Journal of Official Statistics, Vol. 16, No. 2, 2000, pp. 113–131

# Nearest Neighbor Imputation for Survey Data

*Jiahua Chen*[1] *and Jun Shao*[2]

Nearest neighbor imputation is one of the hot deck methods used to compensate for nonresponse in sample surveys. Although it has a long history of application, few theoretical properties of the nearest neighbor imputation method are known prior to the current article. We show that under some conditions, the nearest neighbor imputation method provides asymptotically unbiased and consistent estimators of functions of population means (or totals), population distributions, and population quantiles. We also derive the asymptotic variances for estimators based on nearest neighbor imputation and consistent estimators of these asymptotic variances. Some simulation results show that the estimators based on nearest neighbor imputation and the proposed variance estimators have good performances.

*Key words:* biases; hot deck; quantiles; sample means; variance estimation.

## 1. Introduction

Imputation is commonly applied to compensate for nonresponse in sample surveys (Kalton 1981; Sedransk 1985; Rubin 1987). The nearest neighbor imputation (NNI) method is used in many surveys conducted at Statistics Canada, the U.S. Bureau of Labor Statistics, and the U.S. Census Bureau, and this trend will continue because of the availability of a computer software, the Generalized Edit and Imputation System, which provides a simple way of performing NNI (Cotton 1991; Rancourt, Särndal, and Lee 1994; Kovar, Whitridge, and MacMillan 1998). Let us begin with an introduction of the NNI method in the simplest case. Consider a bivariate sample $(x_1, y_1),\ldots,(x_n, y_n)$ and suppose that $r$ of the $n$ $y$-values are observed (respondents), the rest of $m = n - r$ $y$-values are missing (nonrespondents), and all $x$-values are observed. For simplicity we assume that $y_{r+1},\ldots,y_n$ are missing. The NNI method imputes a missing $y_j$, $r+1 \le j \le n$, by $y_i$, where $1 \le i \le r$ and $i$ is the nearest neighbor of $j$ measured by the $x$-variable, i.e., $i$ satisfies

$$|x_i - x_j| = \min_{1\le l\le r} |x_l - x_j| \tag{1.1}$$

If there are tied $x$-values, then there may be multiple nearest neighbors of $j$ and $i$ is randomly selected from them. Thus, if $x$ is a categorical variable, then the NNI method

[1] University of Waterloo, Department of Statistics and Actuarial Science, Waterloo, Ontario N2L 3G1, Canada. E-mail: jhchen@uwaterloo.ca
[2] University of Wisconsin-Madison, Department of Statistics, 1210 West Dayton Street, Madison, WI 53706-1685, U.S.A. E-mail: Shao@Stat.wisc.edu

**Acknowledgments:** The research of J. Chen was supported by Natural Sciences and Engineering Research Council of Canada. The research of J. Shao was supported by National Science Foundation Grants DMS-9504425 and DMS-9803112 and National Security Agency Grants MDA904-96-1-0066 and MDA904-1-0032. The authors would like to thank an associate editor and two referees for their valuable comments and suggestions.

imputes nonrespondents in category $x$ by a random sample from the respondents in the same category, which is the same as the well-known random hot deck imputation method. Our study, however, focuses on the case where $x$ has a continuous (or nearly continuous) distribution. In applications, NNI is often carried out by first dividing the sample into several ''imputation classes'' and then applying (1.1) within each imputation class (see Section 2).

The NNI method has some nice features. First, it is a hot deck method in the sense that nonrespondents are substituted by a value of the same variable from a respondent of the same file; the imputed values are actually occurring values, not constructed values, and they may not be perfect substitutes, but are unlikely to be nonsensical values. Second, the NNI method may be more efficient than other hot deck methods such as the mean imputation and the random hot deck imputation, since the NNI makes use of auxiliary information provided by the $x$-values and is a nonrandom imputation method (in the sense that nonrespondents are imputed by deterministic values, given the $y$-respondents and $x$-values). Third, the NNI method does not use an explicit model relating $y$ and $x$ and, hence, it is expected to be more robust against model violations than methods based on explicit models, such as ratio imputation and regression imputation. Finally, one of the results in the current article shows that the NNI method provides asymptotically valid distribution and quantile estimators, which is a superiority over the mean, ratio, and regression imputation methods which do not lead to valid distribution and quantile estimators.

Although the NNI method has been used for a long time, we cannot find any theoretical result regarding the validity of NNI in the literature. It is natural to ask the following questions that are crucial for applications of NNI. First, is a point estimator based on the data imputed by NNI (which is called an NNI estimator henceforth) an unbiased estimator of the population parameter? If not, what is the size of the bias? Empirical results (see, e.g., Section 4 or Fay 1996) show that the bias of the NNI sample mean is negligible; in fact, Rancourt, Särndal, and Lee (1994) stated that ''normally, NN imputation yields point estimates with small or negligible bias, assuming that a linear relationship exists between the variable of interest $y$ and the concomitant variable $x$ used for nearest neighbor identification.'' But this claim was not supported by any theoretical result in general. Second, what are the size and the form of the variance of an NNI estimator? Clearly, this relates to variance estimation for NNI estimators, another important task in analyzing survey data.

The purpose of this article is to answer these two questions by providing some theoretical results for the biases and variances of NNI estimators such as the sample mean, functions of sample means of estimated totals, and sample quantiles.

In Section 2, under some regularity conditions on the distribution of the $x$-variable and the response mechanism, we show that the bias of the NNI sample mean (and, hence, any smooth function of sample means) is asymptotically negligible, not only in the case where variables $y$ and $x$ are linearly related (thus the claim in Rancourt, Särndal, and Lee (1994) is verified) but also in the case where almost no assumption concerning the model between $y$ and $x$ is imposed. As a corollary of our result, the empirical distribution and the quantile estimators based on the data imputed by NNI are asymptotically unbiased for the distribution of $y$ and its quantiles, which is a superiority of the NNI method over the mean, ratio, and regression imputation methods. Note that the random hot deck imputation method also

provides valid distribution and quantile estimators, but the NNI distribution and quantile estimators are more efficient.

In Section 3 we derive an approximate formula for variances of NNI estimators. Using this formula and some assumptions concerning the model that relates $y$ and $x$, we can obtain asymptotically unbiased and consistent variance estimators for NNI estimators. The variance formula also enables us to compare the efficiency of an NNI estimator with that of another estimator obtained by using mean imputation or random hot deck imputation.

In Section 4 we examine empirically the finite sample performances of the NNI sample mean and our proposed variance estimator, using a population that is a real data set from the 1988 Current Population Survey (Valliant 1993).

## 2. The Biases of NNI Estimators

Let $\mathcal{P}$ be a finite population containing indices $1,\ldots,N$, and let $\mathcal{S}$ be a sample of size $n$ selected without replacement from $\mathcal{P}$, according to some sampling plan. For each unit $i$, there are characteristics of interest, $x_i$, $y_i$, $z_i$, etc. We assume that the values $x_i$, $y_i,\ldots$, are random variables from a superpopulation, such that $\{x_i, y_i, \ldots\}$ and $\{x_j, y_j, \ldots\}$ are independent for $i \neq j$. For a given variable $y$, let $a$ be the response indicator for $y$ (i.e., for the $i$th unit, $a_i = 1$ if $y_i$ is a respondent and $a_i = 0$ otherwise). Throughout the article, we make the following assumption:

**Assumption A.** The finite population is divided into $K$ imputation classes such that within each imputation class, $(x_i, y_i, a_i)$'s are iid and $P(a_i = 1|x_i, y_i) = P(a_i = 1|x_i)$. NNI is carried out within each imputation class.

Imputation classes are usually constructed using a categorical variable whose values are observed for all sampled units; for example, if $\mathcal{S}$ is a stratified sample, then strata or unions of strata are often used as imputation classes. The assumption on the response probability $P(a = 1|x, y)$ means that the response indicator $a$ is independent of $y$, given $x$. This is called ''unconfounded response mechanism'' by Lee, Rancourt, and Särndal (1994), which is required for the validity of many popular imputation methods such as the mean, ratio, regression, and random hot deck imputation methods. Within an imputation class, if $F$ is the marginal distribution of $x$ and $p = P(a = 1) \in (0, 1)$, then

$$P(x \leq t|a = 1) = P(a = 1|x \leq t)F(t)/p = F_1(t) \tag{2.1}$$

and

$$P(x \leq t|a = 0) = P(a = 0|x \leq t)F(t)/(1 - p) = F_0(t) \tag{2.2}$$

This means that within an imputation class and conditional on $a_i$'s, $x_i$'s may have two different distributions according as whether $a_i = 1$ or 0; $F_1 = F_0 = F$ if $a_i$'s are independent of $x_i$'s.

In this article, we focus on continuous $F_1$ and $F_0$. When $x$ is discrete, NNI behaves like random hot deck imputation whose properties are well-known (e.g., Rubin 1987). The problem when the distribution of $x$ is a mixture of a continuous distribution and a discrete distribution can be treated using the results in this article and the results for random hot deck imputation.

### 2.1. Simple random sampling with one imputation class

To study the biases of NNI estimators, we start with the simplest case where $\mathcal{S}$ is a simple random sample (srs) and $K = 1$ (a single imputation class). More complex cases are considered in Section 2.2.

Without loss of generality we assume that $\mathcal{S} = \{1, \ldots, n\}$, $a_i = 1$ for $1 \le i \le r$ and $a_i = 0$ for $r + 1 \le i \le n$. Under the superpopulation model, assumption A, and the srs sampling design, conditional on the number of respondents $r$, $\{(y_1, x_1), \ldots, (y_r, x_r)\}$ and $\{(y_{r+1}, x_{r+1}), \ldots, (y_n, x_n)\}$ are independent sets of iid random vectors from two possibly different distributions.

For $r + 1 \le j \le n$, let $\tilde{y}_j$ denote the value imputed by NNI according to (1.1). Then the NNI sample mean is

$$\bar{y}_{\text{NNI}} = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \tilde{y}_i\right) = \frac{1}{n}\sum_{i=1}^{r}(1 + d_i)y_i \tag{2.3}$$

where $d_i$ is the number of times that unit $i$ is used as a donor, $1 \le i \le r$. Note that $\sum_{i=1}^{r} d_i = n - r = m$, the number of missing $y$-values. Let $x_{(1)} \le \ldots \le x_{(r)}$ denote the ordered values of $x_1, \ldots, x_r$ and $d_{(i)}$ be the $d$-value corresponding to $x_{(i)}$. For continuous $F_1$ and $F_0$,

$$d_{(i)}|r, x_1, \ldots, x_r \sim \text{binomial } (m, \pi_i) \tag{2.4}$$

with

$$\pi_i = F_0\left(\frac{x_{(i+1)} + x_{(i)}}{2}\right) - F_0\left(\frac{x_{(i)} + x_{(i-1)}}{2}\right)$$

$i = 1, \ldots, r, x_{(0)} = -\infty$ and $x_{(r+1)} = +\infty$, since, conditional on $r$, $x_1, \ldots, x_r$ are iid with $F_1$ in (2.1) and $x_{r+1}, \ldots, x_n$ are iid with $F_0$ in (2.2).

Before we state a general result for the bias of $\bar{y}_{\text{NNI}}$, let us consider two examples. Throughout this article, expectations (conditional or unconditional) are with respect to sampling and the superpopulation model.

**Example 1.** Symmetric $F_1$ and $F_0$. Assume that

$$E(y|x) = \alpha + \beta x \tag{2.5}$$

where $\alpha$ and $\beta$ are unknown parameters, and that $F_1$ and $F_0$ are symmetric about $E(x)$. Then $\bar{y}_{\text{NNI}}$ is exactly unbiased, i.e.,

$$E(\bar{y}_{\text{NNI}}|r) = E(y) \tag{2.6}$$

We now prove (2.6). Under Model (2.5),

$$E(y) = \alpha + \beta E(x)$$

Under (2.5) and assumption A,

$$E(\bar{y}_{\text{NNI}}|r) = \alpha + \frac{\beta}{n} E\left[\sum_{i=1}^{r}(1 + d_i)x_i|r\right]$$

since $E(d_i y_i|r) = E[E(d_i y_i|x_1, \ldots, x_r)|r] = E[d_i E(y_i|x_i)|r] = E[d_i(\alpha + \beta x_i)|r]$. Hence, it suffices to show that

$$E\left(\sum_{i=1}^{r} d_i x_i|r\right) = E\left(\sum_{i=1}^{r} d_{(i)} x_{(i)}|r\right) = mE(x) \tag{2.7}$$

Without loss of generality we assume $E(x) = 0$. Then $x_{(1)}, x_{(2)}, \ldots$ have the same joint distribution as that of $-x_{(r)}, -x_{(r-1)}, \ldots$, given $r$. For $2 \le i \le r-1$, by (2.4) and the fact that $F_0(-t) - F_0(-s) = F_0(s) - F_0(t)$,

$$\begin{aligned} E(d_{(i)}x_{(i)}|r) &= mE\left\{x_{(i)}\left[F_0\left(\frac{x_{(i+1)} + x_{(i)}}{2}\right) - F_0\left(\frac{x_{(i)} + x_{(i-1)}}{2}\right)\right]|r\right\} \\ &= -mE\left\{x_{(r-i+1)}\left[F_0\left(\frac{-x_{(r-i)} - x_{(r-i+1)}}{2}\right) - F_0\left(\frac{-x_{(r-i+2)} - x_{(r-i+1)}}{2}\right)\right]|r\right\} \\ &= -mE\left\{x_{(r-i+1)}\left[F_0\left(\frac{x_{(r-i+2)} + x_{(r-i+1)}}{2}\right) - F_0\left(\frac{x_{(r-i)} + x_{(r-i+1)}}{2}\right)\right]|r\right\} \end{aligned}$$

Thus,

$$E\left(\sum_{i=2}^{r-1} d_{(i)}x_{(i)}|r\right) = -E\left(\sum_{i=2}^{r-1} d_{(r-i+1)}x_{(r-i+1)}|r\right) = -E\left(\sum_{i=2}^{r-1} d_{(i)}x_{(i)}|r\right)$$

and this expectation must be 0. Similarly, $E[d_{(1)}x_{(1)} + d_{(r)}x_{(r)}|r] = 0$. Hence

$$E\left(\sum_{i=1}^{r} d_{(i)}x_{(i)}|r\right) = 0$$

This proves (2.7) and thus (2.6) holds.

In survey problems, however, the distribution of $x_i$'s is seldom symmetric. If $F_1$ and $F_0$ are not symmetric, the next example shows that $\bar{y}_{\mathrm{NNI}}$ is biased even when (2.5) holds and $F_1 = F_0$.

**Example 2.** Exponential $F_1 = F_0$. Assume linear model (2.5) and that $F_1 = F_0 = F$ is the exponential distribution with mean 1. To study the bias of $\bar{y}_{\mathrm{NNI}}$, we need to evaluate the expectation on the left-hand side of (2.7). By (2.4),

$$m^{-1}E(d_{(i)}x_{(i)}|r) = E(B_i + A_{i-1} - A_i|r), \quad i = 2, \ldots, r-1$$

where $B_i = \Delta_i e^{-\Delta_i/2} e^{-x_{(i-1)}}$, $A_i = x_{(i)} e^{-x_{(i)}} e^{-\Delta_{i+1}/2}$, and $\Delta_i = x_{(i)} - x_{(i-1)}$
Also

$$m^{-1}E(d_{(1)}x_{(1)}|r) = E(x_{(1)} - x_{(1)}e^{-\Delta_2/2}|r) = r^{-1} - E(A_1|r)$$

and

$$m^{-1}E(d_{(r)}x_{(r)}|r) = E(x_{(r)}e^{-(x_{(r)}+x_{(r-1)})/2}|r) = E(B_r + A_{r-1}|r)$$

Hence

$$E\left(\sum_{i=1}^{r} d_{(i)}x_{(i)}|r\right) = \frac{m}{r} + m\sum_{i=2}^{r} E(B_i|r) \tag{2.8}$$

Using the fact that $\Delta_1, \Delta_2, \ldots,$ are independent and all have exponential distributions, we

obtain that

$$E(B_i|r) = \frac{1}{r+1}\left[1 - \frac{1}{(2r-2i+3)^2}\right] \tag{2.9}$$

By (2.8) and (2.9) and the fact that $E(x) = 1$,

$$\begin{aligned} E(\bar{y}_{\text{NNI}}|r) - E(y) &= \frac{\beta}{n}\left[E\left(\sum_{i=1}^{r} d_{(i)}x_{(i)}|r\right) - mE(x)\right] \\ &= \frac{\beta m}{n}\left[\frac{1}{r} + \sum_{i=2}^{r} E(B_i|r) - 1\right] \\ &= \frac{\beta m}{n}\left[\frac{r^2+1}{r(r+1)} - \frac{1}{r+1}\sum_{i=0}^{r-2}\frac{1}{(2i+3)^2} - 1\right] \\ &= -\frac{\beta m}{n(r+1)}\left[\frac{r-1}{r} + \sum_{i=0}^{r-2}\frac{1}{(2i+3)^2}\right] \end{aligned} \tag{2.10}$$

That is, $\bar{y}_{\text{NNI}}$ is biased unless $\beta = 0$. If $\beta > 0$, $\bar{y}_{\text{NNI}}$ has a negative bias; otherwise $\bar{y}_{\text{NNI}}$ has a positive bias.

What is the size of the bias of $\bar{y}_{\text{NNI}}$? By (2.10) and the fact that $\sum_{i=0}^{\infty}(2i+3)^{-2} = \pi^2/8 - 1$, we have $E(\bar{y}_{\text{NNI}}|r) - E(y) = O(r^{-1})$, i.e., conditional on $r$, the bias of $\bar{y}_{\text{NNI}}$ is of order $r^{-1}$. Note that $r^{-1} \approx p^{-1}n^{-1}$ for large $n$, where $p = P(a = 1)$. Hence, unconditionally, $\bar{y}_{\text{NNI}}$ is also asymptotically unbiased.

Example 2 shows that $\bar{y}_{\text{NNI}}$ may be biased but the bias is asymptotically negligible in a very special case. The following result shows that this is true in general.

**Theorem 1.** Suppose that (i) assumption A holds; (ii) there exist constants $M_1 < M_2$ and $C$ ($M_1$ and $M_2$ may be $\pm\infty$) such that the function $\psi(x) = E(y|x)$ is a monotone function when $x < M_1$ or $x > M_2$, and $|\psi(x) - \psi(x')| \le C|x - x'|$ when $x, x' \in [M_1, M_2]$; (iii) the marginal distribution of $x$ has a density, $E|x|^3 < \infty$, and $E|\psi(x)|^3 < \infty$; and (iv) the response probability $P(a = 1|x)$ satisfies

$$\inf_{x\in\mathcal{D}} P(a = 1|x) > 0, \tag{2.11}$$

where $\mathcal{D}$ is the support of the marginal distribution of $x$. Then

$$E(\bar{y}_{\text{NNI}}|r) - E(y) = o_p(n^{-1/2}) \tag{2.12}$$

From (2.12), $\bar{y}_{\text{NNI}}$ is asymptotically unbiased for the superpopulation mean $E(y)$. Let $\bar{Y}$ be the finite population mean for $y$-values. Then (2.12) also implies that $\bar{y}_{\text{NNI}}$ is asymptotically unbiased for the finite population mean $\bar{Y}$, since $E(\bar{Y}) = E(y)$. The result in the next section shows that the asymptotic variance of $\bar{y}_{\text{NNI}}$ is of order $O(n^{-1})$. Thus, the asymptotic mean squared error of $\bar{y}_{\text{NNI}}$ is $O(n^{-1})$ and $\bar{y}_{\text{NNI}} = \bar{Y} + O_p(n^{-1/2})$.

The function $\psi(x) = E(y|x)$ is unknown and its form is unspecified. $\psi$ can be a linear function given by (2.5), or completely unknown (nonparametric), i.e., the NNI method requires no model between variables $x$ and $y$. Apart from the moment condition, the condition on the function $\psi$ is very weak. It is satisfied for most practical $\psi$ functions (e.g., polynomial functions).

Condition (2.11) roughly means that there are some $y$-respondents for every $x$-value. It

is satisfied in most practical problems and is almost necessary for the validity of any imputation method: intuitively, if $P(a = 1|x) = 0$ for $x$ in a region $\mathcal{D}_1 \subset \mathcal{D}$, then we do not have any information on the $y$-variable as long as $x$ is in $\mathcal{D}_1$.

**Proof of Theorem 1.** Under assumption A,

$$E(\bar{y}_{\mathrm{NNI}}|r) - E(y) = \frac{1}{n} E\left\{ \sum_{i=1}^{r} d_i \psi(x_i) - mE[\psi(x)|a = 0]|r \right\} \tag{2.13}$$

since $E(d_i y_i|r) = E[E(d_i y_i|x_1, \ldots, x_r)|r] = E[d_i \psi(x_i)|r]$. For notational simplicity we now assume $x_{(i)} = x_i$ in this proof. Let $E_d$ be the conditional expectation of $d_i$'s, given $r$ and $x_1, \ldots, x_r$. Then $E[d_i \psi(x_i)|r] = E[E_d(d_i)\psi(x_i)|r] = E[m\pi_i(x_i)|r]$. Hence the bias in (2.13) is the expectation of

$$\frac{m}{n}\left\{ \sum_{i=1}^{r} \pi_i \psi(x_i) - E[\psi(x)|a = 0] \right\} = \frac{m}{n} \sum_{i=1}^{r} \int_{(x_i + x_{i-1})/2}^{(x_i + x_{i+1})/2} [\psi(x_i) - \psi(t)] dF_0(t) \tag{2.14}$$

where $x_{r+1} = \infty, x_0 = -\infty$, and $F_0$ is defined by (2.2). It is shown in the Appendix that

$$E\left[ \left| \sum_{i=1}^{r} \pi_i \psi(x_{(i)}) - E[\psi(x)|a = 0] \right| |r \right] = o_p(n^{-1/2}) \tag{2.15}$$

which implies that the asymptotic bias in (2.13) is $o_p(n^{-1/2})$. This proves result (2.12).

The most commonly used estimators in surveys are functions of several sample means or estimated totals. Using Theorem 1 and Taylor's expansion, we can prove the following result.

**Corollary 1.** Let $\bar{\boldsymbol{y}}_{\mathrm{NNI}}$ be a vector of NNI sample means (as estimators of the vector $\bar{\boldsymbol{Y}}$ of population means) and let $g$ be a given differentiable function. Then $g(\bar{\boldsymbol{y}}_{\mathrm{NNI}})$ is asymptotically unbiased for $g(\bar{\boldsymbol{Y}})$.

Let $I_{y_i}(t)$ be the indicator function of $y_i$. Replacing $y_i$ by $I_{y_i}(t)$ in Theorem 1 ($\psi(x) = P(y \le t|x)$), we obtain another useful corollary.

**Corollary 2.** The empirical distribution based on the data imputed by NNI,

$$\hat{F}(t) = \frac{1}{n}\left[ \sum_{i=1}^{r} I_{y_i}(t) + \sum_{i=r+1}^{n} I_{\tilde{y}_i}(t) \right]$$

is asymptotically unbiased for the finite population distribution,

$$F(t) = \frac{1}{N} \sum_{i=1}^{N} I_{y_i}(t)$$

Consequently, the NNI sample $q$th quantile, $\hat{F}^{-1}(q)$, is asymptotically unbiased for the finite population $q$th quantile $F^{-1}(q), 0 < q < 1$.

### *2.2. Stratified Sampling with K Imputation Classes*

Suppose that $\mathcal{P}$ is stratified into $H$ strata and $n_h$ units are sampled from stratum $h$ according to some probability sampling plan, $h = 1, \ldots, H$. We still assume $\mathcal{S} = \{1, \ldots, n\}$, $a_i = 1$

for $1 \le i \le r$ and $a_i = 0$ for $r+1 \le i \le n$. Under stratified sampling, the sample mean in the case of no nonrespondents is of the form

$$\bar{y} = \sum_{i=1}^{n} w_i y_i$$

(a Horvitz-Thompson type estimator), where $w_i$ is the survey weight associated with $y_i$ so that $E_s(\bar{y} - \bar{Y}) = 0$ ($E_s$ is the expectation with respect to $\mathcal{S}$) and $\bar{Y}$ is the finite population mean. If an srs is sampled from each stratum, for example, then $w_i = N_h/(n_h N)$, where $N_h$ is the number of population units in stratum $h$.

Results in Section 2.1 (Theorem 1 and Corollaries 1 and 2) still hold under stratified sampling with $K$ imputation classes (see assumption A). A sketched proof is given as follows. Let $\mathcal{P}_k$ denote the $k$th imputation class, $M_k$ be the size of $\mathcal{P}_k$, and assume $(y_i, x_i)$ are iid from a superpopulation such that $\psi_k(x) = E(y|x)$ within $\mathcal{P}_k$. Thus, we have

$$E(\bar{Y}) = \sum_{k=1}^{K} \frac{M_k}{N} E[\psi_k(x)]$$

Let

$$d_{ij} = \begin{cases} 1 & i \text{ is the nearest neighbor of } j \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

Then $E[d_{ij}\psi_k(x_{(i)})|r] = E[\pi_{ik}\psi_k(x_{(i)})|r]$, where $\pi_{ik}$ is the same as $\pi_i$ in (2.4) except that $x_{(i)}$'s are the ordered $x$-values in the $k$th imputation class and $F_0$ should be replaced by the distribution of $x$ in the $k$th imputation class. From (2.15) in the proof of Theorem 1,

$$E\left[\sum_{i \in S_k, i \le r} \pi_{ik}\psi_k(x_{(i)})|r\right] - E[\psi_k(x)|a=0] = o_p(n_k^{-1/2})$$

where $S_k = S \cap \mathcal{P}_k$, $n_k$ is the number of units in $S_k$, and $n_k \to \infty$ is assumed. Note that

$$\begin{aligned} \bar{y}_{\text{NNI}} &= \sum_{k=1}^{K}\left(\sum_{i \in S_k, i \le r} w_i y_i + \sum_{j \in S_k, j > r} w_j \tilde{y}_j\right) \\ &= \sum_{k=1}^{K}\left(\sum_{i \in S_k, i \le r} w_i y_i + \sum_{j \in S_k, j > r} w_j \sum_{i \in S_k, i \le r} d_{ij} y_i\right) \\ &= \sum_{k=1}^{K} \sum_{i \in S_k, i \le r}\left(w_i + \sum_{j \in S_k, j > r} w_j d_{ij}\right) y_i \end{aligned} \tag{2.17}$$

Hence

$$\begin{aligned}
E(\bar{y}_{\mathrm{NNI}}|r) &= \sum_{k=1}^{K} E\left[\sum_{i\in S_k, i\le r}\left(w_i + \sum_{j\in S_k, j>r} w_j d_{ij}\right) y_i | r\right] \\
&= \sum_{k=1}^{K}\left\{\left(\sum_{i\in S_k, i\le r} w_i\right) E[\psi_k(x)|a=1] + \left(\sum_{j\in S_k, j>r} w_j\right) E\left[\sum_{i\in S_k, i\le r} \pi_{ik}\psi_k(x_i)|r\right]\right\} \\
&= \sum_{k=1}^{K}\left\{\left(\sum_{i\in S_k, i\le r} w_i\right) E[\psi_k(x)|a=1] + \left(\sum_{j\in S_k, j>r} w_j\right) E[\psi_k(x)|a=0]\right\} + o_p(1) \\
&= \sum_{k=1}^{K} \frac{M_k}{N}\{p_k E[\psi_k(x)|a=1] + (1-p_k)E[\psi_k(x)|a=0]\} + o_p(1) \\
&= \sum_{k=1}^{K} \frac{M_k}{N} E[\psi_k(x)] + o_p(1)
\end{aligned}$$

where $p_k = P(a_i = 1)$ for $a_i$'s in the $k$th imputation class and the second-last equality follows from the property of the survey weights $w_i$. Hence $\bar{y}_{\mathrm{NNI}}$ is asymptotically unbiased.

## 3. The Variances of NNI Estimators

It is a common practice to report the survey estimates along with their variance estimates or estimates of coefficient of variation. Having shown that NNI estimators are asymptotically unbiased, in this section we assess the variances of NNI estimators and then derive variance estimators.

### *3.1. Approximate variance formulas*

We first consider $\bar{y}_{\mathrm{NNI}}$ in the simplest case where $S$ is an srs and there is only one imputation class. We adopt the same notation used in Section 2. Let $V(\cdot)$ and $V(\cdot|\cdot)$ denote the variance and conditional variance, respectively, with respect to sampling and the superpopulation in Assumption A. Using the argument of conditioning, we obtain that

$$V(\bar{y}_{\mathrm{NNI}}) = \frac{1}{n^2} E\left[\sum_{i=1}^{r}(1+d_i)^2 V(y_i|x_i)\right] + \frac{1}{n^2} V\left[\sum_{i=1}^{r}(1+d_i)\psi(x_i)\right] \tag{3.1}$$

The first term on the right-hand side of (3.1) is simple and its order is $O(n^{-1})$. For assessing and estimating variances, we need an explicit (approximate) formula for the second term on the right-hand side of (3.1). As in Section 2, we first consider two examples.

**Example 3.** Uniform $F_1$ and $F_0$. Assume model (2.5) and that $F_1 = F_0 = F$ is the uniform distribution on [0,1]. Then, the second term on the right hand-side of (3.1) is

$$\frac{\beta^2}{n} E\left[\frac{1}{12} + \frac{2m(r-3)}{n(r+1)(r+2)(r+3)} + \frac{10m(m-1) - m(r+4)}{n(r+1)(r+2)(r+3)(r+4)}\right] = \frac{\beta^2}{12n} + O\left(\frac{1}{n^3}\right)$$

where $E$ is the asymptotic expectation. Details are omitted.

**Example 4.** Exponential $F_1$ and $F_0$. Assume model (2.5) and that $F_1 = F_0 = F$ is the exponential distribution having mean 1. Then, the second term on the right-hand side of (3.1) is

$$\frac{\beta^2}{n} + O\left(\frac{\log n}{n^2}\right)$$

Details are omitted.

In both examples, the second term on the right-hand side of (3.1) satisfies

$$\frac{1}{n^2} V\left[\sum_{i=1}^{r}(1+d_i)\psi(x_i)\right] = \frac{V[\psi(x)]}{n} + o\left(\frac{1}{n}\right) \tag{3.2}$$

Although we conjecture that result (3.2) is true in general, it is difficult to prove (3.2) for general $\psi$, $F_1$ and $F_0$. The following result provides an approximate formula for $V(\bar{y}_{\mathrm{NNI}})$.

Let $\xi_n$ and $\zeta_n$ be two sequences of random variables satisfying $\zeta_n = o_p(\xi_n)$. Assume that $E(\xi_n)$ and $V(\xi_n)$ exist. Then the asymptotic mean and variance (see, e.g., Akahira and Takeuchi 1991) of $\xi_n + \zeta_n$ are $E(\xi_n)$ and $V(\xi_n)$, respectively. An asymptotic variance is often an approximation to the exact variance.

**Theorem 2.** Assume that $V(y|x) < \infty$, $E|\psi(x)|^6 < \infty$, and that the conditions in Theorem 1 hold. Then the asymptotic variance of $\bar{y}_{NNI}$ is

$$\frac{1}{n^2} E\left[\sum_{i=1}^{r}(1+d_i)^2 V(y_i|x_i)\right] + \frac{V[\psi(x)]}{n} \tag{3.3}$$

The proof is given in the Appendix.

Combining Theorems 1 and 2, we conclude that the asymptotic mean squared error of $\bar{y}_{\mathrm{NNI}}$ is of order $O(n^{-1})$. This result can be extended to the case of stratified sampling and $K$ imputation classes, the situation described in Section 2.2: using (2.17), the asymptotic variance of $\bar{y}_{\mathrm{NNI}}$ is

$$\sum_{k=1}^{K} E\left[\sum_{i\in S_k, i\le r}\left(w_i + \sum_{j\in S_k, j>r} w_j d_{ij}\right)^2 V(y_i|x_i)\right] + \sum_{k=1}^{K} V\left[\sum_{i\in S_k} w_i \psi_k(x_i)\right] \tag{3.4}$$

which reduces to that in (3.3) if $K = 1$ and $w_i$ is proportional to $n^{-1}$.

When $\bar{y}_{\mathrm{NNI}}$ is considered as an estimator of the finite population mean $\bar{Y}$, one should assess the variation of $\bar{y}_{\mathrm{NNI}}$ by $V(\bar{y}_{\mathrm{NNI}} - \bar{Y})$. If $n/N \to 0$ (although $n_h/N_h \not\to 0$ for some $n_h$'s), $V(\bar{y}_{\mathrm{NNI}} - \bar{Y})/V(\bar{y}_{\mathrm{NNI}}) \to 1$. The case of $n/N$ not being negligible is more complicated and will not be discussed here.

The asymptotic variance of $g(\bar{\boldsymbol{y}}_{\mathrm{NNI}})$ is $\nabla g(\bar{\boldsymbol{Y}})' \boldsymbol{V} \nabla g(\bar{\boldsymbol{Y}})$, where $\bar{\boldsymbol{y}}_{\mathrm{NNI}}$ is a vector of NNI sample means, $\boldsymbol{V}$ is the asymptotic covariance matrix of $\bar{\boldsymbol{y}}_{\mathrm{NNI}}$, and $\nabla g(\bar{\boldsymbol{Y}})$ is the vector of partial derivatives of the function $g$ evaluated at $\bar{\boldsymbol{Y}}$, the vector of finite population means.

### *3.2. Variance estimation*

There are some methods for estimating variances of NNI estimators (Kovar and Chen 1994; Rancourt, Särndal, and Lee 1994; Lee, Rancourt and Särndal 1994 and

1995; Montaquila and Jernigan 1997), but few of them are rigorously justified. Using a model-based approach, we derive in this section some asymptotically valid variance estimators for NNI estimators.

From result (3.4), the asymptotic variance of $\bar{y}_{\mathrm{NNI}}$ consists of two terms. We first consider the term involving $\psi_k$'s. If $\psi_k$'s were known, we could use the following textbook estimator of the variance of $\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k} w_i\psi_k(x_i)$ (e.g., Cochran 1977):

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i\in\mathcal{S}_{h,k}}\left[w_i\psi_k(x_i)-\frac{1}{n_h}\sum_{i\in\mathcal{S}_{h,k}} w_i\psi_k(x_i)\right]^2 \tag{3.5}$$

where $\mathcal{S}_{h,k}$ is $\mathcal{S}$ restricted to the $h$th stratum and $k$th imputation class. When $\psi_k$'s are unknown, we assume that there exists a model on $\psi_k(x) = E(y|x)$ within imputation class $k$. The simplest model is the linear model (2.5), but we may also consider some nonlinear or nonparametric models. Let $\hat{\psi}_k$ be the estimators of $\psi_k$ by fitting one of these models using data $y_1,\ldots,y_r$ and $x_1,\ldots,x_r$. Under some weak conditions $\hat{\psi}_k(x)$ is consistent for $\psi_k(x)$ and substituting $\psi_k$ in (3.5) by $\hat{\psi}_k$ results in a consistent estimator of the variance of $\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k} w_i\psi_k(x_i)$, the second term in (3.4).

Next, consider the first term in (3.4). If we do not know anything about $V(y|x)$, then this term can be estimated by

$$\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k,i\le r}\left(w_i+\sum_{j\in\mathcal{S}_k,j>r} w_j d_{ij}\right)^2[y_i-\hat{\psi}_k(x_i)]^2 \tag{3.6}$$

When there is a model for $V(y|x)$, we may obtain an improved estimator. A model for $V(y|x)$ frequently used in surveys is

$$V(y|x)=\sigma_k^2 v_k(x) \quad \text{in imputation class } k \tag{3.7}$$

where $\sigma_k^2$ is unknown but $v_k(x)$ is a known function, e.g., $v_k(x)=|x|^\delta$. If (3.7) holds, we may use the following estimator of the first term in (3.4):

$$\sum_{k=1}^{K}\hat{\sigma}_k^2\sum_{i\in\mathcal{S}_k,i\le r}\left(w_i+\sum_{j\in\mathcal{S}_k,j>r} w_j d_{ij}\right)^2 v_k(x_i) \tag{3.8}$$

with

$$\hat{\sigma}_k^2=\sum_{i\in\mathcal{S}_k,i\le r}[y_i-\hat{\psi}_k(x_i)]^2\Big/\sum_{i\in\mathcal{S}_k,i\le r} v_k(x_i)$$

Our variance estimator for $\bar{y}_{\mathrm{NNI}}$ is then the sum of the quantities in (3.5) and (3.6) (or (3.5) and (3.8)). In the case of srs and one imputation class, it reduces to

$$\frac{1}{n^2}\sum_{i=1}^{r}(1+d_i)^2[y_i-\hat{\psi}(x_i)]^2+\frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\hat{\psi}(x_i)-\frac{1}{n}\sum_{j=1}^{n}\hat{\psi}(x_j)\right]^2$$

or

$$\frac{\hat{\sigma}^2}{n^2}\sum_{i=1}^{r}(1+d_i)^2 v(x_i)+\frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\hat{\psi}(x_i)-\frac{1}{n}\sum_{j=1}^{n}\hat{\psi}(x_j)\right]^2 \tag{3.9}$$

Using Taylor's expansion, we can obtain consistent variance estimators for NNI estimators of the form $g(\bar{y}_{\text{NNI}})$ (see Corollary 1).

## 4. Some Simulation Results

As a complement to our theory, we present in this section some results from a limited simulation study. We examine the biases and variances of $\bar{y}_{\text{NNI}}$ and its variance estimator in the case of srs and one imputation class. The population distribution used to generate $x_i$'s and $y_i$'s is a real data set from the 1988 Current Population Survey (Valliant 1993), where $x$ is the hours worked per week and $y$ is the weekly wage. A plot of this bivariate distribution is given in Figure 1. Some descriptive statistics for $x$ and $y$ are given as follows:

| | min | 1st quartile | median | mean | 3rd quartile | max | variance | skewness | $N$ |
|---|---|---|---|---|---|---|---|---|---|
| $x$ | 1 | 38 | 40 | 38.32 | 40 | 99 | 125 | −0.274 | 10,841 |
| $y$ | 1 | 192 | 320 | 372.3 | 500 | 999 | 59,006 | 0.881 | 10,841 |

Both marginal distributions of $x$ and $y$ are skewed and censored at the right end.

We consider $n = 100$ or 200. The respondents (for $y$) are generated according to the response probability function

$$P(a = 1|x) = \frac{\exp(\gamma_1 + \gamma_2 x)}{1 + \exp(\gamma_1 + \gamma_2 x)} \tag{4.1}$$

with various $\gamma_1$ and $\gamma_2$. When $\gamma_2 = 0$, respondents are generated with equal probability

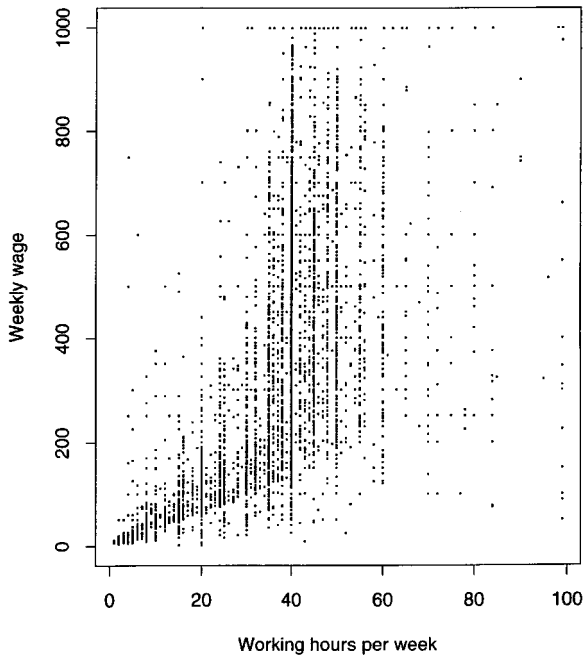


*Fig. 1. 1998 Current Population Survey*

(uniform response); when $\gamma_2 \neq 0$, response rate depends on the value of $x$ (nonuniform response). When uniform response is considered, the response rate is chosen to be between 0.5 and 0.88. Table 1 provides values of $\gamma_1$, $\gamma_2$, the ranges of $P(a = 1|x)$, and the average response rate $\bar{P} = E[P(a = 1|x)]$.

The nonrespondents are imputed by NNI with a single imputation class. The NNI sample mean $\bar{y}_{\text{NNI}}$ is computed according to (2.3). Unlike the NNI sample mean, the use of variance estimator in (3.9) requires a model on $E(y|x)$ and $V(y|x)$. We adopt the following simple but the most commonly used model in sample surveys:

$$E(y|x) = \alpha + \beta x \quad \text{and} \quad V(y|x) = \sigma^2 x \tag{4.2}$$

This is not necessarily the best model for this particular population. In fact, we performed a regression analysis using all data in the population and found that the weighted least squares fitting of model (4.2) yields $\alpha = -51.10$ (with standard error 4.01), $\beta = 11.05$ (with standard error 0.116), residual standard error = 32.63, and multiple $R$-square = 0.45. We also found that a better model for $x$ and $y$ could be obtained by using log-transformations. Nevertheless, we still use model (4.2) in examining the empirical property of the variance estimator for $\bar{y}_{\text{NNI}}$. The variance estimator for $\bar{y}_{\text{NNI}}$ is then computed according to (3.9) with $\hat{\psi}(x) = \hat{\alpha} + \hat{\beta}x, v(x) = x$, and $\hat{\sigma}^2 = \sum_{i \le r}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 / \sum_{i \le r} x_i$, where $\hat{\alpha}$ and $\hat{\beta}$ are the weighted least squares estimators of $\alpha$ and $\beta$ based on the respondents.

Table 2 lists 10,000 Monte Carlo simulation estimates of relative bias (RB) and variance

*Table 1. Parameters in response models*

$P(a = 1|x) = \exp(\gamma_1 + \gamma_2 x)/[1 + \exp(\gamma_1 + \gamma_2 x)]$
$P_- = \min_x P(a = 1|x)$
$P_+ = \max_x P(a = 1|x)$
$\bar{P}$ = average response rate

| Model | $\gamma_1$ | $\gamma_2$ | $P_-$ | $P_+$ | $\bar{P}$ |
|---|---|---|---|---|---|
| 1 | 0 | −0.02 | 0.12 | 0.50 | 0.32 |
| 2 | 0 | −0.01 | 0.27 | 0.50 | 0.41 |
| 3 | 0 | 0.00 | 0.50 | 0.50 | 0.50 |
| 4 | 0 | 0.01 | 0.50 | 0.73 | 0.59 |
| 5 | 0 | 0.02 | 0.50 | 0.83 | 0.68 |
| 6 | 1 | −0.03 | 0.12 | 0.73 | 0.46 |
| 7 | 1 | −0.02 | 0.27 | 0.73 | 0.56 |
| 8 | 1 | −0.01 | 0.50 | 0.73 | 0.65 |
| 9 | 1 | 0.00 | 0.73 | 0.73 | 0.73 |
| 10 | 1 | 0.01 | 0.73 | 0.88 | 0.80 |
| 11 | 2 | −0.04 | 0.12 | 0.88 | 0.61 |
| 12 | 2 | −0.03 | 0.27 | 0.88 | 0.70 |
| 13 | 2 | −0.02 | 0.50 | 0.88 | 0.77 |
| 14 | 2 | −0.01 | 0.73 | 0.88 | 0.83 |
| 15 | 2 | 0.00 | 0.88 | 0.88 | 0.88 |

*Table 2. Empirical results for NNI, mean and random hot deck imputation*

| $n$ | Model | RB($\bar{y}_R$) | RB($\bar{y}_{RHD}$) | RB($\bar{y}_{NNI}$) | $V(\bar{y}_R)$ | $V(\bar{y}_{RHD})$ | $V(\bar{y}_{NNI})$ | RB($\hat{V}$) | SD($\hat{V}$) |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 1 | −0.050 | −0.007 | 0.000 | 1,896.4 | 1,976.4 | 1,998.2 | −0.051 | 666.1 |
| | 2 | −0.022 | −0.003 | 0.001 | 1,452.6 | 1,535.6 | 1,529.8 | −0.012 | 453.6 |
| | 3 | 0.002 | 0.002 | 0.002 | 1,202.4 | 1,274.6 | 1,235.8 | 0.012 | 319.7 |
| | 4 | 0.015 | 0.003 | 0.001 | 993.8 | 1,055.8 | 1,031.5 | 0.019 | 236.7 |
| | 5 | 0.026 | 0.006 | 0.001 | 875.5 | 934.9 | 914.1 | 0.011 | 191.2 |
| | 6 | −0.060 | −0.010 | 0.000 | 1,271.3 | 1,416.8 | 1,459.9 | −0.062 | 388.9 |
| | 7 | −0.032 | −0.006 | 0.000 | 1,068.8 | 1,178.4 | 1,184.5 | −0.040 | 279.6 |
| | 8 | −0.013 | −0.002 | 0.000 | 898.1 | 966.4 | 972.7 | −0.002 | 212.5 |
| | 9 | 0.001 | 0.001 | 0.001 | 781.3 | 844.9 | 840.6 | 0.021 | 169.9 |
| | 10 | 0.008 | 0.002 | 0.000 | 735.4 | 776.1 | 760.4 | 0.018 | 146.4 |
| | 11 | −0.054 | −0.009 | 0.000 | 958.6 | 1,097.8 | 1,109.7 | −0.044 | 258.1 |
| | 12 | −0.031 | −0.005 | 0.001 | 828.8 | 920.9 | 937.8 | −0.024 | 195.2 |
| | 13 | −0.018 | −0.005 | −0.001 | 741.5 | 816.5 | 814.5 | −0.005 | 158.2 |
| | 14 | −0.007 | −0.002 | 0.000 | 719.5 | 776.6 | 759.6 | −0.025 | 138.3 |
| | 15 | −0.001 | −0.001 | −0.001 | 655.5 | 684.8 | 682.3 | 0.009 | 122.4 |
| 200 | 1 | −0.053 | −0.010 | −0.001 | 922.2 | 956.4 | 1,003.2 | −0.044 | 236.9 |
| | 2 | −0.023 | −0.005 | 0.000 | 734.3 | 760.6 | 765.1 | −0.006 | 156.3 |
| | 3 | 0.000 | 0.001 | 0.001 | 587.5 | 621.3 | 612.2 | 0.021 | 110.8 |
| | 4 | 0.016 | 0.004 | 0.000 | 485.5 | 516.4 | 513.4 | 0.032 | 83.7 |
| | 5 | 0.026 | 0.006 | 0.001 | 427.9 | 458.7 | 442.6 | 0.046 | 66.8 |
| | 6 | −0.059 | −0.010 | 0.001 | 636.2 | 701.9 | 711.5 | −0.029 | 138.1 |
| | 7 | −0.033 | −0.006 | 0.001 | 524.3 | 580.1 | 585.7 | −0.024 | 98.2 |
| | 8 | −0.013 | −0.003 | 0.000 | 453.4 | 482.5 | 493.7 | −0.012 | 74.8 |
| | 9 | −0.001 | −0.001 | −0.001 | 386.5 | 419.0 | 407.1 | 0.050 | 59.9 |
| | 10 | 0.007 | 0.000 | −0.001 | 363.9 | 384.1 | 373.7 | 0.034 | 51.2 |
| | 11 | −0.054 | −0.009 | 0.001 | 469.0 | 531.7 | 556.4 | −0.039 | 91.9 |
| | 12 | −0.031 | −0.005 | 0.002 | 424.4 | 473.1 | 483.4 | −0.048 | 69.6 |
| | 13 | −0.016 | −0.003 | 0.001 | 381.9 | 410.9 | 411.1 | −0.013 | 56.5 |
| | 14 | −0.005 | 0.000 | 0.001 | 352.1 | 374.0 | 372.4 | −0.004 | 48.3 |
| | 15 | 0.001 | 0.001 | 0.001 | 334.3 | 347.7 | 343.1 | 0.006 | 43.8 |

$n$ = sample size, $\bar{Y} = 372.3$

of $\bar{y}_{NNI}$, the RB of the variance estimate $\hat{V}$ of $V(\bar{y}_{NNI})$, and the standard deviation (SD) of $\hat{V}$, for different $n$ and response models under consideration.

For reference purposes, we also obtained the RB and variances of two estimators using different imputation methods (Table 2). The first one, $\bar{y}_R$, is obtained by mean imputation, which is the same as the sample mean of responses only. The second one, $\bar{y}_{RHD}$, is obtained by random hot deck imputation. To achieve a more efficient random hot deck imputation, we divide the sample into three sub-imputation classes, according to whether the value of the $x$ is $< 40$, $= 40$, and $> 40$, and perform the random imputation within each class. When a class does not contain any responses, respondents in a neighborhood class are used.

Unlike $\bar{y}_{NNI}$, both $\bar{y}_R$ and $\bar{y}_{RHD}$ are asymptotically biased under the response model (4.1) when $\gamma_2 \neq 0$ (nonuniform response).

The following is a summary of the results in Table 2.

1. The performance of $\bar{y}_{NNI}$ is very good. The population mean in this problem is 372.3

and the RB of $\bar{y}_{\mathrm{NNI}}$ is within 0.2% range. Thus, the bias of $\bar{y}_{\mathrm{NNI}}$ is negligible regardless of the response rate and of whether the response is uniform. This confirms our theoretical result. The variance of $\bar{y}_{\mathrm{NNI}}$ increases as the number of nonrespondents increases.

2. Although model (4.2) is not perfect, the performance of the variance estimator $\hat{V}$ for $\bar{y}_{\mathrm{NNI}}$ is still good. Its relative bias ranges from $-6.2\%$ to 2.1% in the case of $n = 100$ and $-4.4\%$ to 5.0% in the case of $n = 200$. The standard deviation of $\hat{V}$ increases as the number of nonrespondents increases.
3. $V(\bar{y}_{\mathrm{NNI}})$ is about the same as that of $V(\bar{y}_{\mathrm{RHD}})$ and is larger than $V(\bar{y}_{\mathrm{R}})$. However, $\bar{y}_{\mathrm{R}}$ has a small but nonnegligible bias when $\gamma_2 \neq 0$. For example, under model 1, the RB of $\bar{y}_{\mathrm{R}}$ is $-0.05$ which is small but nonnegligible compared with the relative stability of $\bar{y}_{\mathrm{R}} = \sqrt{V(\bar{y}_{\mathrm{R}})}/\bar{Y} = 0.117$ ($n = 100$) or 0.082 ($n = 200$).

## Appendix

### Proof of (2.15)

To study the expectations of the integrals in (2.14), we consider the following four cases.

**Case 1:** the integration limits of the integral in (2.14) are within the interval $[M_1, M_2]$. When $M_1 \le (x_i + x_{i-1})/2$ and $(x_i + x_{i+1})/2 \le M_2$,

$$\left| \int_{(x_i+x_{i-1})/2}^{(x_i+x_{i+1})/2} [\psi(x_i) - \psi(t)] dF_0(t) \right| \le C(x_{i+1} - x_i)[F_0(x_{i+1}) - F_0(x_i)] + C(x_i - x_{i-1})[F_0(x_i) - F_0(x_{i-1})]$$

Let $f_0$ and $f_1$ be the densities of $F_0$ and $F_1$, respectively. Under condition (2.11), it can be shown that $f_0(t)/f_1(t) \le c_0$ for a constant $c_0 > 0$. Then $|F_0(s) - F_0(t)| \le c_0 |F_1(s) - F_1(t)|$. Note that

$$\begin{aligned}
&\sum_{i=1}^{r-1} E\{(x_{i+1} - x_i)[F_1(x_{i+1}) - F_1(x_i)] | r\} \\
&\quad = r(r-1) \int_{t<s} (s-t)[F_1(s) - F_1(t)][1 + F_1(t) - F_1(s)]^{r-2} dF_1(t) dF_1(s) \\
&\quad = r \int t[F_1(t)]^{r-1}[1 - F_1(t)] dF_1(t) - r \int t F_1(t)[1 - F_1(t)]^{r-1} dF_1(t) \\
&\quad + \int t[F_1(t)]^r dF_1(t) - \int t[1 - F_1(t)]^r dF_1(t)
\end{aligned}$$

which is of order $o(r^{-1/2})$ under the finite third order moment condition on $x$. Also, $r^{1/2} \approx p^{-1/2} n^{-1/2}$ for large $n$, where $p = P(a = 1)$. Hence, the sum of the expectations of the integrals in (2.14) with integration limits within $[-M_1, M_2]$ is of order $o_p(n^{-1/2})$.

**Case 2**: the integration limits of the integral in (2.14) are finite and outside $[M_1, M_2]$. Without loss of generality, assume $\psi(t)$ is an increasing function when $t > M_2$. When

$(x_i + x_{i-1})/2 > M_2$ and $(x_i + x_{i+1})/2 < \infty$

$$\left| \int_{(x_i+x_{i-1})/2}^{(x_i+x_{i+1})/2} [\psi(x_i) - \psi(t)] dF_0(t) \right| \le [\psi(x_{i+1}) - \psi(x_i)][F_0(x_{i+1}) - F_0(x_i)]$$
$$+ [\psi(x_i) - \psi(x_{i-1})][F_0(x_i) - F_0(x_{i-1})]$$
$$\le c_0[\psi(x_{i+1}) - \psi(x_i)][F_1(x_{i+1}) - F_1(x_i)]$$
$$+ c_0[\psi(x_i) - \psi(x_{i-1})][F_1(x_i) - F_1(x_{i-1})]$$

Let $\tilde{\psi}(t) = \max[\psi(t), M_2]$. Note that

$$\sum_{i=1}^{r-1} E\left\{[\tilde{\psi}(x_{i+1}) - \tilde{\psi}(x_i)][F_1(x_{i+1}) - F_1(x_i)]|r\right\}$$
$$= r(r-1) \int_{t<s} [\tilde{\psi}(s) - \tilde{\psi}(t)][F_1(s) - F_1(t)][1 + F_1(t) - F_1(s)]^{r-2} dF_1(t) dF_1(s)$$
$$= r \int \tilde{\psi}(t)[F_1(t)]^{r-1}[1 - F_1(t)] dF_1(t) - r \int \tilde{\psi}(t) F_1(t)[1 - F_1(t)]^{r-1} dF_1(t)]$$
$$+ \int \tilde{\psi}(t)[F_1(t)]^r dF_1(t) - \int \tilde{\psi}(t)[1 - F_1(t)]^r dF_1(t)$$

which is of order $o(r^{-1/2})$ under the finite third order moment condition on $\psi(x)$. Hence, the sum of the expectations of the integrals in (2.14) with $(x_i + x_{i-1})/2 > M_2$ and $(x_i + x_{i+1})/2 < \infty$ is of order $o_p(n^{-1/2})$. Similarly, the sum of the expectations of the integrals in (2.14) with $(x_i + x_{i-1})/2 > -\infty$ and $(x_i + x_{i+1})/2 < M_1$ is also of order $o_p(n^{-1/2})$.

**Case 3**: the integration limits of the integral in (2.14) are finite and exactly one of them is inside $[M_1, M_2]$. For $t$ in the interval containing $M_2$,

$$|\psi(t) - \psi(x_i)| \le C|M_2 - x_{i-1}| + |\psi(x_{i+1}) - \psi(M_2)| \le 2CM_2 + |\psi(x_{i+1}) - \psi(M_2)|$$

Since $E(\pi_i) = O(n^{-1})$ and $E[\pi_i|\psi(x_{i+1}) - \psi(M_2)|] = o(n^{-1/2})$ (which can be shown using the same argument used in (2)), the expectation of the integral in (2.14) with integration interval containing $M_2$ is of order $o_p(n^{-1/2})$. Similarly, the expectation of the integral in (2.14) with integration interval containing $M_1$ is also of order $o_p(n^{-1/2})$.

**Case 4**: one of the integration limits of the integral in (2.14) is not finite. Consider

$$\int_{(x_{r-1}+x_r)/2}^{\infty} [\psi(x_r) - \psi(t)] dF_0(t) \tag{A.1}$$

Assume $P(x > M_2) > 0$ (otherwise, it is not necessary to consider the integral with upper integration limit $\infty$). Then $P(x_{r-1} > M_2) \le 1 - r[P(x < M_2)]^{r-1}$. Consequently, we can assume $x_{r-1} > M_2$ because the integration from $x_{r-1}$ to $M_2$ is exponentially small. Thus, the integral in (A.1) is bounded by

$$\int_{x_{r-1}}^{\infty} [\psi(t) - \psi(M_2)] dF_0(t) \le [1 - F_0(x_{r-1})]^{1/2} \left\{ \int_{x_{r-1}}^{\infty} [\psi(t) - \psi(M_2)]^2 dF_0(t) \right\}^{1/2}$$

Its expectation is of order $o_p(n^{-1/2})$ under the moment assumptions on $x$ and $\psi(x)$. Similarly, the expectation of the integral

$$\int_{-\infty}^{(x_1+x_2)/2} [\psi(x_1) - \psi(t)]dF_0(t)$$

is also of order $o_p(n^{-1/2})$.

Combining (1)–(4), we obtain (2.15).

**Proof of Theorem 2**

From (3.1), it suffices to show that the asymptotic variance of $n^{-1}\sum_{i=1}^{r}(1+d_i)\psi(x_i)$ is the second term in (3.3). Let $z_i = \psi(x_i) - E[\psi(x)|a=0]$. For simplicity we assume $x_{(i)} = x_i$. Note that

$$\begin{aligned} V\left[\frac{1}{n}\sum_{i=1}^{r}(1+d_i)\psi(x_i)\right] &= V\left[\frac{1}{n}\sum_{i=1}^{r}(1+d_i)z_i\right] \\ &= V\left\{E_d\left[\frac{1}{n}\sum_{i=1}^{r}(1+d_i)z_i\right]\right\} + E\left\{V_d\left[\frac{1}{n}\sum_{i=1}^{r}(1+d_i)z_i\right]\right\} \end{aligned} \quad \text{(A.2)}$$

where $E_d$ and $V_d$ are the conditional expectation and variance, given $x_1, \ldots, x_r$, in addition to conditioning on $r$, as always. By (2.15) and the fact that $E(z|a=0) = 0$,

$$\sum_{i=1}^{r}\pi_i z_i = o_p(n^{-1/2}) \quad \text{(A.3)}$$

Then

$$E_d\left[\frac{1}{n}\sum_{i=1}^{r}(1+d_i)z_i\right] = \frac{1}{n}\sum_{i=1}^{r} z_i + o_p(n^{-3/2})$$

and its asymptotic variance is

$$V\left(\frac{1}{n}\sum_{i=1}^{r} z_i\right) = \frac{pV[\psi(x)|a=1] + p(1-p)(\mu_1-\mu_0)^2}{n} + o_p\left(\frac{1}{n}\right) \quad \text{(A.4)}$$

where $\mu_v = E[\psi(x)|a=v], v=0,1$. Let $d_{ij}$ be the indicator defined in (2.16).

Then $d_i = \sum_{j=r+1}^{n} d_{ij}$ and

$$
\begin{aligned}
E_d\left(\sum_{i=1}^{r} d_i z_i\right)^2 &= Ed\left[2\sum_{r+1\le j<k\le n}\left(\sum_{i=1}^{r} d_{ij}z_i\right)\left(\sum_{i=1}^{r} d_{ik}z_i\right)\right] + E_d\left[\sum_{j=r+1}^{n}\left(\sum_{i=1}^{r} d_{ij}z_i\right)^2\right] \\
&= m(m-1)\left[E_d\left(\sum_{i=1}^{r} d_{ij}z_i\right)\right]^2 + E_d\left[\sum_{j=r+1}^{n}\left(\sum_{i=1}^{r} d_{ij}z_i\right)^2\right] \\
&= m(m-1)\left[E_d\left(\sum_{i=1}^{r} d_{ij}z_i\right)\right]^2 + mE_d\left(\sum_{i=1}^{r} d_{ij}z_i^2\right) \\
&= m(m-1)\left(\sum_{i=1}^{r} \pi_i z_i\right)^2 + m\left(\sum_{i=1}^{r} \pi_i z_i^2\right)
\end{aligned}
$$

where the second equality holds because $d_{ij}$ and $d_{ik}$ are conditionally independent and have the same distribution, and the third equality holds because $d_{ij}^2 = d_{ij}$ and $d_{ij}d_{lj} = 0$ for $i \neq l$ (there is no tied $x_i$'s). Then,

$$
\begin{aligned}
V_d\left[\frac{1}{n}\sum_{i=1}^{r}(1+d_i)z_i\right] &= \frac{1}{n^2}\left\{E_d\left(\sum_{i=1}^{r} d_i z_i\right)^2 - \left[E_d\left(\sum_{i=1}^{r} d_i z_i\right)\right]^2\right\} \\
&= \frac{m}{n^2}\sum_{i=1}^{r}\pi_i z_i^2 + \frac{m(m-1)-1}{n^2}\left(\sum_{i=1}^{r}\pi_i z_i\right)^2 \\
&= \frac{m}{n^2}\sum_{i=1}^{r}\pi_i z_i^2 + 0_p\left(\frac{1}{n}\right)
\end{aligned}
$$

(by (A.3)) and its asymptotic mean is

$$
\begin{aligned}
E\left(\frac{m}{n^2}\sum_{i=1}^{r}\pi_i z_i^2\right) &= E\left[\frac{m}{n^2}E(z^2|a=0) + o\left(\frac{m}{n^2}\right)\right] \\
&= \frac{(1-p)V[\psi(x)|a=0]}{n} + o_p\left(\frac{1}{n}\right) \qquad \text{(A.5)}
\end{aligned}
$$

where the first equality follows from the proof of Theorem 1. From (A.2), the asymptotic variance of $n^{-1}\sum_{i=1}^{r}(1+d_i)\psi(x_i)$ is the sum of the quantities in (A.4) and (A.5). The result follows from the fact that

$$
\begin{aligned}
V[\psi(x)] &= E\{V[\psi(x)|a]\} + V\{E[\psi(x)|a]\} \\
&= pV[\psi(x)|a=1] + (1-p)V[\psi(x)|a=0] + p(1-p)(\mu_1-\mu_0)^2
\end{aligned}
$$

## 5. References


Akahira, M. and Takeuchi, K. (1991). On the Definition of Asymptotic Expectations. Asymptotic Theory of Statistical Estimation. In M. Akahira (ed.): Institute of Mathematics, University of Tsukuba, Japan.

Cochran, W.G. (1977). Sampling Techniques, Third Edition, New York: Wiley.

Cotton, C. (1991). Functional Description of the Generalized Edit and Imputation System. Business Survey Division, Statistics Canada.

Fay, R.E. (1996). Replication-Based Variance Estimators for Imputed Survey Data from Finite Populations. Preprint.

Kalton, G. (1981). Compensating for Missing Data. ISR Research Report Series. Ann Arbor: Survey Research Center, University of Michigan.

Kovar, J.G. and Chen, E.J. (1994). Jacknife Variance Estimation of Imputed Survey Data. Survey Methodology, 20, 45–52.

Kovar, J.G., Whitridge, P., and MacMillan, J. (1998). Generalized Edit and Imputation System for Economic Surveys at Statistics Canada. Proceedings of the Section on Survey Research Methods, American Statistical Association, 690–695.

Lee, H., Rancourt, E., and Särndal, C.E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. Journal of Official Statistics, 10, 231–243.

Lee, H., Rancourt, E., and Särndal, C.E. (1995). Variance Estimation in the Presence of Imputed Data for the Generalized Estimation System. Proceedings of the Section on Survey Research Methods, American Statistical Association, 384–389.

Montaquila, J.M. and Robort W. Jernigan, R.W. (1997). Variance Estimation in the Presence of Imputed Data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 273–277.

Rancourt, E., Särndal, C.E., and Lee, H. (1994) Estimation of the Variance in the Presence of Nearest Neighbor Imputation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 888–893.

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

Sedransk, J. (1985). The Objective and Practice of Imputation. Proceedings of the First Annual Research Conference, U.S. Bureau of the Census, Washington D.C., 445–452.

Valliant, R. (1993). Post-Stratification and Conditional Variance Estimation. Journal of the American Statistical Association, 88, 89–96.

Received April 1998
Revised March 1999