VALID INFERENCES FROM IMPUTED SURVEY DATA

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

KEY WORDS: *Missing data, replication methods, sample surveys, multiple imputation*

Abstract Rubin has offered multiple imputation as a general approach to inference from survey data sets with missing values filled in through imputation. In spite of the considerable scope of work on the subject, the literature on multiple imputation has failed to produce a set of clear and sufficient conditions for the validity of multiple imputation that would justify many of its previous applications. In fact, significant counterexamples to multiple imputation inference are easily produced.

This paper extends previous work of Rao and Shao and of Fay to obtain valid inferences from imputed data sets. The paper includes extensions of Rao and Shao's Biometrika results to both multiple imputation and to *fractionally weighted imputation* under appropriate conditions.

1. INTRODUCTION

One of the significant contributions of multiple imputation (Rubin 1978, 1987) is its emphasis on the quantification of the effect of missing values on inference. The methodology seems of special importance to survey research, where imputation has frequently been used to fill in missing values for item nonresponse. A number of researchers in this and other fields previously recognized that traditional methods of inference from survey data, such as design-based inference, are compromised when imputed values are treated as observed. Until quantitative measures of effect of missing data became available, however, imputed values were almost always treated as observed in variance estimation. Indeed, even with the availability of multiple imputation, inclusion of the effect of missing data in variance estimation continues as the exception rather than the rule in most practice.

Chapter 1 of Rubin (1987) promises much. Although the development first describes multiple imputation from a Bayesian perspective, the combination of simple examples where multiple imputation gives the "right" answer from a frequentist perspective and envisioned applications to four clearly complex problems (Examples 1.1-1.4 on pp. 4-7) leads to the impression that an almost universal tool has been invented. This impression has been enforced by an extensive literature of theoretical work and applications to a number of complex problems.

Consequently, counterexamples to multiple imputation (Fay 1991, 1992), usually in the form of slightly altered versions of the simple examples for which multiple imputation is correct, provided a new perspective on possible limitations of multiple imputation inference. Among other consequences, the counterexamples surfaced a current shortcoming in both the theory and practice of multiple imputation, namely, an absence of a clear characterization of when the technique produces inferences with asymptotically valid frequentist interpretations.

For example, Fay (1992) described an instance where an imputation employed two imputation classes, but the usual direct sample estimates, including the imputed values, were made for both the imputation classes and for two classes cutting across the imputation classes. Multiple imputation inference was shown to be consistent for the first set of inferences but inconsistent for the second. In the second case, multiple imputation overstates the variance of the direct estimates for each class and understates their covariance.

In general, the counterexamples share the feature that the analyst, using direct sample estimates for a domain, conducts an analysis inconsistent with the imputer's original assumptions. These counterexamples each grant the imputer the correct model. The analyst ignores the model's information about the population, and computes instead the direct estimates commonplace in survey practice. The counterexamples also assume that the inference is from a sample to a population; in general, none of the counterexamples show the multiple imputation variance to be wrong for inferences about the effect of missing data if the entire population is observed, except for missing data.

A second flaw in both the theoretical development and practice of multiple imputation has been the absence of a consistent approach to accommodate issues of stratification, clustering, and weighting to compensate for varying probabilities of selection. The apparent hope has appeared to have been that the imputation model would correctly compensate for all of these design factors. Experienced survey practitioners generally recognize, however, that complex sample designs frequently have marked effects on estimation of model parameters when dealing with complex survey data with complete response. There is no convincing reason that models specified for missing data problems will always be immune from such problems.

Previously, Fay (1992) considered the following three estimators and estimators of their variance for a problem with two imputation classes and two additional classes cutting across the imputation classes:

1) Mean imputation, using a jackknifed variance estimator including the effect of estimating the imputation class means, along the general lines of Fay (1991) and Rao and Shao (1992);
2) Multiple imputation, with m=10 imputations, according to Rubin (1987);
3) The single-imputation hot deck, following Rao and Shao (1992), including their variance estimator.

The three estimators are ranked in increasing order of variance.

In the Monte Carlo results, the variance estimates in both cases 1) and 3) appeared virtually unbiased. The multiple imputation variances were equally successful for the overall mean and means of the imputation classes, but unacceptable for the means of the cross-classes.

Because this comparison involved three separate estimators and their variance estimates, multiple imputation could remain the method of choice in applications in which properties of the multiple imputation estimator were preferred, in spite of its inferential deficiencies. Mean imputation, 1), is not a general purpose solution if estimates of distributions as well as totals are of interest. For example, use of mean imputation for earnings would lead to spikes or artificial concentration of the imputed values near the middle of the earnings distribution, distorting the estimated earnings distribution. The single imputation hot deck, 3), while avoiding extreme distortion in the estimated distribution, produces estimates of mean earnings with generally higher variance than the multiple imputation alternative.

This paper revisits this situation and demonstrates that the results of Rao and Shao (1992) have far wider implications than apparent at first glance. Section 2 reviews elements of multiple imputation and the Rao and Shao (1992) results as initially presented. Section 3 introduces *fractionally weighted imputation* as an alternative to multiple imputation. Like multiple imputation, fractionally weighted imputation uses more than one imputation for missing values in order to increase the precision of estimates. As Section 3 will describe further, fractionally weighted imputations are generally *improper* in the sense of Rubin (1987), unlike a well-constructed multiple imputation, which should be *proper*. Also unlike multiple imputation, inference for fractionally weighted imputation on does not employ variability between the different imputations in variance estimation; instead, the Rao-Shao variance formulas extend, without any required modification, to fractionally weighted imputation. This separation of the purposes of variance reduction and variance estimation has highly beneficial effects: for the same number of imputations, fractionally weighted imputation generally produces estimates with smaller variance than multiple imputation. Fractionally weighted imputation also permits construction of confidence intervals based on normal approximations rather than the more complex procedures required with multiple imputation.

The Rao-Shao results are stated for a specific weighted hot-deck, where the probability of selection is proportional to the weight of the "donor." Since unweighted hot decks have appeared more frequently in practice, Section 3 also discusses modifications of the Rao-Shao estimator for the unweighted hot deck, along with consideration of circumstances that would justify its use. These results pertain both to a single imputation hot deck and to fractionally weighted imputation.

In the spirit of this extension of the Rao-Shao results, Section 3 also exhibits a variance estimator for a proper multiply imputed data set. The range of application of this variance estimator is more restricted than the scope of the Rao-Shao variance estimator for fractionally weighted imputation, however.

Section 4 revisits the example of Fay (1992), but with an expanded analysis. Now 4 estimators, instead of the previous 3, are available for consideration, with 2 competing variance estimators for the multiple imputation estimator, including the variance estimator from Section 3. The section summarizes a more extensive set of available Monte Carlo results than presented earlier.

Section 5 turns to a discussion of calculation for the Rao-Shao variance estimator. A possible reading of Rao and Shao (1992) might leave the impression that the required calculations would be onerous and potentially limiting. Section 5 discusses how the calculations can be organized to be carried out effectively, without requiring the creation and permanent retention of supplemental complex data sets. In fact, the calculations can already be implemented through the VPLX program (Fay 1990, 1993a) with relative ease, and modest enhancements to VPLX could further simplify the calculation and thus facilitate general use of the Rao-Shao estimator.

For somewhat more than a decade, a number of its advocates have been inclined to view multiple imputation as the only practical way to represent uncertainty from missing data for complex applications. Recent developments, including the implications of the work of Rao and Shao (1992), have challenged that preeminence. This paper does not systematically review other approaches to missing data, including work of Särndal, Tollifson, and others, which offer even more alternative paths. These comparisons are clearly of future interest but beyond the scope of this effort.

Section 6 concludes with a number of conjectures about further extensions of the Rao and Shao approach. A companion document (Fay 1993b) reports results in more detail than their discussion in Sections 4 and 5 and is available from the author.

In keeping with the intent of the *Proceedings,* this version of the paper departs little from the original presentation, but some clarifications and a few extensions have been added. The most significant of these is to introduce the term *fractionally weighted imputation* to replace *repeated imputation* in the original version, to avoid any confusion with previous definitions of the latter term in Rubin (1987) or elsewhere. Section 6 notes other departures from the original version.

2. SOME ALTERNATIVE APPROACHES TO INFERENCE FROM MISSING DATA

2.1 Multiple Imputation. Multiple imputation (MI) (Rubin 1978, 1987) has been widely discussed elsewhere. For purposes of comparison to other approaches, MI inference for a "hot deck" situation with a single imputation class will be summarized here.

Suppose a simple random sample, y_j, $j=1,\ldots,n$, of size n is drawn from an infinite (or extremely large) population. Suppose further that the values of y_j are observed only for a subset of respondents, $j \in A_r$. Multiple imputation, for $m \geq 2$, provides m imputed values $y^*_{(MI)j\ell}$, $\ell=1,\ldots,m$, for each nonrespondent $j \in A_{nr}$. Assuming that the data are missing at random (e.g., Rubin 1978), inferences about the population mean may be based on first computing m separate estimates $\bar{y}_{(MI)\ell}$, based on the observed values $y^*_{(MI)j\ell} = y_j$, $j \in A_r$, and imputed values $y^*_{(MI)j\ell}$, $j \in A_{nr}$. The MI estimate of the population mean, θ, is

$$\bar{y}_{(MI).} = \sum_{\ell=1}^{m} \bar{y}_{(MI)\ell}/m \qquad (2.1)$$

Multiple imputation provides inferences about the underlying true θ through the approximation

$$\frac{(\theta - \bar{y}_{(MI).})}{\hat{T}^{1/2}} \sim t_\nu \qquad (2.2)$$

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:

$$\hat{T} = \hat{W} + (1 + m^{-1})\hat{B} \qquad (2.3)$$

where

$$\begin{aligned} \hat{W} &= m^{-1} \sum_{\ell=1}^{m} \hat{W}_{*\ell} \\ \hat{W}_{*\ell} &= \frac{1}{(n(n-1))} \sum_{j=1}^{n} (y^*_{(MI)j\ell} - \bar{y}_{(MI)\ell})^2 \end{aligned} \qquad (2.4)$$

estimates the variance of the estimator under complete response, and

$$\hat{B} = (m-1)^{-1} \sum_{\ell=1}^{m} (\bar{y}_{(MI)\ell} - \bar{y}_{(MI).})^2 \qquad (2.5)$$

estimates the between imputation variability. The degrees of freedom, ν, for the t-distribution, t_ν, in (2.2) is estimated by

$$\nu = (m-1)\left(1 + \left(\frac{m}{(m+1)}\right)\frac{\hat{W}}{\hat{B}}\right)^2. \qquad (2.6)$$

Naturally, there are restrictions on the manner in which the multiple imputations, $y^*_{(MI)j\ell}$, $\ell=1,\ldots,m$, $j \in A_{nr}$, are derived. The most obvious choice, repeatedly making independent draws from y_j, $j \in A_r$, is not "proper," in the sense of Rubin (1987), since these draws do not represent the full uncertainty in estimating the data, for purposes of MI variance estimation. Depending on the extent of assumptions about the population distribution of the y's, several choices are available. A useful general approach, called the approximate Bayesian bootstrap (Rubin and Schenker 1986, Rubin 1987, p. 124), for each imputation ℓ:

1) draws a hot deck by drawing r times, with replacement, from the r elements of A_r, and then

2) draws from the hot deck of 1) the n-r values, $y^{*}_{(MI)jl}$, $j \in A_{nr}$, again by sampling with replacement.

(Note for clarification: Rao and Shao (1992, p. 812) slightly misdescribe this procedure, stating that n rather than r values should be drawn at step 1).) Because of its simplicity, the approximate Bayesian bootstrap was used in the results for MI reported in Section 4.

2.2 Rao and Shao: Jackknife Variance Estimation. Rao and Shao (1992) proposed a modification to the standard design-based stratified jackknife variance formula to provide suitable estimates of missing data uncertainty for a data set with single imputation. The initial motivation (Rao and Shao 1992, p. 812) for studying single imputation was to provide an alternative for large statistical agencies who prefer this approach to multiple imputation for its greater simplicity of data storage and processing.

Under their proposal, the "hot deck" must conform to specified conditions. For example, in the simplest case, a single imputation class and a simple random sampling design, imputations are made through simple random sampling with replacement from the donors. For multi-stage stratified sampling, which may lead to differential probabilities of selection and associated weights, the authors consider the selection of "donors" in the imputation with probabilities proportional to their respective survey weights within the imputation class. Estimates are produced from the singly imputed data set in the normal manner, that is, by using the imputed values as if they were observed for purposes of estimation. The analysis is modified at the point of variance estimation to reflect the uncertainty due to missing data.

For simplicity, the case of simple random sampling with a single imputation class will be described in this section, although appendix A.1 employs their general results for multiple imputation classes under stratified multistage sampling, in extending their results to fractionally weighted imputation.

The estimator of the mean may be written:

$$\bar{y}_{(HD)} = \left(\frac{r}{n}\right)\bar{y}_r + \left(1 - \left(\frac{r}{n}\right)\right)\bar{y}^{*}_{nr} \qquad (2.7)$$

where $\bar{y}_r$ is the respondent mean and $\bar{y}^{*}_{nr}$ is the mean of the imputed values.

The standard jackknife variance formula is:

$$v_{J(1)} = \frac{n-1}{n} \sum_{j=1}^{n} (\bar{y}_{(HD)}(-j) - \bar{y}_{(HD)})^2 \qquad (2.8)$$

where

$$\begin{aligned} \bar{y}_{(HD)}(-j) &= \frac{1}{(n-1)}(n\bar{y}_{(HD)} - y_j) \quad \text{if } j \in A_r \\ &= \frac{1}{(n-1)}(n\bar{y}_{(HD)} - y^{*}_j) \quad \text{if } j \in A_{nr} \end{aligned} \qquad (2.9)$$

represents the mean of y computed by omitting observation j. Thus, (2.9) treats imputed values as if they were observed, and may appropriately be called "naive" for doing so. Rao and Shao modify (2.8) and (2.9) by:

$$v_J = \frac{n-1}{n} \sum_{j=1}^{n} (\bar{y}^{a}_{(HD)}(-j) - \bar{y}_{(HD)})^2 \qquad (2.10)$$

where

$$\begin{aligned} &\bar{y}^{a}_{(HD)}(-j) \\ &= \frac{1}{n-1}[r\bar{y}_r - y_j + \sum_{i \in A_{nr}} (y^{*}_i + \bar{y}_r(-j) - \bar{y}_r)] \qquad (2.11) \\ &\qquad \text{if } j \in A_r \\ &= \frac{1}{n-1}(r\bar{y}_r + (n-r)\bar{y}^{*}_{m} - y^{*}_j) \quad \text{if } j \in A_{nr} \end{aligned}$$

and where $\bar{y}_r(-j) = (r\bar{y}_r - y_j)/(r-1)$. In other words, if $j \in A_{nr}$, then (2.11) is computed in the same way as (2.9), by omitting the imputed value for j. If $j \in A_r$, then y_j is omitted and the imputed values are adjusted to reflect y_j's influence on the mean of the imputed values. Rao and Shao establish the consistency of this variance estimate, both for the single imputation class, as shown, and for multiple classes.

To summarize the strategy in their proof, Rao and Shao show

1) that the modification in (2.9) is the one required to estimate variances under mean imputation; and
2) the amount by which the variance using the hot deck values, (2.7), exceeds the variance of the estimator using mean imputation is asymptotically equivalent to the increase in expected value for (2.10), when going from mean imputation to the single imputation hot deck.

(The direct applicability of replication methods for variance estimation for 1) was observed by Fay 1991, although without taking the critical step 2) by Rao and Shao to extend replication to the hot deck.) Thus, the required correction to (2.9) to produce an appropriate variance estimator for mean imputation fortuitously yields an appropriate variance estimator for the single imputation hot deck.

3. EXTENSIONS OF THE RAO-SHAO VARIANCE ESTIMATOR FRACTIONALLY WEIGHTED AND MULTIPLE IMPUTATION

3.1 Fractionally weighted imputation. Fractionally weighted imputation (FWI) resembles MI in most respects but may be distinguished 1) by the manner in which the imputations are made, and 2) estimation and analysis of the resulting data set.

FWI differs from MI by drawing imputations from the full set of donors in the manner of the original hot deck. For example, as noted in Section 2.1, the approximate Bayesian bootstrap is one of the available methods to produce the variation among multiply imputed sets assumed by the MI analysis. If the first step of the approximate Bayesian bootstrap is skipped, imputations appropriate for fractionally weighted imputation (FWI) result. In other words, FWI is the process, in this instance, of producing *m* imputed values for each missing case, simply by repeating the hot deck selections independently, with replacement. FWI is generally improper, from the point of view of MI, since the FWI imputations do not fully provide the variation required by MI variance estimation.

The FWI estimator assigns a fractional weight to each imputed value. For example, if the analysis of a complete data set would be unweighted, then each of the *m* imputed values should receive the weight 1/*m*, and observed values receive a weight of 1. More generally, the *m* imputed values should divide the original weight for the case equally. Thus, each imputed value receives a fractional weight. This represents a fundamental difference from MI: for FWI, a single, weighted analysis of the data set is envisioned, instead of *m* separate analyses in (2.1). The distinction between (2.1) and fractional weighting collapses for linear estimators, but differences would occur with nonlinear estimators.

As noted in Section 2.2, the Rao-Shao (RS) variance estimator (2.10)-(2.11) begins as a variance estimator under mean imputation but extends to the single imputation hot deck. Further extension of their estimator to FWI, under the conditions of their proof, is virtually immediate. Appendix A.1 discusses this extension.

The RS variance estimator does not employ variation among the *m* different imputed sets in estimating the variance from (2.10). Instead, each of the *m* imputed values is given a weight of m^{-1} times the weight of the imputed case, as if the imputed values had become a cluster of observations. The RS variance calculation is performed once on these weighted estimates, instead of the *m* separate variance calculations required by MI. The RS variance estimator, when applied to FWI, therefore incorporates two features of variance estimation for complex samples - weighting and clustering - that are often incorrectly treated by general purpose statistical software but are always handled by any general software designed for analysis of data from complex samples. In other words, the MI variance estimator built upon variance expressions familiar in the context of simple random samples; the RS variance estimator employs concepts familiar in analysis of complex surveys.

Because the effect of missing data is incorporated in the variance calculation as a whole, instead of isolated as in (2.4) for MI, it is generally unnecessary to reference a *t*-distribution to obtain adequate approximations for construction of confidence intervals.

3.2 Unweighted Hot Deck. The RS estimator was developed for weighted imputation, even though the unweighted hot deck is far more common. Rao and Shao (1992, p. 816) note the bias of the unweighted hot deck and discuss the weighted hot deck for purposes of its overall consistency, under their assumptions about missingness. This consistency does not necessarily extend to estimates for subdomains cutting across the imputation classes, however, and many of the other applications of imputed values in survey analyses. Consequently, most analyses rest implicitly on stronger modeling assumptions than Rao and Shao considered.

The author expects that the popularity of the unweighted hot deck will continue into the future, although applications of the weighted hot deck should begin to appear with greater regularity. Consequently, it is worthwhile to note the extension of Rao and Shao's general approach to the unweighted hot deck, both for single imputation and FWI versions. Appendix A.2 discusses this extension in detail. Because both weighted and unweighted quantities are involved in the calculation, the statement of conditions becomes somewhat more involved. Nonetheless, the conditions appear quite mild. In general, the variance estimator is based on using the survey weights in most parts of the calculation, except for terms of $\bar{y}_r(-j) = (r\bar{y}_r - y_j)/(r-1)$, and the full sample mean, which are computed on an unweighted basis. Again, the variance estimator can be motivated by constructing the appropriate variance estimator for mean imputation, which in this case involves the unweighted mean.

One could employ the stratified jackknife to assess, for a given sample, the degree of evidence about whether the weighted and unweighted analyses differ significantly from each other in expected value. Accumulation of empirical evidence on this question across a series of applications could help inform practice on this question.

3.3 Variance Estimation for Multiple Imputation.

The original form of the RS variance estimator does not succeed at capturing all of the variance of the MI estimator, because it does not successfully capture the additional variation usually added by proper imputation, such as the variance arising from step 1) of the approximate Bayesian bootstrap. In some cases, however, it is possible to add additional terms to (2.10) to accomplish this end. Specifically, in some cases, including the examples studied in Section 4, MI differs from FWI by an increase variance due to step 1) of the approximate Bayesian bootstrap, or other equivalent adjustments to represent the full uncertainty in the data, and this component of variation simply increases the overall variance as a separate additive piece. In fact, this extension is not true to the same level of generality as the previous extensions of the RS variance estimator. This section discusses this possibility for only the case of simple random sampling; Appendix A.3 discusses the issues in extending this approach to the general case.

The strategy is to supplement the jackknife replicates in the RS variance estimator by additional replicates to represent the increased variance of MI compared to FWI. For simplicity, one such version, which employs the original n replicate estimates from the RS estimator plus n additional replicates, will be described here.

$$v_{(MI)J} = \frac{n-1}{n} \Big[\sum_{j=1}^{n} (\bar{y}^{a}_{(MI).}(-j) - \bar{y}_{(MI).})^2 + \left(\frac{1}{m}\right) \sum_{j=1}^{n} (\bar{y}^{a'}_{(MI).}(-j) - \bar{y}_{(MI).})^2 \Big] \qquad (3.1)$$

where

$$\begin{aligned} \bar{y}^{a}_{(MI)}(-j) &= \frac{1}{n-1} \Big[r\bar{y}_r - y_j + \left(\frac{1}{m}\right) \sum_{i \subset A_{nr}} \sum_{t=1}^{m} (y^{*}_{(MI)it} + \bar{y}_r(-j) - \bar{y}_r) \Big] && \text{if } j \in A_r \\ &= \frac{1}{n-1} \Big(r\bar{y}_r + (n-r)\bar{y}^{*}_{m} - \left(\frac{1}{m}\right) \sum_{t=1}^{m} y^{*}_{(MI)jt} \Big) && \text{if } j \in A_{nr} \end{aligned} \qquad (3.2)$$

parallels the terms in (2.11), and

$$\begin{aligned} \bar{y}^{a'}_{(MI).}(-j) &= \bar{y}_{(MI).} + \left(\frac{1}{n-1}\right) \sum_{i \subset A_{nr}} (\bar{y}_r(-j) - \bar{y}_r) && \text{if} \\ &= \bar{y}_{(MI).} && \text{if } j \in A_{nr} \end{aligned} \qquad (3.3)$$

In general, this approach assumes that the additional variance added by step 1) of the approximate Bayesian bootstrap and its equivalents is asymptotically omitted from the RS variance estimator. This condition holds for simple random sampling but may fail in some applications to complex designs. Appendix A.3 discusses this issue further.

4. A COMPARISON OF IMPUTATION PROCEDURES AND VARIANCE ESTIMATORS

Following the earlier example in Fay (1991), we consider imputation for a problem with two imputation class, s and t, of equal sizes in the underlying population. We consider in addition two tabulation classes, a and b, independently cutting across the imputation classes, again of equal size in the population. In other words, the imputation class variable and tabulation variable divide the population into 4 equal-sized cells.

The new results reported here compare 4 estimators and 5 variance estimators, again ranked by generally increasing order of variance of the estimator:

1) Mean imputation, using a jackknifed variance estimator including the effect of estimating the imputation class means, as before;
2) Fractionally weighted imputation, with m = 5 and the RS variance estimator, (2.10);
3) Multiple imputation, with m = 5 imputations, the MI variance estimator, (2.3), and the modified RS variance estimator (3.1); and
4) The single-imputation hot deck, using the RS

variance estimator, (2.10).

The simulations were performed by first drawing the distributions of *a, b, r,* and *s,* and drawing the response from the probability of response and the observed values for respondents from the presumed distribution. The analysis examined three different distributions for y_r: normal, Bernoulli, and χ_1^2. The expected number of total respondents was 70 for three different populations: n=100 with response probability p=.7; n=140, p=.5; and n=350, p=.2. A second set of populations with the same p's was also studied. Each evaluation used 20,000 repetitions.

The performance of the 4 estimators was evaluated in terms of their actual variance. The performance of the variance estimators were assessed both in terms of 1) their relative bias in estimating the true variance of the estimator and 2) the actual coverage probability for nominal 95% two-sided confidence intervals. Average lengths of confidence intervals were also measured. The intervals were based on (2.2) and (2.6) for MI, and the normal approximation (i.e., 1.96) for the other 4 estimator/variance estimator combinations.

The degree of bias in variance estimation and confidence coverage are both important issues for survey research applications. It is particularly true of Federal practice that variance measures are often published in the form of averages or generalizations. For example, the usually reported standard error for the unemployment rate from the Current Population Survey is based on average values over several months. Thus, seriously biased variance estimators, even if they should exhibit acceptable performance in constructing confidence intervals in direct application, do not serve the purpose of variance generalization or averaging well. On the other hand, close agreement between nominal and actual coverage probabilities is also a desirable feature, especially in applications where the variance estimate is used directly without reliance on a generalization.

Tables 1, 2, and 3 abstract results available in full detail from the author. Table 1 assesses performance for estimation for one of the two imputation classes. Multiple imputation is consistent in this case. The first three comparisons consider progressively lower response rates for normal data, but the patterns with increasing nonresponse are similar for other distributions. With moderately high response, r=.7, estimates using the hot deck have a considerably higher variance than the other alternatives. Both FWI and MI have a bit higher variance than mean imputation, with FWI having only a slight edge over MI. For very low response, r=.2, however, FWI still is within sight of mean imputation, while the MI estimator shows no measured advantage over the hot deck. For the same three comparisons, all variance estimators show reasonable properties, although there is a slight deterioration in performance of MI confidence coverage for low response. (This pattern was consistently repeated for the binomial and chi-squared populations as well, although these results are not shown in the table.)

The remaining comparisons in Table 1 largely repeat the lessons from the normal population, with r=.7. The table shows little to choose among the variance estimators. In all cases, confidence coverage is reduced by the skewness of the distributions, but this effect is reduced with increasing sample size.

Table 2 presents results for the mean of a cross-class not used in the imputation. In this case, the MI variance estimator is not consistent, and the MI variance estimates are upwardly biased by about 20-25%. MI confidence coverage is not severely harmed for nonnormal populations, however, and in some cases the effect of skewness in the population offsets the bias in the MI variance estimate to produce approximately the correct coverage. On the other hand, the fact that this offsetting occurs in some instances offers only a bit more than the accuracy of a stopped clock, which is correct twice a day. Better confidence coverage would tend not to occur for normal or nearly normal populations, and thus the advantage to MI could only occur for sufficiently small samples. Furthermore, the example has not managed to show MI at its worst; the effects of the inconsistency of the MI variance estimator could be made worse, for example, by increasing the number of tabulation classes.

Table 3 shows the effect on inferences about differences in means for cross-classes. Because the symmetry of the Monte Carlo setup removed the impact of skewness, the 4 consistent variance estimators exhibit exemplary confidence coverage. In marked contrast, confidence intervals based on the MI variance estimator are punishingly conservative. (Although not shown, the performance of MI further deteriorates with decreasing response.)

Although the preceding comparisons would make FWI estimator the clear favorite over MI, the consistent performance of the modified RS variance estimator for the MI estimator, in contrast to the unreliable performance of the MI variance estimator, should help to promote an objective assessment of MI. Without attempting to account for the phenomenon, the author notes his own perception that a mystique has grown around MI. The emphasis on

Bayesian motivation has perhaps dissuaded some researchers from the realization that the whole approach could be assessed from a frequentist perspective, in fact clarifying properties that had remained obscured in many previous discussions. The simulations here show that the MI estimator does in fact have a variance that can be estimated consistently in cases where the MI variance estimator is inconsistent, and that, when both variance estimators are available, there appears to be no objective reason to favor the MI variance estimator over the modified RS variance estimator.

5. COMPUTATIONAL CONSIDERATIONS

In general, Rao and Shao's approach permits the use the single hot deck for tabulation in the same way as before, but requires modification of the usual replicate values for variance estimation. Since the hot deck described by Rao and Shao differs somewhat from most practice, either practice will have to change to meet theory, or the current theory will have to be extended to meet practice (as this paper has partially accomplished). In addition, past experience indicates that the availability of convenient algorithms will affect the degree to which this research influences survey practice. Consequently, this section describes a general computational approach and a specific implementation in order to encourage the practical use of the RS estimator.

Although there are many mathematically equivalent computational strategies to implement the RS variance estimator, some are more convenient than others. For example, one approach, to create permanent sets of replicate estimates modified in the required manner, will probably represent a hindrance to use of this method. A preferable strategy is to create modified replicates on an "as needed" basis. This second approach permits analytic flexibility without requiring prespecification of each potential use of the imputed values.

The VPLX program (Fay 1990, 1993a) is written in portable FORTRAN 77 for the analysis of complex survey data through replication. The current VPLX can implement the full RS variance estimator, including applications to multistage designs, if the user specifies the RS variance estimator through a series of commands in the VPLX syntax. A standard run consists of a CREATE step to establish the basic replicates, a TRANSFORM step in which the user, through a series of standard functions implements the RS adjustments to the replicates, and a DISPLAY step to compute the variances. Without the TRANSFORM step, the naive variance estimates would result. A new version of VPLX, available approximately November, 1993, will implement the RS adjustment in the TRANSFORM step as a single function call, simplifying even complex applications from the user's perspective.

A VPLX implementation of the RS variance estimator now requires about a page (i.e., about 50 lines) of commands, including comments. Generally, much more complex applications require only somewhat more commands. For example, approximately 3 pages of commands, again including comments, 1) read observed and mean, repeated, and hot deck imputations from a file, 2) establish related variables squaring the original variates and grouping them into 4 categorial intervals, 3) cross-classify the variates by imputation class and tabulation class, 4) perform the RS adjustments to the replicates and compute differences of class means, 6) calculate and write 126 estimates, estimated standard errors, and 6 covariance matrices to a file.

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

REFERENCES

Fay, R. E. (1990), "VPLX: Variance Estimation for Complex Samples," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria, VA, pp. 266-271.

______ (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.

______ (1992), "When Are Inferences from Multiple Imputation Valid?" *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria, VA, pp. 227-232.

______ (1993a), "VPLX: Variance Estimation for Complex Samples, Program Documentation," unpublished Census Bureau report.

______ (1993b), "Valid Inferences from Imputed Survey Data, Supplemental Results," unpublished Census Bureau report.

Krewski, D., and Rao, J. N. K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods," *Annals of Statistics*, 9, 1010-1019.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John