BIASES AND VARIANCES OF SURVEY ESTIMATORS BASED ON NEAREST NEIGHBOR IMPUTATION

Jiahua Chen, University of Waterloo, Jun Shao, University of Wisconsin-Madison
Jiahua Chen, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ont N2L 3G1, Canada

Key Words: 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 and the U.S. Census Bureau. A computer software, the Generalized Edit and Imputation System, provides a simple way of performing NNI. Although favoured by many users, there has little been done on the theoretical properties of the NNI method. In particular, it is not clear that under what conditions the NNI method is asymptotically unbiased. There has been no theoretically justified variance estimators. In this paper, we derive a bound for the asymptotic bias, obtain the asymptotic variance of the NNI method and construct its estimator which is shown to be good via simulation.

For brevity, we only discuss the NNI method in the simplest case. More detailed discussion can be found from Chen and Shao (1997). Consider a bivariate sample $(x_1, y_1), ..., (x_n, y_n)$ and suppose that the first 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. The NNI method imputes a missing y_j, $r+1 \leq j \leq n$, by y_i $(1 \leq i \leq r)$ such that

$$|x_i - x_j| = \min_{1 \leq l \leq r} |x_l - x_j|. \qquad (1.1)$$

If there are tied x-values, i may be randomly selected from them.

The NNI method has some nice features. First, it is a hot deck method in the sense that nonrespondents are substituted by respondents from the same variable; the imputed values are actually occurring values, not constructed values. Second, the NNI method may be more efficient than other hot deck methods. Third, the NNI method does not use an explicit model relating y and x. Finally, one of the results in the current paper shows that the NNI method provides asymptotically valid distribution and quantile estimators.

2. The Biases of NNI Estimators

Consider a size N finite population with (x_i, y_i, a_i) being the characteristics of the ith unit. We assume $a_i = 1$ if y_i is a respondent and $a_i = 0$ otherwise, and let $\mathcal{A}$ be the vector of a_i, $1 \leq i \leq N$.

Assumption A. All units (x_i, y_i, a_i)'s are iid realizations from a super-population and $P(a_i = 1|x_i, y_i) = P(a_i = 1|x_i)$.

The assumption requires the response probability $P(a = 1|x, y)$ depends on x_i, y_i through x component only. 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. If F is the marginal distribution of x and $p = P(a = 1)$, then

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

and

$$P(x \leq t|a = 0) = P(a = 0|x \leq t)F(t)/(1 - p) = F_0(t). \qquad (2.2)$$

This means that conditional on a_i's, x_i's may have two different distributions.

Assume the sampled units are $\mathcal{S} = \{1, 2, \ldots, n\}$. For $r+1 \leq j \leq n$, 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, $1 \le i \le r$. Let $x_{(1)}, \dots, x_{(r)}$ be order statistics. Under assumption A,

$$d_{(i)}|x_1, ..., x_r, \mathcal{S}, \mathcal{A} \sim \text{binomial}(m, \pi_i) \tag{2.4}$$

with $\pi_i = F_0\left(\frac{x_{(i+1)}+x_{(i)}}{2}\right) - F_0\left(\frac{x_{(i)}+x_{(i-1)}}{2}\right)$, $i = 1, ..., r$, where $x_{(0)} = -\infty$ and $x_{(r+1)} = +\infty$.

Let us first consider two interesting special cases.

Example 1. Symmetric F_1 and F_0. Assume that

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

where α and β are unknown parameters, and that $F_1 = F_0 = F$ and F is symmetric. Then $\bar{y}_{\text{NNI}}$ is exactly unbiased, i.e.,

$$E\left(\bar{y}_{\text{NNI}} - \bar{Y}\right) = 0, \tag{2.6}$$

where E is the expectation with respect to x and y, given $\mathcal{S}$ and $\mathcal{A}$, and $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ is the finite population mean.

Its proof can be done by observing the symmetry of $d_{(i)}x_{(i)} + d_{(r-i+1)}x_{(r-i+1)}$.

Thus, $\bar{y}_{\text{NNI}}$ is exactly unbiased if x_i is symmetrically distributed. In survey problems, however, the distribution of x_i's is seldom symmetric. If F_1 and F_0 are not symmetric, we expect that $\bar{y}_{\text{NNI}}$ is biased.

Example 2. Assume linear model (2.5) and that $F_1 = F_0 = F$ and F is the exponential distribution with mean 1. We have

$$E\{\bar{y}_{NNI} - \bar{Y}\} \tag{2.7}$$

$$= -\frac{\beta m}{n(r+1)}\left[\frac{r-1}{r} + \sum_{i=0}^{r-2}\frac{1}{(2i+3)^2}\right]. \tag{2.8}$$

That is, $\bar{y}_{\text{NNI}}$ is biased unless $\beta = 0$.

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

Theorem 1. Suppose that (i) assumption A holds; (ii) there exist nonnegative constants M and C (M may be $+\infty$) such that the function $\psi(x) = E(y|x)$ is a monotone function when $|t| > M$ and $|\psi(t) - \psi(s)| \le C|t-s|$ when $|t| \le M$ and $|s| \le M$; (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

$$\inf_{x \in \mathcal{D}} P(a = 1|x) > 0, \tag{2.9}$$

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

$$E(\bar{y}_{\text{NNI}} - \bar{Y}) = o(n^{-1/2}), \tag{2.10}$$

where the expectation is with respect to x, y, $\mathcal{S}$ and $\mathcal{A}$.

Remark 1. The assumption on ψ is very general. The NNI method requires almost no model between variables x and y. Condition (2.9) roughly means that there are some y-respondents for every x-value. 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 $\mathcal{D}_1$.

Remark 2. From (2.10), $\bar{y}_{\text{NNI}}$ is asymptotically unbiased for the population mean $\bar{Y}$. 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 $\bar{y}_{\text{NNI}}$ is a $\sqrt{n}$-consistent estimator of $\bar{Y}$.

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 immediately conclude that $g(\bar{\boldsymbol{y}}_{\text{NNI}})$ is asymptotically unbiased for $g(\bar{\boldsymbol{Y}})$ when g is a differentiable function.

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

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

is asymptotically unbiased for the finite population distribution, $F(t) = N^{-1}\sum_{i=1}^{N} I_{y_i}(t)$.

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

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 again consider $\bar{y}_{\text{NNI}}$ in the simplest case where $\mathcal{S}$ is an srs, $n/N \to 0$. Using the argument of conditioning, we obtain that

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

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). Like in Section 2, we first consider two interesting examples.

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

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

Example 4. Exponential F_1 and F_0. Assume model (2.5) and that $F_1 = F_0 = F$ and F is the exponential distribution having mean 1. Then, the second term on the right hand side of (3.1) is $n^{-1}\beta^2 + O(n^{-2}\log n)$.

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

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

Although we conjecture that result (3.2) is true in general, it is difficult to prove (3.2) for general ψ, F_1 and F_0. The following result provides an approximate formula for $V(\bar{y}_{\text{NNI}})$. See Chen and Shao (1997) for details.

Theorem 2. Under the conditions in Theorem 1, the asymptotic variance of $\bar{y}_{\text{NNI}}$ is

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

3.2. Variance Estimation

There are some methods for estimating variances of NNI estimators, but none of them is theoretically justified, which is perhaps the reason why these methods did not perform well in simulation studies (Kovar and Chen 1994, Rancourt, Särndal and Lee 1994, Lee, Rancourt and Särndal 1994 and 1995). Using a model-assisted approach, we derive in this section some asymptotically valid variance estimators for NNI estimators. We assume that $n/N \to 0$.

From result (3.3), the asymptotic variance of $\bar{y}_{\text{NNI}}$ consists of two terms. We first consider the term involving ψ. If ψ were known, we could use the following text book estimator of the variance of $\sum_{i\in\mathcal{P}} w_i\psi(x_i)$ (e.g., Cochran 1977):

$$\frac{n}{n-1}\sum_{i\in\mathcal{P}}\left[w_i\psi(x_i)-\frac{1}{n}\sum_{i\in\mathcal{P}}w_i\psi(x_i)\right]^2. \quad (3.4)$$

When ψ is unknown, we assume there exists a model $\psi(x) = E(y|x)$ for the population. The simplest model is the linear model (2.5), but we may also consider some nonlinear or nonparametric models. Let $\hat{\psi}$ be the estimators of ψ by fitting one of these models, using data $y_1, \dots, y_r$ and

$x_1, \dots, x_r$. Under some weak conditions $\hat{\psi}(x)$ is consistent for $\psi(x)$. Substituting $\psi(x)$ in (3.4) by $\hat{\psi}$ results in a consistent estimator of the second term in (3.3).

Next, consider the first term in (3.3). If we don't know anything about $V(y|x)$, then this term can be estimated by

$$\sum_{i\in\mathcal{P}, i\leq r} \left(w_i + \sum_{j\in\mathcal{P}, j>r} w_j d_{ij} \right)^2 [y_i - \hat{\psi}(x_i)]^2, \tag{3.5}$$

where $d_{ij} = 1$ if i is the nearest neighbor of j, and $d_{ij} = 0$ otherwise. 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 $V(y|x) = \sigma^2 v(x)$ where σ^2 is unknown but $v(x)$ is a known function, e.g., $v(x) = |x|^\delta$. In the case of srs and one imputation class, (3.5) reduces to

$$\frac{\hat{\sigma}^2}{n^2} \sum_{i=1}^{r} (1 + d_i)^2 v(x_i) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \left[\hat{\psi}(x_i) - \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}(x_i) \right]^2 \tag{3.6}$$

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 1988 Current Population Survey (Valliant 1993), where x is the hours worked per week and y is the weekly wage.

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

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

with various γ_1 and γ_2 (see Table 1). When $\gamma_2 = 0$, respondents are generated with equal probability (uniform response); when $\gamma_2 \neq 0$, response rate depends on the value of x (non-uniform response). When uniform response is considered, the response rate is chosen to be between 0.5 and 0.88.

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 (3.6) 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:

$$E(y|x) = \alpha + \beta x \qquad \text{and} \qquad V(y|x) = \sigma^2 x. \tag{4.1}$$

The variance estimator for $\bar{y}_{\text{NNI}}$ is then computed according to (3.6) with $\hat{\psi}(x) = \hat{\alpha} + \hat{\beta} x$, $v(x) = x$, and $\hat{\sigma}^2 = \sum_{i\leq r}(y_i - \hat{\alpha} - \hat{\beta} x_i)^2 / \sum_{i\leq r} x_i$, where $\hat{\alpha}$ and $\hat{\beta}$ are the weighted least squares estimators of α and β based on the respondents.

Table 1 lists 1,000 Monte Carlo simulation estimates of the mean values of $\bar{y}_{\text{NNI}}$ and its variance estimate $\widehat{V}$, the variance of $\bar{y}_{\text{NNI}}$, the relative bias of $\widehat{V}$, and the standard deviation of $\widehat{V}$ for different values of n, γ_1, and γ_2. The ranges of the response rate $P(a = 1|x)$ are also given. The following is a summary of the results in Table 1.

1. The performance of $\bar{y}_{\text{NNI}}$ is very good. The population mean in this problem is 372.3 and the relative bias of $\bar{y}_{\text{NNI}}$ ranges from −1.1% to 0.4%. Thus, the bias of $\bar{y}_{\text{NNI}}$ is negligible regardless of the nonresponse rate and whether the nonresponse is uniform. This confirms our theoretical result. The variance of $\bar{y}_{\text{NNI}}$ increases as the number of nonrespondents increases, but does not depend on whether the nonresponse is uniform or not.

2. Although model (4.1) is not perfect, the performance of the variance estimator $\widehat{V}$ for $\bar{y}_{\text{NNI}}$ is still good. Its relative bias ranges from −11.1% to 5.4% in the case of $n = 100$ and −6.4% to 8.6% in the case of $n = 200$. The standard deviation of $\widehat{V}$ increases as the number of nonrespondents increases, but does not depend on whether the nonresponse is uniform or not.

References

Chen, J. and Shao, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. Technical report stat-97-02. Department of Statistics and Actuarial Science, University of Waterloo.

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

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). Jackknife variance estimation of imputed survey data. *Survey Methodology*, **20**, 45-52.

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.

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*, Wiley, New York.

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

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, **88**, 89-96.

Table 1. Empirical estimates of $E(\bar{y}_{\text{NNI}})$, $V(\bar{y}_{\text{NNI}})$, $E(\widehat{V})$, relative bias (RB) of $\widehat{V}$, and standard deviation (SD) of $\widehat{V}$, based on 1,000 simulations

n	γ_1	γ_2	P_-	P_+	$E(\bar{y}_{\text{NNI}})$	$V(\bar{y}_{\text{NNI}})$	$E(\widehat{V})$	RB($\widehat{V}$)	SD($\widehat{V}$)
100	2	0.00	0.88	0.88	372.4	728.9	690.5	−0.053	123.3
		−0.01	0.73	0.88	372.4	757.9	740.7	−0.023	134.5
		−0.02	0.50	0.88	372.0	866.5	809.1	−0.066	158.3
		−0.03	0.27	0.88	373.6	925.4	915.2	−0.011	189.5
		−0.04	0.12	0.88	372.3	1102.6	1056.6	−0.042	250.6
	1	0.00	0.73	0.73	373.6	842.4	853.3	0.013	164.8
		0.01	0.73	0.88	373.9	738.0	778.1	0.054	144.8
		−0.01	0.50	0.73	372.2	1039.9	969.7	−0.068	208.2
		−0.02	0.27	0.73	373.4	1151.3	1142.6	−0.008	275.4
	0	0.00	0.50	0.50	372.9	1248.5	1235.6	−0.010	319.4
		0.01	0.50	0.73	371.8	1070.7	1052.6	−0.017	239.7
		0.02	0.50	0.83	372.8	1028.5	914.1	−0.111	185.9
		−0.01	0.27	0.50	369.9	1558.8	1490.1	−0.044	467.3
		−0.02	0.12	0.50	368.0	2026.8	1852.3	−0.086	639.3
200	2	0.00	0.88	0.88	371.8	346.3	344.7	−0.005	43.8
		0.01	0.88	0.95	371.6	320.3	328.9	0.027	39.1
		0.02	0.88	0.98	372.6	301.4	317.7	0.054	37.3
		−0.01	0.73	0.88	372.7	358.6	371.3	0.035	48.0
		−0.02	0.50	0.88	372.3	401.2	405.2	0.010	54.1
	1	0.00	0.73	0.73	372.5	412.3	431.9	0.047	60.1
		0.01	0.73	0.88	372.3	354.1	384.6	0.086	49.5
		0.02	0.73	0.95	372.8	332.4	358.9	0.080	46.4
		−0.01	0.50	0.73	372.9	521.6	488.2	−0.064	76.8
		−0.02	0.27	0.73	372.5	608.0	572.2	−0.059	100.0
	0	0.00	0.50	0.50	373.0	589.9	621.9	0.054	107.0
		0.01	0.50	0.73	371.5	532.4	524.4	−0.015	85.9
		0.02	0.50	0.88	372.0	434.0	461.8	0.064	66.8
		−0.01	0.27	0.50	372.8	752.9	760.3	0.010	154.3
		−0.02	0.12	0.50	371.6	936.5	964.0	0.029	230.3

n = sample size $P_- = \min_x P(a=1|x)$ $P_+ = \max_x P(a=1|x)$

$P(a=1|x) = \exp(\gamma_1+\gamma_2 x)/[1+\exp(\gamma_1+\gamma_2 x)]$