Statistica Sinica **18**(2008), 281-297

CONFIDENCE INTERVALS BASED ON SURVEY DATA WITH NEAREST NEIGHBOR IMPUTATION

Jun Shao and Hansheng Wang

University of Wisconsin–Madison and Peking University

Abstract: Nearest neighbor imputation (NNI) is a popular method used to compensate for item nonresponse in sample surveys. Although previous results showed that the NNI sample mean and quantiles are consistent estimators of the population mean and quantiles, large sample inference procedures, such as asymptotic confidence intervals for the population mean and quantiles, are not available. For the population mean, we establish the asymptotic normality of the NNI sample mean and derive a consistent estimator of its limiting variance, which leads to an asymptotically valid confidence interval. For the quantiles, we obtain consistent variance estimators and asymptotically valid confidence intervals using a Bahadur-type representation for NNI sample quantiles. Some limited simulation results are presented to examine the finite-sample performance of the proposed variance estimators and confidence intervals.

Key words and phrases: Bahadur representation, hot deck, mean, quantiles, variance estimation.

1. Introduction

Consider a bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$ with observed $y_1, \ldots, y_r$ (respondents), missing $y_{r+1}, \ldots, y_n$ (nonrespondents), and observed $x_1, \ldots, x_n$. In sample surveys, imputation is commonly applied to compensate for this type of item nonresponse (Sedransk (1985), Kalton and Kasprzyk (1986) and Rubin (1987)). The nearest neighbor imputation (NNI) method imputes a missing y_j by y_i, where $1 \leq i \leq r$ and i is the nearest neighbor of j measured by the x-variable, i.e., i satisfies $|x_i - x_j| = \min_{1 \leq l \leq r} |x_l - x_j|$. We focus on the case where x has a continuous distribution so that there are no tied x-values. In practice, if there are tied x-values due to reasons such as rounding, NNI can be applied by randomly selecting a j from the nearest neighbors from tied x-values. Also, NNI is often carried out by first dividing the sample into several "imputation classes" and then finding nearest neighbors within each imputation class.

The NNI method has some nice features. First, it imputes a nonrespondent by a respondent from the same variable; the imputed values are actually occurring values, not constructed values and, while they may not be perfect substitutes, they are unlikely to be nonsensical. Second, the NNI method may be more efficient than other popular methods, such as mean imputation and random hot deck imputation, when the x-variable provides useful auxiliary information. Third, the NNI method does not assume a parametric regression model between y and x and, hence, it is more robust against model violations than methods such as ratio imputation and regression imputation that are based on a linear regression model. Finally, under some conditions, NNI estimators (i.e., estimators calculated using standard formulas and treating nearest neighbor imputed values as observed data) are asymptotically valid not only for moments of the y-variable, but also for the distribution and quantiles of the y-variable, which is an advantage over other non-random imputation methods (such as the mean, ratio, and regression imputation) that lead to valid moment estimators only.

There are other nonparametric imputation methods (see, e.g., Cheng (1994) and Wang and Rao (2002)) that are more efficient than NNI, although they impute nonrespondents by constructed values. However, the NNI method has a long history of applications in such surveys as Census 2000 and the Current Population Survey conducted by the U.S. Census Bureau (Farber and Griffin (1998) and Fay (1999)), the Job Openings and Labor Turnover Survey and the Employee Benefits Survey conducted by the U.S. Bureau of Labor Statistics (Montaquila and Ponikowski (1993)), and the Unified Enterprise Survey, the Survey of Household Spending, and the Financial Farm Survey conducted by Statistics Canada (Rancourt (1999)). In these agencies, it is unlikely that NNI will be replaced by another nonparametric imputation method in the near future.

Therefore, a theoretical study of the properties of NNI is important. Although NNI is the same as regression imputation using k-nearest neighbor regression with $k = 1$, the existing theoretical (asymptotic) results for k-nearest neighbor regression (see, e.g., Härdle (1990)) are all for the case where $k \to \infty$ as the sample size increases. Theoretical studies of the NNI methodology started with Lee, Rancourt and Särndal (1994) and Rancourt (1999), who showed some properties of NNI estimators when y_i and x_i are assumed to follow a simple linear regression model. The first theoretical work on NNI under a general nonparametric setting is Chen and Shao (2000), who established the consistency of NNI estimators such as the sample mean and sample quantile; Chen and Shao (2001) investigated jackknife variance estimators for the sample mean. In practice, statistical inference, such as setting an approximate confidence interval for a population parameter, is often needed. Asymptotic results on confidence intervals based on survey data with NNI, however, are not available, and the purpose of this paper is to fill this gap.

The most important population parameter in surveys is the population mean (or a function of several population means). Although empirical results (e.g., Chen and Shao (2001)) showed that a confidence interval of the form

$$\text{NNI sample mean} \pm z_\alpha \sqrt{\text{variance estimator for NNI sample mean}} \tag{1}$$

works well, where α is a fixed nominal confidence level and z_α is the $100(1-\alpha/2)$th normal percentile, the use of (1) lacks statistical justification, since it has *not* been shown that the interval in (1) is asymptotically valid in the sense that

$$P(\text{confidence interval covers the true population parameter}) \to 1-\alpha \qquad (2)$$

(under some limiting process as the sample size increases). After an introduction to some notation and assumptions in Section 2, we establish (2) in Section 3 by first showing the asymptotic normality of NNI sample means, and then finding a variance estimator in (1) that is consistent for the variance in the limiting distribution of the NNI sample mean.

Estimation or inference on population quantiles has become more and more important in modern survey statistics (Rao, Kovar and Mantel (1990) and Francisco and Fuller (1991)). For income variables, for example, the median income or other quantiles are as important as the mean income. In children with cystic fibrosis, the 10th percentiles of height and weight are important clinical boundaries between healthy and possibly nutritionally compromised patients (Kosorok (1999)). Although Chen and Shao (2000) showed that NNI sample quantiles are consistent for population quantiles, variance estimation for NNI sample quantiles was not discussed. Note that the jackknife cannot be directly applied to sample quantiles (Efron (1982)). In Section 4, using some Bahadur-type representations for NNI sample quantiles, we show the asymptotic normality of NNI sample quantiles, and provide consistent variance estimators for NNI sample quantiles and asymptotically valid confidence intervals (in the sense of (2)) for population quantiles.

To complement the theoretical results, some simulations are presented in Section 5 to examine the performance of the proposed estimators and confidence intervals. Some discussion is in the last section.

2. Notation and Assumption

This section introduces some notation and general assumptions used throughout the paper. Let $\mathcal{P}$ be a finite population containing M units indexed by i. A sample $\mathcal{S}$ of size n is taken from $\mathcal{P}$ according to some sampling design. Let w_i be the survey weight for unit i, the inverse of the probability that unit i is selected. For any set of values $\{z_i : i \in \mathcal{P}\}$,

$$E_s\left(\sum_{i\in\mathcal{S}} w_i z_i\right) = \sum_{i\in\mathcal{P}} z_i \qquad (3)$$

(i.e., $\sum_{i\in\mathcal{S}} w_i z_i$ is a Horvitz-Thompson-type estimator), where E_s is the expectation with respect to sampling. Although w_i is defined for any $i \in \mathcal{P}$, we only need w_i for $i \in \mathcal{S}$ in applications.

We consider one-stage sampling without clusters. Some discussion of cluster sampling and multistage sampling is given in the last section.

To consider asymptotics, we assume that the finite population $\mathcal{P}$ is a member of a sequence of finite populations indexed by ν. All limiting processes in this paper are understood to be as $\nu \to \infty$. As $\nu \to \infty$, the population size M and the sample size n increase to infinity. In sample surveys, the following two regularity conditions on w_i's are typically imposed:

$$\max_{i \in \mathcal{P}} \frac{n w_i}{M} \le b_0, \tag{4}$$

$$\frac{n}{M^2} \mathrm{Var}_s \Big(\sum_{i \in \mathcal{S}} w_i \Big) \le b_1, \tag{5}$$

where b_0 and b_1 are some positive constants, and Var_s is the variance with respect to sampling. Condition (4) ensures that none of the weights w_i is disproportionately large (see Krewski and Rao (1981)). Condition (5) means that $\mathrm{Var}_s(\sum_{i \in \mathcal{S}} w_i/M)$ is at most of the order n^{-1}. Conditions (4)-(5) are satisfied for stratified simple random sampling designs.

Note that (3) and (5) imply that $\sum_{i \in \mathcal{S}} w_i/M \to_p 1$, where $\to_p$ is convergence in probability as $\nu \to \infty$. Furthermore, $E_s\left(\sum_{i \in \mathcal{S}} w_i^4\right) = \sum_{i \in \mathcal{P}} w_i^3 \le (M b_0/n)^3$ under condition (4). Hence, it follows from the Liapunov Central Limit Theorem that

$$\Big(\sum_{i \in \mathcal{S}} \frac{w_i}{M} - 1 \Big) \Big/ \sqrt{\mathrm{Var}_s \Big(\sum_{i \in \mathcal{S}} \frac{w_i}{M} \Big)} \to_d N(0,1), \tag{6}$$

where $\to_d$ is convergence in distribution as $\nu \to \infty$.

Let (x, y) be a bivariate characteristic from a given unit in the population, where y is the main variable of interest and x is a covariate. Let a be the response indicator, i.e., $a = 1$ if y is observed and $a = 0$ if y is a nonrespondent. Note that we define a for every unit in $\mathcal{P}$ (see, e.g., Shao and Steel (1999)). For unit $i \in \mathcal{P}$, we denote (x, y, a) by (x_i, y_i, a_i). To study asymptotic validity of NNI, we need some assumptions.

Assumption A. Each finite population $\mathcal{P}$ is divided into K (a fixed integer) imputation classes $\mathcal{P}_k$, $k = 1, \ldots, K$, such that the population and sample sizes of each imputation class increase to infinity and, within imputation class k, the (x_i, y_i, a_i) are independent and identically distributed (i.i.d.) from a superpopulation with $P(a_i = 1 | x_i, y_i, k) = P(a_i = 1 | x_i, k)$, and $P(a_i = 1 | k) = p_k > 0$. Sampling is independent of the superpopulation and the (x_i, y_i, a_i) from different imputation classes are independent. NNI is carried out within each imputation class.

Throughout this paper, the probability, expectation, and variance with respect to sampling and the superpopulation in Assumption A are denoted by P,

E, and Var , respectively. When $\mathcal{S}$ is not present, however, P, E, and Var reduce to the probability, expectation, and variance with respect to the superpopulation only. For example, $E(y|x)$ is the conditional expectation of y given x with respect to the superpopulation; $E(\sum_{i\in\mathcal{S}} y_i)$ is the expectation with respect to both sampling and the superpopulation. Furthermore, i.i.d. is always with respect to the superpopulation.

The assumption on the response probability means that the response indicator a is independent of y, given the covariate x and the imputation class k. That is, within an imputation class, the response mechanism is covariate-dependent (Little (1995)) or unconfounded (Lee, Rancourt and Särndal (1994)), an assumption made for the validity of many other popular imputation methods. Although the (x_i, y_i, a_i) within an imputation class are assumed i.i.d., the response mechanism is still not completely at random, since $P(a = 1|x)$ depends on the covariate x. In particular, the conditional distribution of (x, y) given $a = 1$ may be different from the conditional distribution of (x, y) given $a = 0$.

Imputation classes are usually constructed using a categorical variable whose values are observed for all sampled units; for example, under stratified sampling, strata or unions of strata are often used as imputation classes. Each imputation class should contain a large number of sampled units. When there are many strata of small size, imputation classes are often obtained through poststratification (Valliant (1993)) and/or by combining small strata.

The superpopulation assumption on (x, y, a) within each imputation class is natural, since NNI requires some exchangeability of units within each imputation class. The NNI method is a model-based approach, rather than a design-based approach. However, the model assumption is nonparametric (i.e., we only assume that (x, y, a)'s are i.i.d. within each imputation class) and is much weaker than a parametric linear model assumption on the conditional mean of y given the covariate x, which is typically assumed for a regression-type imputation.

Let $\psi(x)$ be a continuous, bounded, and strictly increasing function of x. Then $E(y|x) = E(y|\psi(x))$. Since the NNI method recovers information about nonrespondents using x-covariates through $E(y|x)$, without loss of generality we can always apply the transformation $\psi(x)$. Hence, we assume in the rest of this paper that x is bounded, say $-\infty < x_- \leq x \leq x_+ < \infty$. However, we do not need to know the values of x_- and x_+.

Assumption B. For any fixed imputation class k, $P(a = 1|x, k) > 0$ for all $x \in [x_-, x_+]$ and is a continuous function of x; conditional on a, x has a bounded and continuous Lebesgue density $f_{k,a}$; $E(y|x, k)$ is Lipschitz continuous of x.

In imputation class k, let $\mathcal{R}_k$ be the set of indices of y-respondents, $\bar{\mathcal{R}}_k$ be the set of indices of nonrespondents, and $\mathcal{S}_k = \mathcal{R}_k \cup \bar{\mathcal{R}}_k$. Define $\mu_k(x) = E(y|x, k)$, $\mu_{k,a} = E(y|k, a)$, $\mu_k = E(y|k) = p_k\mu_{k,1} + (1-p_k)\mu_{k,0}$, and $\mu = E(y) =$

$\sum_{k=1}^K (M_k/M)\mu_k$, where M_k is the size of $\mathcal{P}_k$ and $M = \sum_k M_k$. Conditional on $\mathcal{R}_k$ and $\mathcal{X}_k = \{x_i : i \in \mathcal{R}_k\}$ (the covariates from the respondents), let

$$q_{k,i} = P\left(|x - x_i| = \min_{j\in\mathcal{R}_k} |x - x_j| \middle| a = 0, k, \mathcal{R}_k, \mathcal{X}_k, \mathcal{S}_k \right)$$

be the probability that $i \in \mathcal{R}_k$ will be selected as the nearest neighbor for a nonrespondent within $\bar{\mathcal{R}}_k$, where P is with respect to x conditional on $a = 0, k, \mathcal{R}_k, \mathcal{X}_k$, and $\mathcal{S}_k$.

When there is only one imputation class ($K = 1$), the subscript k on $\mathcal{S}_k$, $\mathcal{R}_k$, $\mathcal{X}_k$, p_k, μ_k, $\mu_{k,a}$, $q_{k,i}$, $f_{k,a}$, etc., will be dropped.

3. Confidence Intervals for Means

Without nonresponse, the superpopulation mean μ and the finite population mean $\bar{Y} = M^{-1}\sum_{i\in\mathcal{P}} y_i$ are estimated by a Horvitz-Thompson estimator $\sum_{i\in\mathcal{S}} w_i y_i/M$. After NNI, our estimator of μ or $\bar{Y}$ is

$$\hat{\mu} = \frac{1}{M}\sum_k \left(\sum_{i\in\mathcal{R}_k} w_i y_i + \sum_{i\in\bar{\mathcal{R}}_k} w_i \tilde{y}_i \right) = \sum_k \frac{M_k}{M}\left(\sum_{i\in\mathcal{R}_k} \bar{w}_{k,i} y_i + \sum_{i\in\bar{\mathcal{R}}_k} \bar{w}_{k,i}\tilde{y}_i \right), \quad (7)$$

where $\tilde{y}_i$ is the imputed value for the nonrespondent y_i, $i \in \bar{\mathcal{R}}_k$, and $\bar{w}_{k,i} = w_i/M_k$ when $i \in \mathcal{S}_k$. The estimator $\hat{\mu}$ will be referred to as the NNI sample mean, although it is a weighted average of respondents and imputed values. In some cases, M is unknown. From (6), $\hat{M} = \sum_{i\in\mathcal{S}} w_i$ is a consistent estimator of M. We can estimate μ or $\bar{Y}$ by a ratio estimator

$$\frac{1}{\hat{M}}\sum_k \left(\sum_{i\in\mathcal{R}_k} w_i y_i + \sum_{i\in\bar{\mathcal{R}}_k} w_i \tilde{y}_i \right) = \frac{\hat{\mu}}{\frac{\hat{M}}{M}}.$$

An example is the (ratio) estimator given by (20) in Section 4. The asymptotic property of this estimator can be obtained using (6), the result for $\hat{\mu}$, and the delta-method. Similarly, if the parameter of interest is a differentiable function of several population means, the point estimator is the same function of sample means and its asymptotic property can be derived using the delta-method. Thus, in what follows, we focus on the asymptotic property of $\hat{\mu}$.

Note that, conditional on $\mathcal{S}$ and all observed (y_i, x_i), imputed values within imputation class k are i.i.d. taking the value y_i with probability $q_{k,i}$, $i \in \mathcal{R}_k$. Hence,

$$E(\hat{\mu}|\text{sample and respondents}) = \sum_k \frac{M_k}{M}\left(\sum_{i\in\mathcal{R}_k} \bar{w}_{k,i} y_i + \sum_{i\in\bar{\mathcal{R}}_k} \bar{w}_{k,i} \sum_{i\in\mathcal{R}_k} q_{k,i} y_i \right).$$

Applying part (iii) of the following lemma to each imputation class, we conclude that $\sum_{i\in\mathcal{R}_k} q_{k,i}y_i$ converges to $\mu_{k,0}$ (the mean of y-nonrespondents in imputation class k), which shows how NNI recovers information about y-nonrespondents using y-respondents and x-values, under Assumptions A−B.

Lemma 1. *Suppose that Assumptions A-B hold with $K = 1$.*

(i) *If g is a function of x with $E[g(x_i)]^2 < \infty$ then, for any $m = 1, 2, \ldots,$*

$$E\left[r^{m-1}\sum_{i\in\mathcal{R}} q_i^m g(x_i)\right] - \frac{(m+1)!}{2^m}E\left[\frac{g(x_i)f_0^{m-1}(x_i)}{f_1^{m-1}(x_i)}\middle| a_i = 0\right] \to 0, \qquad (8)$$

where r is the size of $\mathcal{R}$ (the number of respondents).

(ii) *If $E[g(x_i)]^4 < \infty$, then, for any $m = 1, 2, \ldots,$*

$$E\left\{r^{m-1}\sum_{i\in\mathcal{R}} q_i^m g(x_i) - \frac{(m+1)!}{2^m}E\left[\frac{g(x_i)f_0^{m-1}(x_i)}{f_1^{m-1}(x_i)}\middle| a_i = 0\right]\right\}^2 \to 0. \qquad (9)$$

(iii) *If $E(y_i^{4l}) < \infty$ for a positive integer l, then*

$$\sum_{i\in\mathcal{R}} q_i y_i^l \to_p E(y_i^l|a_i = 0). \qquad (10)$$

The proof of Lemma 1 is in Shao and Wang (2007). The following is a heuristic argument on why NNI and any other type of regression imputation can use the value of x to impute a missing y and produce an almost unbiased estimator of μ, under Assumption A. Assume that $K = 1$ and $\mu(x)$ is a known function. Then a missing y_i is imputed as $\mu(x_i)$ and the resulting estimator of μ is $\tilde{\mu} = \sum_{i\in\mathcal{S}}[\bar{w}_i a_i y_i + \bar{w}_i(1 - a_i)\mu(x_i)]$. Under Assumption A,

$$\begin{aligned}
E(\tilde{\mu}|x_1, \ldots, x_n) &= \sum_{i\in\mathcal{S}} \bar{w}_i[E(a_i y_i|x_i) + E((1 - a_i)\mu(x_i)|x_i)]\\
&= \sum_{i\in\mathcal{S}} \bar{w}_i[E(a_i|x_i)E(y_i|x_i) + E(1 - a_i|x_i)\mu(x_i)]\\
&= \sum_{i\in\mathcal{S}} \bar{w}_i\mu(x_i)\\
&= E\left(\sum_{i\in\mathcal{S}} \bar{w}_i y_i\middle| x_1, \ldots, x_n\right),
\end{aligned}$$

where the second equality follows from $P(a_i = 1|x_i, y_i) = P(a_i = 1|x_i)$ in Assumption A. Hence, $\tilde{\mu}$ has the same asymptotic mean as $\sum_{i\in\mathcal{S}} \bar{w}_i y_i$, the estimator without nonresponse.

The following result is fundamental for any inference method based on normal approximation.

Theorem 1. *Assume Assumptions A–B and, within each imputation class k, conditions* (4)–(5). *Assume further that $E(y_i^8) < \infty$. Then*

$$\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \rightarrow_d N(0,1) \tag{11}$$

for some $\sigma > 0$, where $\rightarrow_d$ is convergence in distribution, unconditionally with respect to the superpopulation model in Assumption A and sampling.

Proof. Let $\hat{\mu}_k = \sum_{i\in\mathcal{R}_k} \bar{w}_{k,i} y_i + \sum_{i\in\bar{\mathcal{R}}_k} \bar{w}_{k,i}\tilde{y}_i$. Then $\hat{\mu} = \sum_k (M_k/M)\hat{\mu}_k$. Since variables are independent across imputation classes and imputation is carried out within each imputation class, the $\hat{\mu}_k$ are independent. Hence, it suffices to show result (11) for each $\hat{\mu}_k$. We now drop the subscript k with the understanding that the rest of the proof is for a single fixed imputation class. Let $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{X}$ be defined in Section 2, and let $\mathcal{Y} = \{y_i : i \in \mathcal{R}\}$. Then $E(\tilde{y}_i|\mathcal{Y},\mathcal{X},\mathcal{R},\mathcal{S}) = \sum_{i\in\mathcal{R}} q_i y_i$. Define $\tilde{e}_i = \tilde{y}_i - E(\tilde{y}_i|\mathcal{Y},\mathcal{X},\mathcal{R},\mathcal{S})$. Consider the decomposition

$$\hat{\mu} - \mu = Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6,$$

where $Q_1 = \sum_{i\in\bar{\mathcal{R}}} \bar{w}_i\tilde{e}_i$, $Q_2 = \sum_{i\in\mathcal{R}} \bar{w}_i[y_i - \mu(x_i)] + (1-p)\sum_{i\in\mathcal{R}} q_i[y_i - \mu(x_i)]$, $Q_3 = \sum_{i\in\mathcal{R}} \bar{w}_i[\mu(x_i) - \mu_1]$, $Q_4 = (\mu_1 - \mu_0)\sum_{i\in\mathcal{S}} \bar{w}_i(a_i - p)$, $Q_5 = \mu\left(\sum_{i\in\mathcal{S}} \bar{w}_i - 1\right)$, and $Q_6 = \left[\sum_{i\in\bar{\mathcal{R}}} \bar{w}_i - (1-p)\right]\sum_{i\in\mathcal{R}} q_i[y_i - \mu(x_i)] + \sum_{i\in\bar{\mathcal{R}}} \bar{w}_i\left[\sum_{i\in\mathcal{R}} q_i\mu(x_i) - \mu_0\right]$. By repeatedly applying Lemma 1 in Schenker and Welsh (1988), (11) follows from

$$\begin{aligned}
P(\sqrt{n}Q_1 \le \sigma_1 t|\mathcal{Y},\mathcal{X},\mathcal{R},\mathcal{S}) &\rightarrow \Phi(t) \ \text{ a.s.}, && (12)\\
P(\sqrt{n}Q_2 \le \sigma_2 t|\mathcal{X},\mathcal{R},\mathcal{S}) &\rightarrow \Phi(t) \ \text{ a.s.}, && (13)\\
P(\sqrt{n}Q_3 \le \sigma_3 t|\mathcal{R},\mathcal{S}) &\rightarrow \Phi(t) \ \text{ a.s.}, && (14)\\
P(\sqrt{n}Q_4 \le \sigma_4 t|\mathcal{S}) &\rightarrow \Phi(t) \ \text{ a.s.}, && (15)\\
P(\sqrt{n}Q_5 \le \sigma_5 t) &\rightarrow \Phi(t), && (16)\\
\sqrt{n}Q_6 &\rightarrow_p 0, && (17)
\end{aligned}$$

for any real t, where $P(\cdot|\mathcal{A})$ denotes the conditional probability given $\mathcal{A}$, Φ is the standard normal distribution function, σ_i's are some nonnegative parameters, and $\sigma^2 = \sigma_1^2 + \cdots + \sigma_5^2$. The proofs of (12)–(17) can be found in Shao and Wang (2007).

We now consider a variance estimator for $\hat{\mu}$ that is a simplified version of the partially adjusted jackknife variance estimator in Chen and Shao (2001):

$$v_n = \sum_{k=1}^{K} \frac{1}{m_k(m_k-1)M^2} \sum_{j\in\mathcal{S}_k} \left(m_k w_j \tilde{y}_j - \bar{y}_k\right)^2, \tag{18}$$

where m_k is the size of $\mathcal{S}_k$, $\bar{y}_k = \sum_{i\in\mathcal{R}_k}(1+d_i^{(k)})w_i y_i$, $d_i^{(k)} = \sum_{j\in\bar{\mathcal{R}}_k}(w_j/w_i)d_{ij}$, $d_{ij} = 1$ if i is the nearest neighbor of j and $d_{ij} = 0$ otherwise, $\tilde{y}_j = y_j + d_j^{(k)} g_j^{(k)} (y_j - (y_{j_{k1}} + y_{j_{k2}})/2)$ if $j \in \mathcal{R}_k$ and $\tilde{y}_j =$ the imputed value of y_j if $j \in \bar{\mathcal{R}}_k$, $g_j^{(k)} = [\sqrt{6(d_j^{(k)})^2 + 6d_j^{(k)} + 4} - 2]/3d_j^{(k)}$ ($g_j^{(k)} = 0$ if $d_j^{(k)} = 0$), and j_{k1} and j_{k2} are the two nearest neighbors of j in $\mathcal{R}_k$. It is shown in Chen and Shao (2001) that $v_n/V_n \to_p 1$, where

$$V_n = E\left[\sum_{i\in\mathcal{R}}(1+d_i)^2\,\bar{w}_i^2 \mathrm{Var}\,(y_i|x_i)\right] + \mathrm{Var}\left[\sum_{i\in\mathcal{S}}\bar{w}_i\mu(x_i)\right]. \tag{19}$$

For the purpose of showing that the confidence interval $[\hat{\mu} - z_\alpha\sqrt{v_n}, \hat{\mu} + z_\alpha\sqrt{v_n}]$ is asymptotically valid for μ, we need to show that $nv_n/\sigma^2 \to_p 1$, because (11) has σ^2/n as the variance of the limiting distribution of $\hat{\mu}-\mu$. Since $v_n/V_n \to_p 1$, this can be achieved by showing $nV_n/\sigma^2 \to 1$, which is the first part of the following theorem. The theorem also shows that $n\mathrm{Var}\,(\hat{\mu})/\sigma^2 \to 1$ and, hence, v_n is consistent for $n\mathrm{Var}\,(\hat{\mu})$. The proof of Theorem 2 can be found in Shao and Wang (2007).

Theorem 2. *Assume the conditions in Theorem* 1. *Then,*
(i) $nv_n/\sigma^2 \to_p 1$,
(ii) $v_n/\mathrm{Var}\,(\hat{\mu}) \to_p 1$, *and*
(iii) $P(\hat{\mu} - z_\alpha\sqrt{v_n} \le \mu \le \hat{\mu} + z_\alpha\sqrt{v_n}) \to 1-\alpha$.

4. Confidence Intervals for Quantiles

Population quantiles are typically estimated by sample quantiles (Rao, Kovar and Mantel (1990) and Francisco and Fuller (1991)). In what follows we use F and f to denote the distribution and density of y with respect to the superpopulation. Note that $f_{k,a}$ or f_a is used in the previous section for the density of x given a.

Let $I_y(t) = 1$ if $y \le t$ and $I_y(t) = 0$ if $y > t$, and let

$$\hat{F}(t) = \frac{1}{\hat{M}}\sum_k\left(\sum_{i\in\mathcal{R}_k} w_i I_{y_i}(t) + \sum_{i\in\bar{\mathcal{R}}_k} w_i I_{\tilde{y}_i}(t)\right) \tag{20}$$

be the empirical distribution, a survey estimator of $F(t) = P(y \le t)$. Even if M is known, $\hat{M}$ in (20) cannot be replaced by M unless $\hat{M} = M$, since $\hat{F}(\infty)$ has to be 1. If sampling is stratified simple random sampling, then $\hat{M} = M$; otherwise, $\hat{M}$ and M may be different. For any fixed $q \in (0,1)$, the qth sample quantile is $\hat{\theta} = \hat{F}^{-1}(q) = \inf\{t : \hat{F}(t) \ge q\}$, a survey estimator of the population quantile $\theta = F^{-1}(q)$.

Replacing y_i by $I_{y_i}(t)$ in Theorem 1, we obtain that, for any fixed t, $\sqrt{n}[\hat{F}(t) - F(t)]$ is asymptotically normal with mean 0. The following is a Bahadur-type representation for $\hat{\theta}$, a result similar to that in Francisco and Fuller (1991). This result together with the asymptotic normality of $\sqrt{n}[\hat{F}(t) - F(t)]$ imply that $\sqrt{n}(\hat{\theta}-\theta)$ is asymptotically normal. The proof is given in Shao and Wang (2007).

Theorem 3. *Assume the conditions in Theorem* 1 *with* y_i *replaced by* $I_{y_i}(t)$ *for any* t. *Assume further that* F *is differentiable at* θ *and* $F'(\theta) = f(\theta) > 0$. *Then*

$$\hat{\theta} - \theta = \frac{F(\theta) - \hat{F}(\theta)}{f(\theta)} + o_p(n^{-\frac{1}{2}}).$$

It follows from Theorems 1−3 and the delta-method that the asymptotic variance of $\hat{\theta} - \theta$ is $V_n(\theta)/f^2(\theta)$, where, for any fixed t,

$$V_n(t) = F^2(t)\mathrm{Var}\,(\hat{M}/M) - 2F(t)\,\mathrm{Cov}\,(\hat{M}/M, \hat{G}(t)) + \mathrm{Var}\,(\hat{G}(t)), \tag{21}$$

and $\hat{G}(t)$ is defined by (20) with $\hat{M}$ replaced by M. In the case where $\hat{M} = M$, $V_n(t) = \mathrm{Var}\,(\hat{G}(t)) = \mathrm{Var}\,(\hat{F}(t))$.

A consistent estimator of $V_n(\theta)/f^2(\theta)$ and a confidence interval for θ satisfying (2) can be constructed in two steps. First, we construct a consistent estimator of $V_n(\theta)$. Second, we use the Bahadur representation.

For any fixed t, let $\bar{I}_k(t) = \sum_{i\in\mathcal{R}_k} w_i(1 + d_i^{(k)})I_{y_i}(t)$, $\xi_j(t) = I_{\tilde{y}_j}(t)$ if $j \in \bar{\mathcal{R}}_k$ and $\xi_j(t) = I_{y_j}(t) + d_j^{(k)} g_j^{(k)}\{I_{y_j}(t) - [I_{y_{j_{k1}}}(t) + I_{y_{j_{k2}}}(t)]/2\}$ if $j \in \mathcal{R}_k$, where $d_j^{(k)}$, $g_j^{(k)}$, $y_{j_{k1}}$ and $y_{j_{k2}}$ are the same as those in (18). Define

$$\begin{aligned}\hat{V}_n(t) = \hat{F}^2(t)\sum_{k=1}^{K}\frac{1}{m_k(m_k-1)M^2}\sum_{j\in\mathcal{S}_k}(m_k w_j - \hat{M}_k)^2 \\ -2\hat{F}(t)\sum_{k=1}^{K}\frac{1}{m_k(m_k-1)M^2}\sum_{j\in\mathcal{S}_k}\left[m_k w_j \xi_j(t) - \bar{I}_k(t)\right](m_k w_j - \hat{M}_k) \\ +\sum_{k=1}^{K}\frac{1}{m_k(m_k-1)M^2}\sum_{j\in\mathcal{S}_k}\left[m_k w_j \xi_j(t) - \bar{I}_k(t)\right]^2,\end{aligned}$$

where $\hat{M}_k = \sum_{i\in\mathcal{S}_k} w_i$. From the proof of Theorem 2, $\hat{V}_n(\theta)/V_n(\theta) \to_p 1$. However, $\hat{V}_n(\theta)$ is not an estimator since θ is unknown. The following lemma, whose proof is given in Shao and Wang (2007), shows that $\hat{V}_n(\hat{\theta})$ is consistent for $V_n(\theta)$.

Lemma 2. *Assume the conditions in Theorem* 1 *with* y_i *replaced by* $I_{y_i}(t)$ *for any* t. *Assume further that* F *is continuous at* θ. *Then* $\hat{V}_n(\hat{\theta})/V_n(\theta) \to_p 1$.

From the Bahadur representation, we propose the Woodruff confidence interval for θ,

$$CI = [\hat{F}^{-1}(q - z_\alpha s_n), \hat{F}^{-1}(q + z_\alpha s_n)], \tag{22}$$

where $s_n = [\hat{V}_n(\hat{\theta})]^{1/2}$, and the variance estimator for $\hat{\theta}$,

$$v_n = \frac{[\hat{F}^{-1}(q + z_\alpha s_n) - \hat{F}^{-1}(q - z_\alpha s_n)]^2}{4z_\alpha^2}. \tag{23}$$

The next theorem establishes the asymptotic validity of CI and v_n. The proof is given in Shao and Wang (2007).

Theorem 4. *Assume the conditions in Theorem* 3. *Assume further that* F *is differentiable in a neighborhood of* θ *and* $f = F'$ *is continuous at* θ. *Then*
(i) $v_n/[V_n(\theta)/f^2(\theta)] \to_p 1$ *and*
(ii) $P(\theta \in CI) \to 1 - \alpha$.

5. Simulation Results

A simulation study was performed to examine the finite sample performance of the proposed variance estimators and confidence intervals. Stratified simple random samples were generated from a population that matches the main characteristics of an aggregated dataset from the 1998 Financial Farm Survey (FFS) published by Statistics Canada (Rancourt (1999)). The FFS is a bi-annual survey collecting information on agriculture operations in Canada. The survey collects information on revenues, expenses, assets, investments, and liabilities for the reference year. Nonrespondents in the survey are imputed by NNI for some variables (Rancourt (1999)). We focus on dairy farms and two variables: net assets (x) and cash income (y). Strata in the FFS are constructed using the size of farm and province (five provinces and ALT, a group of small provinces, with three size classes in each province). These 18 strata are also used as imputation classes and, hence, imputation does not cut across strata. Information about population size, sample size, number of respondents, mean and standard deviation of x and y, and the correlation coefficient between x and y in each stratum are given in Shao and Wang (2007, Table 1). The overall sampling fraction n/M is about 7%.

For each pair (x, y), a y-respondent is generated according to the response probability function

$$P(a = 1|x) = \frac{\exp(\gamma_1 + \gamma_2(x - \mu_x)\sigma_x^{-1})}{1 + \exp(\gamma_1 + \gamma_2(x - \mu_x)\sigma_x^{-1})}$$

for some γ_1 and γ_2. For each pair (γ_1, γ_2), we define a model as follows:

Model	1	2	3	4	5	6	7	8	9	10
γ_1	0.5	0.5	0.5	1.0	1.0	1.0	2.0	2.0	2.0	∞
γ_2	-1.0	1.0	0.0	-1.0	1.0	0.0	-1.0	1.0	0.0	0.0
p	0.607	0.610	0.628	0.700	0.703	0.735	0.846	0.848	0.883	1.000

where $p = E[P(a = 1|x)]$ is the average response probability and Model 10 corresponds to the case of no nonresponse.

We considered the estimation of six different parameters of the distribution of y, the mean, the median, the 10th, 25th, 75th and 95th percentiles. In addition to the NNI, we considered the linear regression imputation (LRI) that assumes a linear model between y and x, the random hot deck imputation (RHD), and the Bayesian bootstrap multiple imputation given by Rubin (1987), with 10 imputations (MI10).

Since $P(a = 1|x)$ depends on x when $\gamma_2 \neq 0$ (Models 1, 2, 4, 5, 7 and 8), RHD and MI10 are biased. On the other hand, when $\gamma_2 = 0$ (Models 3, 6 and 9), $P(a = 1|x)$ is a constant and RHD and MI10 are unbiased. The LRI is biased for percentile estimation. For the estimation of the mean, LRI is biased when $P(a = 1|x)$ depends on x, because the relationship between y and x is nonlinear. When $P(a = 1|x)$ is a constant (Models 3, 6, and 9), LRI is unbiased even though the relationship between y and x is not linear.

Table 1 provides empirical results, based on 2,000 simulations, in terms of the relative bias of the point estimator, the variance of the estimator, the relative bias (RB) of the variance estimator (given in (18) and (23) for NNI), the coverage probability of the confidence interval of the form $\hat{\mu} \pm z_\alpha \sqrt{v_n}$ for the case of sample mean, and Woodruff's confidence interval (22) for the sample quantiles ($1 - \alpha$ is chosen to be 95%), together with the average length of the confidence interval.

The following is a summary of the results in Table 1.

1. The relative bias. For NNI, relative bias is smaller than 1% in absolute value in all cases (it is actually not larger than 0.5% in absolute value for most cases). The relative bias of RHD, MI10, and LRI is smaller than 0.5% in absolute value when $P(a = 1|x)$ is constant (Models 3, 6, and 9), but it is not negligible in cases where $P(a = 1|x)$ depends on x (Models 1, 2, 4, 5, 7, and 8). Although the relative bias in some cases is small (e.g., it is 0.8% for LRI in the estimation of the mean under Model 7), it still leads to a low coverage probability of the associated confidence interval.
2. The variance. Under Models 3, 6, and 9, LRI, RHD, and MI10 are unbiased but NNI is more efficient in terms of the variance because RHD and MI10 do not use the covariate information and LRI assumes a linear relationship between y and x; the variance of MI10 is smaller than that of RHD for estimation of the mean, but it is larger for estimation of percentiles. In situations where LRI, RHD, and MI10 are biased, their variances are sometimes smaller than that of NNI, but having a small variance is not necessarily an advantage when a point estimator is biased.
3. Variance estimators. The proposed variance estimator for NNI performs well in the estimation of mean and median. For the estimation of other percentiles, it overestimates, especially for the estimation of extreme percentiles.

Table 1. Average Results based on 2,000 Simulations

		Performance of point estimator								Relative bias of variance estimator (%)				Performance of confidence interval							
		Relative bias (%)				Variance/1000								Coverage prob (%)				Length/1000			
θ	$\mathcal{M}$	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10
μ	1	-0.01	1.60	-3.31	-3.33	229	258	355	298	6.50	9.18	-3.98	-5.20	95.4	62.4	15.4	9.8	0.96	1.01	1.14	1.06
	2	0.17	1.72	3.89	3.90	227	247	253	210	0.92	0.67	-3.05	-2.96	94.3	54.2	2.3	1.4	0.93	0.96	0.97	0.90
	3	0.01	0.11	-0.01	0.00	216	221	297	253	0.59	-1.82	1.86	-0.24	95.2	94.3	95.3	94.8	0.91	0.91	1.08	1.00
	4	-0.04	1.28	-2.37	-2.36	204	226	294	237	3.30	-0.43	-3.20	1.19	94.3	70.0	36.2	29.6	0.90	0.93	1.04	0.97
	5	0.08	1.24	3.18	3.19	204	207	225	180	-1.22	-3.45	-3.16	2.33	94.6	70.0	6.5	3.3	0.88	0.87	0.91	0.85
	6	0.02	0.07	0.00	0.01	189	192	258	211	1.75	-1.40	-2.82	1.38	94.6	94.1	94.8	94.6	0.86	0.85	0.98	0.91
	7	0.00	0.80	-0.99	-0.99	169	186	201	183	3.62	-2.43	5.78	4.18	95.6	82.9	80.9	79.9	0.82	0.83	0.90	0.86
	8	0.01	0.57	1.81	1.82	172	165	178	151	1.41	1.37	3.45	9.27	95.2	88.2	40.8	35.4	0.82	0.80	0.84	0.80
	9	0.04	0.06	0.04	0.03	183	181	207	188	-7.20	-8.83	-6.56	-5.91	94.6	93.8	94.2	94.8	0.81	0.79	0.86	0.83
	10	-0.03				165				-5.86				94.3				0.77			
q_{50}	1	0.05	1.35	-3.31	-3.35	654	543	709	768	4.76	18.72	4.42	-19.26	93.2	88.3	52.8	47.6	1.57	1.55	1.66	1.54
	2	0.10	0.35	3.36	3.38	506	539	531	579	7.07	-12.05	2.82	-20.97	93.7	90.8	41.0	34.6	1.41	1.32	1.42	1.32
	3	0.05	-0.62	0.02	-0.01	509	440	594	631	7.73	3.55	4.70	-17.18	93.9	90.4	93.9	89.7	1.42	1.27	1.52	1.41
	4	0.05	1.18	-2.29	-2.32	564	490	596	646	-1.85	5.39	2.04	-20.35	93.2	86.9	68.8	62.4	1.42	1.39	1.50	1.39
	5	0.07	0.11	2.79	2.75	413	419	433	471	7.90	-5.81	8.97	-15.84	94.1	91.4	48.0	44.7	1.29	1.21	1.32	1.22
	6	0.10	-0.39	0.08	0.04	449	401	509	539	3.28	-5.23	1.16	-18.06	93.2	91.6	93.2	90.6	1.31	1.19	1.38	1.29
	7	0.08	0.82	-0.92	-0.92	402	399	439	427	3.52	-2.37	2.13	-6.14	93.4	88.4	88.7	86.7	1.24	1.20	1.29	1.22
	8	0.04	-0.08	1.62	1.61	353	332	384	393	4.40	0.72	1.61	-11.32	93.9	92.9	74.2	70.6	1.17	1.12	1.21	1.14
	9	0.08	-0.16	0.09	0.08	350	328	362	371	4.69	1.01	8.45	-3.30	93.8	93.6	95.0	93.2	1.17	1.11	1.21	1.16
	10	0.03				318				0.04				94.4				1.09			
q_{25}	1	-0.12	1.67	-5.07	-5.03	721	577	1139	1337	12.37	47.28	4.37	-26.48	93.9	89.3	50.8	45.8	1.72	1.77	2.09	1.93
	2	0.08	3.61	5.39	5.45	777	699	712	788	15.82	-35.70	3.53	-21.24	94.4	40.1	27.0	23.2	1.81	1.28	1.64	1.53
	3	-0.06	1.23	-0.09	-0.10	684	553	913	999	13.84	-8.74	9.62	-15.85	94.8	84.5	94.3	90.3	1.69	1.37	1.92	1.78
	4	-0.05	1.47	-3.55	-3.55	663	577	933	1044	5.24	25.54	5.23	-22.25	93.1	89.4	68.9	62.9	1.60	1.64	1.91	1.75
	5	0.11	2.61	4.54	4.51	718	582	610	675	6.37	-26.68	9.31	-17.18	92.8	59.0	35.3	32.2	1.67	1.25	1.57	1.45
	6	0.04	0.95	-0.05	-0.07	625	504	781	837	9.07	-4.68	7.29	-15.40	94.4	88.4	94.0	90.9	1.58	1.33	1.76	1.63
	7	0.04	1.03	-1.46	-1.47	538	511	689	683	6.20	12.38	3.83	-5.73	93.8	90.4	88.6	86.2	1.46	1.46	1.63	1.55
	8	-0.01	1.09	2.63	2.63	540	460	519	516	11.33	-7.12	6.66	-3.31	93.8	85.3	66.8	63.0	1.49	1.26	1.43	1.36
	9	0.06	0.49	0.04	0.05	510	458	566	575	9.87	1.92	9.96	-1.77	95.0	92.5	94.8	93.2	1.44	1.32	1.52	1.45
	10	-0.02				463				5.25				94.9				1.35			

θ: the parameter of interest; q_a = the ath percentile of y.

$\mathcal{M}$: the model for simulation.

NNI: nearest neighbor imputation.

LRI: linear regression imputation.

RHD: random hot deck imputation.

MI10: Bayesian bootstrap multiple imputation with 10 imputations.

Table 1 (continued)

		Performance of point estimator								Relative bias of				Performance of confidence interval							
		Relative bias (%)				Variance/1000				variance estimator (%)				Coverage prob (%)				Length/1000			
θ	$\mathcal{M}$	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10	NNI	LRI	RHD	MI10
q_{75}	1	0.07	1.52	-2.60	-2.61	779	753	756	836	5.68	-10.54	5.41	-19.54	92.4	79.9	59.2	53.9	1.71	1.58	1.71	1.59
	2	0.08	-1.44	2.42	2.49	501	408	510	623	8.10	23.94	16.19	-20.43	93.4	75.8	56.2	47.3	1.41	1.36	1.47	1.37
	3	-0.01	-1.31	-0.02	0.00	557	475	643	713	7.17	8.85	3.05	-21.96	93.0	79.6	92.6	88.2	1.48	1.39	1.56	1.45
	4	0.01	1.24	-1.79	-1.81	602	581	604	652	6.65	-6.69	7.82	-15.91	93.3	82.2	72.6	66.7	1.52	1.42	1.55	1.43
	5	-0.04	-1.16	1.91	1.91	454	371	497	517	0.94	22.26	0.67	-16.60	92.9	82.3	64.2	59.5	1.30	1.30	1.36	1.27
	6	0.07	-0.91	0.02	0.04	499	442	559	572	0.69	1.71	-1.55	-18.27	92.8	85.9	92.2	89.5	1.36	1.29	1.42	1.32
	7	0.08	0.80	-0.66	-0.68	448	452	460	484	-0.20	-10.31	0.98	-13.63	92.7	85.7	89.1	87.2	1.28	1.22	1.31	1.25
	8	-0.03	-0.54	1.07	1.07	378	350	401	394	1.97	10.98	2.42	-5.46	93.6	90.7	81.6	79.1	1.19	1.20	1.23	1.18
	9	-0.03	-0.44	-0.02	-0.01	386	354	403	416	3.15	5.61	3.67	-8.39	92.9	91.6	93.2	91.3	1.21	1.18	1.24	1.19
	10	0.00				330				0.88				93.3				1.11			
q_{10}	1	0.02	2.63	-5.80	-5.78	1225	1225	1703	1935	10.98	45.91	12.93	-16.57	93.6	89.8	66.2	62.0	2.24	2.56	2.64	2.45
	2	0.88	11.90	10.10	10.10	2106	1269	1534	1728	15.97	24.51	22.90	-10.20	92.4	9.5	26.6	22.4	2.98	2.40	2.60	2.39
	3	0.12	5.36	0.02	0.02	1470	1188	1926	2124	4.36	30.39	12.20	-14.75	93.4	66.4	93.7	89.7	2.44	2.39	2.80	2.59
	4	0.06	2.04	-4.14	-4.11	1143	1139	1476	1626	13.39	29.98	15.54	-11.93	93.2	90.4	79.3	73.7	2.14	2.34	2.49	2.30
	5	0.53	9.43	8.43	8.38	1731	1188	1397	1604	10.53	11.54	17.05	-13.50	90.9	20.2	35.0	31.9	2.65	2.20	2.43	2.26
	6	0.19	3.99	0.08	0.10	1191	1148	1656	1693	8.76	19.51	6.22	-11.88	93.8	74.8	92.2	89.2	2.29	2.25	2.53	2.35
	7	0.18	1.26	-1.74	-1.74	973	1068	1189	1193	13.27	10.37	14.16	1.19	94.2	90.6	91.7	89.3	2.02	2.08	2.23	2.11
	8	0.27	5.15	5.09	5.07	1260	1092	1257	1309	1.15	3.13	7.42	-8.05	93.3	56.8	62.8	59.8	2.27	2.04	2.22	2.11
	9	0.23	1.89	0.20	0.20	1095	1072	1245	1282	8.35	8.59	5.76	-6.42	93.2	87.2	92.4	91.3	2.09	2.07	2.20	2.10
	10	0.18				937				10.17				93.7				1.95			
q_{95}	1	-0.31	1.45	-3.78	-3.77	5462	4477	8518	9410	12.45	62.08	-7.18	-29.23	92.7	93.0	66.1	61.3	4.73	5.15	5.37	4.98
	2	-0.04	4.44	4.43	4.44	7777	4996	5624	6119	8.42	-14.68	20.51	-9.07	91.5	38.7	52.8	46.0	5.48	3.91	4.93	4.52
	3	-0.11	1.77	-0.08	-0.10	6211	4422	8298	9459	6.20	5.85	1.35	-24.76	92.4	83.0	91.6	86.9	4.89	4.14	5.54	5.15
	4	-0.17	1.23	-2.67	-2.65	5073	4567	7296	7738	10.76	37.77	-2.17	-20.78	92.3	92.6	77.2	72.8	4.53	4.80	5.11	4.77
	5	-0.02	3.35	3.65	3.63	6741	4503	5633	5713	2.43	-19.17	5.80	-12.65	91.3	52.5	60.8	55.4	4.99	3.64	4.64	4.27
	6	-0.18	1.30	-0.14	-0.10	5229	4074	6914	7089	11.44	11.08	4.46	-12.93	93.4	87.4	93.0	89.2	4.61	4.07	5.14	4.78
	7	0.01	0.78	-1.13	-1.12	4395	4180	5769	5914	10.91	24.89	2.19	-10.89	92.9	91.4	90.6	87.6	4.23	4.37	4.66	4.41
	8	-0.13	1.45	2.03	2.05	5129	3863	4699	4780	7.45	-7.28	10.81	-2.00	92.1	81.1	81.9	77.9	4.48	3.61	4.35	4.14
	9	-0.11	0.57	-0.09	-0.12	4674	4142	5027	5357	5.46	5.58	10.57	-4.86	92.8	91.0	92.8	90.5	4.25	4.00	4.51	4.33
	10	-0.01				3864				15.71				93.8				4.06			

4. The performance of the confidence interval. For NNI, the coverage probability of confidence interval is close to the nominal level of 95% for mean estimation, and is between 91% and 94% for percentile estimation. For mean estimation, the coverage probability of the confidence intervals associated with LRI, RHD, and MI10 is comparable to that of NNI under Models 3, 6, and 9, but can be much lower than the nominal level of 95% when $P(a = 1|x)$ depends on x. For percentile estimation, only RHD has a comparable coverage probability with NNI under Models 3, 6, and 9; the coverage probability for LRI is clearly low due to its bias; the coverage probability for MI10 is much lower than that for RHD.

We conclude that the empirical results are consistent with our theoretical findings. In the cases where LRI, RHD, and MI10 are unbiased, NNI is better than RHD and MI10 when useful covariate information is used; NNI is better than LRI when the relationship between y and x is not linear. When $P(a = 1|x)$ depends on x, NNI is still unbiased but LRI, RHD and MI10 may not be.

6. Discussion

The results in Sections 3 are for the case where μ is the parameter of interest. If the parameter of interest is the finite population mean $\bar{Y}$ instead of μ, then we first need an asymptotic distribution for $\sqrt{n}(\hat{\mu} - \bar{Y})$. Note that $\bar{Y} = \frac{1}{M}\sum_k \Big(\sum_{i\in\mathcal{R}_k} y_i + \sum_{i\notin\mathcal{R}_k, i\in\mathcal{P}_k} y_i\Big)$, and $\{y_i : i \in \mathcal{R}_k\}$ and $\{y_i : i \notin \mathcal{R}_k, i \in \mathcal{P}_k\}$ are independent. Hence it follows from the Central Limit Theorem and our Theorem 1 that $\sqrt{n}(\hat{\mu} - \bar{Y}) = \sqrt{n}(\hat{\mu} - \mu) + \sqrt{n}(\mu - \bar{Y})$ is asymptotically normal with mean 0, but with a variance that may be different from the one in (11). Since $\mu - \bar{Y} = O_p(M^{-1/2})$, the limiting variances of $\sqrt{n}(\hat{\mu} - \bar{Y})$ and $\sqrt{n}(\hat{\mu} - \mu)$ are the same if $n/M \to 0$. Hence, if $n/M \to 0$, the variance estimator v_n in (18) is still consistent and the confidence interval for $\bar{Y}$ of the form $\hat{\mu} \pm z_\alpha\sqrt{v_n}$ satisfies (2). A similar discussion applies to the estimation of quantiles. If n/M is not negligible, then our variance estimators may not be consistent and an extra effort is needed to derive consistent variance estimators.

The main difficulty for extending our results to cluster sampling or multistage sampling is that the independence of (x_i, y_i, a_i)'s (Assumption A) does not hold, since values within a cluster are typically dependent. In a few steps in our proof, we apply a result for order statistics of i.i.d. x_i's that is not available for general dependent x_i's. If the cluster sizes in cluster sampling (or the first stage cluster sizes in multistage sampling) are bounded by a fixed integer, then our proof can be modified to establish similar asymptotic results. For general cases, however, further research is needed.

NNI can be applied when the covariate x is multivariate, but the study of its properties faces the curse of dimensionality, a problem for many other nonparametric imputation methods. In consequence, NNI with a multivariate x may not be efficient. As an alternative, we are working on a method for finding a linear combination of the multivariate x with which to define neighbors.

Acknowledgements

Jun Shao's research is partially supported by NSF Grant DMS-0404535 and SES-0705033. Hansheng Wang's research is partially support by NSFC Grant 10771006. The authors would like to thank referees for their helpful comments.

References

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *J. Official Statist.* **16**, 113-132.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *J. Amer. Statist. Assoc.* **96**, 260-269.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89**, 81-87.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM, Philadelphia.

Farber, J. E. and Griffin, R. (1998). A comparison of alternative estimation methodologies for Census 2000. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 629-634.

Fay, R. E. (1999). Theory and application of nearest neighbor imputation in Census 2000. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 112-121.

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

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, UK.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing data. *Survey Methodology* **12**, 1-16.

Kosorok, M. R. (1999). Two-sample quantile tests under general conditions. *Biometrika* **86**, 909-921.

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.

Lee, H., Rancourt, E. and Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *J. Official Statist.* **10**, 231-243.

Little, R. J. (1995). Modeling the dropout mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90**, 1112-1121.

Montaquila, J. M. and Ponikowski, C. H. (1993). Comparison of methods for imputing missing responses in an establishment survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 446-451.

Rancourt, E. (1999). Estimation with nearest neighbor imputation at Statistics Canada. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 131-138.

Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365-375.

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

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

Sedransk, J. (1985). The objective and practice of imputation, *Proceedings of the First Annual Research Conference*, 445-452, Bureau of the Census, Washington D.C.

Shao, J. and Steel, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *J. Amer. Statist. Assoc.* **94**, 254-265.

Shao, J. and Wang, H. (2007). Confidence intervals based on survey data with nearest neighbor imputation. Technical Report 1137 (http://www.stat.wisc.edu), Department of Statistics, University of Wisconsin-Madison.

Valliant, R. (1993). Poststratification and conditional variance estimation, *J. Amer. Statist. Assoc.* **88**, 89-96.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896-924.

Department of Statistics, University of Wisconsin−Madison, 1300 university Ave, Madison, WI 53706, U.S.A.

E-mail: shao@stat.wisc.edu

Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing, 100871, P. R. China.

E-mail: hansheng@gsm.pku.edu.cn

(Received October 2005; accepted June 2007)