Statistica Sinica **9**(1999), 361-384

INFERENCE WITH SURVEY DATA IMPUTED BY HOT DECK WHEN IMPUTED VALUES ARE NONIDENTIFIABLE

Yinzhong Chen and Jun Shao

University of Wisconsin-Madison

Abstract: Hot deck imputation for nonrespondents is often used in surveys. It is a common practice to treat the imputed values as if they are true values, and compute survey estimators and their variance estimators using standard formulas. The variance estimators, however, have serious negative biases when the rate of nonresponse is appreciable. Methods such as the multiple imputation and the adjusted jackknife have been proposed to obtain improved variance estimators. However, multiple imputation requires that multiple data sets be generated and maintained and that the imputation procedure be proper; the adjusted jackknife requires "flags" to identify imputed values. In many practical problems there is only a single imputed data set with unknown response status (no identification flag). In this paper we derive some asymptotically design-consistent inference procedures in the situation where a stratified multistage sampling design is used to collect survey data; hot deck imputation is applied to form a single imputed data set; the imputed values are nonidentifiable; and the survey estimators under consideration are functions of sample means, sample quantiles, and sample low income proportions.

Key words and phrases: Identification flags, item nonresponse, low income proportion, sample mean and quantile, single imputation, stratified multistage sampling, uniform response.

1. Introduction

Most survey data are incomplete due to nonresponse. We consider item nonresponse which occurs when a sampled unit cooperates in the survey but fails to provide answers to some of the questions. Imputation techniques (which insert values for nonrespondents) are commonly used to compensate for missing data because of various practical (not necessarily statistical) reasons (Kalton (1981), Sedransk (1985)). We focus on the hot deck imputation method described in Rao and Shao (1992) which inserts missing values by a random sample from the respondents. An advantage of using this hot deck imputation method is that it preserves the distribution of item values so that valid estimators that depend on the entire distribution of item values (e.g., the sample quantiles) can be obtained based on the imputed data set. This important property is not shared by some deterministic imputation methods such as the mean imputation, the ratio imputation, and the regression imputation.

It is a common practice to treat the imputed values as if they are true values, and then make inference using standard formulas. If imputation is suitably carried out, survey estimators of population parameters, computed by using the imputed data and standard formulas, are asymptotically valid; their variance estimators, however, have serious negative biases when the proportion of nonrespondents is appreciable, because standard formulas for variance estimation do not account for the inflation in variance due to missing data and/or imputation. Consequently, inference based on these variance estimators can be very misleading.

There exist two types of methods which provide better variance estimators:

(1) Rubin (1978) and Rubin and Schenker (1986) proposed the multiple imputation method which requires several independent imputations and computes variance estimators using the variabilities among the imputed data sets.

(2) There are methods based on some adjustments which account for the inflation in the variance due to nonresponse and/or imputation, e.g., the adjusted jackknife method (Rao and Shao (1992)), the adjusted linearization method (Rao (1993)), and the bootstrap method (Shao and Sitter (1996)). These methods provide asymptotically design-consistent variance estimators and work for both single and multiple imputation, but require identification flags to locate imputed values.

In this paper we focus on the situation where both types of methods discussed above are not applicable; that is, the situation where we only have a single imputed data set and we do not know which sampled units are imputed values (no identification flag).

Multiple imputation requires multiple imputations and some extra spaces to maintain multiple data sets and, therefore, is not appreciated by many practical users. It also requires that the imputation method be "proper" (i.e., it satisfies conditions 1-3 in Rubin (1987), pp. 118-119). However, some commonly used imputation methods (including the hot deck method) are not proper, but are simple, asymptotically valid, and more efficient than proper imputation methods. In some complex situations, it is hard to find a proper imputation method (Fay (1991) and (1993), Rao (1996)). These are the reasons why we study inference methods that work for a single imputed data set.

The reason why we condsider the case of unknown response status is the following. Many public data sets do not carry identification flags for imputed values. Note that adding identification flags is the same as adding a response indicator variable to the data set. When we have multivariate data and item nonresponse, identification flags have to be added for all items, which nearly doubles the size of the original data set and is not easy to handle in large scale surveys. Another situation in which imputed values are not identifiable is when confidentiality edit is applied to the data set for confidential reasons (Griffin, Navarro and Flores-Baez (1991)). One type of confidentiality edit is done by selecting a portion of the data and interchanging them with a random sample from the rest of data. If we treat the selected data for interchange as "nonrespondents" and the rest of data as "respondents", then the "imputed values" cannot be identified.

Under a general stratified multistage sampling design (see Section 2), we propose some inference methods (variance estimators and confidence intervals) based on two types of most commonly used estimators in surveys: (1) the sample mean or a function of sample means (Section 3), (2) the sample quantiles (Section 4). The sample low income proportion, a statistic that is important for income studies, is also considered in Section 4. Design-consistency of our proposed procedures are established under a usual asymptotic framework. Some simulation results are given in Section 5.

2. The Population and Sampling Design

Consider a population with L strata and N_h first-stage units in the hth stratum, $h = 1, \ldots, L$. Suppose that $n_h \geq 2$ first-stage units are sampled from stratum h without replacement, independently across the strata. Within the hth stratum, the (h, i)th first-stage unit is selected with probability $p_{hi} > 0$, $i = 1, \ldots, N_h$. We focus on the common case where L is large and the n_h are bounded by a fixed integer. We assume that the first-stage sampling fraction $\sum_h n_h / \sum_h N_h$ is negligible. If the first-stage units are clusters, then a second-stage sample, a third-stage sample,..., may be selected within each cluster, and the samples are selected independently across the clusters. We do not specify the number of stages and the sampling methods used after the first-stage sampling. For simplicity, we shall index the ultimate units in a first-stage cluster by using a single index, i.e., unit (h, i, j) is the jth ultimate unit in the ith first-stage cluster of stratum h, $i = 1, \ldots, n_h$, $h = 1, \ldots, L$. Item values for unit (h, i, j) are denoted by y_{hij}, z_{hij}, etc. This sampling design is called the stratified multistage sampling plan.

We adopt the design-based approach; that is, we do not use any model assumption on the values y_{hij}, z_{hij},... All probabilities and expectations are with respect to repeated sampling and/or random imputation.

Let A be the index set of all sampled units and let w_{hij} be the survey weight associated with the (h, i, j)th sampled ultimate unit. The survey weights are constructed so that when there is no nonrespondent,

$$\hat{Y} = \sum_A w_{hij} y_{hij}$$

is an unbiased estimator of the population total Y on any item y, where $\sum_A$ denotes the summation over all indices that are in A. Since in multistage sampling the total number of ultimate units M is often unknown, the population mean $\bar{Y} = Y/M$ is estimated by a ratio estimator

$$\bar{y} = \hat{Y}/\hat{M}, \tag{2.1}$$

where

$$\hat{M} = \sum_A w_{hij}.$$

A usual framework for the development of asymptotic theory is provided by the concept of a sequence of populations $\{\mathcal{P}_\nu, \nu = 1, 2, \ldots\}$, where each $\mathcal{P}_\nu$ contains L_ν strata. The population under consideration is then viewed as a member of this sequence of populations. Note that L, N_h, w_{hij}, and y_{hij} depend on ν, but ν is omitted for simplicity. All limiting processes are understood to be as $\nu \to \infty$.

Imputation is usually carried out separately in several imputation classes which form a partition of the whole population. The imputation classes are constructed according to the value of a categorical variable observed for all the sampled units. Within each imputation class, the sampled units respond to an item y with nearly the same probability p_y (see, e.g., Schenker and Welsh (1988), Section 4), although p_y may be different for different items and/or different imputation classes. Within an imputation class, imputation is usually done by cutting across strata and clusters. Thus, imputation can still be carried out even when some strata or clusters have no respondent within an imputation class.

3. Variance Estimation for Functions of Sample Means

In this section, we consider variance estimation for the sample mean (2.1) for an item or a function of sample means for several items.

3.1. Univariate case with uniform response

We start with the simplest case where we consider only one item, y, and there is only one imputation class (the sampled units respond with the same probability $p_y > 0$). Let

$$A_r = \{(h, i, j) : \ y_{hij} \text{ is observed}\}$$

and

$$A_m = \{(h, i, j) : \ y_{hij} \text{ is missing}\}.$$

Suppose that missing y_{hij} are imputed by y^*_{hij}, $(h, i, j) \in A_m$. Define $y^*_{hij} = y_{hij}$ if $(h, i, j) \in A_r$. Treating $\{y^*_{hij}, (h, i, j) \in A\}$ as the true data set and using the standard formula (2.1), we estimate $\bar{Y}$ by

$$\bar{y}^* = \sum_A w_{hij} y^*_{hij} \Big/ \hat{M}. \tag{3.1}$$

We focus on the following hot deck imputation (Rao and Shao (1992)): $\{y^*_{hij} : (h,i,j) \in A_m\}$ is an i.i.d. sample from the respondents, where each y_{hij}, $(h,i,j) \in A_r$, is selected with probability proportional to its weight w_{hij}. Under this hot deck imputation, $\bar{y}^*$ in (3.1) is asymptotically unbiased, consistent, and asymptotically normal (Rao and Shao (1992)).

A variance estimator for $\bar{y}^*$ calculated based on the standard formula (e.g., Cochran (1977), Krewski and Rao (1981)) is

$$v^* = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\zeta^*_{hi} - \bar{\zeta}^*_h)^2, \tag{3.2}$$

where

$$\zeta^*_{hi} = \frac{1}{\hat{M}} \sum_j w_{hij}(y^*_{hij} - \bar{y}^*), \qquad \bar{\zeta}^*_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \zeta^*_{hi},$$

and $\sum_j$ is the summation over j with $(h,i,j) \in A$.

Note that both $\bar{y}^*$ and v^* can be computed according to (3.1)-(3.2) without knowing which sampled units are imputed values. However, v^* has a serious negative bias if the response rate p_y is low (Rubin (1987), Rao and Shao (1992)).

Let E_* and V_* be the asymptotic expectation and variance with respect to the randomness in the imputation process, and let E and V be the asymptotic expectation and variance with respect to the repeated sampling from the population and the response mechanism. Then

$$V(\bar{y}^*) = V(E_* \bar{y}^*) + EV_*(\bar{y}^*) = V(\bar{y}_r) + EV_*(\bar{y}^*), \tag{3.3}$$

where

$$\bar{y}_r = E_*(\bar{y}^*) = \sum_{A_r} w_{hij} y_{hij} \Big/ \sum_{A_r} w_{hij}$$

and

$$V_*(\bar{y}^*) = \frac{1}{\hat{M}^2} \sum_{A_m} w^2_{hij} \sum_{A_r} w_{hij}(y_{hij} - \bar{y}_r)^2 \Big/ \sum_{A_r} w_{hij}. \tag{3.4}$$

It follows from (3.3)-(3.4) that if we can identify which units are imputed values, then a substitution estimator of $V(\bar{y}^*)$ is

$$v_r + V_*(\bar{y}^*),$$

where v_r is the usual variance estimator for $\bar{y}_r$ by treating the respondents $\{y_{hij}, (h,i,j) \in A_r\}$ as the whole data set. Such an estimator is asymptotically consistent. However, neither v_r nor $V_*(\bar{y}^*)$ can be computed when imputed values are not identifiable. Therefore, we have to consider some alternatives.

It can be shown that

$$E(v^*) \approx p_y^2 V(\bar{y}_r) + EV_*(\bar{y}^*) \tag{3.5}$$

(see the proof of Theorem 1). Suppose that a consistent estimator, $\hat{p}_y$, of the response probability p_y is available (e.g., $\hat{p}_y =$ the sample proportion of respondents). Then, by (3.5), an estimator of $V(\bar{y}^*)$ can be obtained if an estimator of $EV_*(\bar{y}^*)$ can be found. For this purpose, we define

$$u^* = \frac{1-\hat{p}_y}{\hat{M}^3}\sum_A w_{hij}^2 \sum_A w_{hij}(y_{hij}^* - \bar{y}^*)^2, \tag{3.6}$$

which can be computed without identifying the imputed values. In view of

$$E_*(u^*) \approx V_*(\bar{y}^*) \tag{3.7}$$

(see the proof of Theorem 1) and (3.3) and (3.5), we obtain the following estimator of $V(\bar{y}^*)$:

$$v_S^* = \hat{p}_y^{-2}v^* + (1-\hat{p}_y^{-2})u^*. \tag{3.8}$$

This estimator is the same as v^* if $\hat{p}_y = 1$ (no nonrespondent).

In the special case where the sampling design is one stage and simple random sampling (no stratification), u^* reduces to $\frac{m(n-1)}{n^2}v^*$ and

$$v_S^* = \frac{n^2}{r^2}v^* + \Big(1-\frac{n^2}{r^2}\Big)\frac{m(n-1)}{n^2}v^* \approx \Big(\frac{n}{r}+\frac{m}{n}\Big)v^*, \tag{3.9}$$

where r is the number of respondents, m is the number of nonrespondents, and $n = r+m$. This is the correct variance estimator for $\bar{y}^*$ given in (2.1) of Rao and Shao (1992).

The proposed estimator in (3.8) may take negative values, although $v_S^* > 0$ is always true in the special case of one stage simple random sampling (see (3.9)). However, $v_S^* > 0$ holds for large sample sizes (Theorem 1) and for moderate sample sizes as well (in view of (3.5) and (3.7)). In our simulation study presented in Section 5 ($L = 32$ and $n = 75$), v_S^* is always positive in 10,000 simulation runs.

The following result shows that v_S^* is consistent.

Theorem 1. *Assume that*

C1. $n^{1+\delta}\sum_h\sum_i E|r_{hi} - E(r_{hi})|^{2+\delta} = O(1)$ *for some fixed* $\delta > 0$, *where* $r_{hi} = \sum_j \tilde{w}_{hij}a_{hij}^y y_{hij}$, $\sum_j \tilde{w}_{hij}a_{hij}^y$, *or* $\sum_j \tilde{w}_{hij}$, $\tilde{w}_{hij} = w_{hij}/M$, $n = \sum_h n_h$, *and*

$$a_{hij}^y = \begin{cases} 1 & \textit{if } y_{hij} \textit{ is observed} \\ 0 & \textit{if } y_{hij} \textit{ is missing} \end{cases}$$

C2. n(*covariance matrix of* $\sum_{A_r}\tilde{w}_{hij}y_{hij}$, $\sum_A \tilde{w}_{hij}$ *and* $\sum_{A_r}\tilde{w}_{hij}$) *has eigenvalues bounded away from* 0 *and* ∞;

C3. $\sum_A \tilde{w}_{hij}|y_{hij} - \bar{Y}|^{2+\delta} = O_p(1)$ *for some* $\delta > 0$, *where* $\bar{Y} = Y/M$ *is the population mean for item* y;

C4. $n(\max_{h,i}\sum_j \tilde{w}_{hij}) = O_p(1)$ *and* $n/N \to 0$, *where* $N = \sum_h N_h$.

Then

$$\frac{v_S^*}{V(\bar{y}^*)} \rightarrow_p 1.$$

The proof of Theorem 1 is given in the Appendix.

3.2. Multivariate case with uniform response

Survey data are usually multivariate, i.e., each ultimate unit has a vector of responses. We focus on the two-dimensional case: (y_{hij}, z_{hij}) is the response of the (h,i,j)th ultimate unit if it responds to both item y and item z. Extensions of our results to three or more dimensional cases are straightforward.

If there is no nonrespondent, then the population mean vector $(\bar{Y}, \bar{Z})$ is estimated by $(\bar{y}, \bar{z})$, where $\bar{z}$ is calculated according to (2.1) with y_{hij} replaced by z_{hij}. Note that the same survey weight w_{hij} is applied to both items y and z.

In practice a sampled ultimate unit cooperates in the survey but often fails to provide answers to some (not all) of the questions. This is referred to as item nonresponse. Define

$$A_{mm} = \{(h,i,j) \in A : \text{ both } y_{hij} \text{ and } z_{hij} \text{ are missing}\},$$

$$A_{rr} = \{(h,i,j) \in A : \text{both } y_{hij} \text{ and } z_{hij} \text{ are observed}\},$$

$$A_{rm} = \{(h,i,j) \in A : y_{hij} \text{ is observed but } z_{hij} \text{ is missing}\},$$

and

$$A_{mr} = \{(h,i,j) \in A : y_{hij} \text{ is missing but } z_{hij} \text{ is observed}\}.$$

Then all these four subsets of A may be nonempty and have appreciable sizes.

If the imputation is carried out jointly, i.e., for any unit in $A_{rm} \cup A_{mr} \cup A_{mm}$, its y and z values are imputed by using (y_{hij}, z_{hij}), $(h,i,j) \in A_{rr}$, irrespective of whether both y and z values are missing or only one of these values is missing, then the extension of the results in Section 3.1 to the multivariate case is trivial: We only need to view y_{hij} as a vector and change the squares to vector products in appropriate places. However, using joint imputation we throw away the data in $A_{rm} \cup A_{mr}$, which is not desirable. Furthermore, A_{rr} may be of a small size (which is more serious when we have higher dimensional data). Because of these considerations, in practice imputation is often carried out marginally, i.e., missing y values are imputed using the respondents y_{hij} with $(h,i,j) \in A_{rr} \cup A_{rm}$, missing z values are imputed using z_{hij} with $(h,i,j) \in A_{rr} \cup A_{mr}$, and the y and z values are imputed independently.

Marginal imputation is simple and does not require any model assumption (between y and z variables). A limitation of the marginal imputation is that it does not preserve the relation between the y and z variables so that we cannot estimate any parameter which measures how y and z are related (e.g., the correlation coefficient between the two variables). In this paper we focus on the situation where the parameter of interest is $\theta = g(\bar{Y}, \bar{Z})$, a function of the population mean vector. In such cases marginal imputation provides an asymptotically valid estimator of θ. Note that many parameters of interest in survey problems are functions of population means (e.g., $\theta = \bar{Y}/\bar{Z}$). Our method of deriving a correct variance estimator can also be applied to cases where imputation are not done marginally. But a careful study of the imputation method under consideration is required and it will not be discussed in this paper.

We still assume that there is only one imputation class and denote the response probabilities to items y and z by p_y and p_z, respectively. For item y, let $v_y^* = v^*$ in (3.2), and $u_y^* = u^*$ in (3.6). For item z, let z_{hij}^*, $\bar{z}^*$, $\hat{p}_z$, v_z^*, and u_z^* be analogs to y_{hij}^*, $\bar{y}^*$, $\hat{p}_y$, v_y^*, and u_y^*, respectively. A naive estimator of the variance-covariance matrix of $(\bar{y}^*, \bar{z}^*)$, calculated based on the standard formula, is

$$\boldsymbol{v}^* = \begin{pmatrix} v_y^* & v_{yz}^* \\ v_{yz}^* & v_z^* \end{pmatrix},$$

where

$$v_{yz}^* = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\zeta_{hi}^{y*} - \bar{\zeta}_h^{y*})(\zeta_{hi}^{z*} - \bar{\zeta}_h^{z*}),$$

$$\zeta_{hi}^{y*} = \frac{1}{\hat{M}} \sum_j w_{hij}(y_{hij}^* - \bar{y}^*), \quad \bar{\zeta}_h^{y*} = \frac{1}{n_h} \sum_{i=1}^{n_h} \zeta_{hi}^{y*}$$

and

$$\zeta_{hi}^{z*} = \frac{1}{\hat{M}} \sum_j w_{hij}(z_{hij}^* - \bar{z}^*), \quad \bar{\zeta}_h^{z*} = \frac{1}{n_h} \sum_{i=1}^{n_h} \zeta_{hi}^{z*}.$$

Similar to the estimator in (3.2), this estimator is inconsistent when $p_y < 1$ or $p_z < 1$.

A multivariate analog of v_S^* in (3.8) is

$$\boldsymbol{v}_S^* = \hat{\boldsymbol{p}}^{-1} \boldsymbol{v}^* \hat{\boldsymbol{p}}^{-1} + \boldsymbol{u}^* - \hat{\boldsymbol{p}}^{-1} \boldsymbol{u}^* \hat{\boldsymbol{p}}^{-1} \tag{3.10}$$

where

$$\hat{\boldsymbol{p}} = \begin{pmatrix} \hat{p}_y & 0 \\ 0 & \hat{p}_z \end{pmatrix} \quad \text{and} \quad \boldsymbol{u}^* = \begin{pmatrix} u_y^* & 0 \\ 0 & u_z^* \end{pmatrix}.$$

The consistency of $\boldsymbol{v}_S^*$ in (3.10) can be established similarly to the univariate case (Theorem 1). If we estimate $\theta = g(\bar{Y}, \bar{Z})$ by $\hat{\theta}^* = g(\bar{y}^*, \bar{z}^*)$, then a linearization variance estimator for $\hat{\theta}^*$ is

$$[\nabla g(\bar{y}^*, \bar{z}^*)]^t \boldsymbol{v}_S^* \nabla g(\bar{y}^*, \bar{z}^*),$$

where ∇g is the vector of partial derivatives of g.

3.3. K imputation classes

As we discussed in Section 2, imputation is usually carried out separately in several (say K) imputation classes. Within the kth class, the sampled units respond to item y and item z with the probabilities $p_y^k > 0$ and $p_z^k > 0$, respectively, $k = 1, \ldots, K$.

Assume that an imputation class label is added to each sampled unit, which is the case when imputation classes are constructed according to the value of a categorical variable observed for all the sampled units. Let $\boldsymbol{v}_S^{*k}$ be the variance estimator calculated according to (3.10) but based on the data in the kth imputation class, $k = 1, \ldots, K$. Then an asymptotically consistent variance estimator for $(\bar{y}^*, \bar{z}^*)$ is

$$\boldsymbol{v}_S^* = \sum_{k=1}^{K} \boldsymbol{v}_S^{*k}.$$

Extensions of our method to more general non-uniform response cases rely on whether we can obtain an explicit asymptotic formula for $V(\bar{y}^*, \bar{z}^*)$ and whether we can use statistics such as $\boldsymbol{v}^*$ and $\boldsymbol{u}^*$ to provide consistent estimators of the unknown quantities in $V(\bar{y}^*, \bar{z}^*)$. These extensions have to be handled case by case and will not be further discussed here.

4. Inference Based on Quantiles

For a given population $\mathcal{P}$, the population distribution for a given item y is defined to be

$$F(x) = \frac{1}{M} \sum_{(h,i,j) \in \mathcal{P}} I_{y_{hij}}(x),$$

where $I_y(x)$ is the indicator function of the set $\{y \le x\}$. If there is no missing datum, a customary estimator of $F(x)$ is

$$\hat{F}(x) = \sum_A w_{hij} I_{y_{hij}}(x) \Big/ \sum_A w_{hij}.$$

Suppose now that there are nonrespondents which are imputed by using the hot deck imputation method described in Section 3.1. We still assume that imputation is carried out independently in K imputation classes and the response probability is $p_y > 0$ for all units within an imputation class. For a concise presentation we assume $K = 1$ throughout this section. The extensions of the results to the case of any fixed K are straightforward.

Based on the imputed data set $\{y_{hij}^*, (h,i,j) \in A\}$, an estimator of $F(x)$ is

$$\hat{F}^*(x) = \sum_A w_{hij} I_{y_{hij}^*}(x) \Big/ \sum_A w_{hij}.$$

Asymptotic properties of $\hat{F}^*(x)$ for any fixed x can be derived from the results in Section 3.1 with y_{hij}^* replaced by $I_{y_{hij}^*}(x)$.

In studies of income shares or wealth distributions, an important class of population characteristics is the p-th quantile of F defined as $\theta = F^{-1}(p) = \inf\{x : F(x) \geq p\}$, $p \in (0,1)$. Another important parameter is the proportion of low income economic families. Let $\mu = F^{-1}(\frac{1}{2})$ be the population median family income. Then, the population low income proportion can be defined as $\rho = F(\frac{1}{2}\mu)$, ($\frac{1}{2}\mu$ is called the poverty line; see Wolfson and Evans (1990)).

Based on the imputed data set, survey estimators of θ (with a fixed p) and ρ are the sample p-th quantile and the sample low income proportion defined by

$$\hat{\theta}^* = (\hat{F}^*)^{-1}(p) \qquad \text{and} \qquad \hat{\rho}^* = \hat{F}^*(\tfrac{1}{2}\hat{\mu}^*), \tag{4.1}$$

respectively, where $\hat{\mu}^* = \hat{\theta}^*$ with $p = \frac{1}{2}$, the sample median.

4.1. Bahadur representation and asymptotic normality

We first establish a Bahadur representation which relates the sampling behavior of $\hat{\theta}^* - \theta$ to that of $F(\theta) - \hat{F}^*(\theta)$. Similar results for the case of no missing datum can be found in Francisco and Fuller (1991) and Shao and Wu (1992). We still adopt the asymptotic framework given in Section 2. Recall that there is a sequence of populations indexed by ν and quantiles F, θ, ρ, $\hat{F}^*$, $\hat{\theta}^*$, and $\hat{\rho}^*$ depend on ν but the index ν is omitted for simplicity.

Theorem 2. *Assume* C4 *and*

C5. *There is a sequence of functions* $\{f = f_\nu : \nu = 1, 2, \ldots\}$ *such that*

$$0 < \inf_\nu f(\theta) \leq \sup_\nu f(\theta) < \infty$$

and for any $\delta_\nu = O(n^{-1/2})$,

$$\lim_{\nu\to\infty} \Big[\frac{F(\theta + \delta_\nu) - F(\theta)}{\delta_\nu} - f(\theta)\Big] = 0.$$

Then

$$\hat{\theta}^* = \theta + \frac{F(\theta) - \hat{F}^*(\theta)}{f(\theta)} + o_p\Big(\frac{1}{\sqrt{n}}\Big) \tag{4.2}$$

and

$$\frac{\hat{\theta}^* - \theta}{\sigma_\nu(\theta)/f(\theta)} \to N(0,1) \ \text{ in distribution,} \tag{4.3}$$

where $\sigma_\nu^2(x)$ denotes the asymptotic variance of $\hat{F}^*(x)$ for any fixed x.

The proofs of Theorem 2 and the following lemma (which is used in the proof of Theorem 2) are given in the Appendix.

Lemma 1. *Assume* C4 *and* C5. *Then*

$$\sup_{|x-\theta|\leq cn^{-1/2}} |H_\nu(x)| = o_p(n^{-1/2})$$

for any constant $c > 0$, where $\theta = F^{-1}(p)$ and $H_\nu(x)$ is defined in (A.20).

We now turn to studying the asymptotic distribution of the sample low income proportion $\hat{\rho}^*$ in (4.1).

Theorem 3. *Assume that* C4 *holds and that* C5 *holds when $\theta = \mu$ (the population median) and $\theta = \frac{1}{2}\mu$. Then*

$$\hat{\rho}^* - \rho = \hat{F}^*(\tfrac{1}{2}\mu) - F(\tfrac{1}{2}\mu) + \tfrac{f(\frac{1}{2}\mu)}{2f(\mu)}[F(\mu) - \hat{F}^*(\mu)] + o_p\Big(\tfrac{1}{\sqrt{n}}\Big) \tag{4.4}$$

and

$$(\hat{\rho}^* - \rho)/\gamma_\nu \to N(0,1) \quad \textit{in distribution,} \tag{4.5}$$

where

$$\gamma_\nu^2 = \sigma_\nu^2(\tfrac{1}{2}\mu) + \sigma_\nu^2(\mu)\Big[\tfrac{f(\frac{1}{2}\mu)}{2f(\mu)}\Big]^2 - \sigma_\nu(\tfrac{1}{2}\mu, \mu)\tfrac{f(\frac{1}{2}\mu)}{f(\mu)}, \tag{4.6}$$

$\sigma_\nu^2(x)$ is the asymptotic variance of $\hat{F}^*(x)$, and $\sigma_\nu(x,t)$ is the asymptotic covariance between $\hat{F}^*(x)$ and $\hat{F}^*(t)$.

Proof. From Lemma 1 and (4.2),

$$U_\nu^* = \hat{F}^*(\tfrac{1}{2}\hat{\mu}^*) - \hat{F}^*(\tfrac{1}{2}\mu) - F(\tfrac{1}{2}\hat{\mu}^*) + F(\tfrac{1}{2}\mu) = o_p\Big(\tfrac{1}{\sqrt{n}}\Big). \tag{4.7}$$

Note that

$$\hat{\rho}^* - \rho = \hat{F}^*(\tfrac{1}{2}\mu) - F(\tfrac{1}{2}\mu) + F(\tfrac{1}{2}\hat{\mu}^*) - F(\tfrac{1}{2}\mu) + U_\nu^*.$$

Hence result (4.4) follows from (4.7) and

$$\begin{aligned} F(\tfrac{1}{2}\hat{\mu}^*) - F(\tfrac{1}{2}\mu) &= \tfrac{1}{2}f(\tfrac{1}{2}\mu)(\hat{\mu}^* - \mu) + o_p\Big(\tfrac{1}{\sqrt{n}}\Big) \\ &= \tfrac{f(\frac{1}{2}\mu)}{2f(\mu)}[F(\mu) - \hat{F}^*(\mu)] + o_p\Big(\tfrac{1}{\sqrt{n}}\Big) \end{aligned}$$

(by C4-C5 and (4.2)). Result (4.5) can be established by applying the same argument used in the proof of (4.3) and by noting that γ_ν^2 in (4.6) is the asymptotic variance of the quantity on the right side of (4.4).

4.2. Variance estimation and confidence intervals

We first focus on the sample quantiles. Result (4.3) shows that $\hat{\theta}^*$ is asymptotically normal with asymptotic mean θ and asymptotic variance $\sigma_\nu^2(\theta)/f^2(\theta)$. In inference we need to either obtain a variance estimator for $\hat{\theta}^*$ or construct a confidence interval for θ.

In the case of no missing datum, we usually start with the construction of a consistent estimator $\hat{\sigma}_\nu^2(x)$ for the asymptotic variance of $\hat{F}(x)$ with a fixed x (Francisco and Fuller (1991)). Then, we estimate the asymptotic variance of $\hat{F}(\theta)$ by $\hat{\sigma}_\nu^2 = \hat{\sigma}_\nu^2(\hat{\theta})$, $\hat{\theta} = \hat{F}^{-1}(p)$. Using the idea of Woodruff (1952) and the estimator $\hat{\sigma}_\nu^2$, we can obtain the following approximate level $1-2\alpha$ confidence interval for θ:

$$C_\nu = [\hat{F}^{-1}(p - z_\alpha \hat{\sigma}_\nu),\ \hat{F}^{-1}(p + z_\alpha \hat{\sigma}_\nu)], \tag{4.8}$$

where z_α is the $(1-\alpha)$th quantile of the standard normal distribution. A consistent estimator for the asymptotic variance of $\hat{\theta}$ is then obtained by equating the interval in (4.8) to a normal theory interval.

Because of the existence of imputed values, the above procedure does not produce correct variance estimators or confidence intervals. However, we only need to modify the estimator $\hat{\sigma}_\nu^2(x)$, using the same idea in Section 3.1. Let $\hat{\sigma}_\nu^{*2}(x) = v_S^*$ in (3.8) with y_{hij}^* replaced by $I_{y_{hij}^*}(x)$. Then, by Theorem 1,

$$\frac{\hat{\sigma}_\nu^{*2}(x)}{\sigma_\nu^2(x)} \to_p 1 \tag{4.9}$$

for any fixed x.

Theorem 4. *Assume* C4 *and* C5. *Then*

(i) $\hat{\sigma}_\nu^{*2}/\sigma_\nu^2(\theta) \to_p 1$, *where* $\hat{\sigma}_\nu^{*2} = \hat{\sigma}_\nu^{*2}(\hat{\theta}^*)$*;*

(ii) $P\{\theta \in C_\nu^*\} \to 1-2\alpha$, *where*

$$C_\nu^* = [(\hat{F}^*)^{-1}(p - z_\alpha \hat{\sigma}_\nu^*), (\hat{F}^*)^{-1}(p + z_\alpha \hat{\sigma}_\nu^*)]. \tag{4.10}$$

Proof. The result in part (i) is not a direct consequence of result (4.9). We need to show that $\hat{\sigma}_\nu^{*2}(x)$ has a continuity property for x near θ. That is,

$$n[\hat{\sigma}_\nu^{*2}(\theta + \delta_\nu) - \hat{\sigma}_\nu^{*2}(\theta)] = O_p(\delta_\nu) \tag{4.11}$$

for any $\delta_\nu = O(n^{-1/2})$. Result (4.11) can be proved using the same argument as that in the proof of Theorem 5 in Shao (1994). The result in part (ii) can be proved by using the result in part (i) (see the proof of Theorem 6 in Shao (1994)).

By equating the interval C_ν^* in (4.10) to a normal theory interval based on (4.3), an estimator of the asymptotic variance of $\hat{\theta}^*$, $\sigma_\nu^2(\theta)/f^2(\theta)$, can be obtained as

$$v_W^*(\alpha) = \Big[\frac{(\hat{F}^*)^{-1}(p + z_\alpha \hat{\sigma}_\nu^*) - (\hat{F}^*)^{-1}(p - z_\alpha \hat{\sigma}_\nu^*)}{2z_\alpha}\Big]^2. \tag{4.12}$$

It is not easy to choose the value α in (4.12). In terms of some limited empirical evidence, $\alpha = 0.05$ is suggested by Sitter (1992).

An alternative estimator of $\sigma_\nu^2(\theta)/f^2(\theta)$ is obtained by directly estimating $\sigma_\nu^2(\theta)$ and $f^{-2}(\theta)$:

$$v_S^* = \hat{\sigma}_\nu^{*2}(\hat{\theta}^*)\Big[\frac{\hat{F}^*(\hat{\theta}^* + n^{-1/2}) - \hat{F}^*(\hat{\theta}^* - n^{-1/2})}{2n^{-1/2}}\Big]^{-2}. \tag{4.13}$$

By Theorem 4, both $v_W^*(\alpha)$ and v_S^* are consistent.

Next, we consider the estimation of γ_ν^2 in (4.6), the asymptotic variance of the sample low income proportion. Examining (4.6), a substitution estimator of γ_ν^2 is

$$\hat{\gamma}_\nu^{*2} = \hat{\sigma}_\nu^{*2}(\tfrac{1}{2}\hat{\mu}^*) + \hat{\sigma}_\nu^{*2}(\hat{\mu}^*)[\hat{f}(\tfrac{1}{2}\hat{\mu}^*)\hat{f}^{-1}(\hat{\mu}^*)/2]^2 - \hat{\sigma}_\nu^*(\tfrac{1}{2}\hat{\mu}^*,\, \hat{\mu}^*)\hat{f}(\tfrac{1}{2}\hat{\mu}^*)\hat{f}^{-1}(\hat{\mu}^*),$$

where

$$\hat{f}(x) = \sqrt{n}[\hat{F}^*(x + n^{-1/2}) - \hat{F}^*(x - n^{-1/2})]/2$$

is an estimator of $f(x)$,

$$\hat{\sigma}_\nu^*(x,t) = \hat{p}_y^{-2} v^*(x,t) + (1 - \hat{p}_y^{-2})u^*(x,t),$$

is an estimator of $\sigma_\nu(x,t)$,

$$v^*(x,t) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\zeta_{hi}^* - \bar{\zeta}_h^*)(x)(\zeta_{hi}^* - \bar{\zeta}_h^*)(t),$$

$$\zeta_{hi}^*(x) = \frac{1}{\hat{M}} \sum_j w_{hij}[I_{y_{hij}^*}(x) - \hat{F}^*(x)], \qquad \bar{\zeta}_h^*(x) = \frac{1}{n_h}\sum_{i=1}^{n_h} \zeta_{hi}^*(x),$$

$$u^*(x,t) = \frac{1 - \hat{p}_y}{\hat{M}^3} \sum_A w_{hij}^2 \sum_A w_{hij}(I_{y_{hij}^*} - \hat{F}^*)(x)(I_{y_{hij}^*} - \hat{F}^*)(t),$$

and $\hat{\sigma}_\nu^{*2}(x) = \hat{\sigma}_\nu(x,x)$ is the same as that in (4.9).

Theorem 5. *Assume that* C4 *holds and that* C5 *holds with $\theta = \mu$ and $\theta = \frac{1}{2}\mu$. Then*

$$\frac{\hat{\gamma}_\nu^{*2}}{\gamma_\nu^2} \to_p 1.$$

Proof. We have shown the consistency of $\hat{\sigma}_\nu^{*2}(\frac{1}{2}\hat{\mu}^*)$ and $\hat{\sigma}_\nu^{*2}(\hat{\mu}^*)$. Following the proof of Theorem 4, we can show that

$$n[\hat{\sigma}_\nu^*(\tfrac{1}{2}\hat{\mu}^*,\, \hat{\mu}^*) - \sigma_\nu(\tfrac{1}{2}\mu,\, \mu)] = o_p(1).$$

Then the result follows from the consistency of $\hat{f}(\frac{1}{2}\hat{\mu}^*)$ and $\hat{f}(\hat{\mu}^*)$ under C4 and C5.

Based on Theorems 3 and 5, an approximate level $1 - 2\alpha$ confidence interval for ρ is $[\hat{\rho}^* - z_\alpha \hat{\gamma}_\nu^*, \hat{\rho}^* + z_\alpha \hat{\gamma}_\nu^*]$.

5. Simulation Results

In this section we present the results from a simulation study comparing the true asymptotic variance and our variance estimator in the stratified one stage simple random sampling case.

The population we used is similar to those in Kovar, Rao and Wu (1988) and Sitter (1992). There are $L = 32$ strata in the population. In the hth stratum, the y-values of the population were generated according to

$$y_{hi} \overset{i.i.d.}{\sim} N(\bar{Y}_h, \sigma_h^2), \quad i = 1, \ldots, N_h,$$

where the population parameters N_h, $\bar{Y}_h$, and σ_h are given in Table 1.

Table 1. Population parameters and sample sizes

h	N_h	$\bar{Y}_h$	σ_h	h	N_h	$\bar{Y}_h$	σ_h
1	38	8.6	4.00	17	34	8.6	0.25
2	38	8.7	4.00	18	34	8.4	0.25
3	38	8.5	4.00	19	34	8.5	0.25
4	38	8.3	4.00	20	34	8.8	0.25
5	38	8.9	4.00	21	34	8.4	0.25
6	38	8.8	4.00	22	22	8.7	1.00
7	38	8.2	4.00	23	22	8.6	1.00
8	38	8.6	4.00	24	22	8.5	1.00
9	38	8.6	4.00	25	22	8.4	1.00
10	38	8.4	4.00	26	22	8.8	1.00
11	38	8.4	4.00	27	22	8.9	1.00
12	34	8.5	0.25	28	22	8.3	1.00
13	34	8.1	0.25	29	22	8.2	1.00
14	34	8.4	0.25	30	22	8.9	1.00
15	34	8.3	0.25	31	22	8.4	1.00
16	34	8.6	0.25	32	22	8.6	1.00

After the population was generated, a simple random sample of size n_h was drawn from stratum h, independently across the 32 strata. Two sets of sample sizes n_h were considered: (1) $n_h = 3$ when $h = 1, \ldots, 11$ and $n_h = 2$ when $h = 12, \ldots, 32$; (2) $n_h = 6$ when $h = 1, \ldots, 11$ and $n_h = 4$ when $h = 12, \ldots, 32$. (The purpose of using the second set of sample sizes is to see the effect of large sample sizes, as suggested by a referee.) After the samples were generated, the respondents $\{y_{hi}, (h, i) \in A_r\}$ were obtained by assuming that the sampled units responded with equal probability p_y; and the missing values $\{y_{hi}, (h, i) \in A_m\}$ were imputed by taking an i.i.d. sample from $\{y_{hi}, (h, i) \in A_r\}$, with selection probability $w_{hi}/\sum_{A_r} w_{hi}$ for y_{hi}, $(h, i) \in A_r$, where the survey weight $w_{hi} = w_h = N_h/n_h$ in this special case. This process was repeated 10,000 times in the simulation.

All the computations were done in a UNIX at the Department of Statistics, University of Wisconsin-Madison, using IMSL subroutines GENNOR, IGNUIN and GENUNF for random number generations.

5.1. Inference based on the sample mean $\bar{y}^*$

In each simulation iteration, we calculated the following statistics based on the imputed data set:

$$\bar{y}_h^* = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}^*, \quad \bar{y}^* = \sum_{h=1}^{L}\frac{N_h}{N}\bar{y}_h^*$$

(note that $\hat{M} = N$ in the stratified one stage case),

$$v^* = \sum_{h=1}^{L}\frac{w_h^2 n_h}{N^2(n_h-1)}\sum_{i=1}^{n_h}(y_{hi}^* - \bar{y}_h^*)^2,$$

$$u^* = \Big(1 - \frac{r}{n}\Big)\sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{w_h^2}{N^2}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{w_h}{N}\Big(y_{hi}^* - \bar{y}^*\Big)^2,$$

and v_S^* according to (3.8) with $\hat{p}_y = r/n$, where r is the number of respondents.

Table 2 lists, for some values of p_y, the variance of $\bar{y}^*$ (approximated by the sample variance of the 10,000 simulated values of $\bar{y}^*$), and the relative bias (RB) and mean square error (MSE) of v_S^* (based on 10,000 simulated values of v_S^*). In addition, Table 2 also lists the empirical coverage probabilities (NCP and TCP) of 95% confidence intervals, where NCP is the coverage probability of the confidence interval obtained by treating $(\bar{y}^* - \bar{Y})/\sqrt{v_S^*}$ as the standard normal random variable, whereas TCP is the coverage probability of the confidence interval obtained by treating $(\bar{y}^* - \bar{Y})/\sqrt{v_S^*}$ as the t-random variable with r degrees of freedom.

The results in Table 2 indicates that the variance estimator v_S^* performs well. Its relative bias is under 3% for all cases considered. The coverage probabilities of the confidence intervals are close to the nominal level, especially for the case of larger sample size.

Table 2. Simulation result for the sample mean

Sample Sizes	p_y	$V(\bar{y}^*)$	RB(%)	MSE	NCP(%)	TCP(%)
$n_h = 3,\quad h = 1,\ldots,11$	0.4	0.2864	0.99	0.1913	93.42	94.35
$n_h = 2,\quad h = 12,\ldots,32$	0.5	0.2368	-2.18	0.1277	93.43	94.36
	0.6	0.1888	0.09	0.0928	94.26	94.78
	0.7	0.1592	-2.02	0.0687	93.91	94.48
	0.8	0.1308	-0.17	0.0510	94.13	94.57
	0.9	0.1072	0.80	0.0386	94.25	94.63
$n_h = 6,\quad h = 1,\ldots,11$	0.4	0.1410	-0.05	0.0591	94.36	94.77
$n_h = 4,\quad h = 12,\ldots,32$	0.5	0.1111	2.55	0.0413	94.91	95.26
	0.6	0.0918	1.05	0.0301	94.86	95.24
	0.7	0.0769	1.23	0.0222	94.85	95.22
	0.8	0.0647	-0.20	0.0167	94.73	94.94
	0.9	0.0552	-2.62	0.0127	94.20	94.36

5.2. Inference based on the sample median $\hat{\theta}^*$

For the sample median, we computed $\hat{\sigma}_\nu^{*2}(x) = v_S^*$ with y_{hi}^* replaced by $I_{y_{hi}}^*(x)$,

$$\hat{F}^*(x) = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{w_h}{N} I_{y_{hi}^*}(x),$$

and the v_S^* defined in (4.13).

Table 3 lists, for some values of p_y, the asymptotic variance of $\hat{\theta}^*$, and the RB and MSE of v_S^*. Table 3 also lists the empirical coverage probabilities (NCP) of the 95% confidence interval

$$C_\nu^* = [(\hat{F}^*)^{-1}(p - z_{.025}\hat{\sigma}_\nu^*),\ (\hat{F}^*)^{-1}(p + z_{.025}\hat{\sigma}_\nu^*)], \tag{5.1}$$

and the empirical coverage probabilities (TCP) of the interval obtained by replacing $z_{0.05}$ in (5.1) with the 97.5% percentile of the t-distribution with r degrees of freedom.

Table 3. Simulation result for the sample median

Sample Sizes	p_y	$V(\hat{\theta}^*)$	RB(%)	MSE	NCP(%)	TCP(%)
$n_h = 3, \quad h = 1, \ldots, 11$	0.4	0.0374	4.37	0.0291	91.73	92.96
$n_h = 2, \quad h = 12, \ldots, 32$	0.5	0.0279	3.59	0.0195	92.77	93.98
	0.6	0.0219	1.23	0.0147	93.54	94.11
	0.7	0.0167	−2.17	0.0116	94.25	94.42
	0.8	0.0131	−8.30	0.0087	94.36	95.13
	0.9	0.0095	−4.93	0.0066	95.02	95.34
$n_h = 6, \quad h = 1, \ldots, 11$	0.4	0.0154	4.05	0.0140	93.73	93.96
$n_h = 4, \quad h = 12, \ldots, 32$	0.5	0.0118	−0.49	0.093	93.77	93.98
	0.6	0.0088	−1.41	0.0062	94.00	94.21
	0.7	0.0072	−7.91	0.0045	94.50	94.82
	0.8	0.0051	0.33	0.0033	94.66	95.10
	0.9	0.0039	0.52	0.0024	95.32	95.52

The results in Table 3 indicates that the variance estimator v_S^* performs well. Its relative bias is under 5% except for two cases where the relative bias is about 8%. When $p_y = 0.4$, the performances of confidence intervals are not very good in the smaller sample size case, but are acceptable in the larger sample size case.

We also computed the RB and MSE for the variance estimator v_W^* in (4.12). But its performance is not as good as v_S^*. Details are not reported here.

Acknowledgement

The research was partially supported by National Sciences Foundation Grant DMS-9504425 and National Security Agency Grant MDA904-96-1-0066.

Appendix

Proof of Theorem 1. By C2 and

$$v_S^* = \frac{p_y^2}{\hat{p}_y^2} \frac{(v^* - u^*)}{p_y^2} + u^*,$$

we only need to show that

$$n[u^* - EV_*(\bar{y}^*)] \to_p 0 \tag{A.1}$$

and

$$n[v^* - u^* - p_y^2 V(\bar{y}_r)] \to_p 0. \tag{A.2}$$

Under C3 and C4,

$$\begin{aligned} E_*(u^*) &= \frac{(1-\hat{p}_y)M^2}{\hat{M}^2} \Big(\sum_A \tilde{w}_{hij}^2\Big)\Big(1 - \sum_{A_m} \tilde{w}_{hij}^2\Big) \sum_{A_r} w_{hij}(y_{hij} - \bar{y}_r)^2 \Big/ \sum_{A_r} w_{hij} \\ &= \frac{(1-\hat{p}_y)M^2}{\hat{M}^2} \Big(\sum_A \tilde{w}_{hij}^2\Big) \sum_{A_r} w_{hij}(y_{hij} - \bar{y}_r)^2 \Big/ \sum_{A_r} w_{hij} + o_p\Big(\frac{1}{n}\Big). \end{aligned}$$

By (3.4) and the fact that $\hat{M}/M \to_p 1$ and

$$(1-\hat{p}_y)\Big(\sum_{A}\tilde{w}_{hij}^2\Big)\Big/\sum_{A_m}\tilde{w}_{hij}^2 \to_p 1,$$

we conclude that

$$n[E_*(u^*)-V_*(\bar{y}^*)] \to_p 0. \tag{A.3}$$

Also, by C3, $n[V_*(\bar{y}^*)-EV_*(\bar{y}^*)] \to_p 0$. Hence (A.1) follows from

$$n[u^*-E_*(u^*)] \to_p 0. \tag{A.4}$$

Note that

$$\begin{aligned} n[u^*-E_*(u^*)] &= O_p(1)\Big[\sum_{A_n}\tilde{w}_{hij}(y_{hij}^*-\bar{y}^*)^2 - E_*\sum_{A_n}\tilde{w}_{hij}(y_{hij}^*-\bar{y}^*)^2\Big] \\ &= O_p(1)\Big[\sum_{A_n}\tilde{w}_{hij}(y_{hij}^*-\bar{Y})^2 - E_*\sum_{A_n}\tilde{w}_{hij}(y_{hij}^*-\bar{Y})^2 \\ &\quad -(\bar{y}^*-\bar{Y})^2+E_*(\bar{y}^*-\bar{Y})^2\Big]. \end{aligned}$$

Since $\bar{y}^*-\bar{Y}\to_p 0$ and $E_*(\bar{y}^*-\bar{Y})^2 = V_*(\bar{y}^*)+(\bar{y}_r-\bar{Y})^2 = o_p(1)$, (A.4) follows from

$$\sum_{A}\tilde{w}_{hij}(y_{hij}^*-\bar{Y})^2 - E_*\sum_{A}\tilde{w}_{hij}(y_{hij}^*-\bar{Y})^2 \to_p 0,$$

which follows from the Law of Large Numbers (e.g., Krewski and Rao (1981), Lemma 1) under C3. This proves (A.1).

For (A.2), it suffices to show that

$$n[E_*(v^*-u^*)-p_y^2V(\bar{y}_r)] \to_p 0 \tag{A.5}$$

and

$$n[v^*-u^*-E_*(v^*-u^*)] \to_p 0. \tag{A.6}$$

Define

$$\begin{aligned} u_{hi} &= \sum_{j:(h,i,j)\in A}\tilde{w}_{hij}a_{hij}^y y_{hij}, & \bar{u}_h &= \frac{1}{n_h}\sum_{i=1}^{n_h}u_{hi}, \\ v_{hi} &= \sum_{j:(h,i,j)\in A}\tilde{w}_{hij}a_{hij}^y, & \bar{v}_h &= \frac{1}{n_h}\sum_{i=1}^{n_h}v_{hi}. \end{aligned}$$

A straightforward calculation shows that

$$\begin{aligned} V(\bar{y}_r) = \frac{1}{p_y^2}\Big[&V\Big(\sum_{A}\tilde{w}_{hij}a_{hij}^y y_{hij}\Big)+\bar{Y}^2V\Big(\sum_{A}\tilde{w}_{hij}a_{hij}^y\Big) \\ &-2\bar{Y}Cov\Big(\sum_{A}\tilde{w}_{hij}a_{hij}^y y_{hij},\sum_{A}\tilde{w}_{hij}a_{hij}^y\Big)\Big] \end{aligned}$$

and

$$\begin{aligned}E_*(v^*-u^*) &= \frac{M^2}{\hat{M}^2}\Big[\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(u_{hi}-\bar{u}_h)^2+\bar{y}_r\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(v_{hi}-\bar{v}_h)^2\\ &\quad -2\bar{y}_r\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(u_{hi}-\bar{u}_h)(v_{hi}-\bar{v}_h)\Big]+V_*(\bar{y}^*)-E_*u^*\\ &=\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(u_{hi}-\bar{u}_h)^2+\bar{Y}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(v_{hi}-\bar{v}_h)^2\\ &\quad -2\bar{Y}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(u_{hi}-\bar{u}_h)(v_{hi}-\bar{v}_h)+o_p\Big(\frac{1}{n}\Big),\end{aligned}$$

where the last equation follows from (A.3), $M/\hat{M}\to_p 1$ and $\bar{y}_r-\bar{Y}\to_p 0$. Hence (A.5) follows from

$$n\Big[\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(u_{hi}-\bar{u}_h)^2-V\Big(\sum_A\tilde{w}_{hij}a_{hij}^y y_{hij}\Big)\Big]\to_p 0,$$

$$n\Big[\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(v_{hi}-\bar{v}_h)^2-V\Big(\sum_A\tilde{w}_{hij}a_{hij}^y\Big)\Big]\to_p 0,$$

and

$$n\Big[\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(u_{hi}-\bar{u}_h)(v_{hi}-\bar{v}_h)-Cov\Big(\sum_A\tilde{w}_{hij}a_{hij}^y y_{hij},\sum_A\tilde{w}_{hij}a_{hij}^y\Big)\Big]\to_p 0,$$

which are consequences of C1 and the Law of Large Numbers.

By (A.4), (A.6) follows from

$$n(v^*-E_*v^*)\to_p 0. \tag{A.7}$$

Let

$$\gamma_{n1}=\frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\Big[\sum_j\tilde{w}_{hij}(y_{hij}^*-\bar{y}_r)-\frac{1}{n_h}\sum_{i=1}^{n_h}\sum_j\tilde{w}_{hij}(y_{hij}^*-\bar{y}_r)\Big]^2,$$

$$\begin{aligned}\gamma_{n2}=\frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\Big[&\sum_j\tilde{w}_{hij}(y_{hij}^*-\bar{y}_r)\\ &-\frac{1}{n_h}\sum_{i=1}^{n_h}\sum_j\tilde{w}_{hij}(y_{hij}^*-\bar{y}_r)\Big]\Big[\sum_j\tilde{w}_{hij}-\frac{1}{n_h}\sum_{i=1}^{n_h}\sum_j\tilde{w}_{hij}\Big],\end{aligned}$$

and

$$\gamma_{n3} = \frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\Big(\sum_j \tilde{w}_{hij} - \frac{1}{n_h}\sum_{i=1}^{n_h}\sum_j \tilde{w}_{hij}\Big)^2.$$

Then

$$v^* = \gamma_{n1} + 2\gamma_{n2}(\bar{y}_r - \bar{y}^*) + \gamma_{n3}(\bar{y}_r - \bar{y}^*)^2.$$

Define

$$t^*_{hi} = \sum_j \tilde{w}_{hij}(1-a^y_{hij})(y^*_{hij} - \bar{y}_r) \quad \text{and} \quad \bar{t}^*_h = \frac{1}{n_h}\sum_{i=1}^{n_h} t^*_{hi}.$$

Then

$$\begin{aligned}
\gamma_{n1} &= \frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}[(t^*_{hi} + u_{hi} - \bar{y}_r v_{hi}) - (\bar{t}^*_h + \bar{u}_h - \bar{y}_r\bar{v}_h)]^2 \\
&= \frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}[(u_{hi} - \bar{u}_h) - \bar{y}_r(v_{hi} - \bar{v}_h)]^2 \\
&\quad + \frac{2M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}[(u_{hi} - \bar{u}_h) - \bar{y}_r(v_{hi} - \bar{v}_h)](t^*_{hi} - \bar{t}^*_h) \\
&\quad + \frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(t^*_{hi} - \bar{t}^*_h)^2 \\
&= E_*(v^*) - V_*(\bar{y}^*) + \frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(t^*_{hi} - \bar{t}^*_h)^2 \\
&\quad + \frac{2M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}[(u_{hi} - \bar{u}_h) - \bar{y}_r(v_{hi} - \bar{v}_h)](t^*_{hi} - \bar{t}^*_h),
\end{aligned}$$

where the last equation follows from

$$E_*(v^*) = V_*(\bar{y}^*) + \frac{M^2}{\hat{M}^2}\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}[(u_{hi} - \bar{u}_h) - \bar{y}_r(v_{hi} - \bar{v}_h)]^2.$$

From Rao and Shao (1992),

$$n\Big[\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(t^*_{hi} - \bar{t}^*_h)^2 - V_*(\bar{y}^*)\Big] \to_p 0$$

and

$$n\sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}[(u_{hi} - \bar{u}_h) - \bar{y}_r(v_{hi} - \bar{v}_h)](t^*_{hi} - \bar{t}^*_h) \to_p 0.$$

Thus,

$$n[\gamma_{n1} - E_*(v^*)] \to_p 0. \tag{A.8}$$

By C4,

$$\gamma_{n3} \leq \frac{M^2}{\hat{M}^2} \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \Big(\sum_j \tilde{w}_{hij} \Big)^2 = O_p\Big(\frac{1}{n}\Big),$$

which and (A.8) imply that

$$2|\gamma_{n2}| \leq \gamma_{n1} + \gamma_{n3} = O_p\Big(\frac{1}{n}\Big).$$

Since $E_*\bar{y}^* = \bar{y}_r$ and $V_*(\bar{y}^*) = O_p(\frac{1}{n})$, we have $\bar{y}^* - \bar{y}_r \to_p 0$. This proves (A.7).

Proof of Lemma 1. Let $\eta_{\nu\ell} = \theta + cn^{-3/2}\ell$, $\ell = -n, \ldots, n$. Then

$$\begin{aligned} \sup_{|x-\theta| \leq cn^{-1/2}} |H_\nu(x)| &\leq \max_{-n \leq \ell \leq n} |H_\nu(\eta_{\nu\ell})| + \max_{-n \leq \ell \leq n} |F(\eta_{\nu(\ell+1)}) - F(\eta_{\nu\ell})| \\ &= \max_{-n \leq \ell \leq n} |H_\nu(\eta_{\nu\ell})| + O(n^{-3/2}), \end{aligned}$$

where the last equality follows from C5. Thus, it suffices to show that

$$\max_{-n \leq \ell \leq n} |H_\nu(\eta_{\nu\ell})| = o_p(n^{-1/2}). \tag{A.9}$$

Let

$$G_\nu^r(x) = \frac{1}{M} \sum_{A_r} w_{hij} I_{y_{hij}}(x) \quad \text{and} \quad q_\nu^r = G_\nu^r(\infty).$$

Then

$$E_*[\hat{F}^*(x)] = G_\nu^r / q_\nu^r = \hat{F}_r. \tag{A.10}$$

Let $H_\nu^*(x) = \hat{F}^*(x) - \hat{F}^*(\theta) - \hat{F}_r(x) + \hat{F}_r(\theta)$, and $H_\nu^r(x) = G_\nu^r(x) - G_\nu^r(\theta) - q_\nu^r[F(x) - F(\theta)]$. Then $H_\nu(x) = H_\nu^*(x) + H_\nu^r(x)/q_\nu^r$. Under the uniform response assumption,

$$q_\nu^r - p_y = o_p(1). \tag{A.11}$$

Hence (A.9) follows from

$$\max_{-n \leq \ell \leq n} |H_\nu^*(\eta_{\nu\ell})| = o_p(n^{-1/2}) \quad \text{and} \quad \max_{-n \leq \ell \leq n} |H_\nu^r(\eta_{\nu\ell})| = o_p(n^{-1/2}). \tag{A.12}$$

Let P_* be the probability under the random imputation. Then by Bernstein's Inequality and C4,

$$P_*\Big\{\sqrt{n}|H_\nu^*(\eta_{\nu\ell})| \geq \epsilon\Big\} \leq 2\exp\Big\{-\frac{\epsilon^2 n^{-1}}{2V_*[\hat{F}^*(\eta_{\nu\ell}) - \hat{F}^*(\theta)] + O_p(n^{-3/2})}\Big\}. \tag{A.13}$$

for any $\epsilon > 0$. Let I_{hij} be the indicator of $\{\theta \leq y^*_{hij} \leq \eta_{\nu\ell}\}$ if $\ell > 0$ and the indicator of $\{\eta_{\nu\ell} \leq y^*_{hij} \leq \theta\}$ if $\ell < 0$. Since

$$\begin{aligned} V_*[\hat{F}^*(\eta_{\nu\ell}) - \hat{F}^*(\theta)] &= \frac{1}{\hat{M}^2} \sum_{A_m} w^2_{hij} V_*(I_{hij}) \\ &\leq \frac{1}{\hat{M}^2} \sum_{A_m} w^2_{hij} E_*(I_{hij}) \\ &= \frac{1}{M^2 \hat{M}^2} \sum_{A_m} w^2_{hij} \sum_{A_r} w_{hij} I_{hij} \Big/ \sum_{A_r} w_{hij} \\ &\leq \max_{h,i} \sum_j \tilde{w}_{hij} |\hat{F}(\eta_{\nu\ell}) - \hat{F}(\theta)| / q^r_\nu, \end{aligned}$$

we obtain that

$$\max_{-n \leq \ell \leq n} V_*[\hat{F}^*(\eta_{\nu\ell}) - \hat{F}^*(\theta)] \leq O(n^{-1}) |\hat{F}(\theta + cn^{-1/2}) - \hat{F}(\theta - cn^{-1/2})| = O_p(n^{-3/2})$$

by C4-C5 and (A.11), and

$$P_*\Big\{\sqrt{n} \max_{-n \leq \ell \leq n} |H^*_\nu(\eta_{\nu\ell})| \geq \epsilon\Big\} \leq 4n \exp\Big\{ - \frac{\epsilon^2 n^{-1}}{O_p(n^{-3/2})}\Big\} = o_p(1)$$

by (A.13). This proves the first assertion in (A.12). Similarly, since $E[H^r_\nu(x)] = 0$, we obtain

$$P\Big\{\sqrt{n} |H^r_\nu(\eta_{\nu\ell})| \geq \epsilon\Big\} \leq 2 \exp\Big\{ - \frac{\epsilon^2 n^{-1}}{2V[H^r_\nu(\eta_{\nu\ell})] + O(n^{-3/2})}\Big\} \tag{A.14}$$

for any $\epsilon > 0$, and

$$\begin{aligned} \max_{-n \leq \ell \leq n} V[H^r_\nu(\eta_{\nu\ell})] &= \max_{-n \leq \ell \leq n} \frac{1}{M^2} \sum_{h=1}^{L} \sum_{i=1}^{n_h} V\Big(\sum_j w_{hij} \delta_{hij}\Big) \\ &\leq \max_{-n \leq \ell \leq n} \frac{1}{M^2} \sum_{h=1}^{L} \sum_{i=1}^{n_h} E\Big(\sum_j w_{hij} \delta_{hij}\Big)^2 \\ &\leq \max_{-n \leq \ell \leq n} \max_{h,i} \sum_j \tilde{w}_{hij} E\Big(\frac{1}{M} \sum_A w_{hij} |\delta_{hij}|\Big) \\ &\leq O(n^{-1}) \max_{-n \leq \ell \leq n} |F(\eta_{\nu\ell}) - F(\theta)| \\ &= O(n^{-3/2}), \end{aligned}$$

where $\delta_{hij} = I_{hij} - [F(\eta_{\nu\ell}) - F(\theta)]$ if $\ell > 0$ and $\delta_{hij} = I_{hij} - [F(\theta) - F(\eta_{\nu\ell})]$ if $\ell < 0$. This, and (A.14), imply the second assertion in (A.12).

Proof of Theorem 2. Define $\zeta_\nu(t) = \sqrt{n}[F(\theta + tn^{-1/2}) - \hat{F}^*(\theta + tn^{-1/2})]/f(\theta)$. Then (4.2) is the same as $\sqrt{n}(\hat{\theta}^* - \theta) - \zeta_\nu(0) = o_p(1)$, which is implied by

$$P\{\sqrt{n}(\hat{\theta}^* - \theta) \le t, \zeta_\nu(0) \ge t + \epsilon\} \to 0 \text{ and } P\{\sqrt{n}(\hat{\theta}^* - \theta) \ge t + \epsilon,\, \zeta_\nu(0) \le t\} \to 0 \tag{A.15}$$

for any fixed $t \ne 0$ and $\epsilon > 0$ (Lemma 1 of Ghosh (1971)). Since

$$\{\sqrt{n}(\hat{\theta}^* - \theta) \ge t\} = \{\hat{F}^*(\hat{\theta}^*) \ge \hat{F}^*(\theta + tn^{-1/2})\} = \{\zeta_\nu(t) \ge \eta_\nu(t)\},$$

where $\eta_\nu(t) = \sqrt{n}[F(\theta + tn^{-1/2}) - \hat{F}^*(\hat{\theta}^*)]/f(\theta)$, (A.15) is equivalent to

$$P\{\zeta_\nu(t) \le \eta_\nu(t),\, \zeta_\nu(0) \ge t + \epsilon\} \to 0 \text{ and } P\{\zeta_\nu(t+\epsilon) \ge \eta_\nu(t+\epsilon),\, \zeta_\nu(0) \le t\} \to 0,$$

which is implied by

$$\eta_\nu(t) - t = o_p(1) \tag{A.16}$$

and

$$\zeta_\nu(t) - \zeta_\nu(0) = o_p(1). \tag{A.17}$$

By C5,

$$\sqrt{n}\frac{F(\theta + tn^{-1/2}) - F(\theta)}{f(\theta)} \to t. \tag{A.18}$$

By C4 and the boundedness of $f(\theta)$,

$$\Big|\sqrt{n}\frac{F(\theta) - \hat{F}^*(\hat{\theta}^*)}{f(\theta)}\Big| \le \frac{\sqrt{n}}{f(\theta)}\Big[\frac{1}{M} + \frac{1}{\hat{M}}\max_{h,i,j}\frac{w_{hij}}{M}\Big] = O_p\Big(\frac{1}{\sqrt{n}}\Big). \tag{A.19}$$

Then (A.16) follows from (A.18)-(A.19) and the definition of $\eta_\nu(t)$. Note that

$$\zeta_\nu(0) - \zeta_\nu(t) = \sqrt{n}H_\nu(\theta + tn^{-1/2})/f(\theta),$$

where

$$H_\nu(x) = \hat{F}^*(x) - \hat{F}^*(\theta) - F(x) + F(\theta). \tag{A.20}$$

Then (A.17) is a consequence of Lemma 1 and the proof for (4.2) is completed.

Note that

$$E_*[\hat{F}^*(x)] = \sum_{A_r} w_{hij} I_{y_{hij}}(x) \Big/ \sum_{A_r} w_{hij} = \hat{F}_r(x).$$

Hence

$$(\hat{F}^* - F)(x) = (\hat{F}_r - F)(x) + \sum_{A_m} w_{hij}[I_{y^*_{hij}} - E_*(I_{y^*_{hij}})](x) \Big/ \hat{M}. \tag{A.21}$$

Applying the Central Limit Theorem for stratified sampling (Krewski and Rao (1981), Bickel and Freedman (1984) and using C4 and the delta method, we obtain that

$$[\hat{F}_r(\theta) - F(\theta)]/\sigma_{1\nu}(\theta_\nu) \rightarrow N(0,1) \text{ in distribution,} \tag{A.22}$$

where $\sigma^2_{1\nu}(x)$ is the asymptotic variance of $\hat{F}_r(x)$. Similarly, conditional on y_{hij}, $(h,i,j) \in A$,

$$\sum_{A_m} w_{hij}[I_{y^*_{hij}} - E_*(I_{y^*_{hij}})](\theta)\Big/\hat{M}\sigma_{2\nu}(\theta) \rightarrow N(0,1) \text{ in distribution,} \tag{A.23}$$

where $\sigma^2_{2\nu}(x) = (1-p_y)G_\nu(x)[1-G_\nu(x)]\sum_A w^2_{hij}$ and $G_\nu(x) = \hat{M}\hat{F}(x)/M$. It follows from (A.21)-(A.23) and Lemma 1 in Schenker and Welsh (1988) that

$$\frac{F(\theta) - \hat{F}^*(\theta)}{\sigma_\nu(\theta)} \rightarrow N(0,1) \text{ in distribution,}$$

where $\sigma^2_\nu(x) = \sigma^2_{1\nu}(x) + \sigma^2_{2\nu}(x)$ is the asymptotic variance of $\hat{F}^*(x)$. This and (4.2) imply (4.3).

References

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12**, 470-482.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. Wiley, New York.

Fay, R. E. (1991). A design-based perspective on missing data variance. *Proc. 7th Annual Res. Conf.* Bureau of the Census, Washington, D.C., 429-440.

Fay, R. E. (1993). Valid inferences from imputed survey data. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, 41-48.

Francisco, C. A. and Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Ann. Statist.***19**, 454-469.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42**, 1957-1961.

Griffin, R. A., Navarro, A. and Flores-Baez, L. (1991). Disclosure avoidance for the 1990 census. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 516-521.

Kalton, G. (1981). *Compensating for Missing Data*, ISR research report series, Survey Research Center, University of Michigan.

Kovar, J. G., Rao, J. N. K. and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates *Canadian J. Statist.* **16**, Supplement, 25-45.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* **9**, 1010-1019.

Rao, J. N. K. (1993). Linearization variance estimators under imputation for missing data. Technical Report, Laboratory for Research in Statistics and Probability, Carleton University.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *J. Amer. Statist. Assoc.* **91**, 499-506.

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811-822.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81**, 366-374.

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16**, 1550-1566.

Sedransk, J. (1985). The objective and practice of imputation. In *Proc. First Annual Res. Conf.* 445-452.

Shao, J. (1994). L-statistics in complex surveys. *Ann. Statist.* **22**, 946-967.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *J. Amer. Statist. Assoc.* **91**, 1278-1288.

Shao J. and Wu, C. F. J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Ann. Statist.* **20**, 1571-1593.

Sitter, R. R. (1992). A resampling procedure for complex survey data, *J. Amer. Statist. Assoc.* **87**, 755-765.

Wolfson, M. C. and Evans, J. M. (1990). Statistics Canada low income cut-offs, methodological concerns and possibilities. Discussion Paper, Statistics Canada.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.* **47**, 635-646.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.
E-mail: shao@stat.wisc.edu

(Received May 1997; accepted July 1998)