

## INFERENCE WITH COMPLEX SURVEY DATA IMPUTED BY HOT DECK WHEN NONRESPONDENTS ARE NONIDENTIFIABLE

Y. Chen and J. Shao<sup>1</sup>

### 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 seriously 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, the 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 nonrespondents. 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 nonrespondents are nonidentifiable; and the survey estimators under consideration are functions of sample means, or sample quantiles.

**KEY WORDS:** Identification flags; Item nonresponse; Single imputation; Stratified multistage sampling; Uniform response.

### RÉSUMÉ

L'imputation hot deck pour les non-répondants est souvent utilisée dans les sondages. C'est une pratique usuelle que de traiter les valeurs imputées comme si elles étaient des vraies valeurs, et de calculer les estimateurs et leurs variances estimées par les formules standards. Les estimateurs de variance, cependant, ont de sérieux biais négatifs lorsque le taux de non-réponse est élevé. Les méthodes comme l'imputation multiple et le jackknife ont été proposés pour l'obtention de meilleurs estimateurs de variance. Cependant, l'imputation multiple nécessite que plusieurs bases de données multiples soient générées et maintenues et que les procédures d'imputation soient appropriées. De plus, le jackknife ajusté nécessite des drapeaux indicateurs pour repérer les non-répondants. Cependant, dans la pratique, un des problèmes rencontré est celui où il n'est pas possible de distinguer les non-répondants dans un groupe de données imputées car des drapeaux indicateurs pour ceux-ci n'ont pas été établis. Dans cet article, nous obtenons des procédures d'inférence asymptotiques robustes par rapport au plan dans les situations où un plan de sondage stratifié à plusieurs degrés est utilisé pour la collecte des données d'enquête; l'imputation hot deck est appliquée pour former une base unique de données imputées n'ayant aucun indicateur pouvant identifier des non-répondants; les estimateurs considérés sont des fonctions de la moyenne échantillonnale ou des quantiles de l'échantillon.

**MOTS CLÉS:** Drapeaux indicateurs; non-réponse à une question; imputation simple; échantillonnage stratifié à plusieurs degrés; réponse uniforme.

### 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

---

<sup>1</sup> Yinzhong Chen, and Jun Shao, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706.

set. This important property is not shared by some deterministic imputation methods such as mean imputation, ratio imputation, and 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 seriously 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 nonrespondents.

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 nonrespondents (no identification flag). The reason why we consider the case of unknown response status is the following. Many public data sets do not carry identification flags for nonrespondents. 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 nonrespondents are not identifiable is when confidentiality edit is applied to the data set for confidential reasons (Griffin, Navarro and Flores-Baez, 1991). Confidentiality edit is done by selecting a portion of the data and interchanging them with a random sample from the respondents. If we treat the selected data for interchange as “non-respondents” (which are imputed by a random sample

from “respondents”), then these nonrespondents cannot be identified. The interchanging in confidentiality edit can also be done by using data from another data set, which will be discussed later.

Under a general stratified multistage sampling design (see Section 2), we propose some variance estimators based on two types of most commonly used estimators in surveys: (1) the sample mean or a function of sample means (Section 3); and (2) the sample quantiles (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  $h$ -th stratum,  $h = 1, \dots, L$ . Suppose that  $n_h \geq 2$  first-stage units are sampled from stratum  $h$ , independently across the strata. Within the  $h$ -th stratum, the  $(h, i)$ -th first-stage unit is selected with probability  $p_{hi} > 0$ ,  $i = 1, \dots, N_h$ . 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  $j$ -th ultimate unit in the  $i$ -th first-stage cluster of stratum  $h$ ,  $i = 1, \dots, n_h$ ,  $h = 1, \dots, L$ . 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 focus on the common case where  $L$  is large, all  $n_h$  are small (*i.e.*,  $n_h$  are bounded by a fixed integer), and the sampling fractions  $n_h/N_h$  are negligible.

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 from the population 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

$$\bar{y} = \hat{Y}/\hat{M}, \quad (2.1)$$

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  $\{P_v, v = 1, 2, \dots\}$ , where each  $P_v$  contains  $L_v$  strata. The population under consideration is then viewed as a member of this sequence of populations. Note that  $L, N_h, w_{hij}$ , and  $y_{hij}$ , etc. depend on  $v$ , but  $v$  is omitted for simplicity. We assume that the  $n_h$  are bounded and that  $L \rightarrow \infty$  as  $v \rightarrow \infty$ . All limiting processes are understood to be as  $v \rightarrow \infty$ .

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

$$\bar{y}^* = \sum_A w_{hij} y_{hij}^* / \hat{M}. \quad (3.1)$$

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

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

where

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

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) and (3.2) without knowing which sampled units are nonrespondents. However,  $v^*$  has a seriously 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

$$V(\bar{y}^*) = V(E_* \bar{y}^*) + E V_* (\bar{y}^*) = V(\bar{y}_r) + E V_* (\bar{y}^*), \quad (3.3)$$

where

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

and

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

It follows from (3.3) and (3.4) that if we can identify which units are nonrespondents, 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 nonrespondents are not identifiable. Therefore, we have to consider some alternatives.

It can be shown that

$$E(v^*) \approx p_y^2 V(\bar{y}_r) + E V_* (\bar{y}^*). \quad (3.5)$$

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

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

which can be computed without identifying the nonrespondents. In view of

$$E_*(u^*) = V_*(\bar{y}^*) \quad (3.7)$$

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

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

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

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

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

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. Its proof is omitted.

**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  $\sum_j \tilde{w}_{hij}$ ,  $\tilde{w}_{hij} = w_{hij}/M$ ,  $n = \sum_{h=1}^L n_h$ , and  $a_{hij}^y = 1$  if  $y_{hij}$  is observed and  $= 0$  otherwise;

C2.  $n$  (covariance matrix of  $\sum_A \tilde{w}_{hij} y_{hij}$ ,  $\sum_A \tilde{w}_{hij}$  and  $\sum_A \tilde{w}_{hij}$ ) have eigenvalues bounded away from

0 and  $\infty$ ;

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.  $n (\max_{h,i} \sum_j \tilde{w}_{hij}) = O_p(1)$ .

Then

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

### 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). However, 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 (e.g.,  $\theta = (\bar{Y}/\bar{Z})$ ). In such cases, marginal imputation provides an asymptotically valid estimator of  $\theta$ .

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

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

where

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

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

and

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

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

$$v_S^* = \hat{p}^{-1} v^* \hat{p}^{-1} + u^* - \hat{p}^{-1} u^* \hat{p}^{-1}, \quad (3.10)$$

where

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

If we estimate  $\theta = g(\bar{Y}, \bar{Z})$  by  $\hat{\theta}^* = g(\bar{y}^*, \bar{z}^*)$ , then a variance estimator for  $\hat{\theta}^*$  is

$$[\nabla g(\bar{y}^*, \bar{z}^*)]' v_S^* \nabla g(\bar{y}^*, \bar{z}^*),$$

where  $\nabla_g$  is the vector of partial derivatives of  $g$ .

### 3.3 Multiple imputation classes

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

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  $v_S^{*k}$  be the variance estimator calculated according to (3.10) but based on the data in the  $k$ -th imputation class,  $k = 1, \dots, K$ . Then an asymptotically consistent variance estimator for  $(\bar{y}^*, \bar{z}^*)$  is

$$v_S^* = \sum_{k=1}^K v_S^{*k}.$$

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  $v^*$  and  $u^*$  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.

### 3.4 Confidentiality edit

Confidentiality edit is carried out by selecting a random sample  $\gamma = \{y_{hij}, (h, i, j) \in A_c\}$  from the original sample  $\{y_{hij}, (h, i, j) \in A\}$ ,  $A_c \subset A$ , and replacing the values in  $\gamma$  by  $\{y_{hij}^*, (h, i, j) \in A_c\}$ . If we treat the values in  $\gamma$  as “nonrespondents”, then we have a uniform response mechanism.

If  $\{y_{hij}^*, (h, i, j) \in A_c\}$  is a random sample from the “respondents”  $\{y_{hij}, (h, i, j) \notin A_c\}$ , then the situation is exactly the same as that in Section 3.1 and Theorem 1 applies. If  $\{y_{hij}^*, (h, i, j) \in A_c\}$  is a random sample from another data set, say  $\{x_{hij}, w_{hij}\}$ , then  $\bar{y}^*$  is a valid estimator of  $\bar{Y}$  if and only if  $\bar{X}$ , the population mean for item  $x$ , is nearly the same as  $\bar{Y}$ . If  $\bar{X} \approx \bar{Y}$ , then it can be shown that a consistent estimator of  $V(\bar{y}^*)$  is  $v^*$  in (3.2). Details are omitted.

## 4. INFERENCE BASED ON QUANTILES

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

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

where  $I_y(x)$  is the indicator function of the set  $\{y \leq x\}$ .

If there is no missing datum, a customary estimator of  $F(x)$  is

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

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  $p_v > 0$  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

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

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  $\theta = F^{-1}(p) = \inf\{x: F(x) \geq p\}$ ,  $p \in (0, 1)$ . Based on the imputed data set, a survey estimator of  $\theta$  (with a fixed  $p$ ) is the sample  $p$ -th quantile defined by

$$\hat{\theta}^* = (\hat{F}^*)^{-1}(p). \quad (4.1)$$

We first establish a Bahadur representation which relates the sampling behaviour 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  $v$  and quantiles  $F, \theta, \rho, \hat{F}^*, \hat{\theta}^*$ , and  $\hat{\rho}^*$  depend on  $v$  but the index  $v$  is omitted for simplicity.

**Theorem 2.** Assume C4 and

C5. There is a sequence of functions  $\{f = f_v; v = 1, 2, \dots\}$  such that  $0 < \inf_v f(\theta) \leq \sup_v f(\theta) < \infty$  and for any  $\delta_v = O(n^{-1/2})$ ,

$$\lim_{v \rightarrow \infty} \left[ \frac{F(\theta + \delta_v) - F(\theta)}{\delta_v} - f(\theta) \right] = 0.$$

Then

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

and

$$\frac{\hat{\theta}^* - \theta}{\sigma_v(\theta)/f(\theta)} \rightarrow N(0, 1) \text{ in distribution,} \quad (4.3)$$

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

Result (4.3) shows that  $\hat{\theta}^*$  is asymptotically normal with asymptotic mean  $\theta$  and asymptotic variance  $\sigma_v^2(\theta)/f^2(\theta)$ . In inference we need to either obtain a variance estimator for  $\hat{\theta}^*$  or construct a confidence interval for  $\theta$ .

In the case of no missing datum, we usually start with the construction of a consistent estimator  $\hat{\sigma}_v^2(x)$  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  $\hat{\sigma}_v^2 = \hat{\sigma}_v^2(\hat{\theta})$ ,  $\hat{\theta} = \hat{F}^{-1}(p)$ . Using the idea of Woodruff (1952) and the estimator  $\hat{\sigma}_v^2$ , we can obtain the following approximate level  $1 - 2\alpha$  confidence interval for  $\theta$ :

$$C_v = [\hat{F}^{-1}(p - z_\alpha \hat{\sigma}_v), \hat{F}^{-1}(p + z_\alpha \hat{\sigma}_v)], \quad (4.4)$$

where  $z_\alpha$  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 (4.4) 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  $\hat{\sigma}_v^2(x)$ , using the same idea in Section 3.1. Let  $\hat{\sigma}_v^{*2}(x) = v_s^*$  in (3.8) with  $y_{hij}^*$  replaced by  $I_{y_{hij}^*}(x)$ . Then, by Theorem 1,

$$\frac{\hat{\sigma}_v^{*2}(x)}{\sigma_v^2(x)} \rightarrow_p 1 \quad (4.5)$$

for any fixed  $x$ .

**Theorem 3.** Assume C4 and C5. Then

(i)  $\hat{\sigma}_v^{*2} / \sigma_v^2(\theta) \rightarrow_p 1$ , where  $\hat{\sigma}_v^{*2} = \hat{\sigma}_v^{*2}(\hat{\theta}^*)$ ;

(ii)  $P\{\theta \in C_v^*\} \rightarrow 1 - 2\alpha$ , where

$$C_v^* = [(\hat{F}^*)^{-1}(p - z_\alpha \hat{\sigma}_v^*), (\hat{F}^*)^{-1}(p + z_\alpha \hat{\sigma}_v^*)]. \quad (4.6)$$

By equating the interval  $C_v^*$  in (4.6) to a normal theory interval based on (4.3), an estimator of the asymptotic variance of  $\hat{\theta}^*$ ,  $\sigma_v^{*2}(\theta)/f^2(\theta)$ , can be

obtained as

$$v_w^*(\alpha) = \left[ \frac{(\hat{F}^*)^{-1}(p+z_\alpha \hat{\sigma}_v^*) - (\hat{F}^*)^{-1}(p-z_\alpha \hat{\sigma}_v^*)}{2z_\alpha} \right]^2. \quad (4.7)$$

It is not easy to choose the value  $\alpha$  in (4.7). In terms of some limited empirical evidence,  $\alpha = 0.05$  is suggested by Sitter (1992).

An alternative estimator of  $\sigma_v^2(\theta)/f^2(\theta)$  is obtained by directly estimating  $\sigma_v^2(\theta)$  and  $f^2(\theta)$ :

$$v_s^* = \hat{\sigma}_v^{*2}(\hat{\theta}^*) \left[ \frac{\hat{F}^*(p+n^{-1/2}) - \hat{F}^*(p-n^{-1/2})}{2n^{-1/2}} \right]^{-2}. \quad (4.8)$$

By Theorem 3, both  $v_w^*(\alpha)$  and  $v_s^*$  are consistent.

## 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). There are  $L=32$  strata in the population. In the  $h$ -th stratum, the  $y$ -values of the population were generated according to

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

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

After the population was generated, a simple random sample of size  $n_h$  was drawn from stratum  $h$ , independently across the 32 strata. The sample sizes  $n_h$  are also listed in Table 1. 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 on a UNIX at the Department of Statistics, University of Wisconsin-Madison, using IMSL subroutines GENNOR, IGUNIN and GENUF for random number generations.

Table 1. Population Parameters and Sample Sizes

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

### 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),

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

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

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 indicate 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.

Table 2. Simulation Results for the Sample Mean

$p_y$	$V(\bar{y}^*)$	RB(%)	MSE	NCP(%)	TCP(%)
0.4	0.2864	0.99	0.1913	93.42	94.35
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

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

For the sample median, we computed  $\hat{\sigma}_v^{*2}(x) = v_s^*$  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 (4.8).

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

$$C_v^* = [(\hat{F}^*)^{-1}(p - z_{0.025} \hat{\sigma}_v^*), (\hat{F}^*)^{-1}(p + z_{0.025} \hat{\sigma}_v^*)], \quad (5.1)$$

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.

The results in Table 3 indicate that the variance estimator  $v_s^*$  performs well. Its relative bias is under 4% for all cases considered. The performances of confidence intervals are not as good as in the sample mean case when  $p_y$  is small (which is reasonable since the median is more difficult to estimate), but are still acceptable.

We also computed the RB and MSE for the variance estimator  $v_w^*$  in (4.7). But its performance is not as good as  $v_s^*$ . Details are not reported here.

Table 3. Simulation Results for the Sample Median

$p_y$	$V(\hat{\theta}^*)$	RB(%)	MSE	NCP(%)	TCP(%)
0.4	0.0104	3.30	0.0042	91.77	92.67
0.5	0.0084	0.66	0.0026	92.87	93.62
0.6	0.0068	0.49	0.0018	93.56	94.22
0.7	0.0058	-2.85	0.0013	94.06	94.46
0.8	0.0047	-1.11	0.0010	94.74	95.14
0.9	0.0039	-1.18	0.0007	95.32	95.63

### REFERENCES

- Bickel, P. J., and Freedman, D.A. (1984). "Asymptotic normality and the bootstrap in stratified sampling", *The Annals of Statistics*, 12, 470-482.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition, New York: Wiley.
- Francisco, C.A., and Fuller, W.A. (1991). "Quantile estimation with a complex survey design", *The Annals of Statistics*, 19, 454-469.
- Ghosh, J.K. (1971). "A new proof of the Bahadur representation of quantiles and an application", *The Annals of Mathematical Statistics*, 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, Ann Arbor: 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 Journal of Statistics*, 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", *The Annals of Statistics*, 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., 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*, New York: Wiley.
- Rubin, D.B., and Schenker, N. (1986). "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse", *Journal of the American Statistical Association*, 81, 366-374.
- Schenker, N., and Welsh, A.H. (1988). "Asymptotic results for multiple imputation", *The Annals of Statistics*, 16, 1550-1566.
- Sedransk, J. (1985). "The objective and practice of imputation", in *Proceedings of the First Annual Research Conference*, U.S. Bureau of the Census, 445-452.
- Shao, J., and Sitter, R.R. (1996). "Bootstrap for imputed survey data", *Journal of the American Statistical Association*, 91, 1278-1288.
- Shao, J., and Wu, C.F.J. (1992). "Asymptotic properties of the balanced repeated replication method for sample quantiles", *The Annals of Statistics*, 20, 1571-1593.
- Sitter, R.R. (1992). "A resampling procedure for complex survey data", *Journal of the American Statistical Association*, 87, 755-765.
- Woodruff, R.S. (1952). "Confidence intervals for medians and other position measures", *Journal of the American Statistical Association*, 47, 635-646.