# When Are Inferences from Multiple Imputation Valid?

Robert E. Fay[1]
U.S. Bureau of the Census, Washington, DC 20233-4001

**KEY WORDS:** *jackknife, missing data, nonresponse, survey inference*

**Abstract**: Multiple imputation, as described by Rubin, has seen a wide variety of applications. Counterexamples, presented by Fay (1991), and new methods, such as those of J.N.K. Rao and J. Shao, that can asymptotically disagree with the multiple imputation approach, have raised questions about the validity of multiple imputation. This paper identifies critical restrictions on the practical application of multiple imputation. It also discusses alternatives that can provide asymptotically valid inferences for some of the situations in which multiple imputation fails.

## 1. INTRODUCTION

Rubin (1978, 1987) has proposed multiple imputation as a general technique to represent the increased uncertainty in analysis of survey data from treating imputations for missing data as if they were known. Although relying initially on a Bayesian motivation, he and others have argued that multiple imputation can be used to provide valid inferences from both Bayesian and frequentist perspectives in a variety of contexts.

For example, Clogg *et. al.* (1991) employed multiple imputation to create two augmented Public Use Samples from the 1970 census, of approximately 800,000 cases each. The augmented samples each include five imputations of 1980-equivalent industry and occupation codes from the existing 1970 census codes. The Census Bureau doubly coded a subsample of approximately 125,000 1970 census cases, coded originally by the 1970 coding scheme, by recoding them according to the 1980 scheme. A complex system of hundreds of logistic regression models to predict the 1980 codes on the basis of 1970 codes and other characteristics, derived from fitting the doubly coded sample, formed the basis for the multiply imputed codes. The authors have made the resulting data sets available to researchers with encouragement to use multiple imputation techniques for inference from these data. In other words, analysts are to apply complete data methods to each of the five sets of imputations, and to estimate total variance as the sum of 1) the average variance estimated for the five imputed sets (where, in effect, n=800,000), treating imputed values as if they were known, and 2) variability in the estimates between the five imputed sets (to represent the variance from estimating the 800,000 cases from the subset of n=125,000 doubly coded cases).

There have already been several articles and studies offered in support of the validity of multiple imputation inferences. Consequently, the title of this paper appears to raise a question that has already been answered. In fact, this paper will show how previous work on multiple imputation is not adequate to assure the validity of complex applications, such as that by Clogg *et. al.*, without highly restrictive conditions on the form of the subsequent analysis of the multiply imputed data set.

The focus on this paper will be on limitations of multiple imputations in the context of simple random samples, although these deficiencies are related to other limitations that are evident in the attempt to apply multiple imputation to complex samples. The theory developed for multiple imputation provides relatively little detailed discussion of the interaction between missing data uncertainty and the design-based variance methods typically applied in the analysis of complex surveys. A paper of Fay (1991), initially intended to augment the existing theory of multiple imputation on this point, instead outlined an alternative basis of inference from complex survey data with missing values. In some instances, the new methods give results asymptotically equivalent to those of multiple imputation (assuming an increasing or infinite number of multiple imputations). The paper also presented several counterexamples where this was not the case, however. Some counterexamples involved specific complex sampling situations, while others required only simple random sampling. On the other hand, Fay (1991) proposed specific alternative methods only for a restricted class of missing data procedures, such as mean imputation, and omitted consideration of important other methods, such as the hot deck. Thus, this initial work attempted to cover several issues: the establishment of a new general inferential basis for sample surveys involving missing data; discussion of appropriate methods for mean imputation, including imputing

probabilities for categorical data; and identification of counterexamples where multiple imputation yields inconsistent inferences, both involving complex samples and simple random samples.

Separate work of Rao and Shao (1992), independently of Fay, developed methods to assess the uncertainty of estimates based on a single set of imputations from a hot deck with specific restrictions on the method of selecting the imputed values. The initial motivation of their work was to provide a valid inferential basis for the single-imputation hot deck. They recognized that large survey organizations such as Statistics Canada were likely to continue to prefer singly imputed data sets instead of the more complex multiply imputed versions, so their work focused on providing adequate measures of uncertainty reflecting the effect of missing data without resorting to multiple imputation. Thus, their paper addressed a critical interest not handled by Fay (1991). Rao and Shao modified the stratified jackknife; their findings are pertinent both to simple random samples and to multistage stratified samples. Again, this research was initially motivated by complex sampling problems, but the authors' results have important implications for simple random samples as well.

Although valid inferences for complex samples will be an important topic for additional research, the focus of this paper will be to identify some of the limitations of multiple imputation in the simplest of contexts. This paper investigates missing data uncertainty under conditions studied earlier by Rubin and Schenker (1986), that is, a simple random sample of size $n$ from an infinite population, even though some of the literature, for example, Rubin (1987), also treats the instance of a sample of size $n$ from a finite population of size $N$. Further, the comparisons will be simplified in two additional ways:

i) Section 3 considers only the properties of multiple imputation as if an essentially unlimited number, $m$, of imputations were available, i.e., $m = \infty$. The examples of Section 4 employ a generous $m = 10$ imputations. This choice avoids complexities that occur for small $m$ from limited degrees of freedom to estimate part of the variance; and,

ii) Section 3 considers properties for large $n$, retaining terms only through $O(1/n)$ in expressions for variance. The use of $n = 100$ observations in Section 4 corresponds closely to the behavior predicted by this order of approximation.

Section 2 reviews the theoretical foundations for multiple imputation, both for its properties under special conditions and arguments advanced for its suitability for more general problems. To help fix ideas, Section 3 revisits the last example offered in Fay (1991). In this relatively simple example, it is possible to approximate analytically the multiple imputation inferences as well as to obtain the asymptotic variances and covariances of the same estimators through standard frequentist approximations. The multiple imputation approach implies inferences with a valid frequentist interpretation for the overall sample, but is incorrect for simple subdomain estimates. Examples in section 4 illustrate the differences between multiple imputation and the new approaches.

## 2. Foundations for Multiple Imputation

As noted earlier, there is a considerable literature on multiple imputation. Rubin's book (1987) represents a systematic statement of the underlying theory, and this section will primarily use this source as a point of reference.

Rubin's (1987) first chapter announces the general purpose nature of multiple imputation through four examples, including the coding of industry and occupation later appearing as Clogg *et. al.* (1991). After the preliminaries in the second chapter, the third chapter systematically demonstrates that, under a given Bayesian analysis of the data and missing data mechanism, multiple imputations represent draws from the posterior distribution of the missing data, given the observed. Rubin shows that, for posterior inferences that can be well approximated by a normal distribution, the analysis under the multiple imputation approach yields valid inferences asymptotically equivalent to the full Bayesian analysis.

Chapter 4 assesses the validity of multiple imputation from a frequentist perspective, by a mixture of argument and simulation. The examples of the chapter are based, however, on circumstances in which the Bayesian and frequentist analysis of the complete data would be in virtual agreement in the absence of missing data.

Consequently, one might distinguish four issues in the validity of multiple imputation based on the underlying similarity of the assumptions for the complete data set if no missing data were present:

1) applications to Bayesian inference, when both the imputer and analyst work with the

same Bayesian distribution,

2) applications to frequentist inference when the Bayesian analysis of the complete data case essentially agrees with frequentist inference,
3) applications in which the imputer and analyst have different Bayesian priors, and
4) applications in which the frequentist inference for the complete data is fundamentally different than the Bayesian model employed by the imputer.

Essentially, Rubin recognizes and discusses these issues, although without emphasizing the distinction between 2) and 4).

Much of the literature focuses on the validity of multiple imputation under 1) and 2). Herzog and Rubin (1983) and Rubin and Schenker (1986) assume simple random sampling without covariates, ignorable nonresponse, and a scalar outcome variable. Schenker and Walsh (1988) extended the results by including the effect of covariates in a linear model. Results were stated within the context of the overall model, so that investigations of this sort were not attuned to surfacing the consequences under situations 3) or 4).

### 3. A Simple Counterexample: Proportions

The concluding example in Fay (1991) presumed only simple random sampling, large $n$, and, effectively, $m = \infty$ multiple imputations. The example considers the simple case of estimating a binomial proportion. Suppose $n_1$ out of the $n$ sample cases have reported values, with missing data for the remaining cases, and that the proportion of responses, $r = n_1/n$, remains fixed as $n \to \infty$.

Suppose that an imputer generates multiple imputations by assuming that the underlying proportion $\theta$ is distributed approximately $N(p, pq/n_1)$, where $p$ is the observed proportion for reported cases. For each set, $\ell$, of imputations, the imputer draws a $\theta_\ell$ from this distribution and completes the data set by drawing independent Bernoulli variables with expectation $\theta_\ell$ for each missing case. For each set, let $\hat{\theta}_{*\ell}$ denote the estimated proportion based on the observed and imputed values.

For a finite but large number, $m$, of imputations, multiple imputation provides inferences about the underlying true $\theta$ through $\theta \sim N(\hat{\theta}_*, \hat{T})$ approximately, where $\hat{T}$ denotes the estimated total variance comprised of variance in the completed data set plus variance due to imputation of the missing values:

$$\hat{T} = \hat{W} + (1 + m^{-1})\hat{B} \qquad (2.1)$$

where

$$\hat{\theta}_* = m^{-1} \sum_{\ell=1}^{m} \hat{\theta}_{*\ell}$$

$$\hat{W} = m^{-1} \sum_{\ell=1}^{m} \hat{W}_{*\ell}$$

$$= m^{-1} \sum_{\ell=1}^{m} \frac{\hat{\theta}_{*\ell}(1 - \hat{\theta}_{*\ell})}{n}$$

and

$$\hat{B} = (m - 1)^{-1} \sum_{\ell=1}^{m} (\hat{\theta}_{*\ell} - \hat{\theta}_*)^2$$

As $n, m \to \infty$, Schenker (1988) and Schafer and Schenker (1991) note that $\hat{B}$ will be composed of contributions of approximately $(1 - r)^2\, \hat{\theta}_*(1 - \hat{\theta}_*)/nr$ due to sampling $\theta_\ell$ and approximately $(1 - r)\, \hat{\theta}_*(1 - \hat{\theta}_*)/n$ due to the independent draws for each imputation, while $\hat{W}$ approaches the standard binomial variance, $\hat{\theta}_*(1 - \hat{\theta}_*)/n$. Asymptotically, $\hat{T}$ approaches $\theta(1 - \theta)/n_1$, the binomial variance based on the number of complete cases. Thus, multiple imputation provides the same answer as the frequentist analysis of the problem.

Suppose that an analyst attempts to employ the multiply imputed data set to make inferences about two subdomains. Suppose further that the two subdomains partition the original sample into $n = n_a + n_b$, $n_1 = n_{1a} + n_{1b}$, $nr = n_a r_a + n_b r_b$, etc. Response rates $r_a$ and $r_b$, the underlying population proportion $\theta$, and the relative proportions of $E(n_a)$ and $E(n_b)$ remain fixed as $n \to \infty$.

The analyst forms separate estimates, $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$, for the two subdomains computed using only data from each respective subdomain. For example, $\hat{\theta}_{*a}$ would be computed only from the observed and imputed values in subdomain $a$. Hence, the estimation of $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$ does not exploit a specific assumption, namely that $\theta$ was constant, built into the missing data model; yet, this example, in a simplified form, exemplifies the manner in which survey estimates for subdomains are produced.

If the analyst uses the means of the imputed values as if they were known and computes the naive

estimate of the covariance of $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$, the result, to $O(1/n)$, would be:

$$\mathbf{C_N} = \theta(1-\theta)\begin{pmatrix} \frac{r_a}{n_a} & 0 \\ 0 & \frac{r_b}{n_b} \end{pmatrix}$$

In the absence of complete response, this estimator will systematically understate the true uncertainty.

Suppose that the analyst instead employs the multiple imputation replicates to derive an estimated variance-covariance matrix of $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$. The sum, (2.1), of the within and between components is:

$$\mathbf{C_{MI}} = \theta(1-\theta) \times \begin{pmatrix} \frac{1}{n_a} + \frac{(1-r_a)}{n_a} + \frac{(1-r_a)^2}{nr} & \frac{(1-r_a)(1-r_b)}{nr} \\ \frac{(1-r_a)(1-r_b)}{nr} & \frac{1}{n_b} + \frac{(1-r_b)}{n_b} + \frac{(1-r_b)^2}{nr} \end{pmatrix}$$

A covariance, proportional to $(1-r_a)(1-r_b)$, now appears, which was absent from the naive estimate $C_N$. This covariance derives from $\hat{B}$ of (2.1). The covariance is due to sampling $\theta_\ell$ and using it to impute values for both subgroups concurrently within a single imputation. Thus, multiple imputation recognizes the covariance between $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$ from their shared use of the observed data to estimate the missing data. In fact, however, design-based reasoning identifies three reasons for covariance between $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$:

i) There is a covariance between the observed values for group *a* and the imputed values for group *b*, because the observed values in group *a* partially determine the imputations for group *b*.

ii) There is a covariance between the imputed values for groups *a* and *b*, because they were imputed from the same model and estimated parameters.

iii) Symmetrically with i), there is a covariance between the imputed values of *a* and the observed values of *b*.

In fact, multiple imputation is only properly accounting for the contribution of ii).

When a design-based calculation of the total variance-covariance matrix is performed, either through linearization or replication, the result, to $O(1/n)$, is instead:

$$\mathbf{C_{DB}} = \theta(1-\theta)\begin{pmatrix} \frac{r_a}{n_a} + \frac{(1-r_a^2)}{nr} & \frac{(1-r_a r_b)}{nr} \\ \frac{(1-r_a r_b)}{nr} & \frac{r_b}{n_b} + \frac{(1-r_b^2)}{nr} \end{pmatrix}$$

Differences between the estimated covariances are larger than they might at first seem. For example, if $r_a$ and $r_b$ are both .9, the design-based approach gives 19 times the covariance of multiple imputation. Although less obvious from a quick comparison, the variances of $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$ are each correspondingly less under the design-based approach. Both $\mathbf{C_{MI}}$ and $\mathbf{C_{DB}}$ give the right answer for the variance of $(n_a p_a + n_b p_b)/n$, the overall proportion, however. Hence, the design-based approach identifies the effect of typical missing data treatments is to produce higher covariances among subdomain estimates and lower increases in variance, relative to multiple imputation.

This example can be extended to more than two subgroups with a similar effect, that is, multiple imputation understates the true covariances between subgroups and overstates their individual variances, while obtaining the same answer, asymptotically, as frequentist arguments for the reliability of the estimated overall proportion.

Note that the analyst, in estimating proportions for the two populations separately, does not make the assumption reflected in the original imputation, which was that the binomial proportion, $\theta$, did not depend on subgroup membership. If the analyst had remained entirely consistent with this assumption, then only $\hat{\theta}_*$ would have been produced as the estimate for the subdomain estimates for *a* and *b*, not $\hat{\theta}_{*a}$ and $\hat{\theta}_{*b}$. Since multiple imputation inferences are valid for $\hat{\theta}_*$, it is possible to say that subdomain analysis presents no special problems in this example as long as subsequent analysis employs the same assumptions as the missing data model. The next section provides a less trivial illustration of this phenomenon. On the other hand, such a limitation, if general, imposes severe restrictions on the validity of multiple imputation inferences for complex applications, such as Clogg *et. al.* (1991).

One of the several lessons here is that the validity of multiple imputation can depend on the form of subsequent analysis. In particular, assessing the performance of multiple imputation for the overall proportion or mean or similar global aggregates does not assure its validity for other uses.

This example also illustrates that the decomposition of variance into within and between components is generally unable to reproduce $\mathbf{C}_{DB}$ except when $r_a = r_b = 1$. Then:

$$\mathbf{C}_{DB} - \hat{W} = \theta(1-\theta) \times \begin{pmatrix} \frac{-(1-r_a)}{n_a} + \frac{(1-r_a^2)}{nr} & \frac{(1-r_a r_b)}{nr} \\ \frac{(1-r_a r_b)}{nr} & \frac{-(1-r_b)}{n_b} + \frac{(1-r_b^2)}{nr} \end{pmatrix} - \theta(1-\theta) \begin{pmatrix} \frac{-(1-r_a)}{n_a} & 0 \\ 0 & \frac{-(1-r_b)}{n_b} \end{pmatrix} + \theta(1-\theta) \begin{pmatrix} \frac{(1-r_a^2)}{nr} & \frac{(1-r_a r_b)}{nr} \\ \frac{(1-r_a r_b)}{nr} & \frac{(1-r_b^2)}{nr} \end{pmatrix}$$

The second matrix in the last equation has a nonpositive determinant; it is consequently generally impossible to decompose $\mathbf{C}_{DB}$ along the lines of (2.1). Consequently, the underlying strategy of multiple imputation, namely to first compute variances for the completed data set as if the imputed data were observed and then to add an additional component due to uncertainty in the imputation, is fundamentally flawed for applications of this sort.

## 4. Imputation for Normal Data

The final counterexamples in this section compare multiple imputation with $m = 10$ imputations to the design-based approach for mean imputation and the method of Rao and Shao (1992) for single imputation. Thus, the estimators themselves are different, with mean imputation having the smallest variance, multiple imputation with $m = 10$ having a slightly higher variance, and the estimator based on a single imputation the highest variance. The relevant question is whether the variance estimation strategy in each case, that is, (2.1) for multiple imputation, the jackknife for mean imputation, and a modified jackknife for the single imputation, produce a variance estimate acceptably close to the actual properties its corresponding estimator.

Table 1 presents an example based on 5,000 repetitions, sampling 100 observations from *N(0,1)*. The population is divided into two groups, *a*, and *b*, of equal size, and the probability of response is .7 for each individual, with independent response among individuals.

**Table 1** Comparisons of averaged variance estimates under three procedures, with observed actual variance, 5000 repetitions

| | Mult Imp | DB | Rao+Shao |
|---|---|---|---|
| Var $Y$ | .0146 | .0144 | .0174 |
| | .0146 | .0141 | .0167 |
| Var $Y_a$ | .0280 | .0217 | .0278 |
| | .0224 | .0215 | .0271 |
| Cov $Y_a, Y_b$ | .0014 | .0074 | .0074 |
| | .0071 | .0069 | .0066 |

Note: Multiple imputation estimates are based on $m = 10$ draws, and the estimator is the average over the ten draws; the comparison is between the multiple imputation variance estimate, (2.1), and the observed variance of the estimator. Similarly, the design-based approach employs mean imputation and the jackknife variance estimation for mean imputation; the comparison is between the average jackknife variance estimate and the observed variance using mean imputation. The approach of Rao and Shao produces a variance estimate for the single imputation hot deck based on a modification of the jackknife; the comparison is between the mean of their variance estimator and the actual variation.

The results in Table 1 are analogous to those in the previous section. All three methods give acceptable inferences for the estimation of the overall total, but multiple imputation yields unacceptable results for subdomains, again exaggerating the variance increase while understating the covariance.

The first two examples have the disadvantage of appearing too simple. Imputations frequently are based on observed covariates. In the last example, a second variable forms a dichotomy of the population into classes *s* and *t*. For simplicity, these classes were also assumed of equal size in the population and to be independently distributed across *a* and *b*. Table 2 shows the results of an imputation employing *s* and *t* as imputation classes to impute missing *Y*. As before, membership in *a* and *b* is not considered in the imputation. For example, *s* and *t* could represent two broad occupation groups, considered highly predictive of earnings, *Y*, and *a* and *b* could denote a characteristic not usually employed in the imputation, such as state. Table 2 presents the results.

**Table 2** Comparisons of averaged variance estimates under three procedures, with observed actual variance, 20,000 repetitions

| | Mult Imp | DB | Rao+Shao |
|---|---|---|---|
| Var $Y$ | .0145 | .0145 | .0174 |
| | .0147 | .0142 | .0172 |
| Var $Y_a$ | .0278 | .0219 | .0278 |
| | .0222 | .0214 | .0272 |
| Cov $Y_a$, $Y_b$ | .0014 | .0074 | .0074 |
| | .0075 | .0074 | .0073 |
| Var $Y_s$ | .0291 | .0296 | .0356 |
| | .0298 | .0289 | .0348 |
| Cov $Y_s$, $Y_t$ | .0000 | .0000 | .0000 |
| | .0000 | .0000 | -.0001 |

Multiple imputation produces acceptable results for the overall mean and for means of the imputation classes, but not for subdomains *a* and *b* omitted from the imputation model. As before, multiple imputation may give valid results under restricted conditions that the imputer's model and analyst's estimator sufficiently agree, but the example also indicates that multiple imputation is inappropriate as a general purpose methodology for complex problems or large public use files.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES


Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayes Logistic Regression," *Journal of the American Statistical Association*, 86, 68-78.

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance," *Proceedings of the 1991 Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 429-440.

Herzog, T. N. and Rubin, D. B. (1983), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys," in *Incomplete Data in Sample Surveys*, (W. G. Madow, I. Olkin, and D. B. Rubin, eds.), Academic Press, New York, 185-207.

Rao, J.N.K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, in press.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section*, Washington, DC: American Statistical Association, pp. 20-34.

_____ (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputations for Interval Estimation From Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366-374.

Schafer, J. L., and Schenker, N. (1991), "Variance Estimation with Imputed Means," *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 696-701.

Schenker, N. (1989), "The Use of Imputed Probabilities for Missing Binary Data," *Proceedings of the Fifth Annual Research Conference*, Washington DC: U.S. Bureau of the Census, pp. 133-139.

Schenker, N. and Walsh, A. H. (1988), "Asymptotic Results for Multiple Imputation," *The Annals of Statistics*, 16, 1550-1566.