

5-1-2007

Using the Fractional Imputation Methodology to Evaluate Variance due to hot Deck Imputation in Survey Data

Adriana Pérez

The University of Texas Health Science Center at Houston

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Pérez, Adriana (2007) "Using the Fractional Imputation Methodology to Evaluate Variance due to hot Deck Imputation in Survey Data," *Journal of Modern Applied Statistical Methods*: Vol. 6 : Iss. 1 , Article 23.

DOI: 10.22237/jmasm/1177993320

Available at: <http://digitalcommons.wayne.edu/jmasm/vol6/iss1/23>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Using the Fractional Imputation Methodology to Evaluate Variance due to Hot Deck Imputation in Survey Data

Adriana Pérez

The University of Texas Health Science Center at Houston

This article examines empirically the effect on the variance estimate due to the use of hot deck imputation with a nearest neighbor donor in comparison with the pairwise fractional hot deck imputation methodology in the 1999 Survey of Doctorate Recipients.

Key words: Ignorability, missing at random, item nonresponse, serpentine sorting, nearest neighbor, successive difference replication

Introduction

Imputation is commonly used to deal with nonresponse and incomplete data in surveys. Usually, the use of imputed values as observed values produces appropriate estimates of smooth statistics (totals, means, proportions, etc) as well as non-smooth statistics (quantiles, etc), if the imputation does not cause severe systematic bias. However, the dangers are well known of not correcting the variance estimates to reflect the uncertainty due to missing data. This may lead to larger underestimation as the proportion of imputed values increases when treating the imputed values as observed. Over the years, a number of methods have been suggested in the statistical literature to overcome these issues (Kalton & Kasprzyk, 1986; Brick & Kalton, 1996; Groves et al., 2004).

Among other reasons, imputation techniques typically are not used with survey data because their users are unfamiliar with techniques of analyzing missing data. Due to operational convenience, most of the commonly

used statistical packages do not incorporate adjustments for missing data into their analysis. For simplicity, often an entire observation with one missing variable response is eliminated.

Following Shao and Steel's (1999) description, two general perspectives exist to obtain variance estimators for large complex sample surveys after imputation: design-based and model-assisted perspective (including multiple imputation). Paraphrasing their definitions: the variance estimate in a design-based perspective accounts for repeated sampling from a fixed finite population and uniform nonresponse within an imputation cell.

Using the model-assisted perspective, the variance estimate is with respect to the sample design and response as well as to the model used for the imputation method (Särndal, Swensson, & Wretman, 1992; Shao et al., 1999). Variance estimators under a multiple imputation perspective (Rubin, 1987), are reasonable using Bayesian inference but are not applicable for design-based or deterministic imputation methods (Shao, 2002). The model-assisted perspective variance estimation methods will not be discussed any further.

Several variance estimation methods exist under the design-based perspective after imputation. Two examples are linearization methods (i.e., Taylor series expansions (Chen & Shao, 1997; Chen & Shao, 2000; Kim, 2001))

Adriana Pérez is an Associate Professor of Biostatistics in the School of Public Health. Email: adriana.perez@uth.tmc.edu.

and replication methods (i.e., Jackknife (Rao & Shao, 1992), bootstrap (Shao & Sitter, 1996) and balanced half samples (Lee, Rancourt, & Särndal, 1995; Rao & Shao, 1996; Shao, Chen, & Chen, 1998; Shao & Chen, 1999; Kim, 2001; Kim & Fuller, 2004). Lee, Rancourt and Särndal (2002) discussed the differences between these approaches. All these methods provide adequate estimates. The choice depends on the users, the need for the estimation of variance components, the computational burden, the adaptability of the sampling fraction and the response mechanism (Lee et al., 2002).

This article is focused on the effect on the variance estimates in the 1999 Survey of Doctorate Recipients (SDR) (National Science Foundation, Directorate for Social, & Division of Science Resources Statistics, 2002). In the next section, the 1999 SDR survey methods will be discussed. Next, a description of the aspects of nearest neighbor hot deck imputation method, fractional imputation, successive difference replication method and the effect of multiple weighting stages will be provided. All these methods are used here to evaluate the variance estimates of this survey. This study extends the proposal of pairwise fraction imputation by Kim and Fuller (1999) on the use of variance estimation with pairwise fractional hot deck imputation and the successive difference replication method.

The 1999 Survey of Doctoral Recipients (SDR)

The 1999 SDR is a National Science Foundation (NSF) survey conducted by the U.S. Census Bureau (National Science Foundation et al., 2002). The population of interest for this survey includes individuals who earned a doctoral degree from a United States (U.S.) institution in Science and Engineering (S&E) fields, are less than 76 years old and planned to stay in the U.S. after their degree (US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements

Branch, 2003a). The SDR provides information about demographic and employment characteristics of the nation's science and engineering doctorate holders. The sampling frame consists of the doctorates records file which contains all research doctorate recipients from U.S. universities since 1920 (National Science Foundation et al., 2002).

The 1999 SDR survey sample size was 40,000. The sample was systematically selected from three groups using the probability proportional to size selection methodology. The three groups were the new cohort (doctoral recipients between July 1996 and June 1998), the nearly new cohort (doctoral recipients between July 1992 and June 1996) and the old cohort (doctoral recipients prior to July 1992) (National Science Foundation et al., 2002).

The sampling strata consisted of 240 strata for the old and nearly new cohorts and were defined by demographic group, degree field and sex. The same 240 strata (six of which were empty) defined the sampling strata for the new cohort (US Bureau of the Census et al., 2003a).

Item non-response was observed in this survey in all variables except seven. All seven were critical variables and had to be filled in order for the response to be considered complete. Hence, two imputation methods were used: logical imputation and hot deck imputation. Logical imputation was used when the answer to a question could be determined by the answer to another question either within the same survey year or from a prior survey round (US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch, 2001a). Logical imputation will not be addressed further in this article.

Hot deck imputation was implemented using a nearest neighbor donor. The auxiliary variables selected to identify the pool of donors were determined by prediction models for each

variable in the survey with item nonresponse. A serpentine sorting on the auxiliary variables was implemented to determine the nearest neighbor donor response (US Bureau of the Census et al., 2001a). This survey allowed for the use of information from any one donor a maximum of four times. The missing mechanism and the tentative reasons for missing values in this survey is likely missing at random (Perez, 2003).

Base weights of the 1999 SDR data were computed by the U.S. Census Bureau (US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch, 2001b). To obtain the final weights, the base weights underwent several adjustments to correct for duplicates, frame ineligible, never earned doctorate case and control totals. Included in these weighting adjustments were a non-interview adjustment and a ratio adjustment via a raking methodology (US Bureau of the Census et al., 2001b). Variance estimates were calculated using successive difference replication methods with 160 replicates (US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch, 2003b; Sukasih & Jang, 2003). Point and variance estimates are currently reported using imputation values as observed values (National Science Foundation et al., 2002).

Methodology

Nearest Neighbor Hot Deck Imputation

Hot deck imputation refers to the process where missing responses or items are replaced by values selected from respondents within the same survey. The respondent selected as a donor is chosen by using observable values from auxiliary variables. The 1999 SDR survey used the hot deck imputation method based on imputation cells (US Bureau of the Census et al., 2001a). This means that in using auxiliary variables known for respondents and nonrespondents, the sample was divided into cells. Sorting was performed within each imputation cell and a neighboring case was selected as a donor for each missing value. Then, the missing value was replaced by the selected value within that cell (Chen et al., 2000).

Fractional imputation

Fractional imputation identifies the method where each missing response or item is replaced by several imputed values drawn from the responding values in an imputation cell (Fay, 1996; Kim et al., 1999). Fractional imputation provides an adjustment method for variance estimation in design-based estimators in the presence of missing values (da Silva & Opsomer, 2002).

Fractional imputation estimators were designed to reduce the imputation variance (Kim et al., 2004) by using more than one donor for a recipient and increasing the weight of the donor for each missing item by a value equal to a fraction of the original weight of the missing observation. Respondents who are not donors retain their original weights. Pairwise fractional hot deck imputation is a special case of fractional imputation where two distinct donors are selected for each missing item. The assumption for this method is that there are at least two donors in each imputation cell (Kim et al., 1999).

The Successive Difference Replication Method

The current approach in calculating the 1999 SDR variance estimates is the successive difference replication method (SDRM). Wolter (1984) developed the basic theory of the successive difference method and later Fay and Train (Fay & Train, 1995) extended this theory with replicates generating the SDRM. The variance estimator is calculated based on the squared differences between neighboring sample cases. The SDRM produces variance estimates with a greater number of degrees of freedom than other replication methods. To create the replicates, the SDRM variance estimator uses an orthogonal Hadamard matrix. Because the 1999 SDR used 160 replicates, a 160x160 Hadamard matrix was formed.

Notation

Paraphrasing, Kim and Fuller's (2004) notation: let P be a finite population containing indices $1, \dots, N$. P is stratified into H strata with N_h units in the h -th stratum. $n_h \geq 2$ units are selected following some probability sampling plan called the sampling mechanism.

Let S denote the sample. According to the sampling plan, survey weights $w_i, i \in S$ are constructed. This expectation is in respect to S . Let Y be a variable of interest and $Y = (y_1, y_2, \dots, y_N)$ denotes the population vector. The response mechanism (I) identifies the probability mechanism of the responses obtained in the sample. $I_i = 1$ if y_i is a respondent and $I_i = 0$ otherwise. Let the population characteristic of interest be $\theta_N = \theta(y_1, \dots, y_N)$ and let $\hat{\theta}$ be a linear estimator of θ_N based on the full sample, $\hat{\theta} = \sum_{i \in S} w_i y_i$.

The SDRM variance estimator for $\hat{\theta}$ can be defined without loss of generality as (ignoring the finite population correction factor) in equation (1):

$$V_{SDRM}(\hat{\theta}) = \frac{4}{k} \sum_{r=1}^k (\hat{\theta}^{(r)} - \hat{\theta})^2 \quad (1)$$

where r is the replicate sample ($r = 1, \dots, k$). k is the total number of replicate samples, $\hat{\theta}^{(r)}$ is the r -th replicate of $\hat{\theta}$ and can be written as: $\hat{\theta}^{(r)} = \sum_{i \in S} w_i^{(r)} y_i$, where $w_i^{(r)}$ denotes the replicate weight for the i -th unit of the r -th replicate.

In the imputation procedure, let a_{ij} be the number of times that y_i is used as a donor for the missing y_j . S_R is the set of indices of the sample respondents and S_M is the set of indices of the sample nonrespondents. Let us define $a = \{a_{ij}; i \in S_R, j \in S_M\}$, then the distribution of a is called the imputation mechanism. In addition, when y_i is used as a donor for element j , let w_{ij}^{\bullet} be the fraction of the original weight for element j . w_{ij}^{\bullet} is called the imputation fraction (Fuller & Kim, 2001; Kim et al., 2004). $w_{ii}^{\bullet} = 1$ for $i \in S_R$ and $w_{ii}^{\bullet} = 0$ for $i \in S_M$. The a_{ij} are nonnegative and the sum of the imputation fractions of the donors for a missing item is mandatory to be one:

$\sum_{i \in S_R} a_{ij} w_{ij}^{\bullet} = 1, \forall j \in S$. In the case of a pairwise fractional hot deck imputation, the imputation fractions, w_{ij}^{\bullet} , are equal to 0.5. A linear estimator using fractional hot deck imputation can be written as in equation (2):

$$\hat{\theta}_I = \sum_{i \in S_R} \left(w_i + \sum_{j \in S_M} a_{ij} w_{ij}^{\bullet} w_j \right) y_i \quad (2)$$

The term in parenthesis equation (2) is called the imputation adjustment weight. Kim and Fuller(1999) demonstrated that the linear estimator $\hat{\theta}_I$ is unbiased and consistent under an ignorable response mechanism. These authors also estimated the variance of this fractional hot deck imputation in terms of the imputation cells.

Variance Estimation After Pairwise Fractional Hot Deck Imputation

Extending the idea of variance estimation after imputation (Kim, 2002; Kim et al., 1999), if the imputed values from the pairwise fractional hot deck imputation are treated as true values and apply the successive difference replication method then the variance estimator can be expressed as in equation (3):

$$V_{SDRM,I}(\hat{\theta}) = \frac{4}{k} \sum_{r=1}^k (\hat{\theta}_I^{(r)} - \hat{\theta}_I)^2 \quad (3)$$

where $\hat{\theta}_I^{(r)}$ is the r -th replicate of $\hat{\theta}_I$ and can be

written as
$$\hat{\theta}_I^{(r)} = \sum_{i \in S_R} \left(w_i^{(r)} + \sum_{j \in S_M} a_{ij} w_{ij}^{\bullet} w_j^{(r)} \right) y_i,$$

where $w_i^{(r)}$ denotes the replicate weight for the i -th unit of the r -th replicate and $w_j^{(r)}$ denotes the replicate weight for the j -th unit of the r -th replicate. Because a_{ij} and w_{ij}^{\bullet} are the imputation mechanism and imputation fraction, respectively, they will take on the same value across all replicates. This is to ensure the correct calculation of the imputation adjustment weight.

Effect of Multiple Weighting Stages On Variance Estimation After Imputation

Frequently, multiple stages of weighting adjustments are implemented in survey (Valliant, 2004). The main aim of the weighting plan is to produce final weights that reduce the nonresponse bias in the survey estimates, balance for noncoverage, and adjust sample estimates to control totals. Each stage introduces a different source of variability in an estimator that may perhaps be important to reflect when estimating variances. The advantage of variance estimation through replication is that it can explicitly account for all the stages in estimation by repeating each adjustment separately for each replicate. This concept will be evaluated in this study.

Methods Implemented On 1999 SDR Data

As mentioned previously, this research focuses on variance estimation after imputation of the 1999 SDR. The pairwise fractional hot deck imputation procedure was evaluated and compared to the variance estimates with the ones obtained when treating the imputed values as observed. Five variables were selected: Race, Hispanic, Gender, Citizenship, and Median Basic Annual Salary of the doctoral scientist and engineers. The Woodruff (1952) method was used for calculating the median and its corresponding standard error was estimating using the program described by Gossett et al (2002). Employment status is a variable without missing data that was used in forming estimates for this study. Separate replicates were computed for each variable of interest as the response mechanism differs for each one. Employment, in combination with the aforementioned variables, was used to calculate 19 survey estimates.

After identifying two donors per missing value for each of the variables selected, the imputation adjustment weight was calculated. However, this imputation adjustment weight can be calculated at three stages of the weighting adjustment process: using the base weights, using the weights after the noninterview adjustment or using the final weights (US Bureau of the Census et al., 2001b). It was decided that all three stages should be explored and the corresponding replicates needed for the

SDRM under all three weighting stages were calculated for evaluation purposes. The three weighting stages being evaluated are discussed in Methods B, C and D below. Method A is the nearest neighbor hot deck imputation used in the 1999 SDR, and did not include an imputation weighting adjustment.

- Method A: The original sampling weights based on the one donor hot deck imputation methodology were used and the imputed values were treated as observed values. The imputation weight adjustment was not used in this method.
- Method B: The base weights were used to obtain the imputation adjustment weights. The imputation adjusted weights were then adjusted to include the non-interview and raking adjustments.
- Method C: The base weights were used to obtain the non-interview adjusted weights. The non-interview adjusted weights were then used to determine the imputation adjustment weights. Finally, the raking adjustments were the final weighting step in the weighting process for this method.
- Method D: After applying the non-interview and raking adjustments to the base weights to create the final weights, the final weights were then used to obtain the imputation adjustment weight.

This empirical evaluation will allow for a determination of the stage of the weighting process at which the imputation weighting adjustment should be performed. In addition, it will allow for an evaluation of the impact of using a single hot deck imputation versus a pairwise fractional hot deck imputation.

After the replicates were computed, the point estimates and their corresponding standard errors were obtained. Statistics combining employment status with variables with missing values used the imputation adjustment weight

for the variable with missing values. As an example, when the employed male estimate was formed, the imputation adjustment weight reflected the adjustments due to the gender variable being imputed.

The standard errors (SE) which do not take the imputation adjustment into account (Method A) were compared with the standard errors which take into account the imputation adjustment (Methods B, C and D). To assess this comparison, the relative difference (RD) was used. For example, when comparing method B versus method A the RD is in equation (4):

$$RD = 100\% * \frac{SE_B(\hat{\theta}_I) - SE_A(\hat{\theta}_I)}{SE_A(\hat{\theta}_I)}$$

The RD measures the magnitude of over or under estimation of the alternative method B compared with the current baseline method A. It is important to highlight that all SE are estimates of standard errors instead of true standard errors and furthermore all are subject to sampling errors.

Results

The imputation rates in the 1999 SDR are relatively low and are provided in table 1. Table 1 presents the point estimates for the 19 estimates selected on the doctoral scientists and engineers for methods A through D. As expected due to the low imputation rates, the point estimates did not vary significantly with either method across all the statistics selected.

Table 2 presents the variance estimates with methods A through D; and includes the relative variances comparing each method B, C and D to method A. The results in table 2 suggest that (i) the variance estimator is lower when the pairwise fractional imputation methods is used and (ii) there is no preference on the weighting stage of the adjustments, except for the median of the basic annual salary where a 17% reduction on its variance is obtained using method D.

Conclusion

The purpose of this article was to perform the pairwise fractional hot deck imputation to evaluate the effect on the variance estimates due to the imputation procedure. The use of this method shows a lower variance in comparison to the single hot deck imputation method which treated the imputed values as observed values. This is achieved in most of the variables of interest. Exceptions are Naturalized U.S. citizen and employed Naturalized U.S. citizen. For these exceptions, the relative difference is slight at most (1.1%) when compared with the hot deck imputation method.

Nevertheless, the effort involved may argue that the need of having an imputation adjustment weight for each variable may not have been necessary in this particular survey with its low imputation rates. Interestingly, this empirical evaluation confirms the disadvantage pointed out by Kim (2002) that its computation can be cumbersome for a large dataset such as the 1999 SDR.

There are limitations to the empirical evaluation. i) The dataset does not have a serious missing data problem which does not allow us to determine clearly which method should be preferred under what conditions. ii) Separate replicates were computed for each variable of interest, assuming an independent univariate missingness pattern. Neither the nearest neighbor hot deck nor the pairwise fractional hot deck imputation methods allows incorporation of multivariate missingness variables to estimate their replicates. iii) The true variance of the SDR data is unknown; for that reason this empiric investigation does not quantify the true relative efficiency.

Further investigation is needed on how to obtain an imputation adjustment weight for the entire survey, as well as how to use/obtain imputation adjustment weights for statistics where more than one variable with missing data are required. Monte Carlo simulations identifying the true variance for a pseudo SDR population as well as incorporating several patterns and missing data mechanisms beyond missing completely at random need to be explored.

Table 1. Doctoral scientist and engineers in 1999: Point estimates using four methods. Method A: hot deck imputation using one donor and treating the imputed values as observed values. Method B: pairwise fractional hot deck imputation using the base weight to obtain the imputation adjustment weight. Method C: pairwise fractional hot deck imputation using the noninterview weight to obtain the imputation adjustment weight. Method D: pairwise fractional hot deck imputation using the final weight to obtain the imputation adjustment weight.

Statistic/Variable	Sample Size	IR(%)*	Point Estimates			
			A	B	C	D
Total						
1 All	31,318		626,698	626,699	626,699	626,698
2 Hispanic	1,623	1.89	15,007	14,787	14,787	15,045
Race						
		0.89				
3 White [!]	22,949		508,447	508,859	508,863	508,417
4 African American	1,567		14,179	14,081	14,082	14,182
5 Asian or Pacific Islander	4,847		87,034	86,823	86,818	87,075
6 American Indian/Alaskan Native	332		2,032	2,009	2,011	2,017
Gender						
		0.01				
7 Male	22,432		476,495	476,511	476,511	476,503
8 Female	8,886		150,204	150,188	150,188	150,196
9 Employed Male	19,835		419,869	419,884	419,884	419,876
10 Employed Female	7,910		133,494	133,480	133,480	133,486
Citizenship						
		0.93				
11 Native Born U.S. Citizen	24,837		491,928	491,940	491,927	491,930
12 Naturalized U.S. Citizen	3,676		70,921	70,843	70,851	70,943
13 Non-U.S. Citizen. Permanent Resident	2,124		48,938	48,984	48,981	48,919
14 Non-U.S. Citizen. Temporary Resident	681		14,911	14,921	14,930	14,907
15 Employed Native Born U.S. Citizen	21,794		429,085	429,459	429,454	429,507
16 Employed Naturalized U.S. Citizen	3,243		62,507	62,460	62,461	62,540
17 Employed Non-U.S. Citizen. Permanent Resident	2,045		47,264	47,321	47,318	47,258
18 Employed Non-U.S. Citizen. Temporary Resident	663		14,507	14,527	14,536	14,514
19 Median Basic Annual Salary of Full Time Employed	25,686	4.27	70,000	68,000	68,000	68,000

Note: *: IR: Imputation rate (percentage); ! 'Other' race included with 'White'

Table 2. Doctoral scientist and engineers in 1999: Standard error estimates and relative differences using four methods. Method A: hot deck imputation using one donor and treating the imputed values as observed values. Method B: pairwise fractional hot deck imputation using the base weight to obtain the imputation adjustment weight. Method C: pairwise fractional hot deck imputation using the noninterview weight to obtain the imputation adjustment weight. Method D: pairwise fractional hot deck imputation using the final weight to obtain the imputation adjustment weight.

Statistic/Variable	Standard Error				Relative Difference		
	A	B	C	D	$\frac{B-A}{A}$	$\frac{C-A}{A}$	$\frac{D-A}{A}$
Total							
1 All	732.2	732.1	732.1	732.2	0.00	0.00	0.00
2 Hispanic	427.0	416.4	416.3	421.3	-0.02	-0.03	-0.01
Race							
3 White [!]	1,001.0	992.9	993.8	994.1	-0.01	-0.01	-0.01
4 African American	360.7	350.5	350.4	352.7	-0.03	-0.03	-0.02
5 Asian or Pacific Islander	819.8	814.7	813.8	819.0	-0.01	-0.01	0.00
6 American Indian/Alaskan Native	161.1	160.1	160.0	159.5	-0.01	-0.01	-0.01
Gender							
7 Male	694.5	693.9	693.9	694.2	0.00	0.00	0.00
8 Female	374.8	374.1	374.1	374.4	0.00	0.00	0.00
9 Employed Male	1,164.1	1,162.0	1,162.0	1,163.0	0.00	0.00	0.00
10 Employed Female	689.0	689.1	689.1	689.0	0.00	0.00	0.00
Citizenship							
11 Native Born U.S. Citizen	686.9	682.8	683.0	686.5	-0.01	-0.01	0.00
12 Naturalized U.S. Citizen	856.3	865.6	864.7	857.3	0.01	0.01	0.00
13 Non-U.S. Citizen. Permanent Resident	787.0	784.9	783.6	783.9	0.00	0.00	0.00
14 Non-U.S. Citizen. Temporary Resident	471.3	471.0	471.0	468.3	0.00	0.00	-0.01
15 Employed Native Born U.S. Citizen	1,253.6	1,239.7	1,239.1	1,247.6	-0.01	-0.01	0.00
16 Employed Naturalized U.S. Citizen	873.4	875.9	875.2	872.2	0.00	0.00	0.00
17 Employed Non-U.S. Citizen. Permanent Resident	797.8	791.6	790.3	791.1	-0.01	-0.01	-0.01
18 Employed Non-U.S. Citizen. Temporary Resident	486.5	486.3	486.5	483.8	0.00	0.00	-0.01
19 Median Basic Annual Salary of Full Time Employed	1,519	1,326	1,324	1,266	-0.13	-0.13	-0.17

Note: ! 'Other' race included with 'White'

References

- Brick, J. M. & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.
- Chen, J. & Shao, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. American Statistical Association, 365-370.
- Chen, J. & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16, 113-131.
- da Silva, D. N. & Opsomer, J. D. (2002). Estimation after fractional imputation under a nonparametric response mechanism. <http://www.stat.iastate.edu/preprint/abstracts/2002-01.pdf>.
- Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- Fay, R. E. & Train, G. F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. American Statistical Association, 154-159.
- Fuller, W. A. & Kim, J. K. (2001). *Hot deck imputation for the response model*.
- Gossett, J. M., Simpson, P., Parker, J. G., & Simon, W. L. (2002). How complex can complex survey analysis be with SAS®? In SUGI 27 Proceedings, SAS users group international 27th annual conference Apr 14-17, Orlando, FL. <http://www2.sas.com/proceedings/sugi27/p266-27.pdf>
- Groves, R. M., Fowle F. J. Jr, Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. (1st ed.) New Jersey: John Wiley & Sons, Inc.
- Kalton, G. & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Kim, J. K. (2001). Variance estimation after imputation. *Survey Methodology*, 27, 75-83.
- Kim, J. K. (2002). Variance imputation for nearest neighbor imputation with application to census long form data. American Statistical Association, 1857-1862.
- Kim, J. K. & Fuller, W. A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.
- Kim, J. K. & Fuller, W. A. (1999). Jackknife variance estimation after hot deck imputation. American Statistical Association, 825-830.
- Lee, H., Rancourt, E., & Särndal, C. E. (1995). Variance estimation in the presence of imputed data for the generalized estimation system. American Statistical Association, 384-389.
- Lee, H., Rancourt, E., & Särndal, C. E. (2002). Variance estimation from survey data under single imputation. In Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.), *Survey Nonresponse*. New York: John Wiley & Sons, 315-328.
- National Science Foundation, Directorate for Social, B. a. E. S., & Division of Science Resources Statistics (2002). *Characteristics of doctoral scientists and engineers in the United States: 1999. Detailed statistical tables*. (NSF 02-328 ed.) Arlington, VA.
- Perez A. Missing Data Analysis on the Science Resources Statistics Surveys. Focus on 1999 NSRCG, NSCG and SDR. 7-3-2003. National Science Foundation. Division of Science Resources Statistics. Arlington, VA, Powerpoint presentation.
- Rao, J. N. K. & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J. N. K. & Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer Verlag.
- Shao, J. (2002). Replication methods for variance estimation in complex surveys with imputed data. In Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (Eds.), *Survey Nonresponse*. New York: John Wiley & Sons.

Shao, J. & Chen, Y. (1999). Approximate balanced half sample and related replication methods for imputed survey data. *Sankhya: The Indian Journal of Statistics. Series B*, 61, 187-201.

Shao, J., Chen, Y., & Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J. & Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Shao, J. & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sukasih, A. S. & Jang, D. (2003). Monte Carlo study on the successive difference replication method for non-linear statistics. *American Statistical Association*, 4141-4147.

US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch (2003a). *Sample design for the 1999 survey of doctorate recipients* Washington, DC: US department of commerce.

US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch (2001a). *Imputation specification for the 1999 survey of doctorate recipients* Washington, DC: US department of commerce.

US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch (2001b). *The 1999 Survey of Doctorate Recipients Weighting Plan* (Rep. No. SDR99-WT-1). Washington, DC: US department of commerce.

US Bureau of the Census, Demographic Statistical Methods Division, & Health Surveys and Supplements Branch (2003b). *Generalized variance parameters for the 2001 survey of doctorate recipients. Direct calculation of variance estimates, generalized variance functions, generalized variance parameters, standard errors and their use (SDR01-VAR-3)* Washington, DC: United States department of commerce; Economic and Statistics Administration; US Census Bureau.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Wolter, K. M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79, 781-790.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.