Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Évaluer la couverture des intervalles de confiance en cas de non-réponse. Étude de cas sur la moyenne et les quantiles de revenu dans certaines municipalités de l’Enquête intercensitaire mexicaine de 2015

par Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz
et Guillermina Eslava-Gómez

Date de diffusion : le 6 janvier 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 401-412
Statistique Canada, nº 12-001-X au catalogue

Évaluer la couverture des intervalles de confiance en cas de non-réponse. Étude de cas sur la moyenne et les quantiles de revenu dans certaines municipalités de l'Enquête intercensitaire mexicaine de 2015

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

Résumé

L'article présente une étude comparative de trois méthodes de construction d'intervalles de confiance pour la moyenne et les quantiles à partir de données d'enquête en présence de non-réponse. Ces méthodes, à savoir la vraisemblance empirique, la linéarisation et la méthode de Woodruff (1952), ont été appliquées à des données sur le revenu tirées de l'Enquête intercensitaire mexicaine de 2015 et à des données simulées. Un modèle de propension à répondre a servi à ajuster les poids d'échantillonnage, et les performances empiriques des méthodes ont été évaluées en fonction de la couverture des intervalles de confiance au moyen d'études par simulations. Les méthodes de vraisemblance empirique et de linéarisation ont donné de bonnes performances pour la moyenne, sauf quand la variable d'intérêt avait des valeurs extrêmes. Pour les quantiles, la méthode de linéarisation s'est montrée peu performante; les méthodes de vraisemblance empirique et de Woodruff ont donné de meilleurs résultats, mais sans atteindre la couverture nominale quand la variable d'intérêt avait des valeurs à haute fréquence proches du quantile d'intérêt.

Mots-clés : Estimation de l'intervalle de confiance; vraisemblance empirique; linéarisation; données manquantes au hasard; non-réponse; échantillonnage à deux phases.

1. Introduction

L'Enquête intercensitaire mexicaine de 2015 (MIC2015) réalisée par l'Institut national de statistique, de géographie et d'informatique (*Instituto Nacional de Estadística, Geografía e Informática*, INEGI, 2015) a recueilli des renseignements à l'échelle nationale au moyen d'un plan de sondage probabiliste dans 1 643 municipalités et d'un recensement dans 814 municipalités. Dans la présente étude, nous utilisons les données du recensement correspondant à 441 municipalités de l'État de Oaxaca.

Nous nous concentrons sur la variable d'intérêt *revenu* qui présente un taux de non-réponse d'environ 22,5 %. Si l'on considère les répondants, la distribution du *revenu* présente une forte asymétrie, principalement en raison de la présence de valeurs extrêmes, et certaines valeurs ont une haute fréquence.

L'objectif de l'étude est d'évaluer le taux de couverture empirique des intervalles de confiance (IC) calculés par trois méthodes pour la moyenne de population et les quantiles de population, 0,1; 0,5 et 0,9, dans des données d'enquête en présence de non-réponse. L'échantillonnage à deux phases est utilisé avec un échantillon aléatoire sélectionné à la première phase, tandis que dans la seconde, l'échantillon est divisé en répondants et en non-répondants, compte tenu du profil de non-réponse de *revenu* dans les données du recensement. Un modèle de propension à répondre sert à ajuster les poids de non-réponse.

1. Omar De La Riva Torres, Département de mathématiques, Faculté des sciences, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. Courriel : odelariva@ciencias.unam.mx; Gonzalo Pérez-de-la-Cruz, Département de mathématiques, Faculté des sciences, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. Courriel : gonzalo.perez@ciencias.unam.mx; Guillermina Eslava-Gómez, Département de mathématiques, Faculté des sciences, UNAM, Av. Universidad 3000, Circuito Exterior S/N, México. Courriel : eslava@ciencias.unam.mx.

Pour la moyenne de la population, nous considérons l'estimateur de Hájek et deux méthodes de calcul des IC, soit la vraisemblance empirique (Berger, 2020) et la linéarisation (Särndal, Swenson et Wretman, 1992, sections 5.2 et 5.7). En ce qui concerne les quantiles de la population, nous tenons compte de l'estimateur ponctuel obtenu par interpolation de la fonction de répartition comme dans Woodruff (1952) et Graf et Tillé (2014), et de trois méthodes de calcul des IC : la probabilité empirique (Berger, 2020), la méthode de Woodruff (Woodruff, 1952) et la linéarisation (Deville, 1999). Ces méthodes sont décrites à la section 2; les résultats numériques sont présentés à la section 3 pour les données de la MIC2015 et à la section 4 pour certaines populations simulées. Des conclusions sont données à la section 5.

2. Trois méthodes d'estimation des intervalles de confiance

2.1 Estimation par échantillonnage à deux phases

Nous considérons une population finie $U = \{1, 2, \ldots, N\}$ et un échantillon probabiliste $s \subset U$ de taille fixe n, avec des probabilités d'inclusion du premier et du second ordre π_k et π_{kl}, $k, l \in U$, $k \neq l$. Soit y_k la k^{e} valeur de la variable d'intérêt y, et soit θ_0 un paramètre de population et $\hat{\theta}_0$ un estimateur de θ_0.

Nous supposons que la valeur y_k est disponible pour un sous-ensemble $r \subset s$ seulement. Soit ϕ_k la probabilité de réponse de l'unité k. Soit I_k une variable indicatrice de réponse telle que $I_k = 1$ pour $k \in r$ et $I_k = 0$ pour $k \in s \setminus r$. Nous supposons également qu'il y a un vecteur des variables auxiliaires $\mathbf{x}$ observées pour tous les $k \in s$. Nous formulons l'hypothèse de données manquantes au hasard (MAR pour *missing at random*) :

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

Les probabilités de réponse ϕ_k servent à ajuster les poids de sondage. Nous supposons que le plan de sondage et le mécanisme de réponse sont indépendants, comme dans Berger (2020). À partir de la théorie de l'échantillonnage à deux phases (Särndal et coll., 1992, section 9.3), les poids ajustés pour la non-réponse sont définis comme étant $\pi_k^{*-1} = 1/(\pi_k \phi_k)$ et les probabilités d'inclusion de second ordre comme étant $\pi_{kl}^{*} = \pi_{kl}\, \phi_k\, \phi_l, k, l \in U,\ k \neq l.$

L'intérêt réside dans l'estimation de la moyenne de population $\bar{Y} = \sum_{k \in U} y_k / N$ et du quantile de population donné par

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

où $y_{(i)}$ est la valeur de la i^{e} unité arrangée en ordre croissant, $d = \min\left\{l : qN < N_{(l)}, l = 1, \ldots, N\right\}$ et $N_{(l)} = \sum_{j \in U} I\left(y_j \leq y_{(l)}\right), l = 1, \ldots, N.$ On obtient la formule (2.1) en prenant en compte une interpolation linéaire par morceaux de la fonction de répartition par étapes $F(y) = \sum_{k \in U} I(y_k \leq y)/N,$ où $I\left(y_k \leq y\right) = 1$ quand $y_k \leq y.$

Ces paramètres de population sont estimés respectivement par

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

et

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

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

Ces estimateurs et les IC décrits dans la sous-section suivante sont fondés sur l'hypothèse que les probabilités de réponse ϕ_k sont connues, contrairement à ce que l'on a dans Berger (2020) et Kim et Kim (2007). Cependant, nous utilisons $\hat{\phi}_k$ au lieu de ϕ_k dans les études par simulations, où

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

avec $\hat{\beta}$ qu'on obtient en ajustant une régression logistique au moyen de *s*. Cela donne des estimateurs connus sous le nom d'estimateurs empiriques par double dilatation (Haziza et Beaumont, 2017).

2.2 Méthodes d'estimation des intervalles de confiance

2.2.1 Linéarisation

La méthode de linéarisation se fonde sur l'hypothèse selon laquelle la distribution de $\hat{\theta}_0$ est approximativement normale. Un IC pour θ_0 est

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

où $1-\alpha$ est le niveau de confiance, aussi appelé couverture nominale; voir Särndal et coll. (1992, expression 5.2.3). En pratique, $V(\hat{\theta}_0)$ est estimé. Pour les estimateurs donnés par (2.2) et (2.3), un estimateur de la variance est donné par

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

où $\hat{z}_k = \left(y_k - \hat{\bar{Y}}\right)/\hat{N}$ pour $\hat{\bar{Y}}$ (Särndal et coll., 1992, résultat 5.7.1) et $\hat{z}_k = -\left(I(y_k \le \hat{Y}_q) - q\right)/\left(f(\hat{Y}_q)\hat{N}\right)$ pour $\hat{Y}_q$ (Deville, 1999). La *fonction de densité* f a été obtenue de deux façons, a) au moyen d'un noyau gaussien comme dans Osier (2009) et b) au moyen de la technique du plus proche voisin comme dans Graf et Tillé (2014). Nous présentons les résultats relatifs à a), étant donné que la technique en b) a donné des résultats similaires.

Nous remarquons que l'utilisation de (2.5) avec $\hat{\phi}_k$ au lieu de ϕ_k peut entraîner une surestimation de la variance de l'estimateur empirique par double dilatation et des IC plus grands; voir l'expression (17) dans Kim et Kim (2007) associée aux estimateurs de $\bar{Y}$.

2.2.2 Méthode de la vraisemblance empirique

La méthode de la vraisemblance empirique suppose que θ_0 est la seule solution de l'équation d'estimation $G(\theta) = \sum_{k \in U} g_k(\theta) = 0$ pour une fonction donnée g_k. En particulier, nous utilisons :

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

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

La fonction de log-vraisemblance empirique dans Berger et De La Riva Torres (2016) pour un plan de sondage à un degré sans stratification ni information auxiliaire est

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

où $\{m_k : k \in s\}$ satisfait aux contraintes de plan et de paramètre $\sum_{k \in s} m_k \pi_k = n$ et $\sum_{k \in s} m_k g_k(\theta) = 0$.

En présence de non-réponse, nous utilisons (2.6) en remplaçant $\sum_{k \in s} m_k g_k(\theta) = 0$ par $\sum_{k \in s} m_k I_k g_k(\theta) / \phi_k = 0$. Un IC pour θ_0 est donné par

$$\left[\min\{\theta : \hat{R}(\theta) \le \chi^2_1(\alpha)\}, \ \max\{\theta : \hat{R}(\theta) \le \chi^2_1(\alpha)\} \right], \tag{2.7}$$

où $\hat{R}(\theta) = 2\{\ell_{\max}(\hat{\theta}_0) - \ell_{\max}(\theta)\}$ et $\chi^2_1(\alpha)$ est le $(1-\alpha)$-quantile de la distribution χ^2_1. L'estimateur $\hat{\theta}_0 := \text{argmax}_{\{\theta\}} \ell_{\max}(\theta)$ correspond respectivement à (2.2) et (2.3).

Nous avons calculé (2.7) au moyen d'une méthode de recherche de racine, en calculant $\hat{R}(\theta)$ pour plusieurs valeurs de θ, où l'on obtient $\ell_{\max}(\theta)$ pour une valeur donnée θ par un algorithme de Newton-Raphson modifié comme dans Wu (2004).

2.2.3 Méthode de Woodruff pour les quantiles

La méthode de Woodruff (1952) est fondée sur la fonction de répartition estimée $\hat{F}(y)$. Pour un quantile Y_q, la variance de $\hat{F}(Y_q)$ peut être calculée approximativement au moyen de la méthode de linéarisation en séries de Taylor avec une variable linéarisée $z_k = \left(I(y_k \le Y_q) - q\right) / N$, tandis que la variance est estimée au moyen de (2.5) avec $\hat{z}_k = \left(I(y_k \le \hat{Y}_q) - q\right) / \hat{N}$. Si l'on suppose la normalité de $\hat{F}(Y_q)$ et que l'on utilise (2.4), il est possible de trouver un IC $[c_1, c_2]$ pour $F(Y_q)$, ce qui donne $\left[\hat{F}^{-1}(c_1), \ \hat{F}^{-1}(c_2)\right]$ pour Y_q.

3. Étude empirique fondée sur les données des populations MIC2015_Oax et MIC2015_Oax$_{trunc}$

Le recensement de la population examiné dans les travaux présentés ici comprenait 208 101 habitants avec des réponses complètes dans un vecteur $\mathbf{x}$ de six variables auxiliaires, soit l'*âge*, le *niveau de scolarité*, la *situation d'emploi*, le *genre*, la *langue autochtone* et l'*état matrimonial*. La variable d'intérêt y correspond au *revenu* mensuel.

Une régression logistique avec des interactions binaires a été ajustée aux 208 101 observations, avec une variable de réponse $I_k = 1$ si une valeur de *revenu* était donnée par la personne k, et $I_k = 0$ si elle ne l'était pas, et le vecteur $\mathbf{x}$ de six variables explicatives. Ce modèle a ensuite été appliqué à la population de 161 296 personnes, ce qui correspond à celles avec $I_k = 1$. Cela a donné un ensemble de valeurs de propension à répondre $\phi = (\phi_1, \ldots, \phi_{161\,296})$.

L'ensemble de 161 296 répondants, avec des propensions à répondre $\phi_1, \ldots, \phi_{161\,296}$, est appelé population MIC2015_Oax. La distribution du *revenu* dans cette population est très asymétrique, en partie en raison de la présence de valeurs très grandes. Si nous supprimons les 80 observations dont le *revenu* est supérieur ou égal à 50 000, nous obtenons une population tronquée appelée MIC2015_Oax$_{trunc}$, qui est également utilisée dans nos expériences.

La fonction de répartition par étapes du *revenu* n'a que 913 et 887 sauts respectivement dans chaque population, avec quelques sauts importants aux valeurs de *revenu* qui sont proches des quantiles d'intérêt. En particulier, $y =$ 2 571 représente 7,3 % de la distribution et est très proche du quantile $Y_{0,5}$; $y = 643$ (1,1 %) et $y = 857$ (4,2 %) sont proches de $Y_{0,1}$; tandis que $y =$ 6 429 (2,8 %) et $y =$ 7 000 (1,1 %) sont proches de $Y_{0,9}$.

3.1 Résultats numériques

Pour chaque population, le taux de couverture de l'IC de chaque méthode a été estimé comme suit :

1. Un échantillon aléatoire simple s de taille $n \in \{1\,000; 5\,000\}$ a été sélectionné.
2. Pour chaque unité $k \in s$ avec une propension à répondre ϕ_k, nous avons généré I_k à partir de Bernoulli(ϕ_k).
3. Deux cas ont été examinés : a) réponse complète et b) taux moyen de non-réponse de 22,5 %. Dans ce dernier cas, une régression logistique avec interactions binaires, avec I comme réponse et les six variables explicatives, a été sélectionnée par sélection ascendante au moyen du critère BIC. Les probabilités de réponse estimées $\hat{\phi}_k$ ont été obtenues avec le modèle sélectionné.
4. Les IC à 90 % ont été calculés au moyen de la linéarisation (Lin), de la vraisemblance empirique (VE) et de la méthode de Woodruff (W) pour les quantiles.
5. Les étapes 1 à 4 ont été répétées $M =$ 5 000 fois et le taux de couverture de chaque méthode et de chaque paramètre a été calculé comme étant la proportion des IC qui couvraient la valeur correspondante du paramètre.

Le tableau 3.1 montre les résultats pour n = 5 000. Le tableau 3.2 montre la valeur absolue du biais relatif en pourcentage, $\text{BR} = 100\left(\bar{\hat{\theta}} - \theta_0\right)/\theta_0$ avec $\bar{\hat{\theta}}_0 = \sum \hat{\theta}_{0i}/M$, et de la racine de l'erreur quadratique moyenne relative en pourcentage, $\text{REQMR} = 100\left\{\sum (\hat{\theta}_{0i} - \theta_0)^2/M\right\}^{1/2}/\theta_0$. La figure 3.1 présente la distribution des M = 5 000 estimations pour le scénario de non-réponse pour chaque paramètre; les distributions correspondantes pour le scénario de réponse complète sont qualitativement semblables. Les résultats pour n = 1 000 sont omis, car ils étaient semblables à ceux obtenus avec n = 5 000. À partir des tableaux 3.1 et 3.2 et de la figure 3.1, nous formulons les remarques suivantes :

a) Pour $\bar{Y}$, les méthodes Lin et VE ont des performances similaires : elles sont peu performantes (couverture aussi faible que 72,9 %) pour MIC2015_Oax, et donnent de bonnes performances pour MIC2015_Oax$_{\text{trunc}}$, atteignant le niveau nominal avec des taux d'erreur de queue et une longueur moyenne d'IC similaires. La figure 3.1 a) montre que la distribution de $\bar{\hat{Y}}$ est symétrique pour MIC2015_Oax$_{\text{trunc}}$ et hautement asymétrique pour MIC2015_Oax; cette asymétrie semble être liée aux 80 valeurs de *revenu* extrêmes non présentes dans MIC2015_Oax$_{\text{trunc}}$.

b) Pour les quantiles, la méthode Lin est peu performante avec la durée moyenne de l'IC la plus courte dans les deux scénarios, malgré la surestimation attendue de la variance dans le scénario de non-réponse. Cette méthode repose sur la normalité de $\hat{Y}_q$, mais les figures 3.1 b), c) et d) montrent que la distribution de $\hat{Y}_q$ est loin d'être symétrique et unimodale, avec des modes autour des valeurs de *revenu* à haute fréquence. En particulier pour $Y_{0,1}$, où le taux de couverture est aussi bas que 31,4 %, la distribution de $\hat{Y}_{0,1}$ est multimodale avec une proportion élevée de valeurs qui sont plus éloignées de $Y_{0,1}$ que la moitié de la longueur moyenne de l'IC. Les méthodes VE et W fonctionnent généralement bien, sauf pour $Y_{0,5}$ dans le scénario de réponse complète et pour $Y_{0,9}$ dans MIC2015_Oax. Les faibles couvertures semblent liées à la haute fréquence observée des deux valeurs de *revenu* 2 571 (7,3 %) et 6 429 (2,8 %). La première est très proche de $Y_{0,5}$ = 2 570 et certains IC de $Y_{0,5}$ sont trop étroits quand $\hat{Y}_{0,5}$ < 2 571. La seconde est plus éloignée de $Y_{0,9}$ dans MIC2015_Oax que dans MIC2015_Oax$_{\text{trunc}}$, réduisant la proportion d'IC couvrant $Y_{0,9}$ quand $\hat{Y}_{0,9} \approx$ 6 429, voir la figure 3.1 d), où $Y_{0,9}$ = 6 921 dans MIC2015_Oax et $Y_{0,9}$ = 6 856 dans MIC2015_Oax$_{\text{trunc}}$.

c) Le tableau 3.2 montre que le biais relatif (BR) est petit, inférieur à 3,3 %, pour tous les paramètres. Quand on applique un simple ajustement avec le pourcentage de non-réponse (non présenté ici), le BR est plus grand et toutes les méthodes ont de mauvaises performances. Ces résultats suggèrent que l'utilisation d'un modèle de propension aide à obtenir un BR comparable à celui du scénario de réponse complète. Pour $Y_{0,1}$, l'estimateur empirique par double dilatation est encore moins biaisé que celui associé au scénario de réponse complète; cependant, leur racine carrée de l'EQM relative (REQMR) est comparable et la plus grande parmi celles des paramètres d'intérêt, puisque la distribution des estimateurs est multimodale dans les deux scénarios, voir la figure 3.1 b).

Figure 3.1 Distribution des M = 5 000 estimations de $\bar{Y}$ dans a), et de $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$ dans b) à d) pour le scénario avec une non-réponse moyenne de 22,5 % et n = 5 000. Le graphique supérieur correspond à MIC2015_Oax et le graphique inférieur à MIC2015_Oax$_{trunc}$. Les pointillés indiquent les valeurs de population $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$.

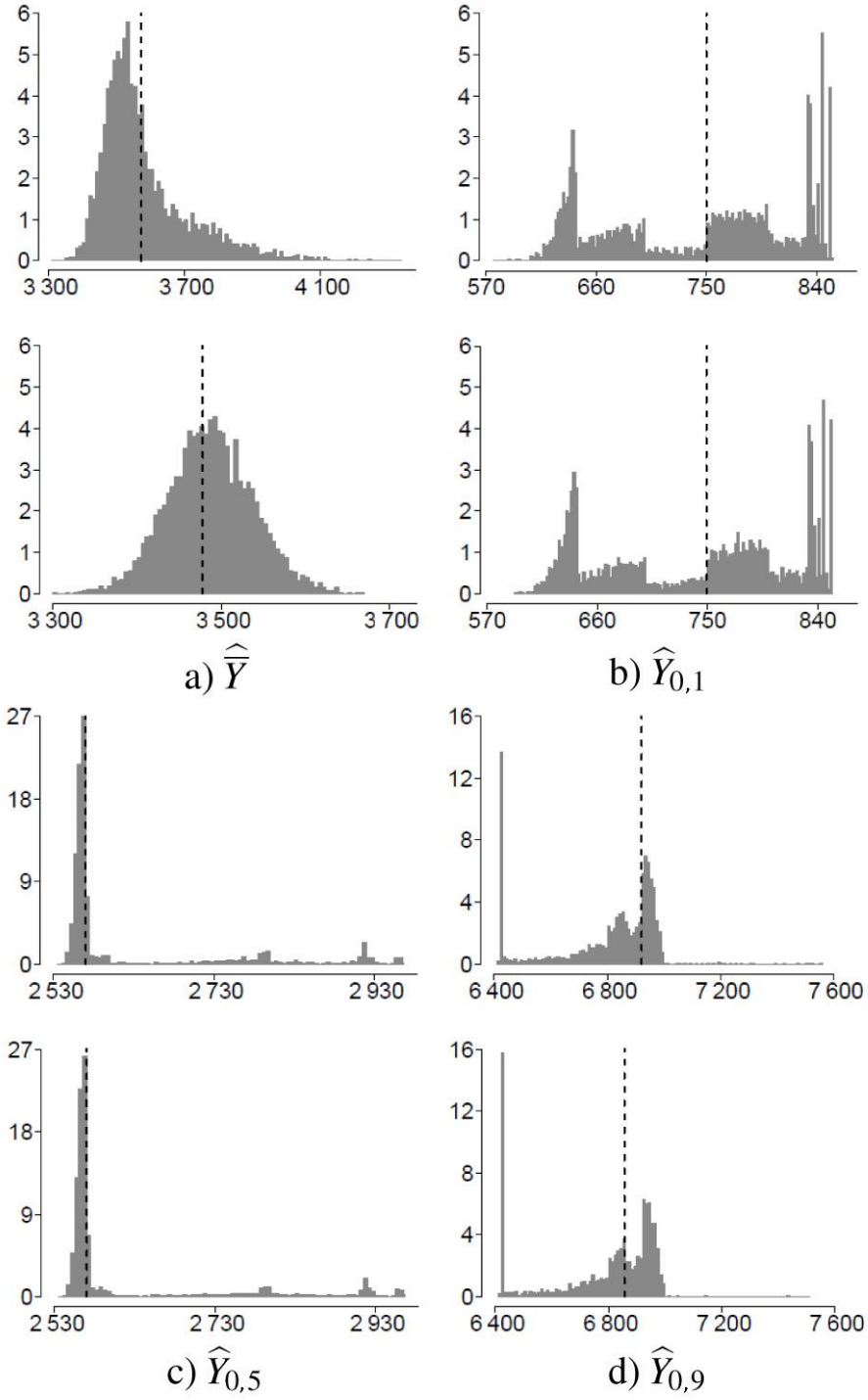

Tableau 3.1
Couvertures des IC à 90 % pour les paramètres $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, pour y = *revenu*. Non-réponse moyenne de 22,5 % (*NR*) et réponse complète (*RC*)

Paramètre θ_0	Méthode	Couverture %		Taux d'err. de queue inférieurs en %		Taux d'err. de queue supérieurs en %		Longueur moyenne des IC	
		NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
MIC2015_Oax									
$\bar{Y}$	VE	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
$Y_{0,1}$	VE	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
$Y_{0,5}$	VE	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
$Y_{0,9}$	VE	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
$\text{MIC2015\_Oax}_{\text{trunc}}$									
$\bar{Y}$	VE	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
$Y_{0,1}$	VE	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
$Y_{0,5}$	VE	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
$Y_{0,9}$	VE	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

* Couvertures et taux d'erreur de queue différant considérablement de 90 % et 5 % respectivement (Feller, 1968, page 182). p - valeur < 5 % MIC2015_Oax : $N = 161\,296$; $\rho = 0{,}08$; $\gamma = 89{,}9$; $\text{MIC2015\_Oax}_{\text{trunc}}$: $N = 161\,216$; $\rho = 0{,}21$; $\gamma = 3{,}48$, où $\rho = \text{corr}(y, \phi)$ et $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Tableau 3.2
Biais relatif (BR) en pourcentage et racine carrée de l'erreur quadratique moyenne relative (REQMR) en pourcentage des estimateurs de $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, basé sur 5 000 échantillons. Non-réponse moyenne de 22,5 % (*NR*) et réponse complète (*RC*)

Population	$\bar{Y}$				$Y_{0,1}$				$Y_{0,5}$				$Y_{0,9}$			
	|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|	
	NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
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
$\text{MIC2015\_Oax}_{\text{trunc}}$	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. Populations simulées

Afin de contrôler l'asymétrie de la distribution de y, la corrélation $\text{corr}(y, \phi)$ et le pourcentage de non-réponse, nous avons simulé deux populations symétriques et deux populations asymétriques de taille $N = 50\,000$ avec une variable d'intérêt y et six variables auxiliaires $x_1, \ldots, x_6$, comme suit.

1. 50 000 échantillons aléatoires simples pour chacune des variables $x_1, \ldots, x_6$ ont été générés indépendamment d'une distribution $N(0,1)$.
2. Les probabilités de réponse ϕ_k ont été obtenues au moyen d'une régression logistique avec $\beta_1 = \cdots = \beta_6 = -0{,}3$ et β_0 choisies de sorte que la non-réponse moyenne soit égale à 24,8 %.
3. Deux paramètres ont été pris en compte pour la distribution de y :
   i) Symétrique. y_k a été généré à partir de $N(\mu, \sigma^2)$, avec $\mu = 1 + 2{,}16\ \text{corr}(y, x) \sum_{j=1}^{6} x_{kj}$ et $\sigma^2 = 4{,}67 * (1 - 6\text{corr}^2(y, x))$, où $\text{corr}(y, x) = \text{corr}(y, x_j)$, j élément de $\{1, \ldots, 6\}$.
   ii) Asymétrique. $y_k = \exp(z_k)$, où z_k a été généré à partir de $N(\mu, \sigma^2)$, avec $\mu = \text{corr}(z, x) \sum_{j=1}^{6} x_{kj}$ et $\sigma^2 = 1 - 6\text{corr}^2(z, x)$, où $\text{corr}(z, x) = \text{corr}(z, x_j)$, $j \in \{1, \ldots, 6\}$.

   Les valeurs $\text{corr}(y, x)$ et $\text{corr}(z, x)$ ont été toutes les deux choisies de façon à ce que $\text{corr}(y, \phi)$ soit à peu près égal à -0,2 ou -0,8.

4.1 Résultats numériques

Le taux de couverture des IC a été calculé comme à la section 3.1, avec $n = 500$ et au moyen d'une régression logistique sans interactions pour obtenir $\hat{\phi}_k$, $k \in s$.

Le tableau 4.1 présente les résultats pour les populations avec $\text{corr}(y, \phi) = -0{,}8$. Le tableau 4.2 montre le $|\text{BR}|$ et la $|\text{REQMR}|$ des estimateurs. Les résultats numériques pour les populations avec $\text{corr}(y, \phi) = -0{,}2$ sont omis, car ils étaient semblables à ceux obtenus pour les populations avec $\text{corr}(y, \phi) = -0{,}8$. Nous faisons les observations suivantes :

a) Les méthodes VE et Lin ont des performances raisonnables similaires pour $\bar{Y}$, bien que les taux d'erreur de queue supérieurs soient plus grands que 5 % dans la population asymétrique.

b) Pour les quantiles, la méthode Lin a la couverture la plus faible et la plus élevée, soit respectivement 85,7 % et 97,9 %. Contrairement à la distribution de $\hat{Y}_q$ illustrée à la figure 3.1, la distribution de $\hat{Y}_q$ pour les populations simulées est symétrique et unimodale dans tous les cas. Les méthodes VE et W donnent de bons résultats, atteignant le niveau nominal dans tous les cas avec des taux d'erreur de queue et une longueur moyenne d'IC comparables.

c) L'ajustement de la pondération dans le scénario de non-réponse aide à obtenir un BR semblable à celui de la réponse complète.

d) Le taux de couverture de l'IC pour chaque méthode est plus élevé dans le scénario de non-réponse que dans le scénario de réponse complète, et dans certains cas, il est également plus élevé que le niveau nominal. Cela pourrait être lié au fait d'avoir traité la valeur $\hat{\phi}_k$ comme étant fixe dans le scénario de non-réponse.

Tableau 4.1
Populations simulées. Couvertures des IC à 90 % pour $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, pour y avec $\mathbf{corr}(y,\phi) = -0{,}8$. Non-réponse moyenne de 24,8 % (*NR*) et réponse complète (*RC*)

Paramètre θ_0	Méthode	Couverture %		Taux d'err. de queue inférieurs en %		Taux d'err. de queue supérieurs en %		Longueur moyenne des IC	
		NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
Asymétrique									
$\bar{Y}$	VE	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
$Y_{0,1}$	VE	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
$Y_{0,5}$	VE	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
$Y_{0,9}$	VE	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
Symétrique									
$\bar{Y}$	VE	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
$Y_{0,1}$	VE	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
$Y_{0,5}$	VE	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
$Y_{0,9}$	VE	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

* Couvertures et taux d'erreur de queue différant considérablement de 90 % et 5 % respectivement (Feller, 1968, page 182). *p* - valeur < 5 % Symétrique : $\gamma =$ 0,02; Asymétrique : $\gamma =$ 6,2; où $\gamma = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3 \Big/ \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{3/2}$.

Tableau 4.2
Biais relatif (BR) en pourcentage et racine carrée de l'erreur quadratique moyenne relative (REQMR) en pourcentage des estimateurs de $\bar{Y}$, $Y_{0,1}$, $Y_{0,5}$ et $Y_{0,9}$, basé sur 5 000 échantillons. Non-réponse moyenne de 24,8 % (*NR*) et réponse complète (*RC*)

Population	$\bar{Y}$				$Y_{0,1}$				$Y_{0,5}$				$Y_{0,9}$			
	|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|		|BR|		|REQMR|	
	NR	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*	*NR*	*RC*
Asymétrique	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
Symétrique	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

Compte tenu de la distribution de y (*revenu*), on a observé que les mauvaises performances d'une méthode dans le scénario de réponse complète correspondaient généralement à de mauvaises performances dans le scénario de non-réponse, même si dans plusieurs cas, le taux de couverture était plus élevé dans ce dernier. Cela suggère que le fait d'avoir traité la variable $\hat{\phi}_k$ comme étant fixe a eu peu

d'effet sur les performances des méthodes par rapport à l'effet des caractéristiques de la distribution de y. Les valeurs extrêmes étaient liées à une faible couverture des IC pour la moyenne des méthodes de vraisemblance empirique et de linéarisation. La présence de valeurs à haute fréquence près d'un quantile d'intérêt avait aussi un effet sur la couverture de ses IC. Cela pourrait s'expliquer par le comportement de la fonction de répartition par étapes, où les sauts dans $F(y)$ et $\hat{F}(y)$ doivent habituellement être de petite taille pour que la méthode de Woodruff donne de bonnes performances (Lohr, 2010, page 390). En général, la méthode de linéarisation a donné des performances médiocres pour les quantiles, tandis que celles de la vraisemblance empirique et de la méthode de Woodruff étaient similaires et meilleures; ce comportement a aussi été observé dans Berger et De La Riva Torres (2016). La méthode Woodruff est simple et facile à mettre en œuvre, mais la méthode de la vraisemblance empirique présente l'avantage d'être utilisable pour des paramètres autres que les quantiles.

Remerciements

Omar De La Riva Torres était soutenu par le Programme de bourses d'études post doctorales de l'Université nationale autonome du Mexique.

Bibliographie

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., et 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). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/1999002/article/4882-fra.pdf.

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

Graf, E., et Tillé, Y. (2014). Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale. *Techniques d'enquête*, 40, 1, 69-88. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014001/article/14000-fra.pdf.

Haziza, D., et 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., et 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. et 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.