Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Assessing the coverage of confidence intervals under nonresponse. A case study on income mean and quantiles in some municipalities from the 2015 Mexican Intercensal Survey

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

by Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz and
Guillermina Eslava-Gómez

Release date: January 6, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 371-382
Statistics Canada, Catalogue No. 12-001-X

Assessing the coverage of confidence intervals under nonresponse. A case study on income mean and quantiles in some municipalities from the 2015 Mexican Intercensal Survey

Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz and Guillermina Eslava-Gómez[1]

Abstract

This note presents a comparative study of three methods for constructing confidence intervals for the mean and quantiles based on survey data with nonresponse. These methods, empirical likelihood, linearization, and that of Woodruff's (1952), were applied to data on income obtained from the 2015 Mexican Intercensal Survey, and to simulated data. A response propensity model was used for adjusting the sampling weights, and the empirical performance of the methods was assessed in terms of the coverage of the confidence intervals through simulation studies. The empirical likelihood and linearization methods had a good performance for the mean, except when the variable of interest had some extreme values. For quantiles, the linearization method had a poor performance, while the empirical likelihood and Woodruff methods had a better one, though without reaching the nominal coverage when the variable of interest had values with high frequency near the quantile of interest.

Key Words: Confidence interval estimation; Empirical likelihood; Linearization; Missing at random; Nonresponse; Two-phase sampling.

1. Introduction

The 2015 Mexican Intercensal survey (MIC2015) conducted by the National Institute of Statistics and Geography (INEGI, 2015) collected information nationwide, using a probability sampling design in 1,643 municipalities and through a census in 814 municipalities. In this study we use the census data corresponding to 441 municipalities from the state of Oaxaca.

We focus on *income* as the variable of interest which exhibits a nonresponse rate of about 22.5%. Considering the respondents, the distribution of *income* has a high skewness mainly due to the presence of extreme values, and shows some values with high frequency.

The objective of this study is to assess the empirical coverage rate of confidence intervals (CI) computed by three methods for the population mean and population quantiles, 0.1, 0.5 and 0.9, in survey data with nonresponse. Two-phase sampling is used with a random sample selected in the first phase, while in the second the sample is split into respondents and nonrespondents considering the nonresponse pattern of *income* in the census data. A response propensity model is used to adjust the weights for nonresponse.

[1] Omar De La Riva Torres, Departamento de Matemáticas, Facultad de Ciencias, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. E-mail: odelariva@ciencias.unam.mx; Gonzalo Pérez-de-la-Cruz, Departamento de Matemáticas, Facultad de Ciencias, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. E-mail: gonzalo.perez@ciencias.unam.mx; Guillermina Eslava-Gómez, Departamento de Matemáticas, Facultad de Ciencias, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. E-mail: eslava@ciencias.unam.mx.

For the population mean, we consider the Hájek estimator and two methods for computing CIs: empirical likelihood (Berger, 2020) and linearization (Särndal, Swenson and Wretman, 1992, Sections 5.2 and 5.7). Concerning the population quantiles, we consider the point estimator obtained by interpolation of the distribution function as in Woodruff (1952) and Graf and Tillé (2014), and three methods for computing CIs: empirical likelihood (Berger, 2020), Woodruff (Woodruff, 1952) and linearization (Deville, 1999). These methods are described in Section 2, the numerical results are presented in Section 3 for the MIC2015 data and in Section 4 for some simulated populations. Some final comments are given in Section 5.

2. Three methods for estimating confidence intervals

2.1 Estimation using two-phase sampling

We consider a finite population $U = \{1, 2, \ldots, N\}$ and a probability sample $s \subset U$ of fixed size $n,$ with first and second-order inclusion probabilities π_k and $\pi_{kl},\ k, l \in U,\ k \neq l.$ Let y_k be the k^{th} value of the variable of interest $y,$ and let θ_0 be a population parameter and $\hat{\theta}_0$ an estimator of $\theta_0.$

We assume that the value y_k is available for a subset $r \subset s$ only. Let ϕ_k denote the response probability for unit $k.$ Let I_k be a response indicator variable such that $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s \setminus r.$ We also assume that there is a vector of auxiliary variables $\mathbf{x}$ observed for all $k \in s.$ We make the missing at random (MAR) assumption:

$$P(I_k = 1 \mid y_k, \mathbf{x}_k) = P(I_k = 1 \mid \mathbf{x}_k) = \phi_k \ \forall\, k \in U.$$

The response probabilities ϕ_k are used for adjusting the design weights. We assume that the sampling design and the response mechanism are independent as in Berger (2020). Borrowing from two-phase sampling theory (Särndal et al., 1992, Section 9.3) the weights adjusted for nonresponse are defined as $\pi_k^{*-1} = 1/(\pi_k \phi_k)$ and the second-order inclusion probabilities as $\pi_{kl}^{*} = \pi_{kl}\, \phi_k\, \phi_l, k, l \in U,\ k \neq l.$

The interest lies in estimating the population mean $\bar{Y} = \sum_{k \in U} y_k / N$ and the population quantile given by

$$Y_q = y_{(d-1)} + \frac{\left(y_{(d)} - y_{(d-1)}\right) \left[qN - N_{(d-1)}\right]}{N_{(d)} - N_{(d-1)}}, \tag{2.1}$$

where $y_{(i)}$ is the value for the i^{th} unit arranged in increasing order, $d = \min\left\{l : qN < N_{(l)}, l = 1, \ldots, N\right\}$ and $N_{(l)} = \sum_{j \in U} I\left(y_j \leq y_{(l)}\right), l = 1, \ldots, N.$ Formula (2.1) is obtained by considering a piecewise linear interpolation of the step distribution function $F(y) = \sum_{k \in U} I(y_k \leq y)/N,$ where $I\left(y_k \leq y\right) = 1$ when $y_k \leq y.$

These population parameters are respectively estimated by

$$\hat{\hat{Y}} = \frac{\sum_{k \in r} y_k / \pi_k^*}{\sum_{k \in r} 1 / \pi_k^*}, \tag{2.2}$$

and

$$\hat{Y}_q = y_{(d-1)} + \frac{\left(y_{(d)} - y_{(d-1)}\right)\left(q\hat{N} - \hat{N}_{(d-1)}\right)}{\hat{N}_{(d)} - \hat{N}_{(d-1)}}, \tag{2.3}$$

where $d = \min\left\{l : q\hat{N} < \hat{N}_{(l)}, l = 1, \ldots, n_r\right\}$, $\hat{N} = \sum_{k \in r} 1/\pi_k^*$, $\hat{N}_{(l)} = \sum_{k \in r} I(y_k \leq y_{(l)}) / \pi_k^*$ and $n_r = \sum_{k \in s} I_k$.

These estimators and the CIs described in the following subsection are based on the assumption that the response probabilities ϕ_k are known, unlike Berger (2020) and Kim and Kim (2007). However, we use $\hat{\phi}_k$ instead of ϕ_k in the simulation studies, where

$$\hat{\phi}_k = \frac{\exp\left(\mathbf{x}_k^{\top}\hat{\beta}\right)}{1 + \exp\left(\mathbf{x}_k^{\top}\hat{\beta}\right)} \ \forall \ k \in s,$$

with $\hat{\beta}$ obtained by fitting a logistic regression using s. This leads to the estimators known as the empirical double expansion estimators (Haziza and Beaumont, 2017).

2.2 Methods for estimating confidence intervals

2.2.1 Linearization

The linearization method relies on the assumption that the distribution of $\hat{\theta}_0$ is approximately normal. A CI for θ_0 is

$$\left[\hat{\theta}_0 - z_{1-\alpha/2}[V(\hat{\theta}_0)]^{1/2},\ \hat{\theta}_0 + z_{1-\alpha/2}[V(\hat{\theta}_0)]^{1/2}\right], \tag{2.4}$$

where $1 - \alpha$ is the confidence level, also known as nominal coverage; see Särndal et al. (1992, expression 5.2.3). In practice $V(\hat{\theta}_0)$ is estimated. For the estimators given by (2.2) and (2.3), a variance estimator is given by

$$\hat{V}(\hat{\theta}_0) = \sum_{k \in r} \sum_{l \in r} \frac{(\pi_{kl}^* - \pi_k^* \pi_l^*)}{\pi_{kl}^*} \frac{\hat{z}_k}{\pi_k^*} \frac{\hat{z}_l}{\pi_l^*}, \tag{2.5}$$

where $\hat{z}_k = \left(y_k - \hat{\hat{Y}}\right) / \hat{N}$ for $\hat{\hat{Y}}$ (Särndal et al., 1992, Result 5.7.1) and $\hat{z}_k = -\left(I(y_k \leq \hat{Y}_q) - q\right) / \left(f(\hat{Y}_q)\hat{N}\right)$ for $\hat{Y}_q$ (Deville, 1999). The *density function* f was obtained in two ways: a) using a Gaussian kernel as in Osier (2009) and b) using the nearest neighbour technique as in Graf and Tillé (2014). We present the results pertaining to a), since the technique in b) led to similar results.

We note that using (2.5) with $\hat{\phi}_k$ instead of ϕ_k might lead to overestimation of the variance of the empirical double expansion estimator and to wider CIs; see the expression (17) in Kim and Kim (2007) associated with the estimators of $\bar{Y}$.

2.2.2 Empirical likelihood method

The empirical likelihood approach assumes that θ_0 is the unique solution of the estimating equation $G(\theta) = \sum_{k \in U} g_k(\theta) = 0$ for a given function g_k. In particular, we use:

i) $g_k(\theta) = y_k - \theta$ for $\theta_0 = \bar{Y}$.

ii) $g_k(\theta) = \rho(y_k, \theta) - q$ for $\theta_0 = Y_q$, where $\rho(y_k, \theta) = I(y_k < y_{(l)}) + I(y_k = y_{(l)})(\theta - y_{(l-1)}) / (y_{(l)} - y_{(l-1)})$ and $l = \min\{j : y_{(j)} > \theta\}$.

The empirical log-likelihood function in Berger and De La Riva Torres (2016) for a one-stage sampling design without stratification or auxiliary information is

$$\ell_{\max}(\theta) = \max_{m_k : k \in s} \left\{ \sum_{k \in s} \log(m_k) : m_k > 0, \sum_{k \in s} m_k g_k(\theta) = 0, \sum_{k \in s} m_k \pi_k = n \right\}, \tag{2.6}$$

where $\{m_k : k \in s\}$ satisfies the design and the parameter constraints $\sum_{k \in s} m_k \pi_k = n$ and $\sum_{k \in s} m_k g_k(\theta) = 0.$

In the presence of nonresponse, we use (2.6) replacing $\sum_{k \in s} m_k g_k(\theta) = 0$ with $\sum_{k \in s} m_k I_k g_k(\theta) / \phi_k = 0.$ A CI for θ_0 is given by

$$\left[\min\{\theta : \hat{R}(\theta) \le \chi_1^2(\alpha)\},\ \max\{\theta : \hat{R}(\theta) \le \chi_1^2(\alpha)\}\right], \tag{2.7}$$

where $\hat{R}(\theta) = 2\{\ell_{\max}(\hat{\theta}_0) - \ell_{\max}(\theta)\}$ and $\chi_1^2(\alpha)$ is the $(1-\alpha)$-quantile of the χ_1^2 distribution. The estimator $\hat{\theta}_0 := \text{argmax}_{\{\theta\}}\, \ell_{\max}(\theta)$ corresponds to (2.2) and (2.3), respectively.

We computed (2.7) using a root search method, calculating $\hat{R}(\theta)$ for several values of $\theta,$ where $\ell_{\max}(\theta)$ for a given value θ was obtained by a modified Newton-Raphson algorithm as in Wu (2004).

2.2.3 Woodruff method for quantiles

The method of Woodruff (1952) is based on the estimated distribution function $\hat{F}(y).$ For a quantile $Y_q,$ the variance of $\hat{F}(Y_q)$ can be approximated using the Taylor linearization method with linearized variable $z_k = (I(y_k \le Y_q) - q) / N,$ while the variance is estimated using (2.5) with $\hat{z}_k = (I(y_k \le \hat{Y}_q) - q) / \hat{N}.$ Assuming normality of $\hat{F}(Y_q)$ and using (2.4), it is possible to find a CI $[c_1, c_2]$ for $F(Y_q),$ which leads to $[\hat{F}^{-1}(c_1),\ \hat{F}^{-1}(c_2)]$ for $Y_q.$

3. Empirical study based on data from the populations MIC2015_Oax and MIC2015_Oax$_{\text{trunc}}$

The population census considered in this work consisted of 208,101 inhabitants with complete responses in a vector $\mathbf{x}$ of six auxiliary variables: *age*, *educational level*, *employment status*, *gender*, *indigenous language*, and *marital status*. The variable of interest y corresponds to the monthly *income*.

A logistic regression with some two-way interactions was fitted to the 208,101 observations, with response variable $I_k = 1$ if an *income* value was given by individual k, and $I_k = 0$ if it was not, and the vector $\mathbf{x}$ of six explanatory variables. This model was then applied to the population of 161,296 individuals, which corresponds to those with $I_k = 1$. This led to a set of response propensity values $\phi = (\phi_1, \ldots, \phi_{161,296})$.

The set of 161,296 respondents, with response propensities $\phi_1, \ldots, \phi_{161,296}$, is referred to as MIC2015_Oax population. The distribution of *income* in this population is highly asymmetric partly due to the presence of some very large values. When removing the 80 observations with *income* larger than or equal to 50,000, we obtain a truncated population referred to as MIC2015_Oax$_{\text{trunc}}$, which is also used in our experiments.

The step distribution function of *income* has only 913 and 887 jumps in each population respectively, with some large jumps at *income* values that are near the quantiles of interest. In particular, $y =$ 2,571 accounts for 7.3% of the distribution and is very close to the quantile $Y_{0.5}$; $y = 643$ (1.1%) and $y = 857$ (4.2%) are close to $Y_{0.1}$; whereas $y =$ 6,429 (2.8%) and $y =$ 7,000 (1.1%) are close to $Y_{0.9}$.

3.1 Numerical results

For each population, the coverage rate of the CI for each method was estimated as follows:

1. A simple random sample s of size $n \in \{1{,}000; 5{,}000\}$ was selected.
2. For each unit $k \in s$ with response propensity ϕ_k, we generated I_k from a $\text{Bernoulli}(\phi_k)$.
3. Two cases were considered: a) full response and b) average nonresponse rate of 22.5%. For the latter, a logistic regression with two-way interactions, with I as the response and the six explanatory variables, was selected by forward selection using the BIC criterion. The estimated response probabilities $\hat{\phi}_k$ were obtained with the selected model.
4. 90% CIs were computed using linearization (Lin), empirical likelihood (EL) and the Woodruff (W) method for quantiles.
5. Steps 1 to 4 were repeated $M =$ 5,000 times and the coverage rate for each method and for each parameter was calculated as the proportion of CIs that covered the corresponding parameter value.

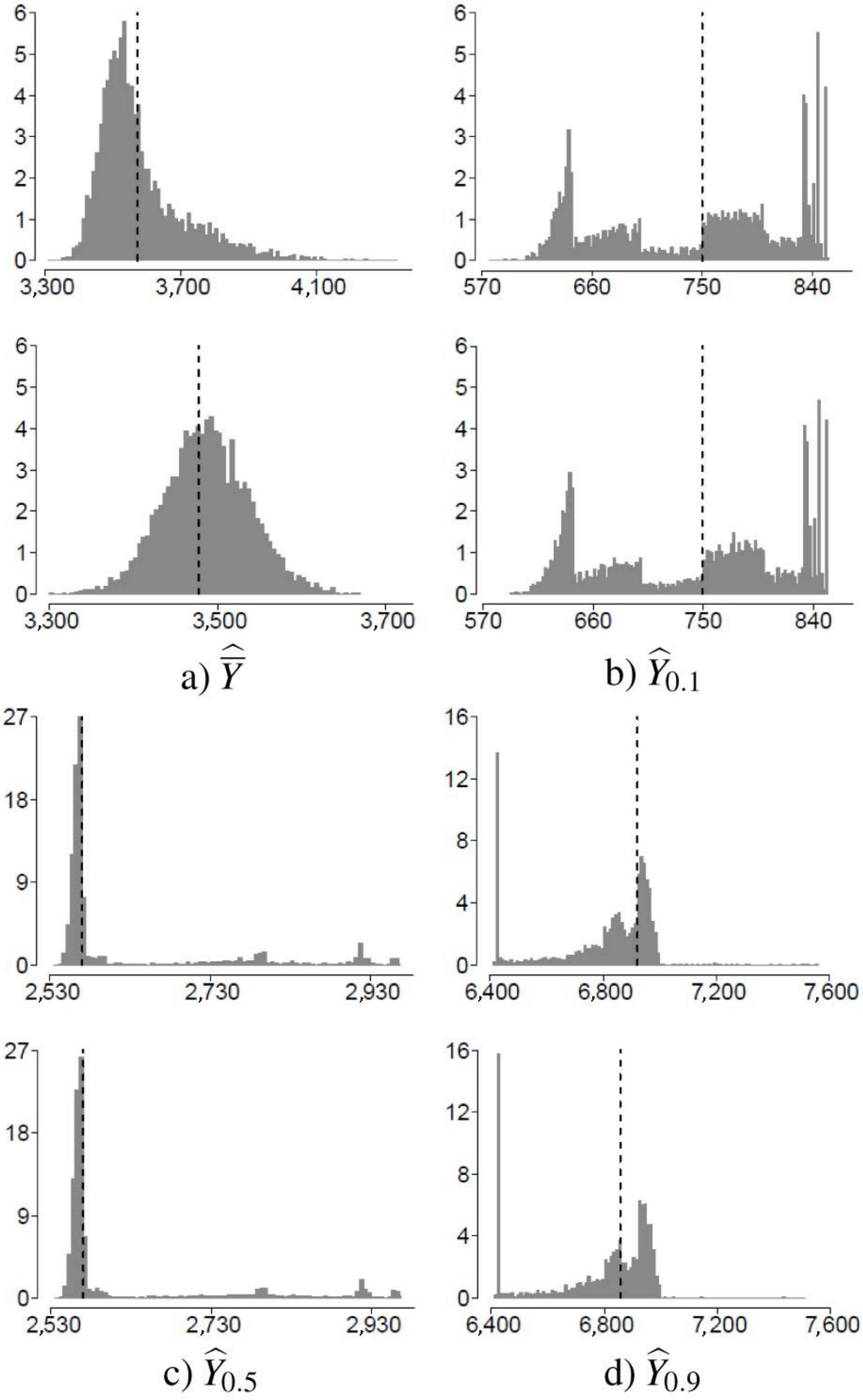
Table 3.1 shows the results for $n = 5{,}000$. Table 3.2 shows the absolute value of the percent relative bias, $\mathrm{RB} = 100\left(\bar{\hat{\theta}} - \theta_0\right) \big/ \theta_0$ with $\bar{\hat{\theta}}_0 = \sum \hat{\theta}_{0i} \big/ M$, and of the percent relative root mean square error, $\mathrm{RRMSE} = 100\left\{ \sum (\hat{\theta}_{0i} - \theta_0)^2 \big/ M \right\}^{1/2} \big/ \theta_0$. Figure 3.1 presents the distribution of the $M = 5{,}000$ estimates for the nonresponse scenario for each parameter; the corresponding distributions for the full response scenario are qualitatively similar. The results for $n = 1{,}000$ are omitted since they were similar to those obtained with $n = 5{,}000$. From Tables 3.1 and 3.2 and Figure 3.1, we make the following remarks:

a) For $\bar{Y}$, Lin and EL methods perform similarly: they have a poor performance (coverage as low as 72.9%) for MIC2015_Oax, and a good one for MIC2015_Oax$_{\text{trunc}}$, reaching the nominal level with similar tail error rates and CI average length. Figure 3.1 a) shows that the distribution of $\hat{\bar{Y}}$ is symmetric for MIC2015_Oax$_{\text{trunc}}$ and highly asymmetric for MIC2015_Oax; this asymmetry seems to be related to the 80 extreme *income* values not present in MIC2015_Oax$_{\text{trunc}}$.

b) For quantiles, Lin method has a poor performance with the shortest CI average length in both scenarios, in spite of the expected overestimation of the variance in the nonresponse scenario. This method relies on the normality of $\hat{Y}_q$, but Figures 3.1 b), c) and d) show that the distribution of $\hat{Y}_q$ is far from being symmetric and unimodal, with modes around *income* values with high frequency. Especially for $Y_{0.1}$, where the coverage rate is as low as 31.4%, the distribution of $\hat{Y}_{0.1}$ is multimodal with a high proportion of values that are farther from $Y_{0.1}$ than half the CI average length. EL and W methods generally perform well, except for $Y_{0.5}$ in the full response scenario and for $Y_{0.9}$ in MIC2015_Oax. The low coverages seem to be related to the observed high frequency of the two *income* values 2,571 (7.3%) and 6,429 (2.8%). The first one is very close to $Y_{0.5} = 2{,}570$ and some of the CIs for $Y_{0.5}$ are too narrow when $\hat{Y}_{0.5} < 2{,}571$. The second one is farther from $Y_{0.9}$ in MIC2015_Oax than in MIC2015_Oax$_{\text{trunc}}$, reducing the proportion of CIs that cover $Y_{0.9}$ when $\hat{Y}_{0.9} \approx 6{,}429$, see Figure 3.1 d), where $Y_{0.9} = 6{,}921$ in MIC2015_Oax and $Y_{0.9} = 6{,}856$ in MIC2015_Oax$_{\text{trunc}}$.

c) Table 3.2 shows that the RB is small, less than 3.3%, for all parameters. When only a simple adjustment with the percentage of nonresponse is applied (not shown in this note), the RB is larger and all the methods have a very poor performance. These results suggest that the use of a propensity model helps to obtain a RB comparable with that of the full response case. For $Y_{0.1}$, the empirical double expansion estimator is even less biased than the one associated with the full response scenario; however their RRMSE are comparable and the largest among those for the parameters of interest, since the distribution of the estimators is multimodal in both scenarios, see Figure 3.1 b).

Figure 3.1 **Distribution of the $M = 5{,}000$ estimates of $\bar{Y}$ in a), and of $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$ in b) to d) for the case with an average nonresponse of 22.5% and $n = 5{,}000$. The upper panel corresponds to MIC2015_Oax and the lower to MIC2015_Oax$_{\text{trunc}}$. The dotted lines indicate the population values $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$.**



a) $\widehat{\bar{Y}}$

b) $\widehat{Y}_{0.1}$

c) $\widehat{Y}_{0.5}$

d) $\widehat{Y}_{0.9}$

Table 3.1
Coverages of 90% CIs for the parameters $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, for y = *income*. Average nonresponse of 22.5% (*NR*) and Full response (*Full*)

Parameter θ_0	**Method**	**Coverage %**		**Lower tail err. rates %**		**Upper tail err. rates %**		**CI average length**	
		NR	***Full***	***NR***	***Full***	***NR***	***Full***	***NR***	***Full***
MIC2015_Oax									
$\bar{Y}$	EL	79.1*	72.9*	4.5	3.8*	16.4*	23.3*	370.8	335.4
	Lin	80.6*	73.8*	0.3*	0.1*	19.1*	26.1*	345.3	309.9
$Y_{0.1}$	EL	91.1*	90.5	6.0*	4.1*	2.9*	5.4	192.8	179.5
	W	90.6	89.8	6.2*	4.2*	3.2*	6.0*	191.2	177.1
	Lin	37.9*	35.1*	28.9*	19.2*	33.3*	45.7*	114.4	89.4
$Y_{0.5}$	EL	88.2*	82.3*	1.6*	0.7*	10.2*	17.1*	274.6	225.8
	W	88.0*	81.0*	1.7*	0.7*	10.3*	18.3*	274.4	224.1
	Lin	79.0*	88.0*	21.0*	12.0*	0.0*	0.0*	152.2	127.3
$Y_{0.9}$	EL	84.0*	83.0*	2.6*	2.7*	13.3*	14.3*	527.2	470.2
	W	86.1*	84.3*	2.5*	2.7*	11.4*	13.0*	533.8	474.2
	Lin	72.3*	73.5*	0.4*	0.1*	27.3*	26.4*	392.8	346.6
MIC2015_Oax$_{\text{trunc}}$									
$\bar{Y}$	EL	90.5	90.6	6.4*	4.4	3.0*	5.0	173.7	147.5
	Lin	90.8*	90.1	5.7*	4.2*	3.5*	5.6*	171.2	145.0
$Y_{0.1}$	EL	89.8	89.9	6.8*	4.0*	3.4*	6.1*	191.7	178.7
	W	89.1*	89.1*	7.0*	4.1*	4.0*	6.8*	190.0	176.2
	Lin	35.5*	31.4*	29.4*	21.0*	35.1*	47.6*	104.9	81.4
$Y_{0.5}$	EL	87.2*	80.4*	1.6*	0.9*	11.2*	18.7*	267.8	218.5
	W	87.1*	79.3*	1.6*	0.9*	11.3*	19.8*	267.7	216.7
	Lin	80.0*	87.4*	20.0*	12.6*	0.0*	0.0*	144.9	121.0
$Y_{0.9}$	EL	90.3	90.1	4.4	4.4*	5.3	5.5	521.3	470.8
	W	92.3*	91.9*	4.3*	4.3*	3.5*	3.8*	528.2	475.3
	Lin	75.6*	77.0*	0.1*	0.1*	24.3*	23.0*	411.7	365.5

$*$ Coverages and tail error rates significantly different from 90% and 5% respectively (Feller, 1968, page 182). p-value < 5%

MIC2015_Oax: $N = 161{,}296$, $\rho = 0.08$, $\gamma = 89.9$; MIC2015_Oax$_{\text{trunc}}$: $N = 161{,}216$, $\rho = 0.21$, $\gamma = 3.48$, where $\rho = \operatorname{corr}(y, \phi)$ and $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Table 3.2
Percent relative bias (RB) and percent relative root mean squared error (RRMSE) of estimators of $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, based on 5,000 samples. Average nonresponse of 22.5% (*NR*) and Full response (*Full*)

Population	$\bar{Y}$				$Y_{0.1}$				$Y_{0.5}$				$Y_{0.9}$			
	|RB|		**|RRMSE|**		**|RB|**		**|RRMSE|**		**|RB|**		**|RRMSE|**		**|RB|**		**|RRMSE|**	
	NR	***Full***	***NR***	***Full***	***NR***	***Full***	***NR***	***Full***	***NR***	***Full***	***NR***	***Full***	***NR***	***Full***	***NR***	***Full***
MIC2015_Oax	0.30	0.01	3.8	3.2	0.58	3.13	10.4	9.9	1.96	0.93	4.8	3.4	1.79	1.68	3.3	3.0
MIC2015_Oax$_{\text{trunc}}$	0.27	0.02	1.5	1.3	0.71	3.23	10.5	10.0	1.87	0.96	4.8	3.5	1.05	0.95	3.0	2.8

4. Simulated populations

In order to control the asymmetry of the distribution of y, the correlation $\operatorname{corr}(y, \phi)$ and the percentage of nonresponse, we simulated two symmetric and two asymmetric populations of size N = 50,000 with a variable of interest y and six auxiliary variables $x_1, \ldots, x_6$, as follows.

1. 50,000 simple random samples for each of the variables $x_1, \ldots, x_6$ were generated independently from a $N(0, 1)$ distribution.
2. The response probabilities ϕ_k were obtained using a logistic regression with $\beta_1 = \cdots = \beta_6 = -0.3$ and β_0 chosen so that the average nonresponse was equal to 24.8%.
3. Two settings were considered for the distribution of y:

   i) Symmetric. y_k was generated from a $N(\mu, \sigma^2)$, with $\mu = 1 + 2.16 \operatorname{corr}(y, x) \sum_{j=1}^{6} x_{kj}$ and $\sigma^2 = 4.67 * (1 - 6 \operatorname{corr}^2 (y, x))$, where $\operatorname{corr}(y, x) = \operatorname{corr}(y, x_j)$, $j \in \{1, \ldots, 6\}$.

   ii) Asymmetric. $y_k = \exp(z_k)$, where z_k was generated from a $N(\mu, \sigma^2)$, with $\mu = \operatorname{corr}(z, x) \sum_{j=1}^{6} x_{kj}$ and $\sigma^2 = 1 - 6 \operatorname{corr}^2 (z, x)$, where $\operatorname{corr}(z, x) = \operatorname{corr}(z, x_j)$, $j \in \{1, \ldots, 6\}$.

   Both $\operatorname{corr}(y, x)$ and $\operatorname{corr}(z, x)$ were chosen so that $\operatorname{corr}(y, \phi)$ was approximately equal to -0.2 or -0.8.

4.1 Numerical results

The coverage rate of the CIs was computed as in Section 3.1, with $n = 500$ and using a logistic regression without interactions to obtain $\hat{\phi}_k$, $k \in s$.

Table 4.1 reports the results for the populations with $\operatorname{corr}(y, \phi) = -0.8$. Table 4.2 shows the $|\mathrm{RB}|$ and $|\mathrm{RRMSE}|$ of the estimators. Numerical results for populations with $\operatorname{corr}(y, \phi) = -0.2$ are omitted since they were similar to those obtained for populations with $\operatorname{corr}(y, \phi) = -0.8$. We make the following observations:

a) EL and Lin methods have a similar reasonable performance for $\bar{Y}$, though the upper tail error rates are larger than 5% in the asymmetric population.

b) For quantiles, Lin method has the lowest and the highest coverage of 85.7 and 97.9% respectively. Unlike the distribution of $\hat{Y}_q$ shown in Figure 3.1, the distribution of $\hat{Y}_q$ for the simulated populations is symmetric and unimodal in all cases. EL and W methods perform well, reaching the nominal level in all cases with comparable tail error rates and CI average length.

c) Weighting adjustment in the nonresponse scenario helps to get a RB similar to that of the full response.

d) The coverage rate of the CI for each method is larger in the nonresponse scenario than in the full response, and in some cases it is also larger than the nominal level. This might be related to the impact of having treated the $\hat{\phi}_k$ as fixed in the nonresponse case.

Table 4.1
Simulated populations. Coverages of 90% CIs for $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, for y with $\mathrm{corr}(y, \phi) = -0.8$. Average nonresponse of 24.8% (*NR*) and Full response (*Full*)

Parameter θ_0	Method	Coverage %		Lower tail err. rates %		Upper tail err. rates %		CI average length	
		NR	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*
Asymmetric									
$\bar{Y}$	EL	90.9*	88.6*	2.2*	5.1	6.9*	6.3	0.50	0.32
	Lin	90.1	88.6*	0.5*	3.2*	9.4*	8.2*	0.48	0.31
$Y_{0.1}$	EL	90.6	89.5	4.0*	4.6	5.4	5.9	0.07	0.07
	W	90.3	89.5	3.6*	4.1*	6.2*	6.4*	0.07	0.07
	Lin	97.9*	97.4*	1.2*	1.5*	1.0*	1.2*	0.10	0.10
$Y_{0.5}$	EL	93.6*	90.5	3.0*	4.4*	3.4*	5.1	0.22	0.19
	W	93.5*	90.4	2.9*	4.4	3.6*	5.1	0.22	0.19
	Lin	92.5*	88.9*	2.9*	5.0	4.6	6.2*	0.21	0.18
$Y_{0.9}$	EL	92.7*	90.3	2.6*	4.1*	4.7	5.7*	1.35	0.91
	W	93.0*	90.4	2.7*	4.6	4.3*	5.0	1.36	0.92
	Lin	87.4*	85.7*	2.6*	4.1*	10.1*	10.2*	1.23	0.87
Symmetric									
$\bar{Y}$	EL	93.6*	90.2	3.3*	5.0	3.1*	4.8	0.40	0.32
	Lin	93.5*	89.9	3.1*	5.2	3.4*	4.9	0.39	0.32
$Y_{0.1}$	EL	91.2*	90.9*	3.6*	3.9*	5.2	5.2	0.59	0.55
	W	91.0*	90.6	3.2*	3.6*	5.8*	5.8*	0.59	0.56
	Lin	88.6*	88.2*	5.8*	5.9*	5.7*	6.0*	0.57	0.53
$Y_{0.5}$	EL	92.8*	90.1	3.3*	4.6	3.9*	5.3	0.46	0.39
	W	92.9*	90.1	3.1*	4.6	4.0*	5.3	0.46	0.39
	Lin	92.4*	90.2	3.4*	4.7	4.2*	5.2	0.46	0.39
$Y_{0.9}$	EL	92.9*	90.4	2.2*	3.6*	4.9	6.0*	0.76	0.54
	W	93.3*	90.3	2.2*	4.2*	4.5	5.4	0.77	0.54
	Lin	90.3	89.1*	2.1*	3.2*	7.6*	7.7*	0.74	0.53

* Coverages and tail error rates significantly different from 90% and 5% respectively (Feller, 1968, page 182). p-value < 5%
Symmetric: $\gamma = 0.02$; Asymmetric: $\gamma = 6.2$; where $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Table 4.2
Percent relative bias (RB) and percent relative root mean squared error (RRMSE) of estimators of $\bar{Y}$, $Y_{0.1}$, $Y_{0.5}$ and $Y_{0.9}$, based on 5,000 samples. Average nonresponse of 24.8% (*NR*) and Full response (*Full*)

Population	$\bar{Y}$				$Y_{0.1}$				$Y_{0.5}$				$Y_{0.9}$			
	|RB|		|RRMSE|		|RB|		|RRMSE|		|RB|		|RRMSE|		|RB|		|RRMSE|	
	NR	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*	*NR*	*Full*
Asymmetric	0.11	0.13	9.0	5.9	0.56	0.53	7.8	7.5	0.00	0.07	6.0	5.7	0.53	0.43	9.7	7.6
Symmetric	0.12	0.09	10.3	9.5	0.98	0.91	10.2	9.7	0.33	0.35	12.7	11.8	0.43	0.34	5.5	4.3

5. Conclusions

Considering the distribution of y (*income*), it was observed that a poor performance of a method in the full response scenario generally corresponded to a poor one in the nonresponse scenario, although in several cases the coverage rate was larger in the latter. This suggests that having treated the $\hat{\phi}_k$ as fixed had little effect on the performance of the methods as compared to the impact of the characteristics of the distribution of y. Extreme values were related to a low coverage of the CIs for the mean for both empirical likelihood and linearization methods. The presence of values with high frequency near a quantile of interest also had an impact on the coverage of its CIs; this might be related to the behavior of the step distribution function, where the jumps in $F(y)$ and $\hat{F}(y)$ are usually required to be small in order to obtain a good performance of the Woodruff method (Lohr, 2010, page 390). In general, the linearization method had a poor performance for quantiles, while the performance of empirical likelihood and of Woodruff were similar and better; this behavior has also been observed in Berger and De La Riva Torres (2016). While Woodruff method is simple and easy to implement, an advantage of the empirical likelihood method is that it can be used for parameters other than quantiles.

Acknowledgements

Omar De La Riva Torres was supported by the Post Doctoral Scholarship Program of the National Autonomous University of Mexico.

References

Berger, Y.G. (2020). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics*, 72, 91-121.

Berger, Y.G., and De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 2, 319-341.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-eng.pdf.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications: Volume I*. New York: John Wiley & Sons, Inc.

Graf, E., and Tillé, Y. (2014). Variance estimation using linearization for poverty and social exclusion indicators. *Survey Methodology*, 40, 1, 61-79. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14000-eng.pdf.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 2, 206-226.

INEGI (2015). *Encuesta Intercensal 2015*. Instituto Nacional de Estadística, Geografía e Informática. https://www.inegi.org.mx/programas/intercensal/2015/.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 4, 501-514.

Lohr, S.L. (2010). *Sampling: Design and Analysis.* Brooks/Cole.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 3, 167-195.

Särndal, C.-E., Swenson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.