Acta Mathematica Sinica, English Series
May, 2010, Vol. 26, No. 5, pp. 863–874
Published online: April 15, 2010
DOI: 10.1007/s10114-010-6271-3
Http://www.ActaMath.com

Acta Mathematica Sinica, English Series
© Springer-Verlag Berlin Heidelberg &
The Editorial Office of AMS 2010

Imputation of Mean of Ratios for Missing Data and Its Application to PPSWR Sampling

Guo Hua ZOU
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Beijing 100190, *P. R. China*
E-mail: *ghzou@amss.ac.cn*

Ying Fu LI
Department of Mathematical Sciences, University of Houston-Clear Lake, Houston,
TX 77058-1098, *USA*
E-mail: *Li@uhcl.edu*

Rong ZHU
Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Beijing 100190, *P. R. China*
E-mail: *rongzhu@amss.ac.cn*

Zhong GUAN
Department of Mathematical Sciences, Indiana University South Bend, South Bend,
IN 46634-7111, *USA*
E-mail: *zguan@iusb.edu*

Abstract In practical survey sampling, nonresponse phenomenon is unavoidable. How to impute missing data is an important problem. There are several imputation methods in the literature. In this paper, the imputation method of the mean of ratios for missing data under uniform response is applied to the estimation of a finite population mean when the PPSWR sampling is used. The imputed estimator is valid under the corresponding response mechanism regardless of the model as well as under the ratio model regardless of the response mechanism. The approximately unbiased jackknife variance estimator is also presented. All of these results are extended to the case of non-uniform response. Simulation studies show the good performance of the proposed estimators.

Keywords imputation, item nonresponse, jackknife variance estimator, non-uniform response, PPSWR sampling

MR(2000) Subject Classification 62D05

1 Introduction

Item nonresponse occurs frequently in sample surveys. For example, in sample survey on transportation, some vehicles may not be found, but their tonnage or seat capacity is known to us. The method for dealing with this problem is usually to impute the missing values of the sampled units. The imputed values are treated as true values and then the standard formulas applicable to complete samples are used.

Received May 23, 2006, Accepted May 18, 2009
Supported by National Natural Science Foundation of China (Grant Nos. 70625004, 10721101 and 70933003)

Commonly used imputation methods include ratio imputation, regression imputation and random imputation. Using these methods, Rao and Sitter [1–2] consider the imputation problem of missing data when the simple random sampling is used and auxiliary information is available, and give the jackknife variance estimator and its linearized version based on the adjusted imputed values; Rao [3] studies the cases of the stratified random sampling and stratified multistage sampling; and Sitter and Rao [4] extend the analysis of [1–2] to the general case where the responses on either the variable of interest or the auxiliary variable or both may be missing; Zou, et al. [5] investigate how the sample rotation method is applied to the case where item nonresponse occurs. In their paper, Zou, et al. [6] suggest an imputation method of the mean of ratios for missing data under uniform response. By utilizing this imputation method, the estimator of the population mean is provided for the case where item nonresponse occurs when the simple random sampling is used. The estimator is shown to be valid under uniform response regardless of the model as well as under the ratio model regardless of the response mechanism. Interestingly, their method can naturally lead to the version of Hartley–Ross' estimator (see [7]) for estimating the population mean under two-phase sampling (see also [8] or [9]). Furthermore, the jackknife variance estimators are given and their approximate design-unbiasedness under uniform response is proved. We should note that the (approximate) design-unbiasedness is the main requirement for a good estimator in survey sampling. A similar property on the approximate design-unbiasedness of variance estimators under uniform response has been observed first in Zou and Feng [10], and then in [5] and [11]. Noting that what Zou, Li and Feng [6] considered is the case where the auxiliary information is complete, Liu et al. [12] extend Zou, Li and Feng's method to the situation of incomplete auxiliary information. On the other hand, Liang, Su and Zou [13] make use of Zou, Li and Feng's estimators to construct the confidence interval for a common mean, which, together with the empirical likelihood method, shows a good performance. In this paper, we apply the imputation method of the mean of ratios for missing data under uniform response to an unequal probability sampling— the probability proportional to size sampling with replacement (PPSWR sampling), and find that it may be the most natural imputation for this sampling. A modification for the imputation approach is made to adapt to the case of non-uniform response.

This paper is organized as follows: Section 2 presents the mean-of-ratios imputation method and its application to the PPSWR sampling under uniform response. The jackknife variance estimator is derived and its approximate design-unbiasedness is proved. Section 3 generalizes these results to the case of non-uniform response. Section 4 provides some simulation results. Some concluding remarks are given in Section 5.

2 Mean-of-Ratios Imputation with Application to PPSWR Sampling

2.1 Imputation Method and Its Application

Let a survey population U consist of N distinct units identified through the labels $i = 1, \dots, N$. A sample s_n with size n is drawn from U by certain sampling design. Suppose that the auxiliary variable, $\mathcal{X}$, is available for each unit of the population, but the variable of interest, $\mathcal{Y}$, is missing for some of the sampled units. Let s_r be the respondent set of size $r\,(\geq 1)$ and s_{n-r} be the nonrespondent set of size $n-r$. In this section, we consider a uniform response mechanism, i.e., independent response across sample units and equal response probability, p.

For the missing $\mathcal{Y}$-values, we consider the following imputation method of the mean of ratios (see [6] or [12]):

$$y_i^* = \Big(\frac{1}{r}\sum_{j\in s_r}\frac{y_j}{x_j}\Big)x_i, \qquad i \in s_{n-r}. \tag{2.1}$$

It can be seen that under the superpopulation model

$$\begin{cases} y_i = \beta x_i + e_i, \\ \varepsilon(e_i) = 0,\ \varepsilon(e_i^2) = \sigma^2 x_i^2,\ \varepsilon(e_i e_j) = 0\ (i \neq j), \end{cases} \tag{2.2}$$

where ε denotes expectation with respect to the model (correspondingly, the following E_d denotes expectation with respect to the design, and E_R denotes expectation with respect to the response mechanism), y_i^* is the best linear predictor of unobserved y_i.

Applying the above imputation method to the PPSWR sampling, the corresponding Hansen–Hurwitz estimator of the population mean $\bar{Y}$ is given by

$$\hat{\bar{Y}}^I_{PPS} = \frac{\bar{X}}{n}\Big(\sum_{i\in s_r}\frac{y_i}{x_i} + \sum_{i\in s_{n-r}}\frac{y_i^*}{x_i}\Big) = \frac{\bar{X}}{r}\sum_{i\in s_r}\frac{y_i}{x_i}, \tag{2.3}$$

where $\bar{X} = X/N = \frac{1}{N}\sum_{i=1}^N X_i$ is the population mean of $\mathcal{X}$-values.

Note that if we use ratio imputation: $y_i^* = \frac{\bar{y}_r}{\bar{x}_r}x_i$ for $i \in s_{n-r}$, then the estimator corresponding to Hansen–Hurwitz estimator is

$$\hat{\bar{Y}}'_I = \frac{\bar{X}}{n}\Big[\sum_{i\in s_r}\frac{y_i}{x_i} + (n-r)\frac{\bar{y}_r}{\bar{x}_r}\Big].$$

From this and the jackknife variance estimator of $\hat{\bar{Y}}^I_{PPS}$ given in the following (2.4), we see that the imputation approach of the mean of ratios is the most natural for the PPSWR sampling.

Applying the mean-of-ratios imputation method to the simple random sampling can naturally lead to the version of Hartley–Ross estimator for estimating the population mean under two-phase sampling. This estimator is design-unbiased (see [6] or [12]). Note that if we use the ratio imputation, then under uniform response, the corresponding imputed estimator $\hat{\bar{Y}}^R_{SRS} = \frac{\bar{y}_r}{\bar{x}_r}\bar{x}_n$ will have the bias of order $O(\frac{1}{n})$.

Noting that conditionally given r, s_r can be regarded as a simple random sub-sample of size r drawn from s_n under uniform response, we can easily obtain the following properties for the imputed estimator $\hat{\bar{Y}}^I_{PPS}$:

Theorem 2.1 (i) *Under the superpopulation model* (2.2) (*where the assumptions on the second moment are unnecessary*), *the estimator $\hat{\bar{Y}}^I_{PPS}$ is the model-unbiased estimator of $\bar{Y}$, regardless of the response mechanism.*

(ii) *The estimator $\hat{\bar{Y}}^I_{PPS}$ is design-unbiased under uniform response, regardless of the underlying model.*

Remark 2.1 From Theorem 2.1, we see that the estimator $\hat{\bar{Y}}^I_{PPS}$ is valid under uniform response regardless of the model as well as under the model (2.2) regardless of the response mechanism.

2.2 Variance of $\hat{\bar{Y}}^I_{PPS}$ and Its Jackknife Estimator

We first give the expression of the variance of $\hat{\bar{Y}}^I_{PPS}$ under uniform response. It can be seen that under uniform response,

$$
\begin{aligned}
V(\hat{\bar{Y}}^I_{PPS}) &= E(\hat{\bar{Y}}^I_{PPS} - \bar{Y})^2 \\
&= E_dE_rE_R\left\{\left[\left(\frac{\bar{X}}{r}\sum_{i\in s_r}\frac{y_i}{x_i} - \frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{x_i}\right) + \left(\frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{x_i} - \bar{Y}\right)\right]^2 \middle| r\right\} \\
&= E_dE_r\left\{\bar{X}^2 \cdot E_R\left[\left(\frac{1}{r}\sum_{i\in s_r}\frac{y_i}{x_i} - \frac{1}{n}\sum_{i\in s_n}\frac{y_i}{x_i}\right)^2 \middle| r\right] + \left(\frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{x_i} - \bar{Y}\right)^2\right\} \\
&= E_d\left\{\bar{X}^2 \cdot \left[E_r\left(\frac{1}{r}\right) - \frac{1}{n}\right] \cdot \frac{1}{n-1}\sum_{i\in s_n}\left(\frac{y_i}{x_i} - \frac{1}{n}\sum_{i\in s_n}\frac{y_i}{x_i}\right)^2 + \left(\frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{x_i} - \bar{Y}\right)^2\right\} \\
&= \frac{1}{N^2} \cdot E_r\left(\frac{1}{r}\right) \cdot \sum_{i=1}^{N} Z_i\left(\frac{Y_i}{Z_i} - Y\right)^2 \\
&= \frac{1}{np} \cdot \frac{1}{N^2}\sum_{i=1}^{N} Z_i\left(\frac{Y_i}{Z_i} - Y\right)^2 + O\left(\frac{1}{n^2}\right),
\end{aligned}
$$

where $Z_i = X_i/X$. Thus, we obtain

Theorem 2.2 *Let $n > 1$. Then the variance of $\hat{\bar{Y}}^I_{PPS}$ under uniform response is given by*

$$
V(\hat{\bar{Y}}^I_{PPS}) = \frac{1}{np} \cdot \frac{1}{N^2}\sum_{i=1}^{N} Z_i\left(\frac{Y_i}{Z_i} - Y\right)^2 + O\left(\frac{1}{n^2}\right).
$$

Motivated by [1–4], [14] and [15], we consider the following adjusted imputed values to obtain a simple jackknife variance estimator of $\hat{\bar{Y}}^I_{PPS}$: For $i \in s_{n-r}$,

$$
y_i^a(j) = \begin{cases} \left(\dfrac{1}{r-1}\displaystyle\sum_{k\in s_r - \{j\}}\frac{y_k}{x_k}\right)x_i, & j \in s_r, \\ \left(\dfrac{1}{r}\displaystyle\sum_{k\in s_r}\frac{y_k}{x_k}\right)x_i, & j \in s_{n-r}, \end{cases}
$$

when the j-th sample unit is deleted.

Based on these adjusted imputed values, the estimator of $\bar{Y}$ can be obtained as

$$
\hat{\bar{Y}}^a_{PPS}(j) = \begin{cases} \dfrac{\bar{X}}{r-1}\displaystyle\sum_{i\in s_r - \{j\}}\frac{y_i}{x_i}, & j \in s_r, \\ \dfrac{\bar{X}}{r}\displaystyle\sum_{i\in s_r}\frac{y_i}{x_i}, & j \in s_{n-r}, \end{cases}
$$

when the j-th sample unit is deleted.

Define the j-th pseudovalue as follows:

$$
\hat{\bar{Y}}_j = n\hat{\bar{Y}}^I_{PPS} - (n-1)\hat{\bar{Y}}^a_{PPS}(j).
$$

A jackknife variance estimator of $\hat{\bar{Y}}^I_{PPS}$ is then given by

$$\begin{aligned} v_J(\hat{\bar{Y}}^I_{PPS}) &= \frac{n-1}{n} \sum_{j\in s_n} [\hat{\bar{Y}}^I_{PPS} - \hat{\bar{Y}}^a_{PPS}(j)]^2 \\ &= \frac{n-1}{n(r-1)} \cdot \frac{\bar{X}^2}{r-1} \sum_{j\in s_r} \left(\frac{y_j}{x_j} - \frac{1}{r} \sum_{i\in s_r} \frac{y_i}{x_i} \right)^2. \end{aligned} \tag{2.4}$$

It can be seen that this estimator is very similar to the standard formula of the variance estimator of Hansen–Hurwitz estimator.

Noting that under uniform response,

$$\begin{aligned} E[v_J(\hat{\bar{Y}}^I_{PPS})] &= E\left[\frac{n(r-1)}{(n-1)r} \cdot v_J(\hat{\bar{Y}}^I_{PPS})\right] + O\left(\frac{1}{n^2}\right) \\ &= E_d E_r E_R \left[\frac{1}{r} \cdot \frac{\bar{X}^2}{r-1} \sum_{j\in s_r} \left(\frac{y_j}{x_j} - \frac{1}{r} \sum_{i\in s_r} \frac{y_i}{x_i} \right)^2 \middle| r \right] + O\left(\frac{1}{n^2}\right) \\ &= E_d \left[E_r\left(\frac{1}{r}\right) \cdot \frac{\bar{X}^2}{n-1} \sum_{j\in s_n} \left(\frac{y_j}{x_j} - \frac{1}{n} \sum_{i\in s_n} \frac{y_i}{x_i} \right)^2 \right] + O\left(\frac{1}{n^2}\right) \\ &= \frac{1}{N^2} \cdot E_r\left(\frac{1}{r}\right) \cdot \sum_{i=1}^{N} Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 + O\left(\frac{1}{n^2}\right) \\ &= V(\hat{\bar{Y}}^I_{PPS}) + O\left(\frac{1}{n^2}\right), \end{aligned}$$

we obtain the following theorem:

Theorem 2.3 *Let $r \geq 2$. Then under uniform response, we have*

$$E[v_J(\hat{\bar{Y}}^I_{PPS})] = V(\hat{\bar{Y}}^I_{PPS}) + O\left(\frac{1}{n^2}\right).$$

Theorem 2.3 shows that the jackknife variance estimator $v_J(\hat{\bar{Y}}^I_{PPS})$ is approximately design-unbiased for large n under uniform response. Its modification

$$v'_J(\hat{\bar{Y}}^I_{PPS}) = \frac{n(r-1)}{(n-1)r} \cdot v_J(\hat{\bar{Y}}^I_{PPS}) = \frac{\bar{X}^2}{r(r-1)} \sum_{j\in s_r} \left(\frac{y_j}{x_j} - \frac{1}{r} \sum_{i\in s_r} \frac{y_i}{x_i} \right)^2 \tag{2.5}$$

is exactly design-unbiased under uniform response. Furthermore, it can be seen that

$$E[v_J(\hat{\bar{Y}}^I_{PPS})] \geq V(\hat{\bar{Y}}^I_{PPS}).$$

This shows that $v_J(\hat{\bar{Y}}^I_{PPS})$ slightly overestimates the variance of $\hat{\bar{Y}}^I_{PPS}$. In other words, the jackknife variance estimator $v_J(\hat{\bar{Y}}^I_{PPS})$ is a somewhat conservative variance estimator under uniform response.

3 Extension to Case of Non-uniform Response

3.1 Modified Imputation Method and Its Application

In this section, we consider non-uniform response mechanism, i.e., independent response across sample units and unequal response probability, p_i for the unit i. We first let p_i be known and denote $q_i = 1 - p_i$. Define the response indicator on y_i as

$$I_i = \begin{cases} 1, & \text{if the unit } i \text{ responds to } y_i, \\ 0, & \text{otherwise.} \end{cases}$$

For the missing $\mathcal{Y}$-values, we adjust the above imputation method of the mean of ratios as follows:

$$y_i^* = \left(\frac{1}{n-r}\sum_{j\in s_r}\frac{q_j y_j}{p_j x_j}\right)x_i, \qquad i \in s_{n-r}. \tag{3.1}$$

It is interesting to note that y_i^* is an approximation of the weighted least squares predictor under the superpopulation model (2.2): In fact, it can be seen that under the superpopulation model (2.2), the weighted least squares estimator of β with the weights $w_i \propto q_i/(p_i x_i^2)$ is given by

$$\hat{\beta} = \frac{\sum_{s_r}\frac{q_i y_i}{p_i x_i}}{\sum_{s_r}\frac{1}{p_i} - r};$$

furthermore, the expectation with respect to the response mechanism of $\sum_{s_r}\frac{1}{p_i}$ is n. On the other hand, the modified imputation can also be regarded as a generalization of the imputation of the mean of ratios under uniform response given in (2.1), as in the case of uniform response, $q_j/p_j = (1-p)/p \approx (n-r)/r$.

Applying the above imputation method to the PPSWR sampling, the corresponding Hansen–Hurwitz estimator of the population mean $\bar{Y}$ is

$$\hat{\bar{Y}}_{PPS}^* = \frac{\bar{X}}{n}\sum_{i\in s_r}\frac{y_i}{p_i x_i} = \frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{p_i x_i}I_i. \tag{3.2}$$

It is not difficult to show that

Theorem 3.1 *The estimator $\hat{\bar{Y}}_{PPS}^*$ is design-unbiased under non-uniform response mechanism.*

Remark 3.1 Theorem 3.1 corresponds to Theorem 2.1 (ii). Under the superpopulation model (2.2), we can use the estimator $\hat{\bar{Y}}_{PPS}^I$, as Theorem 2.1 (i) shows that it is model-unbiased, regardless of the response mechanism.

3.2 Variance of $\hat{\bar{Y}}_{PPS}^*$ and Its Jackknife Estimator

We first give the expression of the variance of $\hat{\bar{Y}}_{PPS}^*$. It can be seen that under non-uniform response,

$$\begin{aligned} V(\hat{\bar{Y}}_{PPS}^*) &= V_d E_R\left(\frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{p_i x_i}I_i\right) + E_d V_R\left(\frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{p_i x_i}I_i\right) \\ &= V_d\left(\frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{x_i}\right) + E_d\left(\frac{\bar{X}^2}{n^2}\sum_{i\in s_n}\frac{q_i y_i^2}{p_i x_i^2}\right) \\ &= \frac{1}{N^2}\cdot\left\{\frac{1}{n}\sum_{i=1}^N Z_i\left(\frac{Y_i}{Z_i} - Y\right)^2 + \frac{1}{n}\sum_{i=1}^N\frac{q_i Y_i^2}{p_i Z_i}\right\}. \end{aligned}$$

So we have

Theorem 3.2 *Under non-uniform response, the variance of* $\hat{\bar{Y}}_{PPS}^*$ *is given by*

$$V(\hat{\bar{Y}}_{PPS}^*) = \frac{1}{N^2}\cdot\left\{\frac{1}{n}\sum_{i=1}^{N} Z_i\left(\frac{Y_i}{Z_i}-Y\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{q_iY_i^2}{p_iZ_i}\right\}.$$

We now give the jackknife variance estimator of $\hat{\bar{Y}}_{PPS}^*$. Define the adjusted imputed values as follows: For $i \in s_{n-r}$,

$$y_i^{*a}(j) = \begin{cases} \left(\dfrac{1}{n-r}\displaystyle\sum_{k\in s_r-\{j\}}\frac{q_ky_k}{p_kx_k}\right)x_i, & j\in s_r,\\ \left(\dfrac{1}{n-r-1}\displaystyle\sum_{k\in s_r}\frac{q_ky_k}{p_kx_k}\right)x_i, & j\in s_{n-r},\end{cases} \tag{3.3}$$

when the j-th sample unit is deleted.

Based on these adjusted imputed values, the estimator of $\bar{Y}$ can be obtained as

$$\hat{\bar{Y}}_{PPS}^{*a}(j) = \begin{cases} \dfrac{\bar{X}}{n-1}\displaystyle\sum_{i\in s_r-\{j\}}\frac{y_i}{p_ix_i}, & j\in s_r,\\ \dfrac{\bar{X}}{n-1}\displaystyle\sum_{i\in s_r}\frac{y_i}{p_ix_i}, & j\in s_{n-r},\end{cases}$$

when the j-th sample unit is deleted.

Define the j-th pseudovalue as follows:

$$\hat{\bar{Y}}_j^* = n\hat{\bar{Y}}_{PPS}^* - (n-1)\hat{\bar{Y}}_{PPS}^{*a}(j).$$

A jackknife variance estimator of $\hat{\bar{Y}}_{PPS}^*$ is then given by

$$\begin{aligned} v_J(\hat{\bar{Y}}_{PPS}^*) &= \frac{n-1}{n}\sum_{j\in s_n}[\hat{\bar{Y}}_{PPS}^* - \hat{\bar{Y}}_{PPS}^{*a}(j)]^2\\ &= \frac{n-1}{n}\left\{\sum_{j\in s_r}\left(\frac{\bar{X}}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i} - \frac{\bar{X}}{n-1}\sum_{i\in s_r-\{j\}}\frac{y_i}{p_ix_i}\right)^2\right.\\ &\quad \left.+\sum_{j\in s_{n-r}}\left(\frac{\bar{X}}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i} - \frac{\bar{X}}{n-1}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right)^2\right\}\\ &= \frac{\bar{X}^2}{n(n-1)}\left\{\sum_{j\in s_r}\left(\frac{y_j}{p_jx_j} - \frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right)^2 + (n-r)\left(\frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right)^2\right\}\\ &= \frac{\bar{X}^2}{n(n-1)}\left\{\sum_{i\in s_r}\frac{y_i^2}{p_i^2x_i^2} - \frac{1}{n}\left(\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right)^2\right\}. \end{aligned} \tag{3.4}$$

It can be seen that this estimator is also similar to the standard formula of the variance estimator of Hansen–Hurwitz estimator. Furthermore, it can be shown that under non-uniform response,

$$\begin{aligned} E[v_J(\hat{\bar{Y}}_{PPS}^*)] &= \frac{\bar{X}^2}{n(n-1)}\left\{E_dE_R\left(\sum_{j\in s_n}\frac{y_j^2}{p_j^2x_j^2}I_j\right) - n\cdot E\left(\frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right)^2\right\}\\ &= \frac{\bar{X}^2}{n(n-1)}\left\{E_d\left(\sum_{j\in s_n}\frac{y_j^2}{p_jx_j^2}\right) - n\cdot\left[V\left(\frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right) + \left(E(\frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i})\right)^2\right]\right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\bar{X}^2}{n-1}\left\{\frac{1}{X^2}\sum_{i=1}^{N}\frac{Y_i^2}{p_iZ_i} - V\left(\frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right) - \frac{Y^2}{X^2}\right\}\\
&= \frac{\bar{X}^2}{n-1}\cdot(n-1)V\left(\frac{1}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right)\\
&= V\left(\frac{\bar{X}}{n}\sum_{i\in s_r}\frac{y_i}{p_ix_i}\right).
\end{aligned}
$$

That is, we obtain

Theorem 3.3 *Let $n > 1$. Then under non-uniform response, we have*

$$E[v_J(\hat{\bar{Y}}^*_{PPS})] = V(\hat{\bar{Y}}^*_{PPS}).$$

Theorem 3.3 shows that the jackknife variance estimator $v_J(\hat{\bar{Y}}^*_{PPS})$ is design-unbiased under non-uniform response.

3.3 Case of Unknown Response Probability

In practice, the response probability p_i is rare to be known. Like Kim and Park [16], we model the response probability p_i by a parametric model $p_i = g(x_i;\theta)$, where g is a known smooth function and θ is an unknown finite-dimensional parameter. An example of such a model is the logistic regression model. The estimator $\hat{\theta}$ of the parameter θ can be obtained by the maximum likelihood approach and we assume it satisfies

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{1}{\sqrt{n}}\sum_{i\in s_n} H(I_i;\theta) + o_p(1),$$

where $H(I_i;\theta)$ has the mean zero and the positive definite variance-covariance matrix (see also [16]).

Let $\hat{p}_i = g(x_i;\hat{\theta})$ be the estimated response probability. Then the corresponding estimator of the population mean $\bar{Y}$ and its jackknife variance estimator are given by

$$\hat{\bar{Y}}^e_{PPS} = \frac{\bar{X}}{n}\sum_{i\in s_r}\frac{y_i}{\hat{p}_ix_i}, \tag{3.5}$$

and

$$v_J(\hat{\bar{Y}}^e_{PPS}) = \frac{\bar{X}^2}{n(n-1)}\left\{\sum_{i\in s_r}\frac{y_i^2}{\hat{p}_i^2x_i^2} - \frac{1}{n}\left(\sum_{i\in s_r}\frac{y_i}{\hat{p}_ix_i}\right)^2\right\}, \tag{3.6}$$

respectively.

For the relationship between the estimators with known and unknown p_i, the Taylor expansion leads to

$$\hat{p}_i^{-1} = p_i^{-1} + (\hat{\theta}-\theta)'\frac{\partial p_i^{-1}(\theta^*)}{\partial\theta}, \tag{3.7}$$

where θ^* is between θ and $\hat{\theta}$. So

$$\hat{\bar{Y}}^e_{PPS} = \hat{\bar{Y}}^*_{PPS} + (\hat{\theta}-\theta)'\cdot\frac{\bar{X}}{n}\sum_{i\in s_r}\frac{\partial p_i^{-1}(\theta^*)}{\partial\theta}\frac{y_i}{x_i}.$$

From this, we see that if $\frac{\partial p_i^{-1}}{\partial \theta}$ is uniformly bounded, then

$$\hat{\hat{Y}}^e_{PPS} = \hat{\hat{Y}}^*_{PPS} + O_p\Big(\frac{1}{\sqrt{n}}\Big).$$

Furthermore, assume that the response probability is bounded below, then (3.7) leads to

$$\begin{aligned}\frac{\bar{X}^2}{n(n-1)}\sum_{i\in s_r}\frac{y_i^2}{\hat{p}_i^2 x_i^2} &= \frac{\bar{X}^2}{n(n-1)}\sum_{i\in s_r}\frac{y_i^2}{x_i^2}\Big[p_i^{-1} + (\hat{\theta}-\theta)'\frac{\partial p_i^{-1}(\theta^*)}{\partial\theta}\Big]^2 \\ &= \frac{\bar{X}^2}{n(n-1)}\sum_{i\in s_r}\frac{y_i^2}{p_i^2 x_i^2} + O_p\Big(\frac{1}{n^{3/2}}\Big).\end{aligned}$$

Thus,

$$\begin{aligned} v_J(\hat{\hat{Y}}^e_{PPS}) &= \frac{\bar{X}^2}{n(n-1)}\sum_{i\in s_r}\frac{y_i^2}{\hat{p}_i^2 x_i^2} - \frac{1}{n-1}(\hat{\hat{Y}}^e_{PPS})^2 \\ &= \frac{\bar{X}^2}{n(n-1)}\sum_{i\in s_r}\frac{y_i^2}{p_i^2 x_i^2} - \frac{1}{n-1}(\hat{\hat{Y}}^*_{PPS})^2 + O_p\Big(\frac{1}{n^{3/2}}\Big) \\ &= v_J(\hat{\hat{Y}}^*_{PPS}) + O_p\Big(\frac{1}{n^{3/2}}\Big).\end{aligned}$$

4 Simulation Studies

In this section, we conduct some simulations to evaluate the performances of the proposed estimators of the population mean and their jackknife variance estimators. The data are generated from the three ratio models which are different only in the auxiliary variables:

$$y_i = 3.9x_i + x_i\varepsilon_i \tag{4.1}$$

with $x_i \sim U(0.1, 2.1)$, $N(1,1)$, and $N(20,16)$, respectively, $\varepsilon_i \sim N(0,1)$, and x_i and ε_i are assumed to be independent.

In the case of uniform response, we set $p = 0.76$; in the case of non-uniform response, the unequal response probability p_i for the unit i follows the logistic model

$$p_i = \frac{\exp(-1+2.3x_i)}{1+\exp(-1+2.3x_i)}. \tag{4.2}$$

These settings are similar to those in [16].

We first generate a finite population with the size of $N = 10,000$ from the model (4.1). Then the samples with $n = 100$ and $n = 500$ are drawn from the finite population by the PPSWR sampling, respectively. We repeat the process $B = 5,000$ times. For the b-th run, denote the estimators of the population mean under the uniform and non-uniform responses as $\hat{\hat{Y}}^{I(b)}_{PPS}$ and $\hat{\hat{Y}}^{*(b)}_{PPS}$, respectively. We calculate the simulated means and variances as follows:

$$E_*(\hat{\hat{Y}}^I_{PPS}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\hat{Y}}^{I(b)}_{PPS}, \quad E_*(\hat{\hat{Y}}^*_{PPS}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\hat{Y}}^{*(b)}_{PPS};$$

and

$$V_*(\hat{\hat{Y}}^I_{PPS}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\hat{Y}}^{I(b)}_{PPS} - \bar{Y})^2, \quad V_*(\hat{\hat{Y}}^*_{PPS}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\hat{Y}}^{*(b)}_{PPS} - \bar{Y})^2.$$

Similarly, the corresponding jackknife variance estimators are calculated as

$$E_*[v_J(\hat{Y}^I_{PPS})] = \frac{1}{B}\sum_{b=1}^{B} v_J^{(b)}(\hat{Y}^I_{PPS}), \text{ and } E_*[v_J(\hat{Y}^*_{PPS})] = \frac{1}{B}\sum_{b=1}^{B} v_J^{(b)}(\hat{Y}^*_{PPS}).$$

Table 1 summarizes the results on the simulated mean, variance and jackknife variance estimate. It can be seen from the table that both of the estimators $\hat{Y}^I_{PPS}$ and $\hat{Y}^*_{PPS}$ are very close to the true population means for the three distributions of auxiliary variable. Also, the jackknife variance estimators perform very well.

Model (population mean)	n	Estimator	Mean	Variance	Jackknife variance estimator
M1	100	$\hat{Y}^I_{PPS}$	4.305	0.01648	0.01610
(4.305)		$\hat{Y}^*_{PPS}$	4.306	0.05491	0.05488
	500	$\hat{Y}^I_{PPS}$	4.305	0.003133	0.003210
		$\hat{Y}^*_{PPS}$	4.304	0.01079	0.01098
M2	100	$\hat{Y}^I_{PPS}$	3.961	0.01359	0.01361
(3.961)		$\hat{Y}^*_{PPS}$	3.961	0.03485	0.03578
	500	$\hat{Y}^I_{PPS}$	3.962	0.002729	0.002717
		$\hat{Y}^*_{PPS}$	3.962	0.007372	0.007154
M3	100	$\hat{Y}^I_{PPS}$	77.82	5.187	5.259
(77.82)		$\hat{Y}^*_{PPS}$	77.83	3.865	3.980
	500	$\hat{Y}^I_{PPS}$	77.82	1.029	1.048
		$\hat{Y}^*_{PPS}$	77.82	0.7839	0.7969

Table 1 Simulated mean, variance, and jackknife variance estimate based on the samples of the sizes $n = 100$ and $n = 500$ when $p = 0.76$ and p_i follows the model (4.2). M1: $x_i \sim U(0.1, 2.1)$; M2: $x_i \sim N(1, 1)$; M3: $x_i \sim N(20, 16)$.

To study the effect of the response probability, we set various response probabilities: $p = 0.5$ for uniform response, and p_i follows

$$p_i = \frac{\exp\{0.3(x_i - \bar{X})\}}{1 + \exp\{0.3(x_i - \bar{X})\}} \tag{4.3}$$

for non-uniform response. The results are presented in Table 2. It is observed that the approximate design-unbiasedness of the proposed estimators still holds. On the other hand, it is also clear that the variances become larger for low response probability. For some other settings of response probability, we obtain similar results but omit them here for saving space.

5 Concluding Remarks

In this paper, we have applied the imputation approach of the mean of ratios for missing data under uniform response to the PPSWR sampling. A modification of the method has been made to adapt to the non-uniform response case. Note that for the estimation of response probability, we have used a parametric model. Clearly, the use of non-parametric approach is also interesting. On the other hand, the auxiliary information considered in this article is complete. Like Liu et al. [12], it is worth extending the results in this paper to the situation of incomplete auxiliary information and this warrants our future research.

Model (population mean)	n	Estimator	Mean	Variance	Jackknife variance estimator
M1	100	$\hat{\bar{Y}}^{I}_{PPS}$	4.309	0.02481	0.02410
(4.305)		$\hat{\bar{Y}}^{*}_{PPS}$	4.311	0.2007	0.1939
	500	$\hat{\bar{Y}}^{I}_{PPS}$	4.305	0.004787	0.004876
		$\hat{\bar{Y}}^{*}_{PPS}$	4.305	0.03875	0.03855
M2	100	$\hat{\bar{Y}}^{I}_{PPS}$	3.961	0.02059	0.02078
(3.961)		$\hat{\bar{Y}}^{*}_{PPS}$	3.965	0.1451	0.1419
	500	$\hat{\bar{Y}}^{I}_{PPS}$	3.962	0.004256	0.004134
		$\hat{\bar{Y}}^{*}_{PPS}$	3.962	0.02734	0.02779
M3	100	$\hat{\bar{Y}}^{I}_{PPS}$	77.82	8.281	8.198
(77.82)		$\hat{\bar{Y}}^{*}_{PPS}$	76.97	106.9	105.8
	500	$\hat{\bar{Y}}^{I}_{PPS}$	77.84	1.596	1.598
		$\hat{\bar{Y}}^{*}_{PPS}$	77.70	19.77	20.85

Table 2 Simulated mean, variance, and jackknife variance estimate based on the samples of the sizes $n = 100$ and $n = 500$ when $p = 0.5$ and p_i follows the model (4.3). M1: $x_i \sim U(0.1, 2.1)$; M2: $x_i \sim N(1, 1)$; M3: $x_i \sim N(20, 16)$.

References

[1] Rao, J. N. K., Sitter, R. R.: Jackknife variance estimation under imputation for missing survey data, Technical Report, No. 214, Carleton University, 1993

[2] Rao, J. N. K., Sitter, R. R.: Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453–460 (1995)

[3] Rao, J. N. K.: On variance estimation with imputed survey data. *J. Amer. Statist. Assoc.*, **91**, 499–506 (1996)

[4] Sitter, R. R., Rao, J. N. K.: Imputation for missing values and corresponding variance estimation. *Canad. J. Statist.*, **25**, 61–73 (1997)

[5] Zou, G., Feng, S., Qin, H.: Sample rotation theory with missing data. *Sci. China Ser. A*, **45**, 42–63 (2002)

[6] Zou, G., Li, Y., Feng, S.: A new imputation method for missing data and its application. The 5th ICSA Statistical Conference, Hong Kong, 2001

[7] Hartley, H: O., Ross, A.: Unbiased ratio estimates. *Nature*, **174**, 270–271 (1954)

[8] Cochran, W. G.: Sampling Techniques, 3rd ed., John Wiley & Sons, New York, 1977

[9] Feng, S., Ni, J., Zou, G.: Survey Sampling—Theory and Methods, The Statistical Publishing House of China, Beijing, 1998

[10] Zou, G., Feng, S.: Sample rotation method with missing data, The 4th ICSA Statistical Conference, Kunming, 1998

[11] Skinner, C. J., Rao, J. N. K.: Jackknife variance estimation for multivariate statistics under hot-deck imputation from common donors. *J. Statist. Plann. Inference*, **102**, 149–167 (2002)

[12] Liu, L., Tu, Y., Li, Y., et al.: Imputation for missing data and variance estimation when auxiliary information is incomplete. *Model Assist. Stat. Appl.*, **1**, 83–94 (2006)

[13] Liang, H., Su, H., Zou, G.: Confidence intervals for a common mean with missing data with applications in AIDS study. *Comput. Statist. Data Anal.*, **53**, 546–553 (2008)

[14] Rao, J. N. K., Shao, J.: Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79**, 811–822 (1992)

[15] Rao, J. N. K.: Developments in sample survey theory: an appraisal. *Canad. J. Statist.*, **25**, 1–21 (1997)

[16] Kim, J. K., Park, H.: Imputation using response probability. *Canad. J. Statist.*, **34**, 171–182 (2006)