Large sample properties of the Midzuno sampling scheme

Guillaume Chauvet

Large sample properties of the Midzuno sampling scheme

Guillaume Chauvet*

September 26, 2018

Abstract

Midzuno sampling enables to estimate ratios unbiasedly. We prove the asymptotic normality for estimators of totals and ratios under Midzuno sampling. We also propose consistent variance estimators.

Keywords: asymptotic normality, consistent variance estimator, coupling.

1 Introduction

Midzuno (1951) proposed a sampling algorithm which enables to select a sample with unequal probabilities, while estimating unbiasedly a ratio. It is

*ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France. E-mail: chauvet@ensai.fr

therefore of interest with a moderate sample size, when the so-called small sample bias may be appreciable. Midzuno sampling has been recently considered in Escobar and Berger (2013) and Hidiroglou et al. (2016), for example.

We introduce a coupling algorithm between Midzuno sampling and simple random sampling, which enables to prove that the Horvitz-Thompson associated to these two procedures are asymptotically equivalent. As a by-product, we obtain a central-limit theorem for the estimator of a total and for the estimator of a ratio. We also prove that variance estimators suitable for simple random sampling are also consistent for Midzuno sampling.

The paper is organized as follows. The notation is introduced in Section 2. The coupling procedure is described in Section 3. It is used in Section 4 to prove the asymptotic normality of total and ratio estimators, and to establish the consistency of the proposed variance estimators. Their behaviour is studied in Section 5 through a simulation study, in case of a small sample size. The proofs are given in Section 6.

2 Notation and assumptions

We consider a finite population U of size N, with a variable of interest y taking the value y_k for the unit $k \in U$. We are interested in estimating the total $Y = \sum_{k\in U} y_k$ or the ratio $R = Y/X$ with $X = \sum_{k\in U} x_k$ and $x_k > 0$ is an auxiliary variable known for any unit $k \in U$.

Let $p_k > 0$ be some probability for unit k, with $\sum_{k\in U} p_k = 1$. If the probabilities are chosen proportional to x_k, we have $p_k = x_k/X$. A sample S of size n is selected according to some sampling design with $\pi_k > 0$ the inclusion probability of unit k. The Horvitz-Thompson (HT) estimator for the total is $\hat{Y} = \sum_{k\in S} \frac{y_k}{\pi_k}$, and the substitution estimator for the ratio is $\hat{R} = \hat{Y}/\hat{X}$, with $\hat{X} = \sum_{k\in S} \frac{x_k}{\pi_k}$.

2.1 Simple random sampling

If the sample is selected by simple random sampling in U, which is denoted as $SI(n;U)$, we obtain $\pi_k^{SI} = n/N$ and the estimators are

$$\hat{Y}_{SI} = \frac{N}{n} \sum_{k\in S_{SI}} y_k \quad \text{and} \quad \hat{R}_{SI} = \frac{\sum_{k\in S_{SI}} y_k}{\sum_{k\in S_{SI}} x_k}. \tag{2.1}$$

The variance of the HT-estimator is

$$V(\hat{Y}_{SI}) = \frac{N(N-n)}{n} S_y^2 \quad \text{with} \quad S_y^2 = \frac{1}{N-1} \sum_{k\in U} \left(y_k - \frac{Y}{N} \right)^2, \tag{2.2}$$

and is unbiasedly estimated by

$$\hat{V}(\hat{Y}_{SI}) = \frac{N(N-n)}{n} s^2_{y,SI} \quad \text{with} \quad s^2_{y,SI} = \frac{1}{n-1} \sum_{k \in S_{SI}} \left(y_k - \frac{\hat{Y}_{SI}}{N} \right)^2 . \quad (2.3)$$

Noting $z_k = y_k - Rx_k$ and $\hat{z}_k = y_k - \hat{R}_\pi x_k$, the linearization variance approximation for $\hat{R}_{SI}$ is

$$V_{lin}(\hat{R}_{SI}) = \frac{N(N-n)}{n\ X^2} S^2_z \quad \text{with} \quad S^2_z = \frac{1}{N-1} \sum_{k \in U} \left(z_k - \frac{\sum_{l \in U} z_l}{N} \right)^2 , \quad (2.4)$$

and the assorted variance estimator is

$$\hat{V}_{lin}(\hat{R}_{SI}) = \frac{N(N-n)}{n\ \hat{X}^2_{SI}} s^2_{\hat{z},SI} \quad \text{with} \quad s^2_{\hat{z},SI} = \frac{1}{n-1} \sum_{k \in S_{SI}} \left(\hat{z}_k - \frac{\sum_{l \in S} \hat{z}_l}{n} \right)^2 \quad (2.5)$$

We prove in Section 4 that $\hat{V}$ and $\hat{V}_{lin}$ are consistent for Midzuno sampling.

2.2 Midzuno sampling

Suppose that the sample S_{MI} is selected by means of the Midzuno (1951) sampling scheme, which is denoted as MI. A first unit (k_1, say) is selected in U with probabilities p_k. A sample S'_{MI} is then selected among the remaining units by $SI(n-1; U \setminus \{k_1\})$. The final Midzuno sample is $S_{MI} = S'_{MI} \cup \{k_1\}$, and the associated inclusion probabilities are

$$\pi^{MI}_k \quad = \quad \frac{n-1}{N-1} + p_k \left(\frac{N-n}{N-1} \right) . \quad (2.6)$$

The main advantage of MI is that $\hat{R}_{MI}$ is exactly unbiased for R if the probabilities p_k are proportional to x_k.

2.3 Assumptions

We work under the asymptotic set-up of Isaki and Fuller (1982), where U is embedded into a nested sequence of finite populations with $n, N \to \infty$. We suppose that the sampling rate is not degenerate, i.e. some constant $f \in]0,1[$ exists s.t. $n/N \to f$. We will consider the following assumptions:

H1: Some constants c_1, C_1 exist, s.t. $0 < c_1 \leq Np_k \leq C_1$ for any $k \in U$.

H2: Some constant M exists, s.t. $N^{-1} \sum_{k \in U} y_k^4 \leq M$.

H3a: Some constant $m_1 > 0$ exists, s.t. $S_y^2 \geq m_1$.

H3b: Some constant $m_2 > 0$ exists, s.t. $S_z^2 \geq m_2$.

3 Coupling procedure

The coupling procedure introduced in Algorithm 1 enables to justify of the closeness between MI and SI, as proved in Proposition 2.

Proposition 1. *The sample S_{SI} in Algorithm 1 is selected by $SI(n;U)$.*

Proposition 2. *Suppose that S_{MI} and S_{SI} are selected by Algorithm 1, and that assumptions (H1)-(H2) hold. Then*

$$E\left[\left(\hat{Y}_{MI} - \hat{Y}_{SI}\right)^4\right] = O(N^4 n^{-4}) \quad \textit{and} \quad E\left[\left(\hat{Y}_{MI} - Y\right)^4\right] = O(N^4 n^{-2}). \quad (3.1)$$

Algorithm 1 Coupling procedure between MI and SI sampling

1. Select some unit (k_1, say) in U with probabilities p_k.

2. Select S'_{MI} by $SI(n-1; U \setminus \{k_1\})$. The MI sample is $S_{MI} = S'_{MI} \cup \{k_1\}$.

3. Select some unit (k_2, say) in $U \setminus S'_{MI}$, with probability n/N for k_1 and $1/N$ otherwise. The SI sample is $S_{SI} = S'_{MI} \cup \{k_2\}$.

The first part of equation (3.1) implies in particular that

$$\left(\sqrt{V(\hat{Y}_{MI})} - \sqrt{V(\hat{Y}_{SI})}\right)^2 \quad = \quad O(N^2 n^{-2}) = o\{V(\hat{Y}_{SI})\}. \tag{3.2}$$

Consequently, $V(\hat{Y}_{MI})$ and $V(\hat{Y}_{SI})$ have asymptotically the same variance.

4 Interval estimation

Theorem 1. *Suppose that assumptions (H1), (H2) and (H3a) hold. Then*

$$\{V(\hat{Y}_{MI})\}^{-0.5}\{\hat{Y}_{MI} - Y\} \quad \longrightarrow_{\mathcal{L}} \quad \mathcal{N}(0,1), \tag{4.1}$$

$$E\left[N^{-2} n \left\{\hat{V}(\hat{Y}_{MI}) - V(\hat{Y}_{MI})\right\}\right]^2 \quad = \quad O(n^{-1}), \tag{4.2}$$

with $\rightarrow_{\mathcal{L}}$ the convergence in distribution, and where $\hat{V}(\hat{Y}_{MI})$ is the SI variance estimator given in (2.3), applied to the sample S_{MI}.

Theorem 1 implies that the HT-estimator is asymptotically normally distributed under MI, and that the SI variance estimator is also consistent for

MI, in the sense that $\{V(\hat{Y}_{MI})\}^{-1}\hat{V}(\hat{Y}_{MI}) \to_{Pr} 1$, with $\to_{Pr}$ the convergence in probability. In particular, the studentized interval

$$\left[\hat{Y}_{MI} \pm u_{1-\alpha}\{\hat{V}(\hat{Y}_{MI})\}^{0.5}\right] \tag{4.3}$$

has an asymptotic coverage of $100(1-2\alpha)\%$, with $u_{1-\alpha}$ the quantile of order $1-\alpha$ of the standard normal distribution.

We now consider ratio estimation. We suppose that the probabilities p_k are defined proportionally to x_k, and we strengthen (H1) as

H1b: Some constants c_1, C_1 exist, s.t. $0 < c_1 \leq x_k \leq C_1$ for any $k \in U$.

Proposition 3. *Suppose that assumptions (H1b) and (H2) hold. Then*

$$E\left[\left\{(\hat{R}_{MI} - R) - X^{-1}(\hat{Z}_{MI} - Z)\right\}^2\right] \;=\; O(n^{-2}). \tag{4.4}$$

This proposition entails in particular the validity of the linearization variance estimation, since it implies that $\{V_{lin}(\hat{R}_{SI})\}^{-1}V(\hat{R}_{SI}) \to 1$ if (H3b) is verified.

Theorem 2. *Suppose that assumptions (H1b), (H2) and (H3b) hold. Then*

$$\{V_{lin}(\hat{R}_{MI})\}^{-0.5}\{\hat{R}_{MI} - R\} \;\longrightarrow_{\mathcal{L}}\; \mathcal{N}(0,1), \tag{4.5}$$

$$E\left|n\left\{\hat{V}_{lin}(\hat{R}_{MI}) - V_{lin}(\hat{R}_{MI})\right\}\right| \;=\; O(n^{-0.5}), \tag{4.6}$$

where $\hat{V}_{lin}(\hat{R}_{MI})$ is the linearization SI variance estimator given in (2.5), applied to the sample S_{MI}.

Theorem 2 implies that the confidence interval $[\hat{R}_{MI} \pm u_{1-\alpha}\{\hat{V}_{lin}(\hat{R}_{MI})\}^{0.5}]$ has an asymptotic coverage of $100(1-2\alpha)\%$.

5 Simulation study

We conducted a small simulation to evaluate the proposed variance estimators with small samples. We generated a population of $N = 100$ units, with auxiliary variable x generated according to a gamma distribution with shape and scale parameters 2 and 5, and we shifted and scaled the values so that x_k lies between 1 and 20. We generated a variable of interest y according to the imputation model $y_k = x_k + \sigma\ \epsilon_k$, with the ϵ_k's generated according to a standard normal distribution, and where σ was chosen so that the coefficient of determination was approximately 0.70.

We repeated $B = 10,000$ times MI with p_k proportional to x_k, and with $n = 20$, 40 or 60. We computed: the relative bias (RB) of the proposed variance estimators $\hat{V}(\hat{Y}_{MI})$ and $\hat{V}_{lin}(\hat{R}_{MI})$, the true variance being approximated by an independent run of $100,000$ simulations; and the error rate of the normality-based confidence intervals with nominal one-tailed error rate of 2.5 % in each tail. The results given in Table 1 indicate that $\hat{V}$ is slightly positively biased with $n = 20$, but the bias decreases quickly when n grows,

as expected. The estimator $\hat{V}_{lin}$ is almost unbiased and the coverage rates are well respected in all cases.

Table 1: Percent relative bias and coverage probabilities

	$n = 20$		$n = 40$		$n = 60$	
	$\hat{V}(\hat{Y}_{MI})$	$\hat{V}_{lin}(\hat{R}_{MI})$	$\hat{V}(\hat{Y}_{MI})$	$\hat{V}_{lin}(\hat{R}_{MI})$	$\hat{V}(\hat{Y}_{MI})$	$\hat{V}_{lin}(\hat{R}_{MI})$
RB (%)	12.1	2.0	4.6	-0.1	2.3	-0.1
Cov. Rate	94.4	94.5	94.9	94.6	94.8	94.3

6 Proofs

6.1 Proof of Proposition 1

We prove that conditionally on k_1, S_{SI} is obtained by $SI(n; U)$. Let $s \subset U$ of size n. If $k_1 \notin s$, then

$$\begin{aligned} Pr(S_{SI} = s) &= \sum_{k \in s} Pr(S'_{MI} = s \setminus \{k\}) Pr(k_2 = k | S'_{MI} = s \setminus \{k\}) \\ &= n \frac{1}{C_{N-1}^{n-1}} \frac{1}{N} = \frac{1}{C_N^n}. \end{aligned}$$

If $k_1 \in s$, then

$$\begin{aligned} Pr(S_{SI} = s) &= Pr(S'_{MI} = s \setminus \{k_1\}) Pr(k_2 = k_1 | S'_{MI} = s \setminus \{k_1\}) \\ &= \frac{1}{C_{N-1}^{n-1}} \frac{n}{N} = \frac{1}{C_N^n}. \end{aligned}$$

6.2 Proof of Proposition 2

Lemma 1. *Under assumption (H1), we have*

$$\frac{|\pi_k^{MI} - \pi_k^{SI}|}{\pi_k^{MI}\pi_k^{SI}} \leq \max(1 - c_1, C_1 - 1)\frac{N-n}{n(n-1)}. \tag{6.1}$$

Lemma 2. *Let a_k denote some characteristic of unit k, and let m be some positive integer. If S_{MI} and S_{SI} are selected by means of Algorithm 1, then*

$$E\left[\left(\sum_{k\in S_{SI}} a_k - \sum_{k\in S_{MI}} a_k\right)^{2m}\right] = \frac{N-n}{N(N-1)}\sum_{k\in U}\sum_{l\in U} p_l(a_k - a_l)^{2m}. \tag{6.2}$$

If in addition the assumption (H1) holds, then

$$E\left[\left\{\left(\sum_{k\in S_{SI}} a_k\right)^2 - \left(\sum_{k\in S_{MI}} a_k\right)^2\right\}^2\right] = O\left(N^{-1}\sum_{k\in U} a_k^4\right). \tag{6.3}$$

Proof. From Algorithm 1, we have $\sum_{k\in S_{SI}} a_k - \sum_{k\in S_{MI}} a_k = a_{k_2} - a_{k_1}$. We obtain successively

$$\begin{aligned} E\left[(a_{k_2} - a_{k_1})^{2m} | k_1, S'_{MI}\right] &= \frac{1}{N}\sum_{k\in U\setminus\{k_1\}} (a_k - a_{k_1})^{2m} 1(k \notin S'_{MI}), \\ E\left[(a_{k_2} - a_{k_1})^{2m} | k_1\right] &= \frac{N-n}{N(N-1)}\sum_{k\in U\setminus\{k_1\}} (a_k - a_{k_1})^{2m}, \end{aligned}$$

which leads to (6.2). The proof of equation (6.3) follows from tedious but straightforward computations. □

We consider the first part of equation (3.1) only, since from $E[(\hat{Y}_{SI} - Y)^4] = O(N^4 n^{-2})$, it implies the second part. From the writing

$$\hat{Y}_{MI} - \hat{Y}_{SI} = \sum_{k\in S_{MI}} \frac{\pi_k^{SI} - \pi_k^{MI}}{\pi_k^{SI}\pi_k^{MI}} y_k + \frac{N}{n}\left(\sum_{k\in S_{MI}} y_k - \sum_{k\in S_{SI}} y_k\right),$$

we obtain

$$E\left[\left(\hat{Y}_{MI}-\hat{Y}_{SI}\right)^4\right] \leq 4E\left[\left(\sum_{k\in S_{MI}}\frac{\pi_k^{SI}-\pi_k^{MI}}{\pi_k^{SI}\pi_k^{MI}}y_k\right)^4\right]+4\frac{N^4}{n^4}E\left[\left(\sum_{k\in S_{MI}}y_k-\sum_{k\in S_{SI}}y_k\right)^4\right] \quad (6.4)$$

From equation (6.2) applied with $a_k = y_k$ and $m = 2$, and since $E[(\sum_{k\in S_{SI}} y_k)^4] = O(n^4)$ (see for example Ardilly and Tillé (2003, equation 2.12)), we have $E[(\sum_{k\in S_{SI}} y_k)^4] = O(n^4)$. By applying Lemma 1, we obtain that the first term in the r.h.s of (6.4) is $O(N^4n^{-4})$. Applying once again equation (6.2), we obtain that the second term in the r.h.s of (6.4) is $O(N^4n^{-4})$, which completes the proof.

6.3 Proof of Theorem 1

We can write

$$\frac{\hat{Y}_{MI}-Y}{\sqrt{V(\hat{Y}_{MI})}} = \sqrt{\frac{V(\hat{Y}_{SI})}{V(\hat{Y}_{MI})}}\left[\frac{\hat{Y}_{SI}-Y}{\sqrt{V(\hat{Y}_{SI})}}+\frac{\hat{Y}_{MI}-\hat{Y}_{SI}}{\sqrt{V(\hat{Y}_{SI})}}\right]. \quad (6.5)$$

From equation (3.2), we have $\{V(\hat{Y}_{MI})\}^{-1}V(\hat{Y}_{SI}) \to 1$ and $\{\sqrt{V(\hat{Y}_{SI})}\}^{-1}\{\hat{Y}_{MI}-\hat{Y}_{SI}\} = o_p(1)$. Equation (6.5) follows from the central-limit theorem for simple random sampling (e.g., Hájek, 1960) and from Slutsky's theorem.

To prove equation (4.2), we simplify the notation as $V(\hat{Y}_{MI}) \equiv V_{MI}$, $\hat{V}(\hat{Y}_{MI}) \equiv \hat{V}_{MI}$, and similarly for SI. We can write

$$\hat{V}_{MI}-V_{MI} = (\hat{V}_{SI}-V_{SI})+(\hat{V}_{MI}-\hat{V}_{SI})+(V_{SI}-V_{MI}).$$

We have $E[(\hat{V}_{SI}-V_{SI})^2] = O(N^4 n^{-3})$, see for example Ardilly and Tillé (2003, ex. 2.21). Also, from Lemma 2, we obtain $E[(\hat{V}_{MI} - \hat{V}_{SI})^2] = O(N^4 n^{-4})$. Finally, from equation (3.2) and since both V_{SI} and V_{MI} are $O(N^2 n^{-1})$, we obtain $(V_{SI} - V_{MI})^2 = O(N^4 n^{-3})$. This completes the proof.

6.4 Proof of Proposition 3

We note $\Delta \equiv (\hat{R}_{MI} - R) - X^{-1}(\hat{Z}_{MI} - Z) = (\hat{X}_{MI}^{-1})(\hat{Z}_{MI} - Z)(X - \hat{X}_{MI})$. From assumption (H1b), we obtain $\hat{X}_{MI} \geq (c_1/C_1)N$, which gives

$$E[\Delta^2] \;\leq\; (C_1/c_1)^2 N^{-2} \sqrt{E[(\hat{Z}_{MI} - Z)^4]} \sqrt{E[(\hat{X}_{MI} - X)^4]}.$$

By applying Proposition 2 to x_k and z_k, we obtain $E[(\hat{Z}_{MI} - Z)^4] = O(n^{-1})$ and $E[(\hat{X}_{MI} - X)^4] = O(N^2 n^{-1})$, which gives the result.

6.5 Proof of Theorem 2

Equation (4.5) follows from Proposition 3 and Slutsky's theorem. To prove equation (4.6), we note $\tilde{V}_{lin}(\hat{R}_{MI}) = N(N-n)s_z^2/(n\ X^2)$. The proof for equation (4.2) is easily adapted to obtain $E[\{\tilde{V}_{lin}(\hat{R}_{MI}) - V_{lin}(\hat{R}_{MI})\}^2] = O(n^{-3})$. Also, we obtain after some algebra $E[|\tilde{V}_{lin}(\hat{R}_{MI}) - V_{lin}(\hat{R}_{MI})|] = O(n^{-1.5})$, which gives the result.

References

Ardilly, P. and Tillé, Y. (2003). *Exercices corrigés de méthodes de sondage*. Ellipses.

Escobar, E. L. and Berger, Y. G. (2013). A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, pages 595–613.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361–74.

Hidiroglou, M. A., Kim, J. K., and Nambeu, C. O. (2016). A note on regression estimation with unknown population size. *Survey Methodology*, 42(1):121.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.*, 77(377):89–96.

Midzuno, H. (1951). On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Stat. Math.*, 3:99–107.