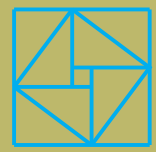
July 5, 2010 · Beijing

Imputation for Missing Data under PPSWR Sampling

Guohua Zou

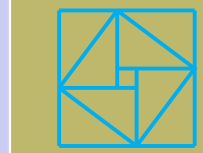
Academy of Mathematics and Systems Science

Chinese Academy of Sciences



Outline

- (一) Background
- (二) Imputation method under PPSWR sampling
- (三) Special case: SRS sampling and uniform response
- (四) Simulation studies
- (五) Future researches

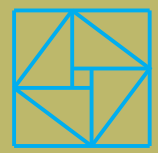


(□) Background

- Item nonresponse: occurs frequently in sample surveys.

  Example In sample survey on transportation, some vehicles may not be found, but their tonnage or seat capacity is known to us.

- Solutions: (1) Increasing response probability; (2) Imputing the missing values of the sampled units.
- Imputation methods: Ratio imputation, regression imputation, random imputation etc.
- Shortcoming: Uniform response (often), simple random sampling
- PPSWR sampling: the sampling with probability proportional to size with replacement which is often used in the first stage of multistage sampling.



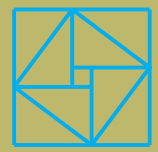
(二) Imputation method under PPSWR sampling

1. Notation

- Survey population U: consist of N distinct units identified through the labels $i = 1, ..., N$.

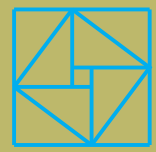
  Suppose that the auxiliary variable, $\mathcal{X}$, is available for each unit of the population, but the variable of interest, $\mathcal{Y}$, is missing for some of the sampled units.

- s_n: sample with size n drawn from U by PPSWR sampling.
- s_r: the respondent set with size $r(\geq 1)$.
- s_{n-r}: the nonrespondent set with size $n - r$.



- Uniform response mechanism: independent response across sample units and equal response probability, p $(q \equiv 1 - p)$.
- Non-uniform response mechanism: independent response across sample units and possibly unequal response probabilities, p_i $(q_i \equiv 1 - p_i)$.
- Response indicator on y_i:

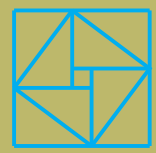
$$I_i = \begin{cases} 1, & \text{if the unit } i \text{ responds to } y_i, \\ 0, & \text{otherwise.} \end{cases}$$



2. Imputation method

We first let p_i be known. For the missing $\mathcal{Y}$–values, we suggest the following imputation method:

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



It is interesting to note that y_i^* is an approximation of the weighted least squares predictor under the following superpopulation model:

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

In fact, it can be seen that under the above superpopulation model, the weighted least squares estimator of β with the weights $w_i \propto q_i/(p_i x_i^2)$ is given by

$$\hat{\beta} = \frac{\sum_{s_r}(q_i y_i)/(p_i x_i)}{\sum_{s_r} 1/p_i - r}.$$

Further, the expectation of $\sum_{s_r} 1/p_i$ with respect to the response mechanism is n.

3. Estimator of population mean and its variance

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

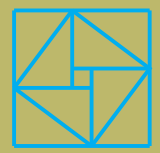
$$\hat{\bar{Y}}^*_{PPS} = \frac{\bar{X}}{n}\sum_{i\in s_r}\frac{y_i}{p_i x_i} = \frac{\bar{X}}{n}\sum_{i\in s_n}\frac{y_i}{p_i x_i}I_i.$$

Theorem 1. The estimator $\hat{\bar{Y}}^*_{PPS}$ is design-unbiased under the non-uniform response mechanism.

Theorem 2. Under the non-uniform response, the variance of $\hat{\bar{Y}}^*_{PPS}$ is given by

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

where $Z_i = X_i/X$.



4. Jackknife variance estimator

Define the imputed values as follows: For $i \in s_{n-r}$,

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

when the j-th sample unit is deleted.

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

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

when the j-th sample unit is deleted.

Define the j-th pseudovalue as follows:

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

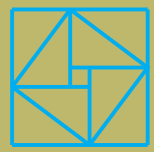
A jackknife variance estimator of $\hat{\bar{Y}}_{PPS}^*$ is then given by

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

Theorem 3. Let $n > 1$. Then under the non-uniform response, we have

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

Theorem 3 shows that the jackknife variance estimator $v_J(\hat{\bar{Y}}_{PPS}^*)$ is design-unbiased under the non-uniform response.



访 □ □ 页

and

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

respectively.

For the relationship between the estimators with known and unknown p_i, under some regularity conditions, we have

$$\hat{\bar{Y}}^e_{PPS} = \hat{\bar{Y}}^*_{PPS} + O_p\left(\frac{1}{\sqrt{n}}\right),$$

and

$$v_J(\hat{\bar{Y}}^e_{PPS}) = v_J(\hat{\bar{Y}}^*_{PPS}) + O_p\left(\frac{1}{n^{3/2}}\right).$$

(三) Special case: SRS sampling and uniform response

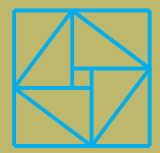
- Imputed value:

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

- Imputed estimator:

$$\hat{\bar{Y}}_{SRS}^m = \bar{u}_r \bar{x}_n + \frac{r(n-1)}{(r-1)n} (\bar{y}_r - \bar{u}_r \bar{x}_r),$$

where $\bar{u}_r = \frac{1}{r} \sum_{j \in s_r} u_j$ with $u_j = y_j / x_j$.



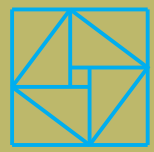
The estimator $\hat{\bar{Y}}_{SRS}^{m}$ is a design-unbiased estimator under uniform response. It is interesting that it is just the version of Hartley-Ross estimator (Hartley and Ross 1954, *Nature*) for estimating the population mean under two-phase sampling.

- Variance:

$$V(\hat{\bar{Y}}_{SRS}^{m}) = \frac{1}{np}[S_Y^2 + (1-p)(2\bar{Y}\bar{U}\bar{X} + \bar{U}^2\bar{T} - \bar{U}^2\bar{X}^2 - 2\bar{U}\bar{Z})] + O\left(\frac{1}{n^2}\right),$$

where $\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$ with $T_i = X_i^2$, and $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ with $Z_i = Y_i X_i$, and

$$S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2.$$



Two approximate Jackknife variance estimators:

$$v'_J(\hat{\bar{Y}}^m_{SRS}) = \frac{1}{n}\bar{u}_r^2 s_x^2(n) + \frac{1}{r}[(\bar{x}_r - \bar{x}_n)^2 s_u^2(r) - 2(\bar{x}_r - \bar{x}_n)s_{yu}(r) + s_y^2(r)] + \left(\frac{1}{r} - \frac{2}{n}\right)\bar{u}_r^2 s_x^2(r) + 2\left(\frac{1}{r} - \frac{1}{n}\right)\bar{u}_r[(\bar{x}_r - \bar{x}_n)s_{ux}(r) - s_{yx}(r)],$$

and

$$v''_J(\hat{\bar{Y}}^m_{SRS}) = \frac{1}{n}\bar{u}_r^2 s_x^2(n) + \frac{1}{r}s_y^2(r) + \left(\frac{1}{r} - \frac{2}{n}\right)\bar{u}_r^2 s_x^2(r) - 2\left(\frac{1}{r} - \frac{1}{n}\right)\bar{u}_r s_{yx}(r).$$

Asymptotic design-unbiasedness:

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

(ii) $E[v'_J(\hat{\bar{Y}}^m_{SRS})] = V(\hat{\bar{Y}}^m_{SRS}) + O\left(\frac{1}{n^2}\right)$.

(iii) $E[v''_J(\hat{\bar{Y}}^m_{SRS})] = V(\hat{\bar{Y}}^m_{SRS}) + O\left(\frac{1}{n^2}\right)$.

Remark: We should note that the (approximate) design-unbiasedness is the main requirement for a good estimator in survey sampling. The approximate design-unbiasedness of the Jackknife variance estimators has been found first in Zou and Feng (1998), and then in Skinner and Rao (2002) and Zou *et al.* (2002). Such a property universally holds from our subsequent analysis.

(四) Simulation studies

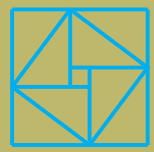
The data are generated from the three ratio models which are different only in the auxiliary variables:

$$y_i = 3.9x_i + x_i\varepsilon_i$$

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

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



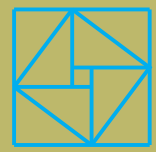
Model (population mean)	n	Estimator	Mean	Variance	Jackknife variance estimator
M1	100	$\hat{\bar{Y}}^{I}_{PPS}$	4.305	0.01648	0.01610
(4.305)		$\hat{\bar{Y}}^{*}_{PPS}$	4.306	0.05491	0.05488
	500	$\hat{\bar{Y}}^{I}_{PPS}$	4.305	0.003133	0.003210
		$\hat{\bar{Y}}^{*}_{PPS}$	4.304	0.01079	0.01098
M2	100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

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 $\hat{\bar{Y}}^{I}_{PPS}$ and $\hat{\bar{Y}}^{*}_{PPS}$ are very close to the true population means for the three distributions of auxiliary variable. Also, the jackknife variance estimators perform very well.

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

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

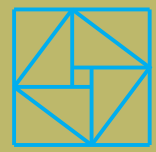
for non-uniform response. The results are presented in Table 2. It is observed that the approximate design-unbiasedness of the proposed estimators still holds. On the other hand, it is also clear that the variances become larger for low response probability. For some other settings of response probability,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



Model (population mean)	n	Estimator	Mean	Variance	Jackknife variance estimator
M1 (4.305)	100	$\hat{\bar{Y}}^{I}_{PPS}$	4.309	0.02481	0.02410
		$\hat{\bar{Y}}^{*}_{PPS}$	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 (3.961)	100	$\hat{\bar{Y}}^{I}_{PPS}$	3.961	0.02059	0.02078
		$\hat{\bar{Y}}^{*}_{PPS}$	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 (77.82)	100	$\hat{\bar{Y}}^{I}_{PPS}$	77.82	8.281	8.198
		$\hat{\bar{Y}}^{*}_{PPS}$	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

(五) Future researches

- Incomplete auxiliary information and multiple auxiliary information
- Estimation of response probability: the use of non-parametric approach.
- Other unequal probability sampling



Thank you!