N}{n}\right]^{-1}$$

Similarly, if $j \in S_1$, the conditional distribution of $X_j^{(1)}$, given that $j \in S_1$ and $N_j^{(1)}$, is the truncated at zero Hypergeometric distribution $\left(N-1, N_j^{(1)}, n\right)$, that is

$$f_{X_j^{(1)}|j\in S_1,N_j^{(1)}}\left(x_j^{(1)}|j \in S_1, n_j^{(1)}\right) = \left[\binom{n_j^{(1)}}{x_j^{(1)}}\binom{N-1-n_j^{(1)}}{n-x_j^{(1)}} \Big/ \binom{N-1}{n}\right]\left[1-\binom{N-1-n_j^{(1)}}{n} \Big/ \binom{N-1}{n}\right]^{-1},$$

whereas if $j \in A_i \in S_A$, the conditional distribution of $X_j^{(A_i)}$, given that $j \in A_i \in S_A$ and $N_j^{(A_i)}$, is the Hypergeometric distribution $\left(N-1, N_j^{(A_i)}, n-1\right)$, that is

$$f_{X_j^{(A_i)}|j\in A_i\in S_A,N_j^{(A_i)}}\left(x_j^{(A_i)}|j \in A_i \in S_A, n_j^{(A_i)}\right) = \binom{n_j^{(A_i)}}{x_j^{(A_i)}}\binom{N-1-n_j^{(A_i)}}{n-1-x_j^{(A_i)}} \Big/ \binom{N-1}{n-1},$$

where we have denoted by $N_j^{(A_i)}$ and $X_j^{(A_i)}$ the variables that count the numbers of venues in the frame and in the sample S_A, respectively, that are linked to person $j \in A_i \in S_A$.

To estimate $N_j^{(k)}$, $k = 1, 2$, and $N_j^{(A_i)}$, $i = 1, \ldots, n$, we will use a latent class approach. Thus, in the case of person $j \in S_2$, we will define the vector of latent indicator variables associated with that person as $C_j^{(2)} = \left(C_{1j}^{(2)}, \ldots, C_{Nj}^{(2)}\right)$, where $C_{lj}^{(2)} = 1$ if $N_j^{(2)} = l$ and $C_{lj}^{(2)} = 0$ otherwise, $l = 1, \ldots, N$. In the case of person $j \in S_1$, his or her associated vector of latent indicator variables is $C_j^{(1)} = \left(C_{0j}^{(1)}, \ldots, C_{N-1j}^{(1)}\right)$, where $C_{lj}^{(1)} = 1$ if $N_j^{(1)} = l$ and $C_{lj}^{(1)} = 0$ otherwise, $l = 0, \ldots, N-1$. Finally, in the case of person $j \in A_i \in S_A$, the vector is $C_j^{(A_i)} = \left(C_{0j}^{(A_i)}, \ldots, C_{N-1j}^{(A_i)}\right)$, where $C_{lj}^{(A_i)} = 1$ if $N_j^{(A_i)} = l$ and $C_{lj}^{(A_i)} = 0$ otherwise, $l = 0, \ldots, N-1$. Notice that $N_j^{(k)}$ determines $C_j^{(k)}$ and conversely; therefore, the problem of estimating $N_j^{(k)}$ is equivalent to that of estimating $C_j^{(k)}$. The same can be indicated about $N_j^{(A_i)}$ and $C_j^{(A_i)}$, $A_i \in S_A$. We will focus on estimating $C_j^{(k)}$, $k = 1, 2$, and $C_j^{(A_i)}$, $A_i \in S_A$. The conditional probability mass function (pmf), $g_{C_j^{(2)}|N,\mu_2,\rho_2}\left(c_j^{(2)}|N,\mu_2,\rho_2\right)$, of $C_j^{(2)}$ given μ_2 and ρ_2 is the one of the Multinomial $\left(1, \left\{g_{N_j^{(2)}|N,\alpha_2,\beta_2}(l|N,\alpha_2^*,\beta_2^*)\right\}_{l=1}^{N}\right)$, whereas the conditional pmf, $g_{C_j^{(1)}|N,\mu_1,\rho_1}\left(c_j^{(1)}|N,\mu_1,\rho_1\right)$, of $C_j^{(1)}$, as well as the

conditional pmf, $g_{C_j^{(A_i)}|N,\mu_1,\rho_1}\left(c_j^{(A_i)}|N,\mu_1,\rho_1\right)$, of $C_j^{(A_i)}$, given μ_1 and ρ_1, are both equal to the one of the same Multinomial $\left(1,\left\{g_{N_j^{(1)}|N,\alpha_1,\beta_1}(l|N,\alpha_1^*,\beta_1^*)\right\}_{l=0}^{N-1}\right)$, where $g_{N_j^{(2)}|N,\alpha_2,\beta_2}$ and $g_{N_j^{(1)}|N,\alpha_1,\beta_1}$ are given by (1) and (2), respectively, and α_k^* and β_k^* are given by the inverse function of (μ_k,ρ_k), $k=1,2$.

Under the assumption that the vectors $\left(X_j^{(k)},C_j^{(k)}\right)$ associated with the r_k people in S_k are mutually independent, we will have that their joint conditional distribution given μ_k and ρ_k is

$$f_{X^{(k)},C^{(k)}|\mu_k,\rho_k}\left(x^{(k)},c^{(k)}|\mu_k,\rho_k\right)=\prod_{j=1}^{r_k}f_{X_j^{(k)}|j\in S_k,C_j^{(k)}}\left(x_j^{(k)}|j\in S_k,c_j^{(k)}\right)g_{C_j^{(k)}|N,\mu_k,\rho_k}\left(c_j^{(k)}|N,\mu_k,\rho_k\right),$$

where $f_{X_j^{(k)}|j\in S_k,C_j^{(k)}}$ is the zero-truncated hypergeometric distribution of $X_j^{(k)}$, $X^{(k)}$ is the vector of variables $X_j^{(k)}$, and $C^{(k)}$ is the matrix whose columns are the vectors $C_j^{(k)}$, $j\in S_k$, $k=1,2$. Also, under the assumption that the vectors $\left(X_j^{(A_i)},C_j^{(A_i)}\right)$ associated with the m people in S_0 are mutually independent, we will have that their joint conditional distribution given μ_1 and ρ_1 is

$$f_{X^{(0)},C^{(0)}|,\mu_1,\rho_1}\left(x^{(0)},c^{(0)}|\mu_1,\rho_1\right)=\prod_{i=1}^{n}\prod_{j=1}^{m_i}f_{X_j^{(A_i)}|j\in A_i\in S_A,C_j^{(A_i)}}\left(x_j^{(A_i)}|j\in S_0,c_j^{(A_i)}\right)g_{C_j^{(A_i)}|N,\mu_1,\rho_1}\left(c_j^{(A_i)}|N,\mu_1,\rho_1\right),$$

where $f_{X_j^{(A_i)}|j\in A_i\in S_A,C_j^{(A_i)}}$ is the hypergeometric distribution of $X_j^{(A_i)}$, $X^{(0)}$ is the vector of variables $X_j^{(A_i)}$, and $C^{(0)}$ is the matrix whose columns are the vectors $C_j^{(A_i)}$, $j\in A_i$, $i=1,\dots,n$.

The joint probability density function of the final distribution of $\left(C^{(0)},C^{(1)},C^{(2)},\mu_1,\mu_2,\rho_1,\rho_2\right)$ is

$$g_{C^{(0)},C^{(1)},C^{(2)},\mu_1,\mu_2,\rho_1,\rho_2|data}\left(c^{(0)},c^{(1)},c^{(2)},\mu_1,\mu_2,\rho_1,\rho_2|data\right)=g_{C^{(0)},C^{(1)},\mu_1,\rho_1|data}\left(c^{(0)},c^{(1)},\mu_1,\rho_1|data\right)g_{\,C^{(2)},\mu_2,\rho_2|data}\left(\,c^{(2)},\mu_2,\rho_2|data\right),$$

where

$$g_{C^{(0)},C^{(1)},\mu_1,\rho_1|data}\left(c^{(0)},c^{(1)},\mu_1,\rho_1|data\right)\propto f_{X^{(1)},C^{(1)}|\mu_1,\rho_1}\left(x^{(1)},c^{(1)}|\mu_1,\rho_1\right)f_{X^{(0)},C^{(0)}|,\mu_1,\rho_1}\left(x^{(0)},c^{(0)}|\mu_1,\rho_1\right)g(\mu_1,\rho_1)$$

and

$$g_{\,C^{(2)},\mu_2,\rho_2|data}\left(\,c^{(2)},\mu_2,\rho_2|data\right)\propto f_{X^{(2)},C^{(2)}|\mu_2,\rho_2}\left(x^{(2)},c^{(2)}|\mu_2,\rho_2\right)g(\mu_2,\rho_2).$$

Therefore, with respect to the final distribution, $(C^{(0)},C^{(1)},\mu_1,\rho_1)$ and $(C^{(2)},\mu_2,\rho_2)$ are independent.

Inferences about the parameters of interest are made by means of Gibbs sampling. Thus, the final conditional distribution of each of the parameters $C_j^{(A_i)}$, $j\in A_i\in S_A$, $C_j^{(k)}$, $j\in S_k$, μ_k and ρ_k, $k=1,2$, given the rest of the parameters is obtained. The procedure is implemented by specifying the number of chains and the length T of each chain, as well as the initial values of the parameters μ_k and ρ_k. Then, at each iteration of the algorithm, a value of each parameter is sampled from its conditional distribution given the most recent values of the rest of the parameters. This allows simulating values of the parameters $N_j^{(k)}$ and $N_j^{(A_i)}$, values of the inclusion probabilities $\pi_j^{(k)}$, $j\in S_k^*$, $k=1,2$, and $\pi_j^{(A_i)}$, $j\in A_i$, $i=1,\dots,n$, and consequently values of the Horvitz-Thompson estimators of the sizes, totals and means. We also simulate values of the multiplicity estimators (Birnbaum and Sirken, 1965) of the sizes, totals and means, as well as values of the estimators of their conditional variances given the values of $N_j^{(k)}$ and $N_j^{(A_i)}$. The idea of computing estimates of the variances the multiplicity estimators is that they are easier to compute than those of the Horvitz-Thompson estimators, and they were used in the computation of the estimates of the variances of the Horvitz-Thompson estimators.

Once the iterations of the Gibbs sampling algorithm were carried out, estimates of the unconditional variances of the Horvitz-Thompson and Multiplicity estimators were computed by using the two-stage formula $\hat{V}\left(\hat{\theta}^{(k)}\right)=V\left[E\left(\hat{\theta}^{(k)}|\widehat{\{N_j^{(k)}\}_{j\in S_k^*}}\right)\right]+E\left[V\left(\hat{\theta}_M^{(k)}|\widehat{\{N_j^{(k)}\}_{j\in S_k^*}}\right)\right]$, where $\hat{\theta}^{(k)}$ denotes either the Horvitz-Thompson or the Multiplicity

estimator of the population parameter θ_k ($that\ is, \tau_k$, Y_k or $\bar{Y}_k$), $V\left[E\left(\hat{\theta}^{(k)}|\widehat{\{\widehat{N}_j^{(k)}\}_{J\in S_k^*}}\right)\right]$ is the sample variance of the simulated values of the estimator $\hat{\theta}^{(k)}$ (after removing the values corresponding to the burn in period), and $E\left[V\left(\hat{\theta}_M^{(k)}|\widehat{\{\widehat{N}_j^{(k)}\}_{J\in S_k^*}}\right)\right]$ is the sample mean of the simulated values of the estimator of the conditional variance of the multiplicity estimator $\hat{\theta}_M^{(k)}$ of θ_k. The estimate of the variance of the estimator $\hat{\theta}$ of the parameter θ of the whole population U is obtained by summing the unconditional variances of the estimators $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$.

4. Monte Carlo study

To observe the performance of the proposed estimators, estimators of their variances and Wald confidence intervals based on those estimators, we carried out a small simulation study. Thus, we used data from the National Longitudinal Study of Adolescent Health (Add Health) collected during the 1994–1995 school year to construct a population. See Harris (2013) for a description of this study. Specifically, data from high school and its feeder middle school in Community 50 were used to construct a population U of $\tau = 2487$ elements divided into subpopulations U_1 and U_2 of sizes $\tau_1 = 1800$ and $\tau_2 = 687$, respectively. The elements in U_1 were grouped into $N = 150$ clusters of sizes m_i, $i = 1, \ldots, N$, whose values were generated from a negative binomial distribution with mean and variance equal to 12 and 24, respectively. A student in U was linked to a cluster if any of the students in the cluster named that student as his or her friend. The variable of interest associated with each student was the number of friends named by him or her. The population totals were $Y_1 = 10162$, $Y_2 = 2631$ and $Y = 12793$; and the population means were $\bar{Y}_1 = 5.65$, $\bar{Y}_2 = 3.83$ and $\bar{Y} = 5.14$. The study was carried out by repeatedly selecting 1000 samples from the population using the sampling design described in Section 2. The size of the initial sample S_A of clusters was $n = 20$. From each sample, inferences about each parameter were obtained by using the Gibbs sampling algorithm with two chains, each one of length 4000 and a burn in period of 2000. The study was carried out using the R software environment for statistical computing (R Core Team, 2022).

Table 4-1
Results of the Monte Carlo study based on 1000 replicated samples selected from an artificial population constructed using data from the National Longitudinal Study on Adolescent Health.

Horvitz-Thompson estimators		Sizes			Totals			Means		
		$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}$	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}$
Estimators of population parameters	Pop. parameter	1800	687	2487	10162	2631	12793	5.65	3.83	5.14
	Mean	1602.8	730.3	2333.1	9556.4	3089.3	12645.7	6.0	4.2	5.4
	Relative bias	-0.11	0.06	-0.06	-0.06	0.17	-0.01	0.06	0.11	0.06
	$\sqrt{\text{Relative } MSE}$	0.12	0.16	0.09	0.08	0.24	0.06	0.06	0.12	0.06
Estimators of standard deviations	Std. deviation	85.2	99.1	144.3	504.6	414.8	711.0	0.08	0.17	0.09
	Mean	144.2	117.8	222.7	916.2	516.9	1215.0	0.11	0.25	0.13
	Relative bias	0.69	0.19	0.54	0.81	0.25	0.71	0.45	0.49	0.41
	$\sqrt{\text{Relative } MSE}$	0.75	0.29	0.60	0.87	0.33	0.76	0.48	0.51	0.44
95% conf. intervals	Coverage prob.	0.78	0.98	0.93	0.96	0.96	0.99	0.08	0.70	0.40
	Relative length	0.31	0.67	0.35	0.35	0.77	0.37	0.08	0.26	0.10
Multiplicity estimators		Sizes			Totals			Means		
		$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}$	$\tilde{Y}_1$	$\tilde{Y}_2$	$\tilde{Y}$	$\tilde{\bar{Y}}_1$	$\tilde{\bar{Y}}_2$	$\tilde{\bar{Y}}$
Estimators of population parameters	Pop. parameter	1800	687	2487	10162	2631	12793	5.65	3.83	5.14
	Mean	1579.5	731.0	2310.4	9581.8	3157.1	12738.8	6.1	4.3	5.5
	Relative bias	-0.12	0.06	-0.07	-0.06	0.20	-0.00	0.08	0.13	0.07
	$\sqrt{\text{Relative } MSE}$	0.13	0.16	0.09	0.08	0.25	0.06	0.08	0.14	0.08
Estimators of standard deviations	Std. deviation	83.1	99.0	142.5	502.3	411.6	704.2	0.09	0.16	0.10
	Mean	149.1	124.0	229.1	945.9	543.4	1248.9	0.11	0.26	0.14
	Relative bias	0.80	0.25	0.61	0.88	0.32	0.77	0.27	0.63	0.37
	$\sqrt{\text{Relative } MSE}$	0.84	0.33	0.66	0.93	0.39	0.82	0.32	0.64	0.40
95% conf. intervals	Coverage prob.	0.75	0.98	0.93	0.98	0.96	0.99	0.02	0.53	0.19
	Relative length	0.33	0.71	0.36	0.35	0.81	0.38	0.08	0.27	0.11

The results of the study are shown in Table 4-1. We can see that the Horvitz-Thompson estimators of the sizes, totals and means presented some bias issues, although most of the values of the relative biases were, in absolute value, less than or close to 0.1, except that of the estimator $\hat{Y}_2$, which was a relatively large value. The values of the square roots of the relative mean square errors of the estimators were, in general, acceptable, that is, they were less than or close to 0.1, except those of the estimators of τ_2 and Y_2, which were somewhat large. Thus, in general, the performance of the Horvitz-Thompson estimators was acceptable. With respect to the estimators of the standard deviations of the Horvitz-Thompson estimators, they presented serious problems of overestimation. In the case of the 95% confidence intervals of the population sizes and totals, they presented acceptable values of the coverage probabilities, except the interval of τ_1, which had a relatively small value of the coverage probability. The relative lengths of these intervals were also acceptable, except those of the intervals of τ_2 and Y_2, which were relatively large. However, the intervals of the population means had very low values of the coverage probabilities. The problem was that their lengths were very small, and this along with the small biases of the point estimators of the means yielded the very low values of the coverage probabilities. Nonetheless, and even considering the very small values of the coverage probabilities of these intervals, we think that they still provide good information about the means because the intervals are very short and close enough to the true values of the means. Finally, with respect to the Multiplicity estimators of the sizes, totals and means, the estimators of their standard deviations and their corresponding 95% confidence intervals, we can say that their performance was similar to, but slightly lower than that of the corresponding Horvitz-Thompson estimators.

Acknowledgements

This research was supported by grant PROFAPI-2022, PRO_A1_014 of the Universidad Autonoma de Sinaloa.

References

Birnbaum, Z.W., and M.G. Sirken (1965), “Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates”, *Vital and Health Statistics*, Ser. 2, No. 11. Washington: Government Printing Office.

Félix-Medina, M.H., and S.K. Thompson (2004), “Combining Cluster Sampling and Link-Tracing Sampling to Estimate the Size of Hidden Populations”, *Journal of Official Statistics*, 20, pp. 19-38.

Félix-Medina, M.H., and P.E. Monjardin (2006), “Combining Link-Tracing Sampling and Cluster Sampling and to Estimate the Size of Hidden Populations: a Bayesian-Assisted Approach”, 32, pp. 187-195.

Félix-Medina, M.H., and P.E. Monjardin (2010), “Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations”, *Journal of Official Statistics*, 26, pp. 603–631.

Félix-Medina, M.H., P.E. Monjardin, and A.N. Aceves Castro (2015), “Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities”, *Survey Methodology*, 41, pp. 349-376.

Félix-Medina, M.H. (2021), “Combining Cluster Sampling and Link-Tracing Sampling to Estimate Totals and Means of Hidden Populations in Presence of Heterogeneous Probabilities of Links”, *Journal of Official Statistics*, 37, pp. 865–905.

Harris, K.M. (2013), “The Add Health Study: Design and Accomplishments”, unpublished report. Available at: htpps://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf.

Lee, J.C., and D.J. Sabavala (1987), “Bayesian Estimation and Prediction for the Beta-Binomial Model”, *Journal of Business and Economic Statistics*, 5, pp. 357-367.

R Core Team (2022), “R: A Language and Environment for Statistical Computing”. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org.

Tourangeau, R. (2014), “Defining hard-to-survey populations”, in R. Tourangeauet et al. (eds.) *Hard-to-Survey Populations*, Cambridge: Cambridge University Press, pp. 3-20.